A Qualitative GIS for Social Media and Big Data

by

Michael E. Martin

M.A., University of British Columbia, 2012
B.Sc. (Hons.), Queen's University, 2009

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
Department of Geography
Faculty of Environment

SIMON FRASER UNIVERSITY
Fall 2017

Approval

Name: **Michael E. Martin**

Degree: **Doctor of Philosophy**

Title: **A Qualitative GIS for Social Media and Big Data**

Examining Committee: **Chair:** Jason Leach
Assistant Professor

Nadine Schuurman
Senior Supervisor
Professor
Department of Geography

Martin Ester
Supervisor
Professor
School of Computing Science

Sarah Elwood
Supervisor
Professor
Department of Geography
University of Washington

Bryan Kinney
Internal Examiner
Associate Professor
School of Criminology

Reuben Rose-Redwood
External Examiner
Professor
Department of Geography
University of Victoria

Date Defended/Approved: December 11, 2017

Abstract

Since the 1990's geographers have called for a qualitative GIScience. While several attempts have been made to achieve a qualitative GIS, limiting factors such as data volume and methods have held the realization of such a system back. However, important changes in the last decade have made it possible to achieve this goal. Social media datasets are available for download that contain coordinate metadata and qualitative data about the experiences of individuals. Big data infrastructures make it possible to harvest, store, and find data expressed on specific phenomena researchers wish to study. Natural language processing methods make it possible to understand the context in which a post or group of posts are authored and extract the geospatial insights therein. GIScience has taken notice of these synergies and is beginning to engage with the data and is producing new insights from social media landscapes. In this dissertation, three articles are presented: 1) a method for producing area based topic models from social media; 2) a methodology for geospatial social media exploration and research, and; 3) a software that implements the methods and methodologies of geospatial social media. These three papers make up a body of research that presents a qualitative GIS from data to analysis to output. In the process, the research reflects critically on the ways in which geospatial social media and big data methods in GIScience are created.

Keywords: Qualitative GIS, Big Data, Social Media, Geographic Information Science, GIS

Dedication

For all those who believed in me along the way

Acknowledgements

Thank you to my supervisor Nadine Schuurman for her help and encouragement both inside and outside the confines of the academic institution.

Thank you to my lab mates, past and present. Jon, Britta, Blake, Tatenda, David, Aateka, and Leah. Grad School is so much more than scholarly output. Thank you to my colleagues and friends in the department Ryan, Jonny, Gitto, Leon, and Krystyna.

Mom, Dad, Lyranda, Bobby, Caitlin, and Victoria. Wherever life takes us, I know you are always there cheering me on. To Lynne, I love you and I am so fortunate to have you beside me, always.

Table of Contents

List of Tables

List of Figures

Chapter 1.

Introduction

1.1. Big Data

Big data has had a big impact on academia. The past five years has witnessed a change from big data as a niche issue in computer science to a new focus for research across the academy. For Geographers big data influenced the study of GIScience the most.

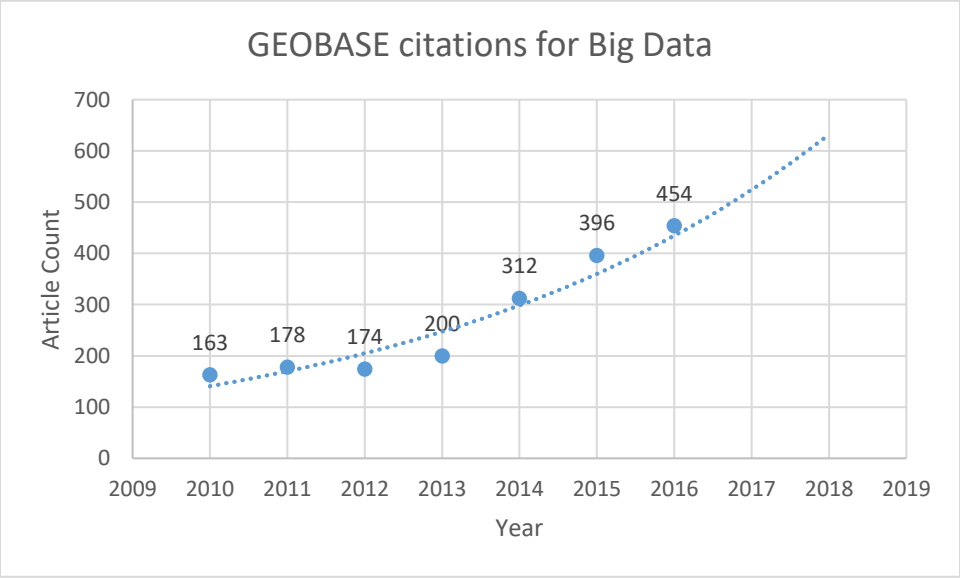


Figure 1.1. Citations matching search for "Big Data" on Geobase, from 2010 to 2016. 2017 already has 431 citations as of early October, and if the yearly rate remains constant will reach more than 500 articles

A GEOBASE search for “Big Data” provides a rough guide for how popular the term has become in the geographical sciences, with 2017 on track to yield more than 500 GEOBASE results.

GIScience has always been focused on the integration of large datasets. The Canadian Geographic System, a precursor of modern GIS, was implemented as a solution to the vast reams of paper maps that bogged down the Canada Land Survey

and as aid for the challenge of analytical overlay in the 1960's (Tomlinson and Boyle, 1981). Since then spatial, temporal, and spectral resolution has continued to increase. The increasingly granular resolutions have required that more efficient methods of geospatial data storage and analysis be created to quantify the natural landscape (Chi et al., 2016).

1.2. Qualitative Data in Geography

Qualitative GIS has been, for some, a paradoxical term. GISystems were designed to integrate geographical information on physical landscapes and represent specific empirical measurements. Qualitative GIS recognises this ability and builds upon it, enabling GIScience to also be capable of incorporating the contextual, situated, and lived experiences when humans interact with the landscape and one another (Elwood, Sarah; Cope, 2009). Representing this using computer data models and structures is possible however, as qualitative information is compatible with numerical representations, too, so long as the signs and symbols of those presentations signify qualitative experiences and concepts, as Palovskaya (2009) demonstrated. At the level of the method, the same is true. Methods that include keywords, statistical processes, or machinations of algorithms utilize numbers to produce output but these numbers represent qualitative information and therefore contribute to a qualitative GIS. Finally, when these data and methods produce output in the form of visualizations, information tables, and text they too contain qualitative information that is interpreted by the reader. These are placed within the context and situated experiences of the data producers, researchers, and ultimate the reader. They are fundamentally a qualitative process. Qualitative GIS has been best positioned as a mixed methods approach that recognises that there are a multitude of ontologies, epistemologies, data, methods, and visualizations that may work in combination or against one another to produce a better understanding of social phenomena (Elwood, Sarah; Cope, 2009; Elwood and DeLyser, 2010; Schuurman and Leszczynski, 2006).

While expansion of quantitative data generated from the physical landscape has continued the tradition of ever-growing, finer resolution information, the effect that big data has had on qualitative information has been radically different.

Although calls for GIS to represent the lived experiences of individuals have been present since the 1990's they have largely gone unanswered (Schuurman, 2000). While attempts have been made to integrate qualitative data and analysis (Elwood, Sarah; Cope, 2009; Jung, 2007; Kwan and Ding, 2008) a major challenge has been the lack of spatial qualitative data and the difficulty in ascertaining it. The amount of spatial information generated in qualitative GIS experiments simply did not compare with those of quantitative projects studying natural landscapes or demographics.

The challenge of qualitative data collection has been ameliorated by changes in the world-wide-web over the last decade. The introduction of Web 2.0 technologies changed the web from acting as a one-way communication medium to a platform for two-way communication (O'Reilly and Battelle, 2009). GIScience took keen interest in the ways that the 'read-write-web' could be used geospatially, especially for advocacy (Okolloh, 2009) and citizen science engagement (Goodchild, 2007). Web 2.0 intersections with geography focused on both the generation of new geographical information in the case of Open Street Map, and in new methods of communicating about place in the case of the participatory geoweb (Johnson et al., 2015).

Participatory geoweb data remained stuck with the problem of what has been termed 'small data' (Sieber and Haklay, 2015). Small data has largely remained separate from quantitative and now big data GIS studies in terms of both data sources and the methods used. Different from their bigger counterparts, these 'small data' projects speak purposefully to the places and spaces in which they are situated. The big data approach to qualitative GIScience has to be different. Instead of collecting data specifically on a subject, big data approaches information collection by aggregating as much data as possible regardless of the research question, then sifts through the collected data to identify patterns and correlations related to a particular inquiry.

1.3. **Big Data, Social Media, GIScience**

Social media has been a key asset for big data qualitative GIS. In particular, Twitter has become a major source for big data investigations because the company offers free access to 1% of its global data stream. While 1% may appear small, Twitter's participation rate is estimated at 500 million postings per day. At a possible 5 million data points per day, a big data researcher can amass a great volume of qualitative

information over a relatively short timespan. Approximately 4% of twitter traffic contains geospatial metadata. While the volume of postings containing location information is higher than the 1% allowed, rate limitations can be worked around by specifying only specific areas to collect information from. For example, this thesis collected 100% of the tweets containing location information in North America. Rate limits aside, Twitter is estimated to be used by 24% of all online adults in the US (Greenwood et al., 2016) and the company reported 319 million global monthly active users in December 2016 (Twitter, 2016).

What Twitter created by releasing a portion of their data to the public is a qualitative data source larger than any available to researchers, and with the global reach the company has (US users make up only 67 of the 319 million reported users), the dataset can be used to study phenomena in a wide variety of locations. Challenges remain, the research outputs from Twitter data may only describe the users of the technology and not all research topics are reflected well in social media discourse. Nonetheless, the data available contains the emotions, ideas, and conversations of its users and due to its accessibility it remains a vast trove of information that can be used for qualitative study. It represents an opportunity for qualitative research methods, methodologies, and software to be written that integrate the lived experiences of individuals directly into spatial analysis.

1.4. An Opening for Qualitative GIS

Social media data that contain location metadata in the form of latitude and longitude offer an opportunity for qualitative GIS to capitalize on. With data volumes similar to – and often rivaling – its quantitative counterparts, it is a critical time to modify, invent and integrate methods for a qualitative GIS. Investigating qualitative data however is not straightforward.

A central component of qualitative data inquiry from interviews and thematic coding is the process of becoming ‘close to the data’. To accomplish this a researcher can listen to, transcribe and/or read qualitative accounts from study participants. Big Data poses great concerns for this methodology, as amount of time required to review all data points collected – even using keywording – far exceeds that of the total time available for study. In a remarkable study Jung (Jung, 2015) attempted to manually

review all Twitter postings (tweets) in their town related to the 2012 US presidential election finding. Not surprisingly, the time required to review the data was onerous. Big data methods have instead sought to find ways to computationally study the data and aggregate the results. The approach of using an algorithm to review and analyze data without user interaction is referred to as an unsupervised approach.

Unsupervised methods are used in other areas of GIScience. Remote sensing utilizes this in land classification systems regularly (Li et al., 2014). However, as social media postings are a product of social relations rather than a set of pixel values, unsupervised methods used on qualitative data have met with skeptical resistance by critical GIScientists (Kwan, 2016). Critical GIScience recognises that while the algorithms that comb through big data are capable of finding correlations that are useful in understanding patterns in data, they are not value neutral (Ricker, 2017). Kitchin and Dodge (2011) explore this concept as Code/Space, explaining that algorithms are constructed by individuals with their own goals and subjectivities imparting their values into the code they write. Stephens (2013a) provides a potent geographic example in her analysis of OpenStreetMap, finding gender bias in not only the data creation mechanism, but also in the review process.

Approaching social media analysis using the tools of GIScience and statistics has been attempted and used by several scholars. Keyword matching (Crooks et al., 2013), heatmaps (Stephens, 2013b), and odds ratio analysis (Zook and Poorthuis, 2014) have been applied to social media datasets providing interesting results, but these methods lack an understanding of the context in which words have been used. Computer science sub-discipline natural language processing (NLP) has become a field of increasing importance because of this criticism and has progressed significantly in the past decade (Bello-Orgaz et al., 2016). As recently as 2014, Eugene Goostman passed the Turing test (Shah et al., 2016). Goostman convinced a panel of experts that he was a human more than 30% of the time when in fact it was a chatbot conceived at Princeton University. While the efficacy of the Turing Test has been called into question as a result of Goostman, it demonstrates the level of sophistication possible from modern NLP systems.

NLP has been used by geographers too. The most common method that geographers have turned their attention to is topic modelling - which finds topics in text

or conversation. It has been used in several applications by GIScientists including obesity and tourism (Ghosh and Guha, 2013; Hao et al., 2010). Sentiment analysis is an up and coming NLP technique that can be used to determine the emotions of text and has been used by geographers in studies on urban poverty (Frank et al., 2013) and infrastructure (Rybarczyk and Melis, 2017). Both topic modelling and sentiment modelling are sensitive to the data they are trained using, further adding to the complexity of their Code/Spaces. A dramatic example of a NLP exercise gone wrong is the Microsoft Tay chatbot released in March 2016. Tay was designed to learn from the tweets sent to it by other twitter users, but after being bombarded by tweets relating to Nazis, Sexism, and hate-speech it began to tweet out similar responses and was shut down within days of going live (Neff and Nagy, 2016).

Clearly there is a tension between the analysis that social media and other forms of qualitative big data make possible and the realities of working with data and algorithms that are non-objective. This thesis reviews these algorithms and processes and attempts to outline not only the ways that algorithms can be non-objective, but to offer suggestions for moving forward within a framework of qualitative GIS.

1.5. How this dissertation builds a qualitative GIS

Qualitative GIS has been a nascent subfield of GIS for more than twenty years. The original calls for a non-quantitative GIS begin at the time of the critical revolution within GIS lead by the work of Pickles (1995), Curry (1994), and Smith (Smith, 1992) who illustrated a non-objectivity that had previously been assumed provided an in the case of the latter, gave an example of the power of GIS when used as a destructive force in the Iraq Gulf War. From these debates (Schuurman, 2000), grew an interest in understanding what a non-objective GIS could look like, and was termed GIS/2 by some (Sieber, 2004). Several attempts emerged from the calls for GIS/2, including the sub-discipline of public participatory GIS (Elwood and Ghose, 2001) and later qualitative GIS (Elwood, Sarah; Cope, 2009). Qualitative GIS called for methods and methodologies that could integrate non-quantitative data directly into spatial environments and software. Efforts ranged from using standard tools with a mixture of community intervention (Knigge and Cope, 2006), to new GIS modules and extensions that integrated thematic coding directly into Esri's ArcGIS (Kwan and Ding, 2008), to reconceptualising GIS operations as more than quantitative (Pavlovskaya, 2009). These efforts to bring

qualitative data and methods into GISystems proved that it was indeed possible, however they suffered from the 'small-data' problems noted earlier (Sieber and Haklay, 2015). However, these two barriers have been ameliorated by the big data avalanche (Miller, 2010).

This thesis capitalizes on the data avalanche by building upon the lessons learned from earlier efforts to conceptualize and build a qualitative GIS. It takes to heart the suggestion in Schuurman's (2000) *Third Wave* of critical GIS, to take a kinder and gentler approach to the critique of emergent methods. This dissertation also takes heed of the advice of Haraway (1988) and attempts to be a part of the construction of the cyborg in order to influence it.

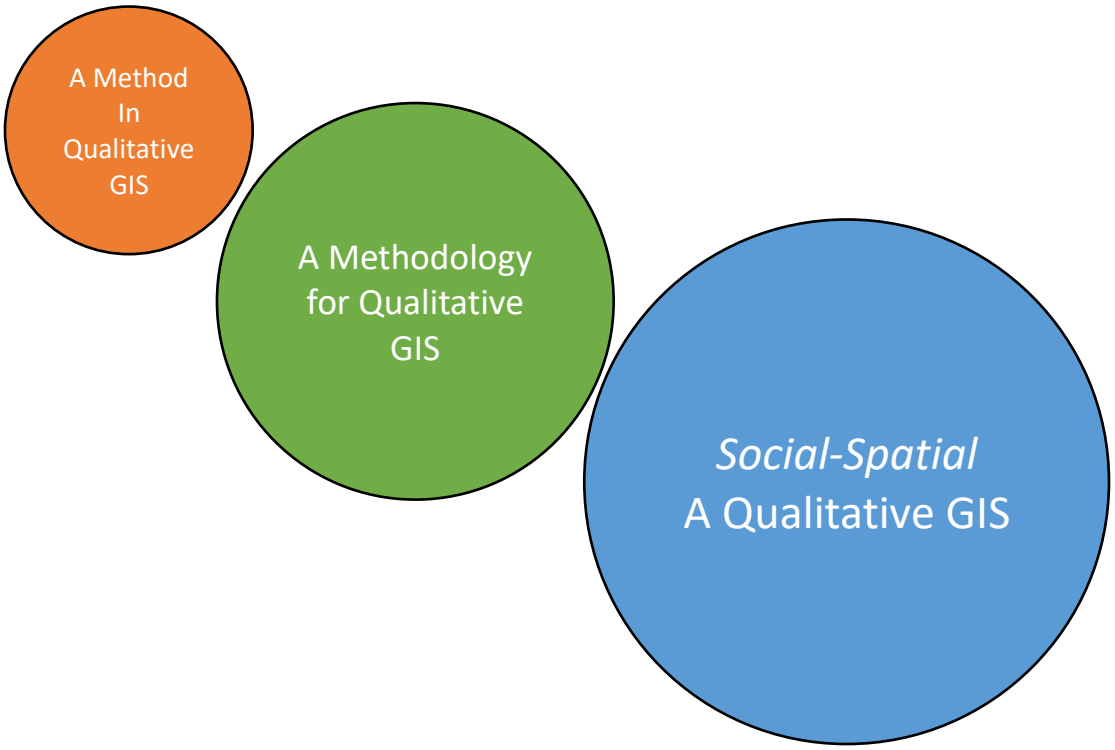


Figure 1.2. Organization of thesis articles of increasing scope from method, to methodology, to software.

This thesis literally and figuratively builds a qualitative GIS through three papers that build on each other. The first paper constructs a single method for mining social media data. It uses a natural language processing technique for topic modelling called Latent Dirichlet Allocation (Blei et al., 2003) and operationalizes it in a way that

geographers are familiar with – areal based boundaries. It visualizes the results of the topic models it produces directly into cartographic space using computer graphics code and matrix mathematics. The code for this method was released open source and in this paper I dissect that ways that it is non-objective in its programming and application.

From this singular method, the second paper increases scope to the level of a generalized social media methodology for qualitative GIS. The progression of increasing scope leads to a different focus. In part, this paper is a review of options for data processing from the point of acquisition through to visualization. However, it is also a critical examination of each method reviewed. It looks at the ways method or process is embodied and non-objective and demonstrates this using a case study of obesity. In this way it not only casts into question assumptions of methodological objectivity, but also supplies evidence of how they may not be objective.

Finally, I turn my focus to the description of a software tool that I created to be used as a qualitative GIS called *Social Spatial*. Social Spatial is an open-ended software that can be used to aid qualitative researchers applying qualitative GIS methodologies. The software, where possible, exposes all parameters used and through doing so offers a prospective user the opportunity to publish the settings and by association the assumptions they made when carrying out their analysis. The program is created using best practices from software development including thorough internal code documentation and an open source publishing using Github[1]. This software is not totally comprehensive, or complete. Rather it is the start of a qualitative GIS that can be continued as standalone development or it can be integrated into other standard GIS software such as QGIS, a popular open source GIS that has gained enormous popularity in the past decade.

This dissertation starts with the description of a method for analysing and representing qualitative big data in GIS. In the second paper, I then describe a range of methods to acquire, analyse, and represent qualitative big data. Finally, I provide an open-source software implementation for qualitative GIS that uses social big data. Each of these steps is an innovative and relevant contribution to the building of a qualitative GIS for social big data and for social scientists in Geography.

[1] http://www.github.com

Chapter 2.

Area based topic modelling and visualization of social media for Qualitative GIS

This paper was published in the Annals of the American Association of Geographers.

Citation Details: Martin M E and Schuurman N (2017) Area-Based Topic Modeling and Visualization of Social Media for Qualitative GIS. *Annals of the American Association of Geographers*: 1–12.

2.1. Abstract

Qualitative GIS has progressed in meaningful ways since early calls for a qualitative GIS in the 1990's. From participatory methods, to the invention of the participatory geoweb, and finally to geospatial social media sources the amount of information available to non-quantitative GIScientists has grown tremendously. Recently, researchers have advanced qualitative GIS by taking advantage of new data sources, like Twitter, to illustrate the occurrence of various phenomena in the dataset geospatially. At the same time, computer scientists in the field of natural language processing have built increasingly sophisticated methods for digesting and analysing large text-based data sources. In this article the authors implement one of these methods, topic modelling, and create a visualization method to illustrate the results in a visually comparative way, directly onto the map canvas. The method is a step towards making the advances in natural language processing available to all GIScientists. The article discusses the ways that geography plays an important part of understanding the results presented from the model and visualization, including issues of place and space.

Keywords

big data, social media analysis, visualization, qualitative GIS, topic model

2.2. **Introduction**

Over the past decade, there have been increasing in-roads in the quest for a truly qualitative GIS (Crooks et al., 2013; Ghosh and Guha, 2013; Jung, 2007, 2015; Zook et al., 2010). While GIS scholars have made progress towards creating such a system, solutions to date have not been successful in reaching a widespread audience. The dearth of widespread integrated qualitative GIS analysis methods and tools has not been without cause however, as qualitative data is challenging to express spatially and methods of qualitative analysis are difficult to integrate with traditional GIS software. In this article, we present a method that supports qualitative geospatial analysis, and provides an example for future research initiatives.

Qualitative GIS has been the goal of non-quantitative and critical scholars for its ability to introduce human experience to maps (Brown and Knopp, 2008; Elwood, 2006; Knigge and Cope, 2006). With the advent of social media, a new and profoundly different source of information is available to researchers that provide an opportunity to represent people using their own voice (Elwood et al., 2013). While other disciplines of science, in particular computing science, continue to make inroads to integrating social information into geospatial products and services, geographers have unique perspectives and methods to contribute. This article is an effort to apply geographic thinking to new geo-social technology creation and to introduce a new method for geographers to use – as a means of expanding the options for qualitative GIS. While calls for a reimagining of GIS in the late 1990's (Harris and Weiner, 1998; Harvey and Chrisman, 1998; John Pickles, 1995; Schuurman, 2000) into a GIS/2 that incorporated the voices of the people it represented has not been possible, incorporating methods that use social media is one way to meet this goal.

We posit that emerging natural language processing techniques and qualitative visualizations are an excellent avenue for interrogating qualitative data. We introduce a method that utilizes social media data to visualize topics present in the geo-social landscape. This method shows gives the user the ability to integrate large amounts of textual information in social media and express the topics contained within on a map surface. This method can be used at any spatial scale using any textual qualitative data with location metadata.

2.3. Literature Review

2.3.1. Qualitative GIS

The foundations of qualitative GIS inquiry are traced to the critical GIS debates of the 1990's (Schuurman, 2000). These debates created a fissure between quantitative and non-quantitative scholars through heated discourse in journals forcing GIS as a research niche to recognise that maps can be used to disrupt or reify power relationships (Harley 1989). Moreover, there were efforts to demonstrate that GIS had largely ignored its potential to represent marginalized people (John Pickles, 1995). As critical GIS as a sub-discipline moved forward, new methods of representing people, places and cultures evolved; indeed participatory methods led to the field of PPGIS and VGI (Chambers, 1994; Goodchild, 2007; Sieber and Johnson, 2013; Zook et al., 2010), and critical conversations that started with a reimagining of GIS for non-qualitative means, known as GIS/2, have led to studies of mixed methods and qualitative GIS scholarship (Elwood, 2008, 2009; Halevy et al., 2009; Sieber, 2004; Yeager and Steiger, 2013).

Qualitative GIS has expanded to include multiple meanings and multiple methodologies (Elwood, 2008). The original calls for GIS to be more than a quantitative tool (Curry, 1994; Harvey et al., 2006; Kwan and Ding, 2008; John Pickles, 1995) have been refined and spatial operations are now understood as more than purely quantitative. Pavlovskaya (2009) argues that overlay, a central component of all GIS, is not a quantitative tool at all but rather a process of qualitative observation. The geographic web (geoweb) (Haklay et al., 2008) refines GIS further as a conduit of qualitative GIS, directly integrating qualitative information from diverse groups of users. Qualitative information increasingly accompanies spatial information in the modern geoweb, and with it the challenge has moved from data integration to data analysis and visualization.

Using qualitative spatial data – beyond raw data presentation in Google Maps mashups (Crampton, 2009; Miller, 2006) – is a difficult task. Analysis of qualitative data in human geography has been improved though the use of computer software (Bazeley

and Jackson, 2013; Richards, 1999), but researchers continue to rely on their intellect to review, think and theorize about the phenomena they observe (Jung, 2015; Woods et al., 2015). Qualitative GIS scholars have made great strides to bridge the gap between GIS and qualitative methods, such as Jung's (2007) CAQ-GIS software for thematic qualitative coding and code clouds (Jung, 2015), Kwan's 3-dimensional (Kwan and Ding, 2008) time-cubes that reveal intersecting lived experiences, and Knigge and Cope's (2006) grounded visualization for iterative community participation in map making.

Analysing qualitative information from social media platforms has received increasing attention from GIS researchers. The interplay and entanglement between qualitative and quantitative methods are an important concern, where analysis of social patterns (Jung, 2015; Shelton et al., 2015) and standard spatial problems (Crooks et al., 2013; Goodchild, 2007; Sieber and Johnson, 2013) are studied using similar datasets. Non-quantitative researchers have increasingly turned to social media as a source for VGI analysis, focusing on what individuals say and where they say it. This has led to better understanding of the role of emotions during elections (Jung, 2015), urban inequality (Shelton et al., 2015), and how gender imbalances are reproduced in geoweb applications (Stephens, 2013a).

Use of social media in social research has not come without criticism. Issues of access and representativeness are key challenges researchers face. The Pew research institute (Duggan, 2015) estimates that only 23 percent of online adults and 20 percent of the general public use Twitter. Twitter appears more popular with Hispanic and Black Americans than White online persons (28 percent,28 percent, and 20 percent, respectively), and most popular with younger 18-49 years of age. Twitter's popularity differs from the other social media by penetration, Facebook commands a user base of 72 percent, with Pinterest at 28 percent and Instagram at 24 percent of online Americans. Gaps in adoption have changed in the last decade however, showing that age and gender are shrinking, however the divide between urban and rural users and higher and lower income households remain (Perrin, 2015). These reports do indicate that there are some voices that are significantly less present in social media postings, and results from using this or any other method utilizing social media should be interpreted with this limitation in mind.

A challenge remains for geographers using social media for analysis to find a way to inductively explore the ideas, themes and topics present within massively aggregated information, while attempting to stay true to the intent of those who produced the information. We know that specific queries organize the results. In essence the problem is allowing information to emerge from the data - rather than asking specific questions that potentially shape the answers - and using this information to encourage further understanding of place.

Lessons from natural language researchers may be a potential pathway for achieving a means of visualizing qualitative data rather than querying it.

2.3.2. Natural Language in Geography

Qualitative researchers face a burden when using social media data. In the age of big data and social media, information volumes have increase by several orders of magnitude over traditional qualitative data analysis. Increased data volumes require new methods, as it has become impossible to continue reading every data point (Jung, 2015). Regular expressions and search terms can make it easier to find and identify social media, but this limits analysis to the data that is specifically searched for, as is the case in Shelton et al. (Shelton et al., 2015). Count metrics gathered using this type of analysis are useful, but it sacrifices the closeness between the researcher and data (Kwan, 2016).

To deal with the tension between remaining close to the meanings hidden in the data and the utility of large data volumes, natural language processing was created to understand the meaning of words in text (Allen 2003). The field has received increasing attention (Asghar et al., 2014; Kim and Chen, 2015; Steiger, de Albuquerque, et al., 2015), with two important branches that greatly increase the capacity of qualitative data coding; sentiment analysis and topic modelling (Asghar et al., 2014; Ohmura et al., 2014). Sentiment analysis is used to label text containing an identifiable emotion within it and geographers have used this technique to identify emotions expressed in differing geographical contexts (Robertson et al., 2015). Topic modelling also aims to label text, but instead of reading emotions it seeks to determine the topics of conversation (Blei et al., 2003). In a geographical context, topic models have been used to find topics at the

level of the city (Bauer et al., 2012) and topics dispersed across a landscape (Slingsby et al., 2007), and for specific topic areas, such as obesity (Ghosh and Guha, 2013; Gore et al., 2015).

The topic model used in this research is based on the latent dirichlet allocation (LDA) model proposed by Blei et al. (2003), cited nearly 16,000 studies on Google Scholar at the time of writing. Since its creation, several other models have been created on top of LDA, such as labeled LDA (L-LDA) (Daniel et al., 2009). L-LDA seeks to generate names for each topic that is generated by an LDA model. In this research we left the interpretation of topic names for the map up to readers. However, while topic modelling is increasingly used for social media analysis, two challenges remain for geographers. First, topic models are complex methods to implement and are sensitive to parameter settings. Standard tools do not currently exist for GIS scholars in either industry or open source GIS. This inaccessibility has kept topic modelling from being fully examined by geographers and as a result a multitude of issues such as modifiable area units, scale, edge effects, landscape, and social process that geographers are well positioned to answer have remained unexplored (Dalton and Thatcher, 2015; Kitchin, 2013). Second, topic models are difficult to express cartographically. Studies that have used topic modeling across geographies have often presented results either in tabulated form (Ghosh and Guha, 2013; Mei et al., 2008; Wang et al., 2007) , or as dispersed clouds of words over general geographies of the city (Slingsby et al., 2007). Instead a method that visualizes the results of topic modelling cartographically would be useful so map readers can easily compare neighbouring areas to one another, illuminating differences and similarities over neighbourhoods.

2.4. **Methods**

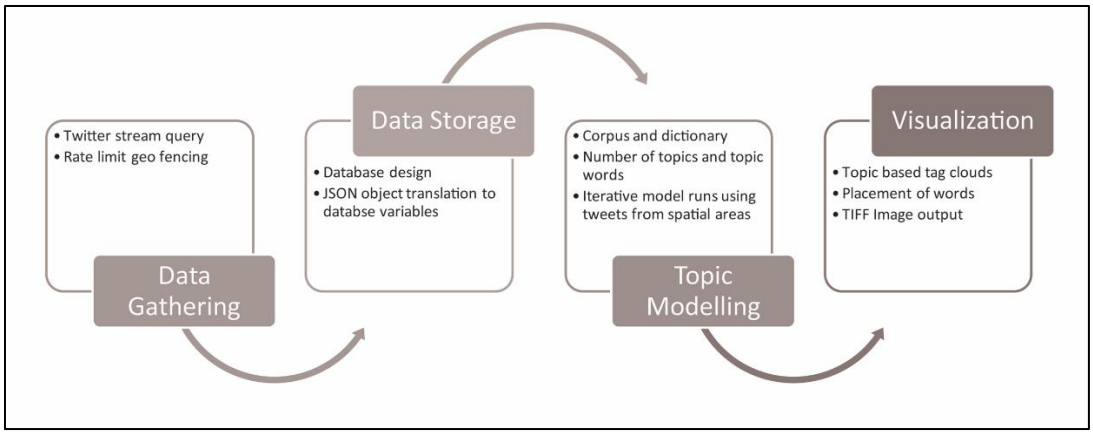


Figure 2.1. Methodological process from data gathering to visualization

Creating an area based topic model requires multiple components of varying complexity. To better understand the process of creating area based topic models, we first review the methodology used in this study involving: 1) gathering social media as a textual base for analysis; 2) analyzing data using topic models; and 3) visualization of results into GIS ready formats. These methodological steps are followed by an explanation of programmatic specific methods.

2.4.1. Gathering Social Media

Maximizing data within rate limitations

Conducting analysis on external social media data requires careful consideration of how to acquire relevant information and store it for analysis. Using Twitter data researchers can connect to a data stream and download large amounts of information, however it is important to consider the exact nature of the data feed in order to achieve optimal data flow.

Twitter imposes a rate limitation on the data they provide. An application using their social media stream can only call for 1 percent of their global traffic – exceeding causes interruptions to the data flow. Additionally, this research project focused explicitly on the 4 percent of Twitter data containing geospatial locations in latitude and longitude

pairs. For this research project a geospatial boundary was placed around amount of information requested from the Twitter data stream. This limited the amount of data requested and kept the database to a manageable size while ensuring a near 100 percent data retrieval rate of geospatially referenced posts.

Data retrieval and storage

The twitter data collected for this study was collected using an HTTP GET request to the Twitter streaming endpoint, and stored in a PostGreSQL database. In the translation from the twitter stream to the database, the data was converted from the provided javascript object notation (JSON) format to the POSTGreSQL table format and included the following variables: name, username, date, time, self-reported location, coordinates (Latitude & Longitude), and tweet text. The table containing the social media data also had spatial and textual indexes to increase efficiency of data retrieval.

2.4.2. Topic Modelling

Following data retrieval, this study focused on creating methods for analyzing the social information in each post. The primary method of analysis used was latent dirichlet allocation (LDA), a form of topic modelling commonly used in natural language processing (Blei et al., 2003). The use of topic models required the development of a twitter dictionary, determining optimal model parameters, and running the model within the context of spatial areas.

Corpus and Dictionary

Natural language processing (NLP) focuses on understanding meanings and attitudes within a body of text. The text analyzed may be one large document, or a collection of documents denoted a *corpus*. In this study, we utilized a corpus of 800 million tweets, generated over a period of 20 months from February 2014 to October 2016 covering the North America. Vancouver neighbourhoods accounted for 690, 337 posts (figure 2.2). From the corpus, a dictionary was created that curtails the number of words considered by the analysis to ensure that output refers to relevant topic words instead of conjunctive and predicate forming words (eg. I, it, he, was, after, into, the,

and, etc.). Two rules were used to remove words from the dictionary; 1) a list of non-topic related words was formed, called stop-words; and 2) any word appearing once in the corpus was ignored. The remaining words used in the corpus after applying these two trimming rules became the dictionary used in the LDA topic model.

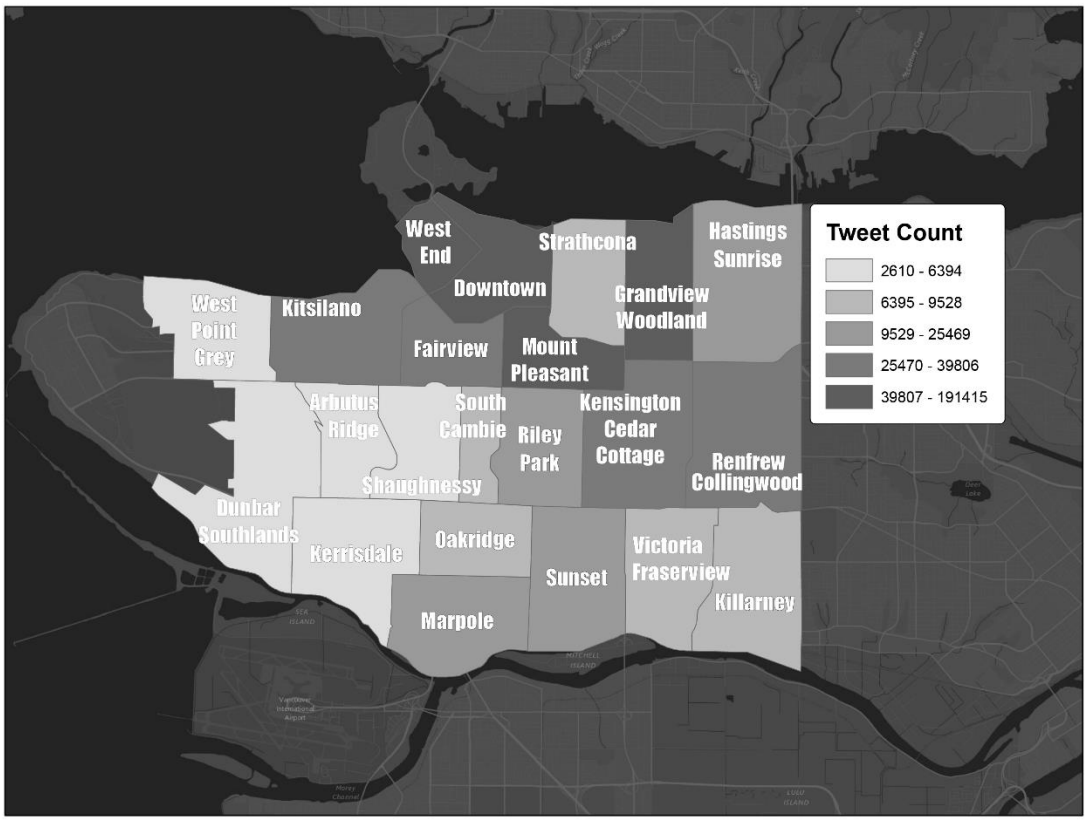


Figure 2.2. Tweet counts per Vancouver Neighbourhood

Running the LDA model for Spatial Areas

Once a corpus and dictionary are set, the LDA model was configured and run. Running the model required two parameters, the number of topics and words per topic, and the number of passes to iterate through the corpus to look for topic words. The number of topics changed the model results, so it was important to ensure the number of topics requested from the model was appropriate. The number of words per topic did not affect the generation of topics, but was an important consideration for output table design. The number of passes for the model to iterate over the corpus is an important consideration for the LDA model, as it directly impacted the predicted probability that any word does exist in a topic at the cost of the computation time.

Several methods exist to introduce spatial context to topic models (Hong et al., 2012; Liu et al., 2015; Yin et al., 2011). This study eschews these, instead focusing on incorporating topic modelling within traditional GIS environments and well-known areal units. This method creates a new topic model and creates independent results for each spatial areal unit in the input landscape (figure 2.3). Approaching topic modelling from this perspective requires a high density of social media posts (alternatively larger spatial areas or longer corpus documents could be used), but ensures that the results of the model can be compared directly with traditional data sources, such as census data or other area based data aggregate. Using areas the results of this method can be visualized onto geographical space, allowing for a visual representation of social qualitative data.

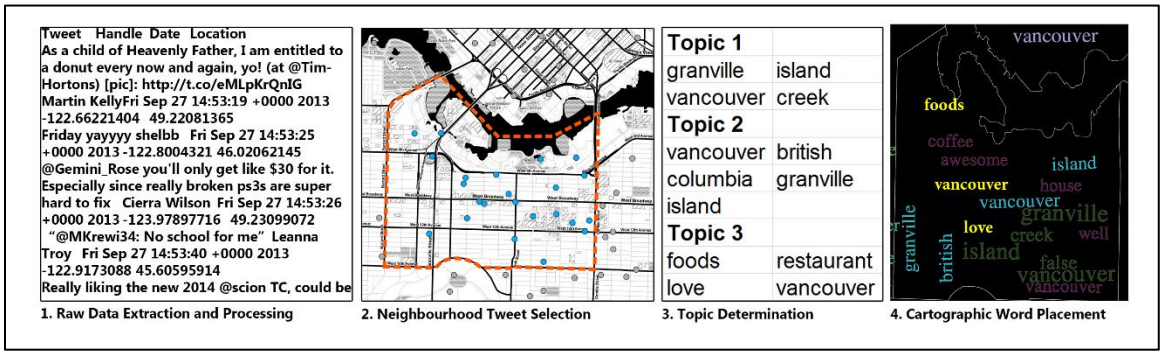


Figure 2.3. Spatial aggregation process from data to visualization. Spatial locations of tweets in neighbourhood tweet selection are simulated using random locations

2.4.3. Visualization of Topic Models

Producing topic model results for spatial areas, while useful, is difficult to understand without the context of a visual output in Cartesian space. This research presents a method for visually conveying the output of the topic models simply, within its spatial context. To do this, tag clouds (often referred to as wordles) have been utilized. However, while tag clouds are typically used with point based data and word counts, this method optimizes the placement and sizing of topic words and families within the areas they represent. To convey the relative probability of specific words to be in specific topic families, font size has been used. To denote the membership of a word within a topic, color has been employed. Across spatial areas font size was preserved, thus if the topic model in one spatial area is unable to predict words as strongly as other areas the words

in the area may not fill the entire space provided. Color was assigned in order of best to least predicted topic families, where the best predicted families will be the same color across spatial areas. Alternatively, font may be used to optimize space filling, and color can be assigned randomly.

2.4.4. Programmatic Implementation

The above methodology was implemented using the python programming language and a conjunction of server request, database and image manipulation modules. Data from Twitter was requested using OAUTH 2.0 credentials using programming and the cURL python module. The data, once received was processed into python objects using the JSON module and reformatted into a format specific for storage into a PostGreSQL database (PostgreSQL, n.d.) with PostGIS extensions (PostGIS, n.d.). Connections to the PostGreSQL database were facilitated by the Psycopg2 module (Varrazzo, 2010). The LDA topic model was implemented using the GENSIM module (Řehůřek and Sojka, n.d.), and the intersection calculation of tweets per spatial area are completed using an input shapefile or PostGIS geometry table, and an SQL command. Completing the intersection command using SQL dramatically increased performance of the algorithm. Finally, the visual tag clouds were initially based on the examples of Nicholas Rougier (Rougier, 2009) and adapted for this project. Generation of output images and manipulation of multi-dimensional arrays required pyCairo (Pycairo, n.d.) and NumPy (NumPy, n.d.), respectively. The output images were saved in TIFF format.

2.5. Results

2.5.1. Topic modelling tool for GIS

The first result of this research project was to create a visual topic model that other researchers will be able to use to carry out their own research on any textual database the covers any geographical area. To this end, the software is available online at http://www.github.com/mikedotonline/VisualTopicModels. Researchers may download

the code therein, and find instructions for running the tool on data of their own. The code consists of two modules, one for topic modelling and another for generating visual representations of the topic models. The two modules can be used in conjunction or separately.

2.5.2. Topic Models

The models run for this project generated five topic families, with five words per topic, for each polygon in the dataset. Each of the twenty-five words generated is accompanied by a probability score that indicates the likelihood that each word is an element of the topic. An example topic for a neighbourhood in Vancouver can be seen in table 1. This model used 50 iterations, producing five topic families, and recorded the top five words. Each topic has been given a name by the author. The example used in table 2.1 illustrates the topic families in the Downtown neighbourhood of Vancouver, an area noted for the presence of the city sports stadium and pubs as well as the Downtown-Eastside (DTES). The presence of the stadium and the restaurants and pubs can point to the abundance of landmarks, sports, and alcohol in the model. Additionally the DTES is known for its low socio-economic status indicators (Bell et al., 2007), and high level of pedestrian injury (Walker et al., 2014). While landmarks are present in many of the neighbourhood topic models, alcohol, and specifically watching sports in alcohol serving establishments are prominent in the Downtown neighbourhood topics. While connections between low-SES, injury and alcohol are established (Bonevski et al., 2014; Burrows et al., 2012; Redonnet et al., 2012; van Oers et al., 1999), it is important to understand that the implications of the results of this topic model warrant further investigation that is outside the scope of this research.

Table 2.1. Topic modelling results for the neighbourhood of the Downtown of Vancouver, Canada. Topics have been given names by the author: Landmarks, Hockey; Soccer; and; Alcohol.

Downtown Topic Model				
Topic 1: Fortune Sound Club	**Topic 2: Landmarks**	**Topic 3: Hockey**	**Topic 4: Soccer**	**Topic 5: Alcohol**
Club	World	Rogers	Vancouver	Drinking
Vancouver	Science	Arena	Whitecaps	IPA
Coffee	Vancouver	Canucks	BC	Vancouver
Fortune	Chinatown	Game	Steamworks	Ale
Sound	Sun	Theater	VWFC	Beer

2.5.3. Image Output

While the results displayed in table 2.1 demonstrate typical output from a single neighbourhood, the spatial topic model was run over the entire geography of Vancouver, Canada. During this run, the model produced a similar table of data for each Vancouver neighbourhood (24 neighbourhoods), and five topic families for each (120 families), and five words in each topic (600 topic words). Each of these words were then written onto the Vancouver landscape, within their respective neighbourhoods. The final output of this yielded an 8MB image, 10000x10000 pixels in size (figure 2.4). Figure 2.6 has subsections of figure 2.4, to provide a closer inspection of the words as they are written.

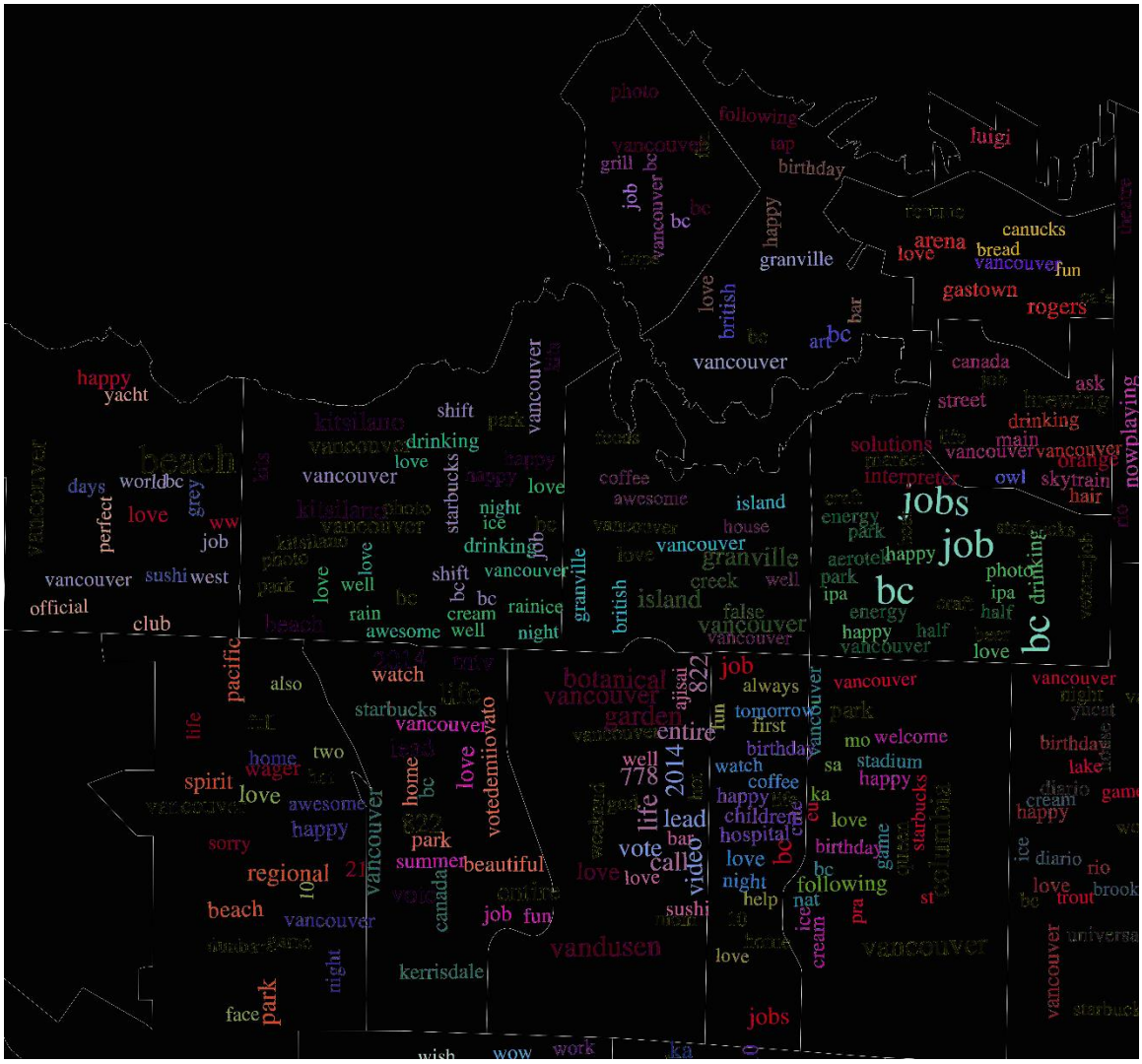


Figure 2.4. A small map-scale visual topic model for the neighbourhoods of Vancouver, Canada.

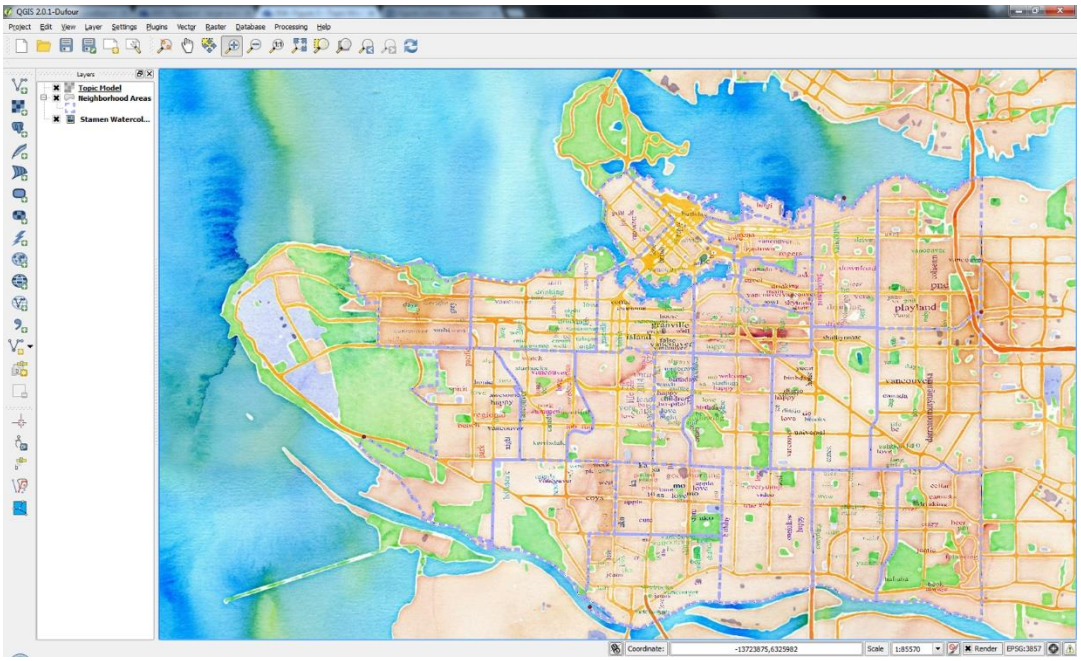


Figure 2.5. **Visual topic model results loaded into QGIS with transparency and OpenStreetMap (www.openstreetmap.org) data for context**

Figure 2.6. **A large map-scale map of the Mount Pleasant and Olympic Village neighbourhoods of Vancouver. The topic models indicate separation of topics by color and relative probability scores by size of word**

2.6. **Discussion**

Topic modelling and visualization are complex procedures. When used correctly, these methods have to ability to provide insight into the ever larger data stores. The complexity of topic modelling and visualization and the decisions that are made during their usage can, however, make it difficult to understand what exactly is shown on the map canvas.

Topic modeling has several input parameters, iterations, number of topics, dictionary, corpus, and stop words (Blei et al., 2003). Adding spatial dimensions introduce more variables that effect model outcome. In the results presented here, the topic models are formed using the social media items collected within. However, it is impossible to know at face value if the topic reference the issues that persons living in the place face, or if the topics are those of a more transient crowd. For example, in the Vancouver neighbourhood including Granville Island (figure 2.6, a very tourist heavy area, the best predicted topic is 'Granville', 'Island', and 'Brewery' and this is not surprising given the crowds of tourists that flock there on vacation and the draw of a brewery. Can we draw the conclusion that social interests of the place are about breweries, or is the effect of tourism drowning out the signal of residents' interests. Similarly, events that promote the use of social media can alter the topics that are posted. For example, in downtown neighbourhood of the convention centre, we see various words relating to specific conferences bubbling up. This, no doubt, can be controlled by integrating temporal controls on topic models, but effective temporal scales and filters are not necessarily evident before running a topic model and viewing the output. Complicating matters further, the filters used in one spatial area may not be appropriate for another. Computer scientists have created methods for checking if topics are trending (Becker et al., 2011; Bolelli et al., 2009), but this adds additional complexity to an already difficult process, and may hide topics important to researchers. In short, geographic topic models require flexibility do deal with a number of scenarios, while still maintaining clarity in their operation.

Geography also plays a role in dividing – or uniting – populations of users and social media posts present in the data. Any area based approach to modelling data will introduce edge effects and the consequence of modifiable area units (Páez and Scott,

2005). Complicating this issue further is the nature of user mobility. Is the content of a tweet influenced by the location that they are tweeting from? This is a central concern for an area based approach, as it entangles the chicken and the egg; did the location bring about the tweet, or do people of this place tweet about that topic in a particular way. These two positions, are ontologically different and make a qualitative analysis of location much more difficult.

Area units, as they unite and divide, beg the question; are they the most relevant (Dalton and Thatcher, 2015)? Neighbourhoods and census divisions logically appear as the most salient places to start, as they are laid out both quantifiably and/or culturally to be homogenously distinct (Morphet, 1993). Computer science has attempted to challenge area based methods, such as using vornoi polygons (Hecht et al., 2011), however frequency of topic words per unit area is frequently used (Ghosh and Guha, 2013; Hao et al., 2010). However, using choropleth maps to show the varied data of qualitative information hides the richness of the language used in each topic.

Areal units also have the effect of hiding the distribution of tweets inside them. In this case study, the number of tweets per area varied from 2,610 to 191,415. A lower number of tweets per area will result in less reliable topics prediction. This makes area unit selection important for two reasons. First, readers of the map need to be made aware of the discrepancy between the highest tweet count areas and the lowest. While adding background shading to indicate the relative tweets was considered for this method, it was ultimately not used at the resulting images became too complex. The second issue concerning tweet counts and areal units is to ensure that the correct scale is used. When the areal unit was smaller than the neighbourhood level, the effect of data density was magnified as some areas had too little data to be adequate predictors of topics. When the area units were much larger than neighbourhoods, the topic model suffered from becoming too generic, offering little information about place. With larger area units, this method may be more appropriate with limiting tweets and topics to a narrower temporal resolution.

Visualizing topic model results engage with core of these concerns and more, as they are made in an effort to convey the information in a particular way to the reader. Choices concerning word color, size, texture, font, positioning, collisions, orientation and others all impact the way that the map will be read by its indented audience (Monmonier,

1996). In producing a geographic visualization of topic modelling results, each of these choices must be kept in mind. Specifically, the connection between a set of visualized words and geography is of primary importance. To visualize this relationship previous studies using topic models have produced topic model results beside the geographies (Kling and Pozdnoukhov, 2012) or to simply show relative frequency of a topic per area (Ghosh and Guha, 2013; Hao et al., 2010). In this research however, the words that make up topic models and their relative probabilities are presented directly on the areas they represent, providing the map reader with both results and context. Drawing words on the map like this requires the reader to be aware of how the map was produced, however that is a concern with all cartographic visualizations (Monmonier, 1996). It is hoped that with a full explanation given in the methods section, the reader will understand the choices made, and the alternatives that could have been used.

2.6.1. Future Work

In creating a visual topic model, we recognize that this is only a first step towards integrating the benefits of natural language processing into the familiar workspaces GIScientists use. It is our hope to integrate the two primary components in this work, topic models the visualization thereof, into a tool that can be integrated with modern GIS, presenting the user with options to tweak and change the parameters of the model and the visualization. Opportunities also exist to make the visualization process more interactive, allowing the cartographer more choice over word placement, and the recipient of the map an ability to interact with the word families displayed, instead of static nature of images currently produced. Additionally, there are a number of other tools that natural language has to offer GIScientists and geographers which we hope to develop and integrate, including theme based topic models (such as looking specifically at health, crime, leisure etc.), and predictive modelling.

2.7. **Conclusion**

In this article we have brought two novel methods to the fore: geographic topic modelling and visualization of qualitative information resulting from topic models. These methods were used to create a topic model for each neighbourhood in Vancouver using

a corpus of social media postings from Twitter users. These geographic topic models have been used as input data for the visualization method developed, rendering topics by color and probability by word size (figure 2.4). These visualizations were saved as image files that can be used with traditional GIS software (figure 2.5). The realization of these methods opens a new avenue for qualitative researchers to probe at big qualitative datasets, where the burden of reading every data point is no longer feasible. Over the past decade, there has been a sustained call to include qualitative techniques in GIS analysis and output. When first articulated, these calls were bordering on utopian – as tools to implement qualitative analysis were largely absent. Since then there has been increasing attention paid to methods of implementation for qualitative GIS. This article offers another tool to achieve the goal of a more inclusive and qualitative GIS environment It is hoped that over time, the methods featured in this article can become standard tools in modern GIS environments. Integration of methods from natural language processing and computer science is particularly well suited to pushing the frontiers of qualitative GIS and providing new avenues of research to be explored.

Chapter 3.

Social Media Analysis for Human Geography and Qualitative GIScience

This paper has been submitted to Geoforum for review

3.1. Abstract

We are in a period of social media data proliferation. These data are inherently geographical which presents opportunities for human geographers and qualitative GIScience researchers. However, many of the computing methods used to analyze social media data are developed in computing science - and black boxed. It is imperative to understand the code-spaces of these methods in order for geographers to utilize the methods and understand their algorithmic bases. This study reviews these methods with the goal of enabling human geographers and qualitative GIScientists to engage with social media data – while understanding the code-spaces the algorithms, data, and researcher embody.

Highlights

- Relevant recent work by big data and critical GIScience scholars is reviewed
- A four stage methodology is presented for researchers who wish to engage with qualitative GIS using social media data
- Critical assessment of the methods used at each stage of the methodology Discussion of the embodied nature of algorithms

Keywords

Big Data, Social Media, Qualitative GIS, Critical Reflexivity, Code-Space

3.2. Introduction

Social media and big data are twin pillars of commerce, surveillance, and community in modern society. Use of social media and big data analytics have become major drivers in national election outcomes (2016 US election), manipulate our sense of wellbeing (Facebook sociological experiments), and even impact our sense of personal privacy (Snowden revelations). Research into big data and social networking has become a major source of inquiry across academia penetrating geography and GIS, as evidenced by its popularity in top academic journals (Kwan, 2016; McNutt, 2016; Neff and Nagy, 2016; Shah et al., 2016; Yin et al., 2017).

Within the sub-discipline of human geography, significant work has been done on the impacts of big data and its relationship in society. Critical scholars have pointed to the ways that it shapes our understandings of the world around us and curates our experiences (Elwood and Leszczynski, 2011). GIScience scholars have looked to big data to identify emergent patterns using the existing spatial analysis toolset. Neighborhood effects, spatial clustering and autocorrelation, and spatio-temporal analysis have been used with social media datasets to illustrate how societies react and interact with natural disasters, epidemics, and social inequality (Crampton et al., 2013; Crooks et al., 2013; Shelton et al., 2015; Zook et al., 2015).

Beyond geography, the digital humanities have studied the effects of big data (Lane, 2017). While the original and traditional grounds for the digital humanities is rooted in literary analysis, they have been experimenting with lexical analysis of large datasets for some time, such as Perseus (Crane et al., 2000) and Transcribe Bentham (Causer and Wallace, 2012) projects that integrate vast quantities of information using linked-data and XML. There is, however, a concern within the discipline that digital humanities ‘missed the boat’ (Prescott 2001, Lane 2016) by becoming bogged down in small or special projects, while other disciplines have focused on building toolkits or engaging in critical debates. Scholars within GIScience have worried about these issues too. Following the great success of large geospatial web applications (Crampton, 2009), geography scholars called for human geographers to engage with big data, or be left behind (Kitchin, 2013). At the same time digital humanists have identified that big data has the potential to rework earlier findings as data volumes increase in size (Jockers, 2013). GIScience, by contrast, has not viewed big data as a means to rebuke findings,

but rather as something that has the potential to examine patterns and phenomena (Elwood et al., 2013).

Qualitative GIS has also looked to big data as a potential data source for new inquiry (Martin and Schuurman, 2017). The push towards a GIS capable of working with non-quantitative data emerged during the science debates of the 1990s when scholars called for non-quantitative methods and critical reflexivity (Harley, 1989; Harvey et al., 2006; John Pickles, 1995; Schuurman, 2000). Qualitative GIScience seeks to identify the data, methods, and cartographic representations that can be used to study social phenomena (Elwood, Sarah; Cope, 2009; Jung, 2015; Kwan and Ding, 2008). While GIScience has begun to interface with big data and social media, it has largely focused on the traditional tools and methods that it is best known for (Poorthuis and Zook, 2015; Zook and Poorthuis, 2014). As Sui and Goodchild pointed out (2011), GIScience must move beyond the study of x,y patterns of place and begin to integrate the ways that social media informs our understandings of *place*.

Qualitative GIS, however, has the potential to integrate methods from other disciplines, such as those learned from the digital humanities and computer science. Computational linguistics has begun to uncover novel methods that deal primarily with the specifically qualitative nature of social data (Larsen et al., 2015; Wang et al., 2007, 2012). A critical moment has emerged in which GIS can integrate the lessons learned from the digital humanities, from computer science and from qualitative GIScience. Big data social media analysis is made possible through the triangulation of methods from these three areas – and can ultimately be used to integrate geographical context into understanding of social media data. However, such analysis requires bridging methods between sub-discipline and disciplines. It also requires a look at these methods through the lens of qualitative research, rather than quantitative. As Pavlovskaya (2009) reimagined, many of the traditional methods of GISience can be interpreted as a qualitative process. As social media data is qualitative in its essence, the approach to its analysis can be done this way, too.

In this article, the authors present a digest of methods for human geographers to interact with social big data for analysis of phenomena. Critical to this approach is its adherence to the advice of Donna Haraway (1988) who offered instruction that - in order to influence a field – such as computer/geographic information science, one must

interact with it as an insider. In so doing we hope that this article may provide an entry point for further uptake and utilization of big social media datasets with a focus on the ways that geography can participate centrally and enhance phenomenological research. Figure 3.1 illustrates steps from data acquisition to communication of output that we discuss at length.

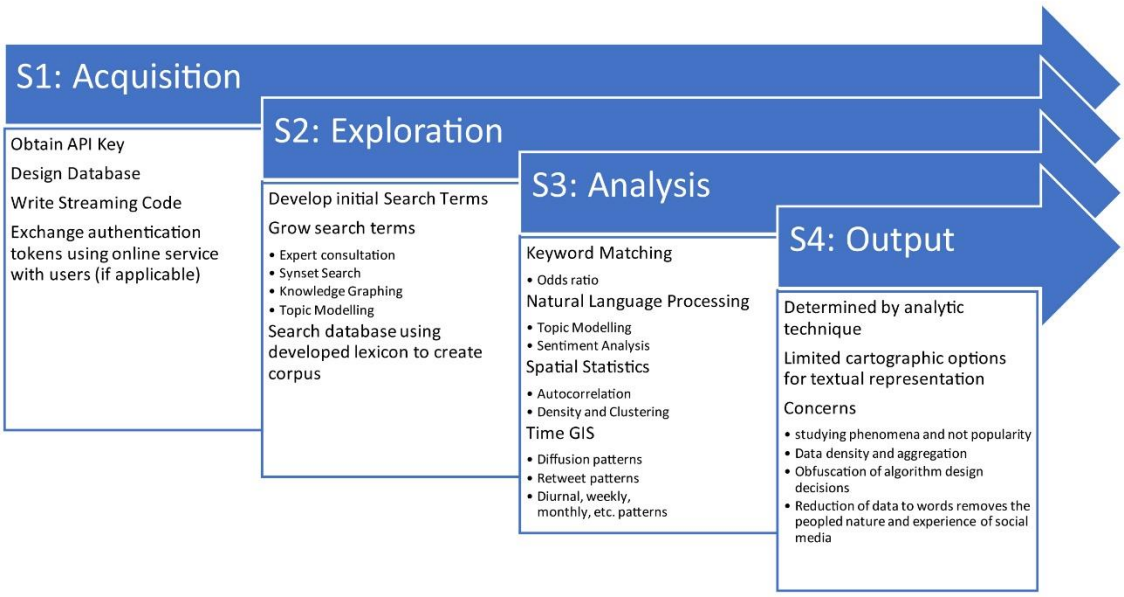


Figure 3.1. The stages of social media research, from acquisition of data to output cartography and figures. Throughout the paper, we offer insights – based on experimentation – that will allow more geographers and GIScientists to integrate these data into their analyses.

3.3. Stages of Social Media based Research

The process of social media research generally follows the pattern of more traditional social research methods. However, because of key differences in big data, important variations to standard procedures and new methods are investigated through this section on the stages of social media research. The volume, velocity, and veracity of social media data require new methods and care, similar to the new methods GIS required over previous mapping methods (Lee and Kang, 2015). For example, while it has always been possible to draw overlays by hand, the Canadian Geographic System made this task far easier and faster to accomplish, over landscape scales that would have previously been impossible (Tomlinson and Boyle, 1981). Computational analysis of social media data is similar. While it is possible to use traditional methods of analysis

with social media postings, the number of postings that are included often necessitate specialized methods of analysis (Jung, 2015).

In this article four stages of social media research are identified: acquisition, exploration, analysis, and representation. Each of these stages are explored here, with a specific focus on integrating geographic elements.

3.3.1. Stage 1: Acquisition

Social media data is an umbrella term that can mean anything from restaurant rating (Yelp), to fitness tracking (Strava), to news sharing (Facebook), and microblogging (Twitter). The commonality of these however is that there is a sharing of personal information, usually self-written text, among a group of people – often the public at large.

Obtaining social media data often requires a researcher to be savvy with three important technologies: an application programming interface (API), a programming language and database server. Often, access to the data also requires financial negotiation, as the data these platforms produce is the prime mechanism through which they make money – including advertising, as advertising is based on data driven placement. In contrary to this, Twitter makes a 1% portion of the data it produces available to researchers for free.

In previous research conducted by the authors (Martin and Schuurman 2017), Twitter was used as a data source and obtaining it required using the API specifications listed on the Twitter developer's website (dev.twitter.com), writing Python code, and creating an indexed table in a PostGIS database. The use of Twitter made collection of data relatively simple, as the information produced is static in time and location. In comparison, with Strava where users post running and cycling activities and other users comment on their activities, data is more complex due to asynchronous generation and collection. Strava also does not provide access to a public data feed like Twitter. To obtain data from Strava, researchers can either contact their data sales division (Strava Metro) and buy anonymized data, or obtain permission to access data from individual users. Purchasing data is a more attractive solution due to the short amount of time it requires from initial interest to data retrieval, however it lacks much of the metadata and social context that can be garnered through consent-based data gathering.

Automating the consent process by writing an application -- often called a bot -- that facilitates access to user data can make the process much faster. Bots make it possible for a researcher to send out invitations to research participants and facilitate account credential handover in a matter of seconds. This can be done using a browser or smartphone (illustrated in Figure 3.2), and the researcher does not need to interact with the participant beyond sending the invitation to participate. Like Strava, most popular social networking sites follow a similar credential exchange paradigm such as used by Facebook and Instagram.

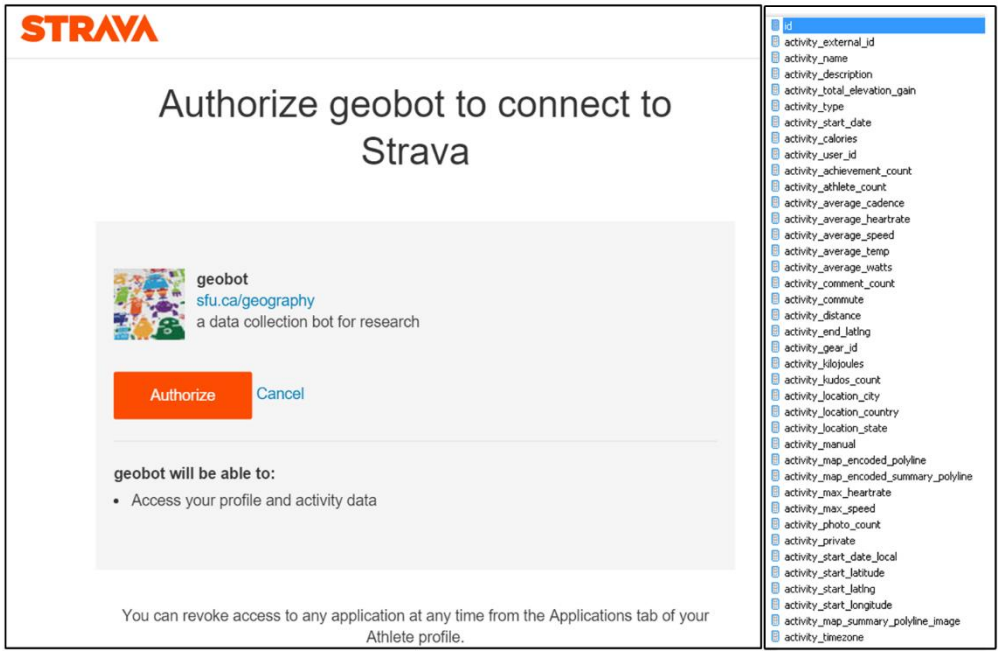


Figure 3.2. Authorizing a data collection bot (left) and the variables that can be harvested (right) on Strava (www.strava.com). This is the basis for an informed-consent data gathering strategy.

Once social media data has been harvested and stored in a database for later use, there are no controls over how long the data can be kept, outside of research ethics approval imposed by academic institutions. It can be analyzed or sold to others verbatim without further contacting the persons who produced the data, raising the concerns of social media researchers (Ko et al., 2010; Wang et al., 2011).

Acquiring social media data has commonalities and differences with standard social research practices. They share the process of discovery of a social phenomena and cohort of participants to investigate and storing that data for later usage. Differences appear however in the techniques used for obtaining that data. Tape recorders and

hours of transcription are instead substituted for time spent writing computer programs that parse streams of information into data structures and negotiate consent digitally.

3.3.2. Stage two: Exploration

Social media data differs from qualitative data capture because it is produced for purposes aside from research. As a result, the vast majority of collected data may be irrelevant to the study at hand and the language used by participants can be both colloquial and opaque in meaning. The words used to express phenomena may be different than the researcher would use, and the opportunity to ask a participant to clarify usage of a word or phrase is rarely possible or practical. Consider the example of researching obesity and the tweet in Figure 3.3. When searching for key data points, the term “#omnomnomivore” may not be the first term that comes to mind, yet it is what signifies this tweet as related to overeating during the holidays.

Figure 3.3. The #Omnomnomnivore’s dilemma. How might health researchers learn to include this (and similar) hashtags into their research?

The challenge of understanding the way a phenomenon is talked about including regional dialects and differences is a key hurdle to address. As research participants are not contacted directly, it is important to use methods that allow researchers to find out

what *they don't know they don't know*. The reason for this focus on finding the ways that phenomena are expressed and determining which keywords used so that they become the search terms used to hone in on source material for later analysis. In essence, the better we understand the right data to feed our social media analysis tools, the more reliable the results of the study will be. Computer scientists will recognize this as attempting to avoid the adage, 'garbage in, garbage out', first identified by Charles Babbage (Babbage, 1864).

Initial data exploration

Initial steps of exploration consist of using the terms that a researcher already knows and searching these against the database of social media. At this point, it is not necessary to limit the search to geographical coordinates, unless the database search traversal times are so great that it is impractical to do so or regional issues are of concern. From these initial search queries, review of the returned material can help to grow the initial term searches, and expand the set of related social media data points in an iterative fashion, similar to a snowball sampling technique. This is illustrated in Table 1.

Table 3.1. From original search terms to an expanded list of search terms using a snowball sampling technique. words with '-' between them are searched as two word combinations, words that were found to be ineffective are removed. The 'feel' keyword proved especially useful, while not initially obvious.

Initial terms: Obesity obese fat fatty overeating overeat bloating bloated unhealthy inactive sedentary sugar fructose lethargy lethargic sloth metabolism tired sad chest-pain
Snowballed terms: Weight cdcobesity giving-up fattest insulin insulin-resistance type-2 preservatives fillers #sugarfree #sugartax diet bariatric pizza hamburgers bernaise cola bmi inflammation T2 #fattytuesday treadmill #healthykids #endchildhoodobesity #strong4life exercise nutrition gobble #weightloss craving cholesterol portion habit unmotivated bored alone KFC glutton gluttony kummerspeck sloth diet-pills tired pokemon-go blood-pressure lazy dietary food-pyramid USDA dorito inactive feel-bloated feel-fat feel-overweight feel-obese

By starting with the terminology that a researcher is familiar with, this method recognizes researchers already have domain specific knowledge that can be built on and allows for research reflexivity (England, 1994). It assumes that researchers are partial subjects with prior knowledge that will shape the initial stages of exploration – as advised by Haraway (1988). By beginning with our understanding and assumptions, the research builds upon limitations in an inductive learning process. For research involving multiple researchers, the convergence, or divergence from starting keywords to new lexicons provides a unique opportunity to observe the braided nature of phenomenological discovery.

Computational methods for building source material

Eventually, researchers will reach a saturation point where they are either unable to find further terms and social media postings using the iterative steps outlined above, or lose patience combing through social media postings beyond a few dozen keywords

(Morse, 1995). While saturation in traditional qualitative research curtails data collection when new thematic threads fail to emerge, it presents a difficulty for qualitative geospatial inquiry. Due to the plenum/field/plural nature of spatial analysis we can see that a high volume of data points across space is essential.

To go beyond the point of researcher saturation, computer methods can identify further terms related to a phenomenon, utilizing the search terms already identified. In this article we will explore a few of the methods to further this, including knowledge graphing, synsets, and topic modelling.

Knowledge Graphing

A knowledge graph is a ontological approach to storing knowledge of the world, regarding both information and objects (Suchanek and Weikum, 2013b). Several knowledge graphs exist as works in progress, the most notable being Wikipedia and Google Knowledge Graph. The information contained in knowledge graphs are created from a multitude of epistemological approaches, with many content creators and data sources. The data may be lists of objects, consumer bands, philosophical text, bibliographic details, or similarly any other information that can be categorized and stored in a graph format. Relevant for social media research is the ability to search the topic keywords against a graph and determine if information adjacent, parent, or child graph nodes expand the set of terms used to search social media in a meaningful way (Suchanek and Weikum, 2013a).

Although multiple knowledge graphs exist and many offer free information retrieval, utilizing them often requires knowledge of computer programming to harvest lists of related items from the graph. An example of which is the Google Knowledge Graph, which provides access to its data through an application programming interface (developers.google.com/knowledge-graph). Google has also integrated its knowledge graph into its core search engine functionality, reducing the burden of writing computer programs to traverse the graph, the search string “types of sausages” in an internet browser returns a list of the types of sausages: Biroldo, Ciauscolo, Ciavár, etc. The Google Knowledge Graph is illustrated in Figure 3.4.

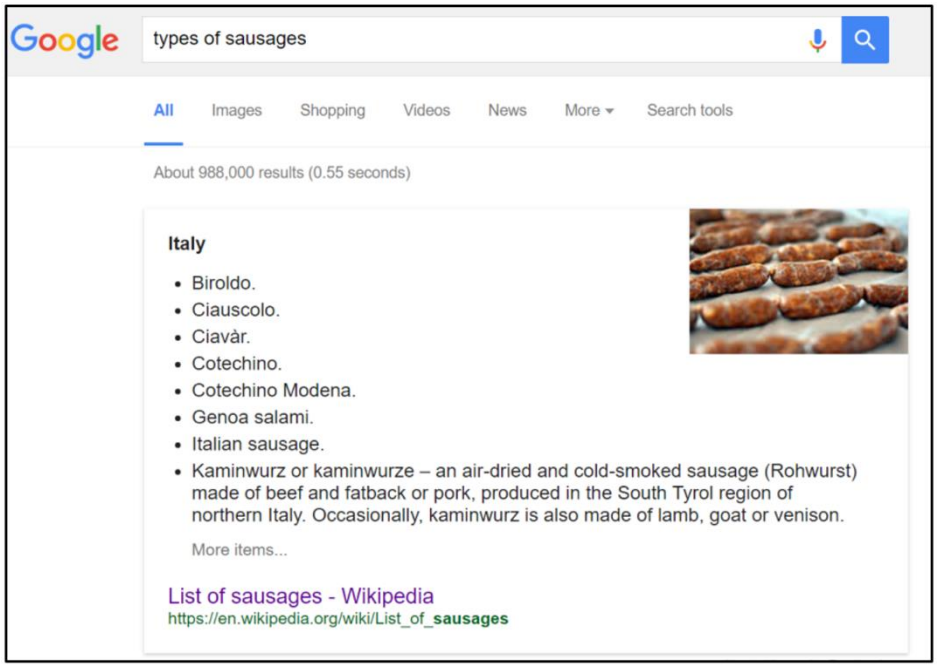


Figure 3.4. Using the Google Knowledge Graph by searching in the browser for 'types of sausages'

Types of sausages may not seem important immediately, however it can help to widen the health geographer's search of healthy or unhealthy foods. For example, Figure 3.5 illustrates that while Biroldo sausage originally had been assumed to be a marker of unhealthy eating (consumption of fatty foods), it instead leads to the discovery of a new hashtag relevant to healthy food relationships '#slowfood'.

Figure 3.5. Using the keyword Biroldo, identified using the Google Knowledge Graph allows the researcher to identify the hashtag #slowfood

Searching for graph items in this way can be effective, but tiresome and the ability to use it effectively can be limited. Using the application programming interface (API) for the Google Knowledge graph, or others, allows for traversal from one item to its parent, child and sibling nodes. While searching for ‘sausage’, the term ‘hotdog’ may emerge and the related types of hotdogs and brands who make them, automatically. Careful application of knowledge graph traversal can vastly increase the amount of search terms available to researchers. Efforts have been made by researchers to implement semi-supervised approaches for automatically finding keywords using the knowledge graph (Huang et al., 2014; Krzywicki et al., 2016; Liu et al., 2013), but an end user software was not identified.

Synsets

Redundancy in language allows communication with nearly infinite nuance. While producing rich discussion, it can mean that social media searches are limited often to only one set of words, and not the many words that have similar meanings. A thesaurus is complementary, but a synset of words is superior as it includes words that are not just equivocal, but are also semantically similar. For example, consider the synset returned from the keyword ‘fatty’ in Figure 3.6, and the linked synset return of ‘roly-poly’, Figure 3.7.

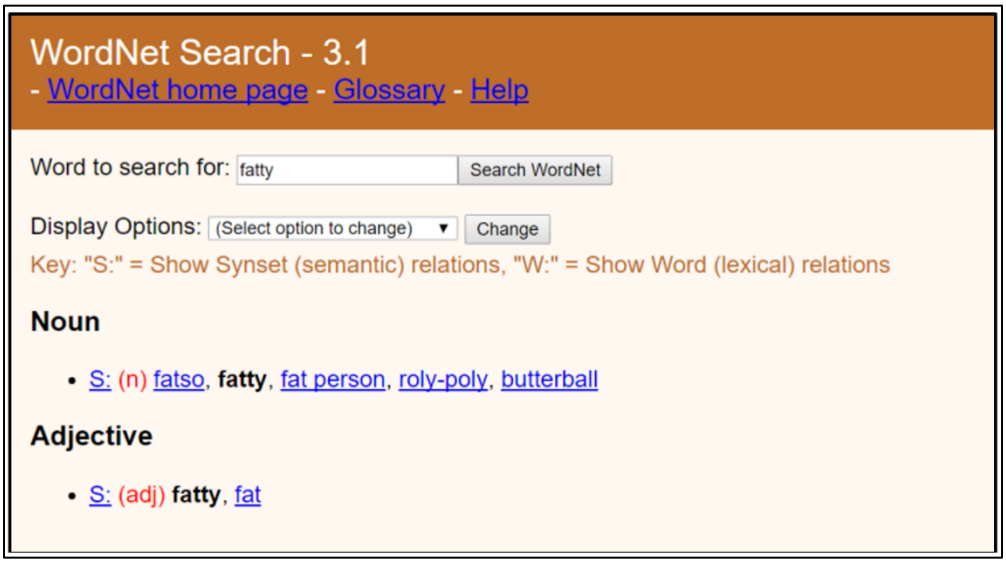


Figure 3.6. Synsets of the word 'fatty' include 'roly-poly' and 'butterball'

Figure 3.7. The synset of the word roly-poly (learned from figure 3.6) include 'dumpy' 'podgy' 'tubby' and 'fatso'

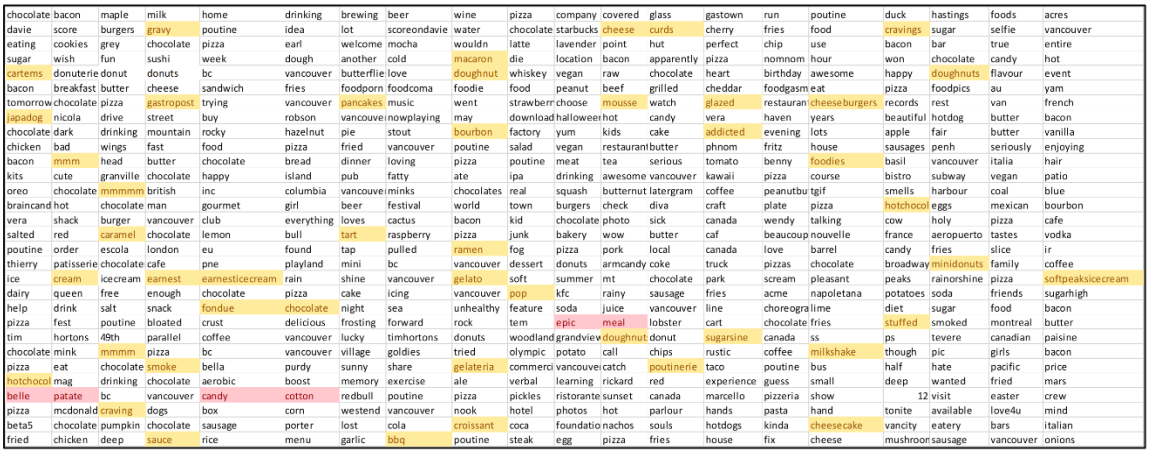
The resource Wordnet (Miller et al., 1990) is a database of synsets, and has been used in a wide array of natural language applications (Caldarola and Rinaldi, 2016; Jayakody, 2016; Rebele et al., 2016). Wordnet provides an API for usage and is integrated into a popular natural language programming codebase NLTK (Bird et al., 2009). Like the knowledge graph, synsets provide a method of exploration that can be done programmatically, and can be used to derive new keywords used to generate source material, such as the emotional tweet in Figure 3.8.

Figure 3.8. Using learned term 'podgy' from wordnet to find new source material from Twitter

Topic Modelling

Topic modelling is a powerful tool that identifies the latent topics present within a set of text documents. It has been successfully adapted to be used on social media postings, and specifically microblogs such as twitter with success (Becker et al., 2011; Ghosh and Guha, 2013; Hong et al., 2012; Wang et al., 2007).

Topic modelling can be used to identify new keywords from social media postings identified from previous explorative steps. While limiting a topic model to only tweets containing already identified terms, it will identify only the topics that these tweets contain. The topics and words generated this way may find word associations that have previously been identifiable. For example, if a topic contained the words, ‘Alcohol, crime, stolen, full, moon’ it might encourage the researcher to see if postings that reference ‘full moon’ are useful in looking at deviant or risky behaviors, regardless of if there is a positive correlation between the moon phases and human activity. Figure 3.9 is an example of a topic model run on unhealthy tweets. While many of the 600 words in the topic model will not be included in the dictionary of terms, some 46 were new and added, expanding the tweets considered in the analysis stage by 59%.



chocolate	bacon	maple	milk	home	drinking	brewing	beer	wine	pizza	company	covered	glass	gastown	run	poutine	duck	hastings	foods	acres
davie	score	burgers	gravy	poutine	idea	lot	scoreondavie	water	chocolate	starbucks	cheese	curds	cherry	fries	food	cravings	sugar	selfie	vancouver
eating	cookies	grey	chocolate	pizza	earl	welcome	mocha	wouldn	latte	lavender	point	hut	perfect	chip	use	bacon	bar	true	entire
sugar	wish	fun	sushi	week	dough	another	cold	macaron	die	location	bacon	apparently	pizza	nomnom	hour	won	chocolate	candy	hot
cartems	donuterie	donut	donuts	bc	vancouver	butterflie	love	doughnut	whiskey	vegan	raw	chocolate	heart	birthday	awesome	happy	doughnuts	flavour	event
bacon	breakfast	butter	cheese	sandwich	fries	foodporn	foodcoma	foodie	food	peanut	beef	grilled	cheddar	foodgasm	eat	pizza	foodpics	au	yam
tomorrow	chocolate	pizza	gastropost	trying	vancouver	pancakes	music	went	strawberr	choose	mousse	watch	glazed	restauran	cheeseburgers	records	rest	van	french
japadog	nicola	drive	street	buy	robson	vancouve	nowplaying	may	download	halloweer	hot	candy	vera	haven	years	beautiful	hotdog	butter	bacon
chocolate	dark	drinking	mountain	rocky	hazelnut	pie	stout	bourbon	factory	yum	kids	cake	addicted	evening	lots	apple	fair	butter	vanilla
chicken	bad	wings	fast	food	pizza	fried	vancouver	poutine	salad	vegan	restaurant	butter	phnom	fritz	house	sausages	penh	seriously	enjoying
bacon	mmm	head	butter	chocolate	bread	dinner	loving	pizza	poutine	meat	tea	serious	tomato	benny	foodies	basil	vancouver	italia	hair
kits	cute	granville	chocolate	happy	island	pub	fatty	ate	ipa	drinking	awesome	vancouver	kawaii	pizza	course	bistro	subway	vegan	patio
oreo	chocolate	mmmmm	british	inc	columbia	vancouve	minks	chocolates	real	squash	butternut	latergram	coffee	peanutbu	tgif	smells	harbour	coal	blue
braincand	hot	chocolate	man	gourmet	girl	beer	festival	world	town	burgers	check	diva	craft	plate	pizza	hotchocol	eggs	mexican	bourbon
vera	shack	burger	vancouver	club	everything	loves	cactus	bacon	kid	chocolate	photo	sick	canada	wendy	talking	cow	holy	pizza	cafe
salted	red	caramel	chocolate	lemon	bull	tart	raspberry	pizza	junk	bakery	wow	butter	caf	beaucoup	nouvelle	france	aeropuerto	tastes	vodka
poutine	order	escola	london	eu	found	tap	pulled	ramen	fog	pizza	pork	local	canada	love	barrel	candy	fries	slice	ir
thierry	patisserie	chocolate	cafe	pne	playland	mini	bc	vancouver	dessert	donuts	armcandy	coke	truck	pizzas	chocolate	broadway	minidonuts	family	coffee
ice	cream	icecream	earnest	earnesticecream	rain	shine	vancouver	gelato	soft	summer	mt	chocolate	park	scream	pleasant	peaks	rainorshine	pizza	softpeaksicecream
dairy	queen	free	enough	chocolate	pizza	cake	icing	vancouver	pop	kfc	rainy	sausage	fries	acme	napoletana	potatoes	soda	friends	sugarhigh
help	drink	salt	snack	fondue	chocolate	night	sea	unhealthy	feature	soda	juice	vancouver	line	choreogra	lime	diet	sugar	food	bacon
pizza	fest	poutine	bloated	crust	delicious	frosting	forward	rock	tem	epic	meal	lobster	cart	chocolate	fries	stuffed	smoked	montreal	butter
tim	hortons	49th	parallel	coffee	vancouver	lucky	timhortons	donuts	woodland	grandview	doughnut	donut	sugarsine	canada	ss	ps	tevere	canadian	paisine
chocolate	mink	mmmm	pizza	bc	vancouver	village	goldies	tried	olympic	potato	call	chips	rustic	coffee	milkshake	though	pic	girls	bacon
pizza	eat	chocolate	smoke	bella	purdy	sunny	share	gelateria	commerci	vancouve	catch	poutinerie	taco	poutine	bus	half	hate	pacific	price
hotchocol	mag	drinking	chocolate	aerobic	boost	memory	exercise	ale	verbal	learning	rickard	red	experience	guess	small	deep	wanted	fried	mars
belle	patate	bc	vancouver	candy	cotton	redbull	poutine	pizza	pickles	ristorante	sunset	canada	marcello	pizzeria	show	12	visit	easter	crew
pizza	mcdonald	craving	dogs	box	corn	westend	vancouver	nook	hotel	photos	hot	parlour	hands	pasta	hand	tonite	available	love4u	mind
beta5	chocolate	pumpkin	chocolate	sausage	porter	lost	cola	croissant	coca	foundatio	nachos	souls	hotdogs	kinda	cheesecake	vancity	eatery	bars	italian
fried	chicken	deep	sauce	rice	menu	garlic	bbq	poutine	steak	egg	pizza	fries	house	fix	cheese	mushroon	sausage	vancouver	onions

Figure 3.9. Topic model results on a tweet dataset created using snowballed terms and knowledge graph results of type of fast food, candy bars and soda brands. Words in yellow and red shading will be included as additional search terms (red are terms that go together such as "epic meal" or "cotton candy"). White terms are those that are already incorporated, or are not useful, such as generic place-names. Original tweet corpus before adding these terms: 10,507 tweets. After incorporating these 46 (plus minor variations) terms: 17,690 tweets

Creating the Corpus

Once all of the above methods have been explored, the researcher can be relatively confident that the list of keywords generated will yield a comprehensive set of social media postings that relate to the phenomena of interest when used as search parameters. In natural language processing, the input set of items for analysis is denoted the 'corpus'.

In this example we have used the social networking service Twitter, however it should be noted that the same methods can be used to identify keywords for numerous other social media sites, such as Facebook or even the comments on activities posted to physical activity sharing site Strava. In addition, the keywords generated here have been focused in only one direction, that of unhealthy eating or relating to how internet users talk about obesity. However, it would be expected that as researchers discover keywords, they may segment them according to multiple themes or research directions. For example, a set of words may be focused on unhealthy eating, but themes of healthy eating or regular exercise may also emerge at the same time. The keywords found in related themes may be useful in comparative analyses as illustrated in Table 2.

Table 3.2. Segmented keywords by emergent themes

Obesity	Unhealthy eating	Emotional Eating	Healthy Eating
sedentary, cholesterol, obesity, obese, insulin -insulin-resistance, type-2, cdcobesity, T2, inflammation, food-pyramid #endchildhoodobesity, #sugartax, blood-pressure, bariatric	Fat, fatty, fattest, overeating, overeat, fructose, sugar, roly-poly, #foodcoma, cola, bernaise, #fattytuesday, bloated, KFC, dorito, pizza,	Feel-bloated, feel-fat, feel-overweight, feel-obese, kummerspeck, feel-lethargic, unmotivated	Boroldo, #strong4life, #wieghtloss, portion-control, nutrition

Place is also an important element of corpus formation, bounding boxes can be placed around any of the previous queries, and is usually necessary at the point of final corpus formulation. By placing a geographic boundary to the corpus, the researchers can attempt to ensure that posts that are used in analysis are sensitive to the effects of place.

3.3.3. Phase Three: Analysis

Once a corpus has been formed, analysis may proceed. Multiple methods exist to analyze the data present and it is up to the researcher to identify the most relevant to apply to their investigation. While is it possible to apply traditional methods of analysis to the corpus, the burden of data often makes it impossible to so (Jung, 2015). Instead, computerized methods may be most useful in these situations. In this paper we illustrate four analytic methods that have been used by geographic researchers thus far; keyword matching and traditional spatial analysis, and natural language processing using topic analysis and sentiment modelling. Each method offers different types of insight into the phenomena studied, and not all may be appropriate for a particular study.

Keyword Matching

Perhaps the most direct method of analysis is keyword matching, using search terms against a corpus of data and organizing the results into thematic categories. For example, the Zook and Poorthuis (2014) have used this technique to study the geographies of beer, Sakaki (2010) identified earthquake victims, and Jung (2015) investigated US election patterns.

Once data had been identified and categorized, standard spatial analysis can be done. Zook and Poorthuis (2014) combined their categorized tweets with odds ratio, showing the relative likelihood a categorized tweet occur at a particular location, Stephens (2013b) used kernel density measures to illustrate hotspots of hate speech (see Figure 3.10), and Kent and Capello (2013) tracked hotspots of wildfire movement.

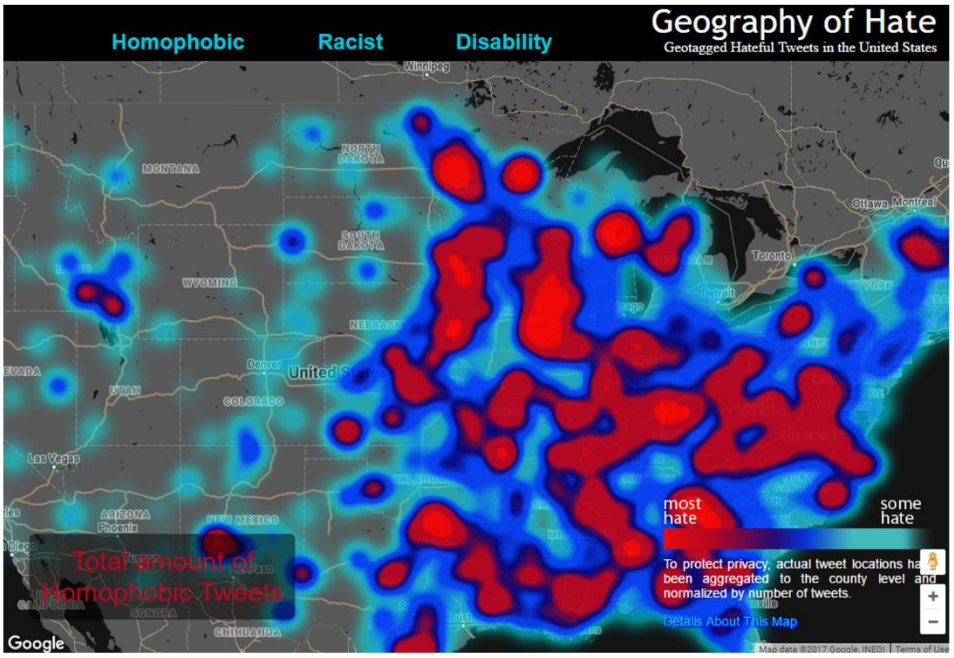


Figure 3.10. Geographies of Hate. Using density measures and keywords, 'Fag, Dyke, Homo, Queer' in Twitter data (http://users.humboldt.edu/mstephens/hate/hate_map.html).

Keyword matching is a useful method for identifying and categorizing social postings, and is easily implementable without requiring special tools. It requires the researcher to create and curate categories however, a potentially tedious process (Jung, 2015). Categorizing data using keywords ignores the context of information, increasing the difficulty of identifying misleading practices such as sarcasm. For example, consider

a keyword map of “foodcoma” vs “kale”, two words that may be categorized into unhealthy and healthy as illustrated in Figure 3.11. The context they are used in can radically change the meaning of the posting. Keyword matching can illustrate the frequency of keyword occurrence, however unless the posts are vetted, they are prone to misinterpretation.

Figure 3.11. Examples of keyword matching for 'food coma' and 'kale' using twitter. These examples illustrate the importance of the context in which a search term is used. Top left illustrates a sincere use of #foodcoma, however it uses it in antithesis. Top right uses the healthy search term ‘Kale’ but in a phenomena known as the ‘humble-brag’. Bottom right uses the keyword kale as allegory, and bottom left uses the term ‘kale’ and its healthy supposition for humour. Perhaps ‘kale’ is too popular a keyword to be helpful.

Using keywords to measure spatial autocorrelation is particularly interesting when overlaying multiple map layers generated from social media on specific topics. Using the example of healthy and unhealthy tweets spatial autocorrelation is high when compared to fast food locations. However, spatial and a-spatial processes can influence how social media posts are located. It is common to have high tweet density at transportation hubs such as bus stations, subway stops, and train stations. At the same time, fast food locations co-occur with transit hubs, due to the high amount of foot traffic and pauses in daily activities associated with waiting for transit. An example can be seen in Figure 3.12. This creates potentially spurious correlations, identifying patterns indicative of healthy and unhealthy conversations located at fast food fast-food restaurants-- or simply that social media usage increases at transit hubs. Separating

study of the technology from how a phenomena is expressed is a delicate and necessary step for social media research (Li et al., 2013).

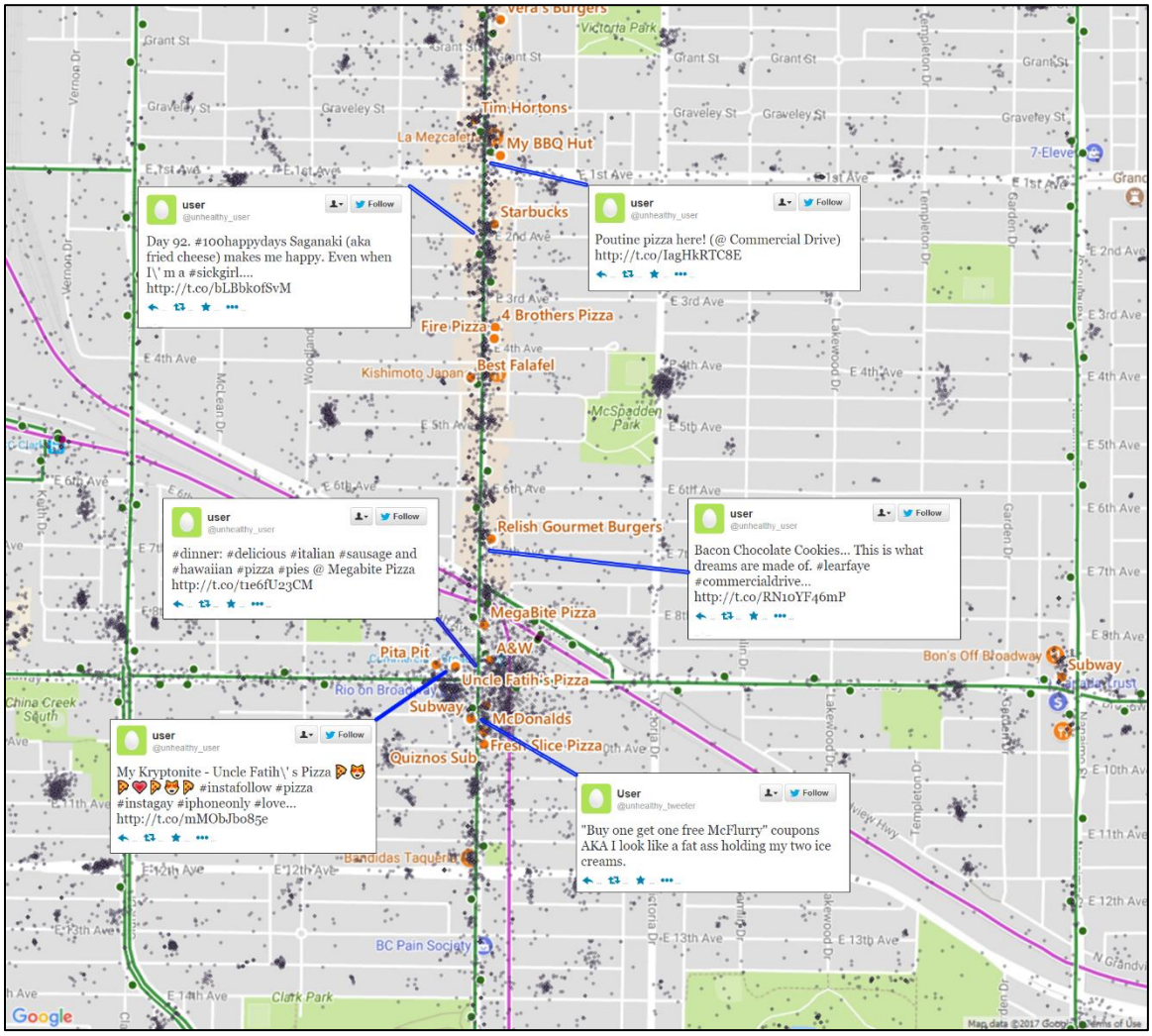


Figure 3.12. Tweet densities, transit, and fast food locations near the Commercial-Broadway transit exchange in Vancouver, Canada. Tweets are represented in black, transit in pink and green, and selected fast food locations in orange. Unhealthy tweets co-occur with the fast food restaurants, but it may also be because fast food locations happen to be near transit hubs, where users eat while waiting for transit.

Utilizing keywords with spatial analysis tools

Regardless of the potential for disingenuous speech and spurious correlation, methods that are already used by spatial scientists can be productive for social media analysis. The most common tools that have been used by geographers are kernel density functions (Li et al., 2013; Lichman and Smyth, 2014; Tsou et al., 2013), and

spatial clustering (Frias-Martinez et al., 2012; Mitchell et al., 2013; Steiger, Westerholt, et al., 2015). These methods have been coupled with keyword matching across a multitude of domains, from the geography of hate speech, to the distribution of beer and wildfire mapping (Kent and Capello, 2013; Shelton et al., 2015; Zook et al., 2015; Zook and Poorthuis, 2014).

Beyond density functions, Crampton et. al. (2013) studied the practice of retweets, where twitter users reproduce the tweets of influential users. An example of re-tweeting, originally provided by Zook (2013) is seen in Figure 3.13. Similarly, Stephens and Poorhuis (Stephens and Poorthuis, 2015) analyze the spatial elements of twitter's social networks, demonstrating networks are quite sensitive to distance.

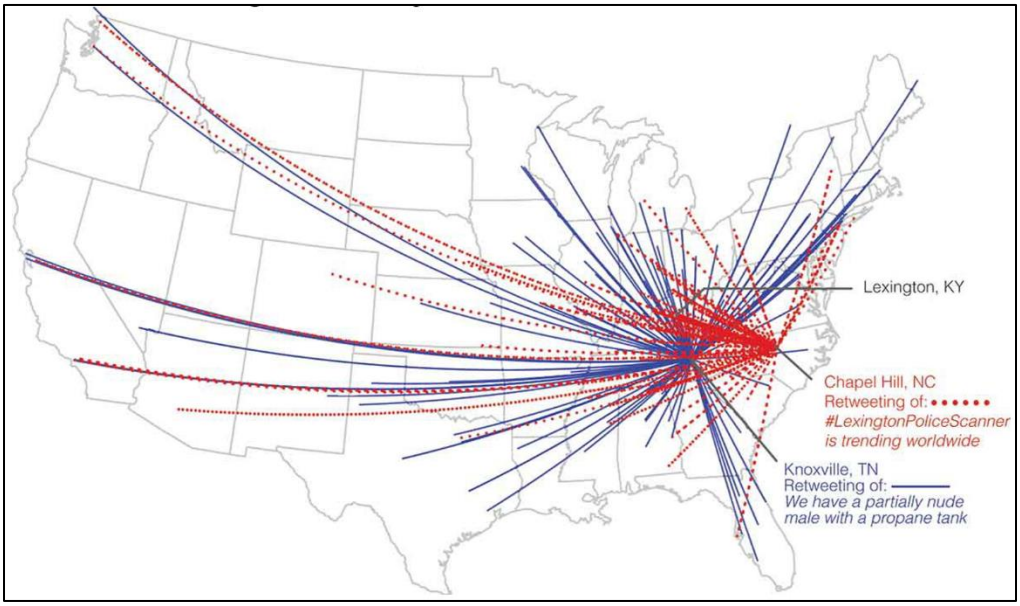


Figure 3.13. A geographic network of retweets from Crampton (2003).

Spatio-temporal relationships have also been a key concern in social media research (Steiger, de Albuquerque, et al., 2015). Social media posts contain a timestamp that can be used to track hashtags, keywords, geospatial locations, and relationships over time. Memes and social movements can be seen gaining traction over countries and worldwide (Ferrara et al., 2013; Kamath et al., 2013). Time analysis are also often used for natural disasters as well, like earthquakes(Crooks et al., 2013; Sakaki et al., 2010) and floods (Herfort et al., 2014). #earthquake is a tag that is often used when earthquakes happen and as Crooks (2013) demonstrated, the extent of

persons affected by a earthquake can be analyzed using social media, identifying the aftershocks beyond the initial event.

Natural Language Processing

Computer Science has much to contribute the field of social media analysis, but perhaps the most useful and powerful is the sub-discipline of natural language processing (NLP). The aim of NLP is to try to understand the meaning of text in its written context (Allen, 2003). NLP is important beyond social media analysis, and is greatly impacting the ways that humans interact with machines. As recently as 2016, a combination of neural networks and NLP techniques led to the first passing of the Turing Test by a chat bot (Shah et al., 2016). While the use of NLP and neural networks has the potential to produce computers capable of impersonating humans, they are only as capable as the data they are trained upon. The artificial intelligence chatbot produced in by Microsoft named *Tay* learned hate speech from Twitter, leading to its shutdown within 24 hours of coming alive (Neff and Nagy, 2016; Vincent, 2016). While natural language methods have large potential, it too can suffer from data quality issues.

Within geographic applications, two uses of NLP are popular. Topic modelling and sentiment analysis. Topic modelling is used to determine the topics present in a body of text, or among bodies of text.

Topic Modelling

Topic modelling consists of several different approaches, but by far the most popular basis for all variants is Latent Dirichlet Allocation (LDA). LDA is a Bayesian method that looks for words that commonly associate with one another over a large collection of text documents (Blei et al., 2003). Following from Blei et al. (2003) many researchers have sought to refine the approach, including geographers (Ghosh and Guha, 2013; Gore et al., 2015; Wang et al., 2007). Topic modelling provides researchers with a way of understanding the dominant topics of text, and in a geographic context, it can be a method for gaining an understanding of what is important in a place. Topic models can be used comparatively across neighborhood areas to see how the context of place and space intersect. Using a twitter database, Martin and Schuurman (2017) use topic modelling and visualization to show how different topics appear in neighborhoods of Vancouver, Canada as shown in Figure 3.14.

Figure 3.14. Topic models in Vancouver, BC from Martin and Schuurman (2017). In the Mount Pleasant neighbourhood containing tourist locations such Granville Island and the False Creek Ferries service are topics, while in the nearby Olympic Village area where several breweries are located topics of Beer and Jobs are prevalent.

Topics models can also be focused on specific phenomena. If a model only makes use of the social media postings that are generated from topic keywords (phase 2), then resulting topics are from the subject matter. The phenomena specific topics models can be bounded areas, similar to Figure 3.14, and differences among neighboring areas can be identified.

Topic models, however, are not without their considerations and troubles when used over geography. Uneven data density has an impact on the reliability of the model. In the case of Vancouver, Canada, data density varies wildly over the city. The number of posts in highly trafficked areas, such as the downtown core are more than ten times those of areas that are dominated by family homes. These neighborhoods can be clearly seen in Figure 3.15.

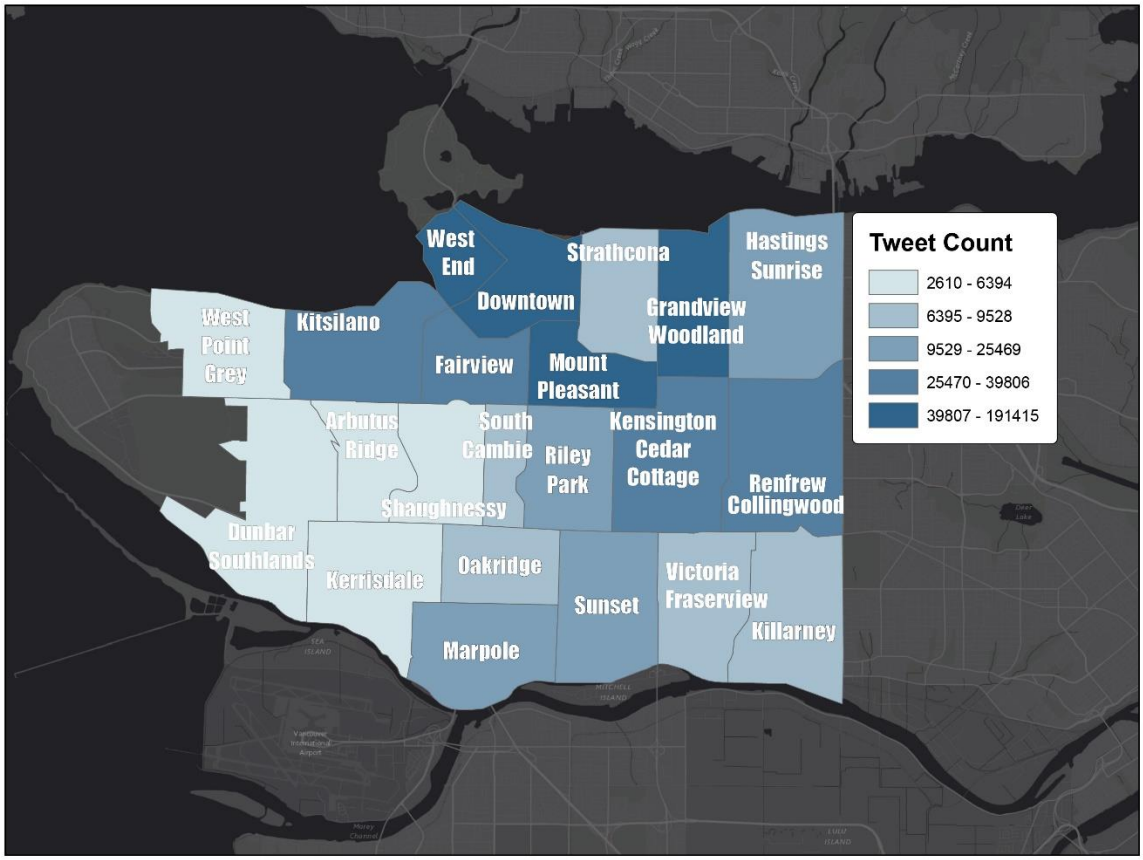


Figure 3.15. Tweet counts of Vancouver from an 8-month period of data collection from Martin and Schuurman (2017).

Topic models also have several controls that affect their output. The model relies on the researcher knowing the correct number of topics to look for and the number of words to fit into those topics. The process of understanding the optimal settings is iterative in nature and requires trial and error, although some topics models are capable of an optimization (Lim et al., 2016; Teh and Jordan, 2010). Topic models are also sensitive to the tradeoff between computing power and time dedicated to running the model (Sukhija et al., 2016). A model that is given time to iterate over data will increase the confidence of topic associations, however it also increases the amount of time the model requires to run. Using distributed computing clusters, is one solution to this problem, but requires access to such facilities (Sukhija et al., 2016). Stop words are also important assumptions that topic models need to consider. When models natural language models are carried out, it is common to remove words from analysis that are unlikely to provide useful output such as ‘a’, ‘it’, ‘the’, or ‘11’. The stops words list may include hundreds of words and their permutations. While some words such as ‘fuck’ or

‘shit’ may be included in the stop words, the also may be important indications of emotion or related to particular sub-cultures (Balasuriya et al., 2016).

Sentiment Analysis

Sentiment modelling is a branch of NLP that seeks to label the emotion of text. While some models are capable of determining specific emotions on long form text, social media implementations are often limited to determining if text is positive or negative in feeling. Part of the problem with sentiment modelling is that it requires a large database of labelled text for training. Often the training information is based in movie reviewers, or restaurant reviews, where each review is accompanied with ratings. The Stanford sentiment analysis technique (Socher et al., 2013) uses Rotten Tomatoes, a movie rating website (www.rottentomatoes.com), and the model created by Agarwal et al. (2011) uses newswire data. Sentiment modelling faces additional challenges when used with social media data, as the short nature of messages do not provide much contextual and emotional cues. However systems exist that have shown promising results (Chikersal et al., 2015).

Geographers have however, attempted to make use of sentiment models, despite the model shortcomings. Specific to the example of healthy and unhealthy eating, Widener and Li (2014) provide a system for marking healthy and unhealthy tweets as positive and negative emotions, and link negative food results with the presence of food deserts. Mitchel et al. (2013) use twitter data to determine a realtime map of happiness in the United States and link this data with correlations to public health issues, such as obesity. Hao et. al. (2011) have used sentiment analysis on broader themes such as popular films, producing a system that focuses on visualization of sentiment in both time and place.

Sentiment analysis continues to be an active area of NLP research, and new tools continue to emerge. No tool exists yet that makes sentiment analysis easily accessible to non-programmers to explore geographic datasets, however some systems are open source and available. The Natural Language Toolkit, a python programming code library has a sentiment analysis module and is perhaps the most accessible entry point (Jongeling et al., 2015).

3.3.4. Stage four: Representation

Reporting the results of social media analysis is an emergent field, and often changes based on the type of analysis that has been conducted, and examples of these output formats has been introduced with each method above. Often however, the challenge is to convey the results of textual content on map surfaces without overdrawing the content of the analysis or using ancillary tables. The advent of interactive displays and the ability to share them using the internet has greatly increased the capacity of maps to display complex data, providing users with the ability to augment visualizations quickly and easily. Interactivity facilitates intuitive data overlay and enables users to continue to focus on the geography of interest. However, the ability to click on a location and view ancillary information on it is not a perfect solution, as it is not conducive to the constant comparison that cartography enables.

Social media is difficult to express cartographically because of the reductive nature of maps. How can sentiment analysis reproduce the raw emotions expressed in tweets pertaining to health crises into a binary of positive-negative sentiment cartographically? Doing so limits understandings of lived experiences. This problem is not limited to sentiment analysis but is generally applicable to any social media analysis. Big data analysis, by definition, reduces information to digestible quanta, and as unsupervised methods rise in popularity the choices that are made throughout are opaque. To combat this, researchers presenting results must not only present the cartographic, or graphic results of a model of analysis output, but also include the postings that typify the analysis results, just as qualitative thematic research presents analysis and key quotes, side-by-side.

While many methods are emerging that could be explored, it would be difficult to discuss more implementations in this paper without lengthy discussion of bias, affect, and ways that cartographic elements inform design and function of maps. In recognition of this, we choose to leave those explorations to the reader, as every implementation is different and requires details and exhaustive effort, which is outside the scope of this article. This initial wayposting contains all the information that is necessary to begin analysis of big data within a GIS or human geography framework. We encourage readers to review the figures and references cited in this article for further exploration.

3.4. **Discussion**

> "Geographers need to grasp the opportunities whilst at the same time tackling the challenges, ameliorating the risks and thinking critically about big data as well as conducting big data studies. Failing to do so could be quite costly as the discipline gets left behind as others leverage insights from the growing data deluge" (Kitchin, 2013)

> "Ambivalence towards the disrupted unities mediated by high-tech culture requires not sorting consciousness into categories of 'clear-sighted critique grounding in a solid political epistemology' versus 'manipulated false consciousness', but subtle understanding of emerging pleasures, experiences, and powers with serious potential for changing the rules of the game." (Haraway, A Cyborg Manifesto, pp.172-3, 1991)

How literate, technologically savvy, do human geographers need to be? Is there a STEMming of competency as computational methods become more necessary to engage in knowledge production? While the GIS-ification of qualitative methods was more of a niche field, the advent of social big data is becoming an increasingly core competency, or standard set of methods. It begs the question of whether we will see more tools such as *Mallet* (McCallum, 2013) that seek to automate methods and build graphical user interfaces. However, as these methods become encoded into software that enable wider access to them, the process and key parameters will be black-boxed in the process described by Latour in Science in Action (Latour, 1987).

At the same time, sociological and specifically geospatial thinkers must continue to engage and be critical of the push to more standardized tools precisely because so much of the output is shaped by the partiality of the programmer, researcher, and participant, as Code-Space teaches us (Kitchin and Dodge, 2011). By black-boxing this process we lose the ability to dig deeply into the how and why we are witness to the results brought out. Black-boxing the process of social media analysis has consequences at each stage of the methodology. While the process of black-boxing GIS methods on quantitative data was relatively easy, lived experiences and phenomenological data do not lend themselves as easily to unsupervised analysis methods.

Building a data source for downstream analysis integrates and obfuscates a wide variety of assumptions. The largest of these is the assumption that the data is a representative sample. While social media usage of online adults in the United States is high among Facebook (70%) Twitter (23%) and Instagram (18%), the segmentation of the population that uses these technologies is not even (Perrin, 2015). Race, gender, and age all influence adoption of social media and differently for each technology (Greenwood et al., 2016). Structural forces also create digital divides, and determine how the technologies are implemented and ultimately used. While persons who have the greatest access to the internet have the ability to use it in the privacy of their homes and on premium devices, those of less economic advantage suffer less private means of access (Dixon et al., 2014; Hargittai, 2010; Wangberg et al., 2008). Libraries have become key points of access for the urban poor, and as public spaces they will influence the ways that technology is interacted with. Kitchin and Dodge (2011) offer the formation of a *code-space*, a framework for integrating the role of place in our interactions with technology.

The most useful data for qualitative analysis convey deep emotional connection, but as spaces/places become more public, the personal and private process of producing reflexive thought is eroded. This is magnified if the social media used is a public posting forum, such as Twitter. Other social media provide opportunities for private dialog, but are difficult for researchers to access large data samples from. Gaining access to private or communications on social media is nearly as difficult, from both the technological (credential exchange server) and interpersonal (participant recruitment, building trust, or ethical clearance) positions.

Ethical considerations become of increasing importance as we become more engaged in the private lives of our subjects. Obtaining the access credentials to social media account (OAUTH2) make it easy to not only to attain the information related to the study at hand, but information that is well beyond the scope of the research project. Indeed, researchers need vast troves of data to employ big data algorithms, being that many algorithms produce rigor through and confidence by integrating many iterations of random samples throughout the dataset. Perhaps a new understanding of research ethics for big data is needed, and geographers have an important part to play in its creation (Boyd and Crawford, 2012). While it is entirely possible to study obesity through

Twitter data (Gore et al., 2015), it is unlikely the users creating the information used in the study gave consent for it to be included.

Exploration of data (stage 2) of data is perhaps the most embodied and positional element of the big data methodology presented in this article. The interpretation of what is included in the cohort of terminology and what is excluded or left out is personal, and/or interpersonal if team members are included. If done poorly, this stage of the research pre-shapes all study findings consequently, the study will say more about the *researchers* than the participants. Just as we have learned how to lie with maps (Monmonier, 1996), we may lie with social media. The volume of information makes it easy to prove our own bias regardless of what our data may indicate. The purpose of this data stage is to hone the data used in the study down to that which is most salient, however it can also exclude data which does not match the epistemic orientation of the researcher, either deliberately or otherwise.

The computational methods suggested at this stage (2) are susceptible to epistemic interference, too. The Google knowledge graph can be used to access nearby nodes of information and knowledge, however the content and quality of this information will be subject to the positionality of those that produced it. This information, that of the producer, is hidden and obfuscated from the researcher. For example, in our search or 'types of sausages' (Figure 3.4) we may only find those that match Italian sausages, as those were the most common in a Western database. This results in a Western data bias. Using the semantic similarity approach of Wordnet, even algorithmically, presents further challenges of recursively finding words that are more and more the same, meaning that it may over emphasize a particular lexicon or epistemic bias of knowing. For example, multiple versions of Wordnet exist, such as EuroWordnet (Alonge et al., 1998) and Medical Wordnet (Smith and Fellbaum, 2004). Using topic analysis methods from a partial dictionary can be affected by geographical bias, as the popularity of a social media is uneven. Using Twitter as an example, we find the highest density of posts on the pacific and eastern coasts, and urban centers (Li et al., 2013). By geographically influencing our dictionary of inclusion, we may marginalize or silence entire populations or regions. The overproduction of knowledge at the metropole may be well understood at the stage of analysis, but perhaps it is even more important to consider its role in shaping our research at the stage of inclusion and exclusion in to the cohort of information integrated in research (Crawford et al., 2014; Kitchin, 2014).

Analysis of social media (stage 3) is an active and rapidly changing mix of methods (Marres and Weltevrede, 2013). As an emerging science there are so many different methods that can be used, each with numerous iterations of differing abilities and suitability to data. While mature disciplines have standard methods that are easy to explain and explore for non-sub-discipline experts, these methods evolve quickly in big data studies. Within topic modelling alone there is such variety of methods that it is not practical for a researcher know enough of each method, and its parameters to be critical of the methods used for every study. While GIScientists are well aware of the appropriateness of Kriging vs Co-Kriging for temperature data, they may not be fluent enough in natural language processing to know the difference between Latent Dirichlet Allocation (Blei et al., 2003), vs Labeled-Latent Dirichlet Allocation (Daniel et al., 2009) vs Latent Semantic Analysis (Landauer et al., 2006). The methods have not coalesced into a standard set of tools that can be discussed coherently among scholars in tangentially related disciplines. This leads to difficultly in knowing what exactly to be critical of in review and publishing.

While the process of utilizing both quantified and qualitative methods and viewpoints is fraught with challenges, this project is an initial roadmap to the mixing of both the positivist and post-positivist elements of computer science and ethnographic enquiry (Kitchin, 2014). Utilizing data science and critical-reflexive methodologies this project aims to signpost the emergence of a platform to mixed methods of quantitative and qualitative GIS. In its essence, this mélange integrates the ability of quantitative methods to dig into the qualitative quality of social media data. Elwood and DeLyser (2010) investigated the difficulties of integrating multiple epistemologies into singular research projects, while recognizing that as researchers engage with reflexive inquiry, they are better equipped to understand their own epistemic and ontological bias. The ability to integrate data and methods that originate from a web of differing world-views makes big data research difficult, and renews a call for ontology-based metadata (Schuurman and Leszczynski, 2006). However, it is our view that while computer science methods for natural language processing are produced from a positivist viewpoint, it is possible to build in a socio-cultural, feminist, Marxist, or other conclusions from the interpretation of the results and by linking the results back to primary sources – indeed a mixed methods approach is possible for qualitative big data research. Indeed, this is the moment at which these possibilities have emerged – especially as there is not

yet an established repertoire of approaches and methods established among GIScientists or human geographers.

The contribution of this paper is to distill these methods at this point in time so that they are accessible to a range of geographers and – in the process – contribute to opening the black box in which so much big data is analyzed.

Chapter 4.

Social Spatial: A Qualitative GIS for Social Big Data Investigations

This paper has been submitted to the International Journal of Geographic Information Science

4.1. Abstract

In this article we present Social Spatial, a qualitative GIS for social media and big data research. This software enables GIScience researchers to build social media corpus that reflects a phenomenon being researched and implement methods that analyse that corpus. Natural Language processing methods are integrated into Social Spatial, and the code framework has been designed to allow for easy integration of further algorithms. The software builds upon the knowledge of the researcher to identify new ways that phenomena are expressed and see where these posts are geospatially. The software was released open-source with thorough internal documentation to a collaborative code repository to encourage others to contribute and make the programming as transparent as possible. Extensive use of settings files was employed to expose parameters – word lists, model coefficients, and stop words to enhance transparency in qualitative social media research methods and the code/spaces they were enacted within without increasing the burden of research documentation.

Keywords

Qualitative GIS, Critical GIS, Big Data, Social Media

4.2. Introduction

There are an increasing abundance of methods originating from computer science that allow GIScientists to push the boundaries of our capacity to capture and integrate big data. They will also potentially help us to process and understand social

phenomena. By integrating computing science approaches to big data integration and analysis, we may finally realize the promises of a qualitative GIS. A GIS that is conversant with qualitative data and that reflect the persons who generate the data themselves in time, space, and place.

GIScience is poised to undertake research and produce tools that better reflect the physical and human landscape than perhaps any other discipline, yet there are concerns that if it does not do so soon, it will suffer the fate of behind left behind (Kitchin, 2013), an echo of what digital humanities has already expressed (Lane, 2017). This article presents software (Figure 4.1) that builds upon the work of GIScientists in the field of big data, focusing on the workflows and natural language methods that may provide the most unique opportunities for GIScience in the last decade. The software is open-source and actively under development. Any reader may download the program, contribute, and produce research using it. It is our belief that this software will contribute a space for further development of algorithmic intelligence for big data geospatial research tools.

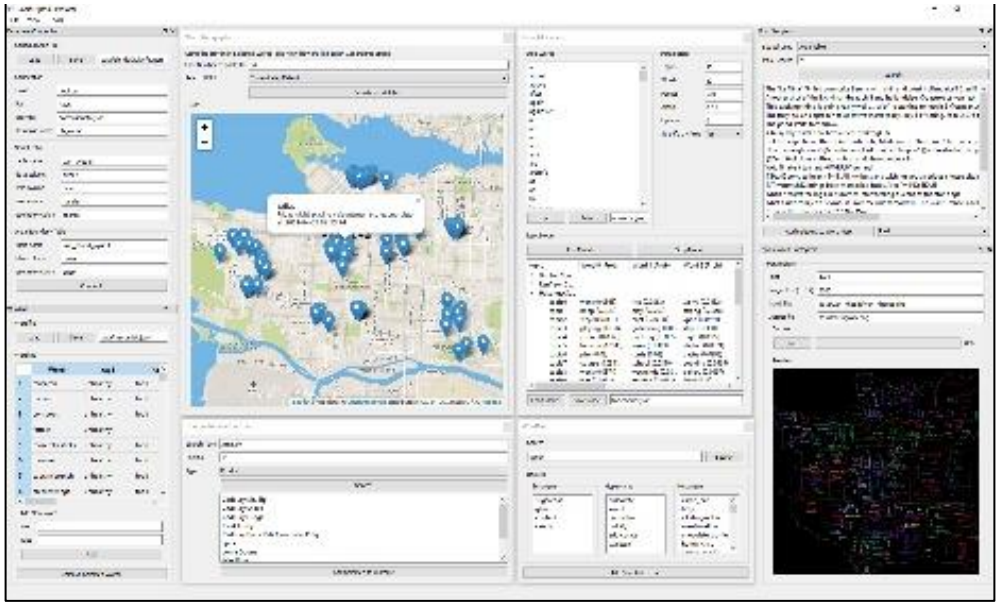


Figure 4.1. Social Spatial interface displaying various modules.

4.3. **Background**

The last five years has seen a burgeoning of big data research in the geographical sciences. Geospatial Web research (Crampton, 2009; Haklay et al., 2008)

and participatory geospatial data research (Crampton et al., 2013; Elwood and Leszczynski, 2011; Johnson et al., 2015) led GIScience directly into the path of social big data. The automatic registration of location metadata through mobile phones has created a deluge of spatial data and new methods for interrogating this information (Lee and Kang, 2015; Yeager and Steiger, 2013). At the same time, critical and qualitative GIS thinkers have followed these developments and contributed towards a better understanding of how these new data mediums might be used (Elwood, 2008; Elwood et al., 2013; Harvey et al., 2006). Perhaps just ahead of its time, Qualitative GIS (2009) by Cope and Elwood exemplifies how critical-qualitative researchers interacted with humanistic data sources, pre-twitter and other open social media data streams. In the early 2000s, scholars explained the advantages (Pavlovskaya, 2009; Schuurman and Leszczynski, 2006) and opportunities (Jung, 2007; Kwan and Ding, 2008) that lay in qualitative analysis with GIS, but in the absence of simple and automated data integration, the burden of data generation and analysis were prohibitive.

As new geospatial social data sources have become available to GIScience researchers, Twitter in particular, more in-depth qualitative research has surfaced. GIScience scholars have used the tools that are already integrated in standard GIS environments to great effect (Crooks et al., 2013; Stephens, 2013b; Zook and Poorthuis, 2014). This work can be grouped by its analytic technique and by its ability to identify the key informants in the study.

Jung (2015) demonstrates the challenge of integrating qualitative data in the face of the avalanche of information that Twitter provides. Jung's work followed the protocols of traditional participant interviews by reading and thematically cataloguing individual tweets relating to the 2008 presidential election. Using a more hands-off approach, Stephens (2013b) generated maps of hate using data from Twitter, thematically displaying words of hate-speech to heat maps (density) in a web browser. This work demonstrates how rapid, simple, and effective qualitative social data can be when expressed through a cartographic medium. Zook (2014) et al. employ the use of odds-ratio analysis of keyworded social media in their investigations of the patterns of various phenomena, including those of American drinking habits[2]. This method of using keyword matching as data selection and spatial analysis of the key data points has been used by

[2] More of their work is accessible at www.floatingsheep.com

GIScience scholars in many applications, including disaster response (Crooks et al., 2013; Sakaki et al., 2010) and obesity (Gore et al., 2015). Disciplines beyond geography and GIScience have engaged in qualitative analysis of geographic social media, too. In particular, computer science has contributed greatly (Cody et al., 2015; Frank et al., 2013; Liu et al., 2015; Mitchell et al., 2013; Wang et al., 2007)

Computer science has a number of sub-disciplines that are actively engaging with social media data, and the geographic metadata that most often accompanies it. For example location recommendation systems determine where social media users may like to go (Liu et al., 2015), predictive analytics have determined where flu epidemics may be emergent (Padmanabhan et al., 2014), and social graphing has predicted networks of relationships that exist in different locations (Backstrom et al., 2010; Liben-Nowell and Kleinberg, 2007). All of these sub-disciplines of computer science challenge the traditional boundaries of GIScience, expanding the methodological possibilities available to GIScientists. For qualitative GIScience and phenomenological GIScience researchers however, one computer science discipline stands out as particularly useful, natural language processing (Allen, 2003).

Natural language processing (NLP) is particularly useful because it is the science of understanding the context of what is being said in a piece of text. Qualitative thinkers have identified this as the most important element of integrating social media analysis in qualitative GIScience (Kwan, 2016). It is imperative to not only find the patterns that exist within the data, but also to understand the situation they are produced within and actors that produce them, which is difficult to determine without contacting social media users. Natural language processing integrates context by identifying elements of text and speech that provides clues about what is being said (Allen, 2003). Within the field, two important methodological techniques stand out as being pertinent to qualitative GIScience, Topic Modelling and Sentiment Analysis.

Topic modelling is the study of determining the latent topics that can be identified within a collection of textual documents. Blei et al. (2003) introduced latent dirichlet allocation, an unsupervised method using a bayesian approach. While Latent dirichlet was not the first method (see latent semantic analysis (Landauer et al., 2006), it quickly became a popular method for analysis of social media, and numerous improvements and augmentations of the method have been published in recent years (Daniel et al.,

2009; Liu et al., 2015; Mei et al., 2008). This method has not gone unnoticed by geographers and GIScientists. Gore et al. (2015) utilized topic modelling to illustrate the trends and topics that are evident in the habits of American fast food eating, and compared these to trends in the obesity epidemic. Gao et al. (2017) utilized topic modelling to in conjunction with foursquare check-in data to determine the topics associated with places of interest and delineate regions those topics occupy. Martin and Schuurman (2017) developed a technique for automating LDA as an area based measure and produced an algorithm that visualizes this data cartographically.

Sentiment analysis is a method that seeks to automatically determine the emotions present in a set of texts (Pang and Lee, 2008). It accomplishes this by using a training set of data that has been pre-tagged with emotions, usually reviews of movies or food (Haughton et al., 2015; Socher et al., 2013). The tagged data is critically important as not only be need it be high quality (e.g. verifiable), but also topically relevant (Korayem et al., 2016; Lu et al., 2016). Even still, most sentiment models are only capable of identifying a set of text as a binary of positive or negative emotion (Cody et al., 2015; Ohmura et al., 2014). Sentiment analysis has been used by geographers, correlating urban infrastructure with positive sentiment (Rybarczyk and Melis, 2017) and that sentiment correlates strongly with wealth and poverty (Frank et al., 2013).

As recent history has shown, there has been an increasing coupling of contemporary computer science methods with GIScience. This coupling, along with the availability of geographic qualitative data, increases the capacity of qualitative GIScience to move beyond the conceptual stage of research and into the primetime methods of everyday research. However, a key ingredient is missing. At the present time, no clear methodologies or tools exist for aspiring qualitative GIScientists to employ.

In previous research, Martin and Schuurman (submitted) present a set of methods for analysing qualitative data in a GIS environment (Figure 4.2). These methods are, however, challenging for scholars who do not have a programming background. In this article a solution for this is offered in the form of a software tool that implements these methods from start to finish. This software has been named Social Spatial.

4.4. **Program Design**

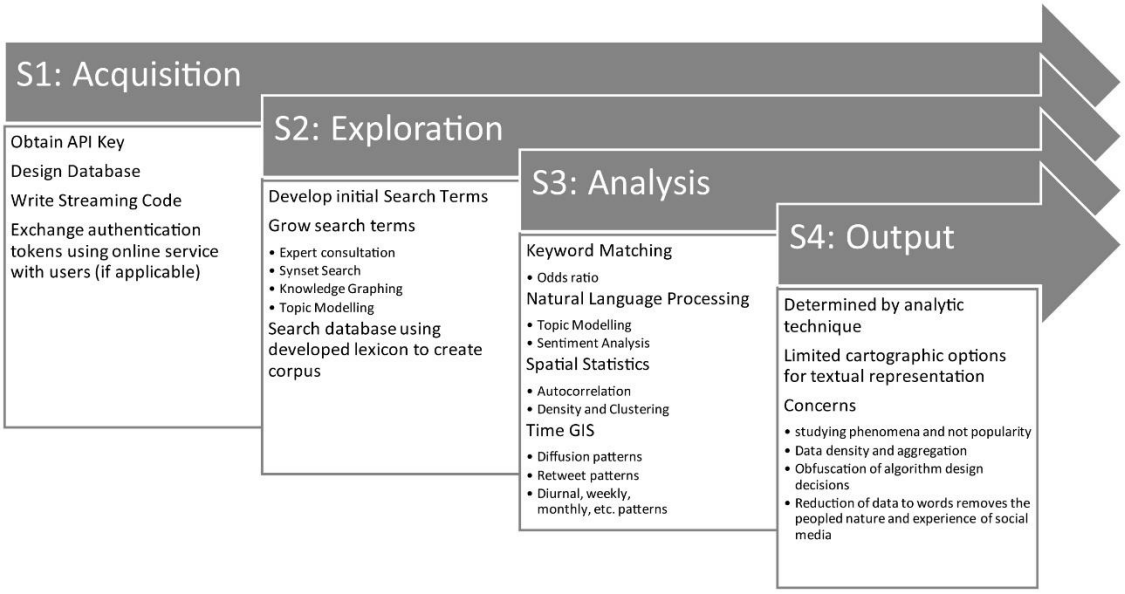


Figure 4.2. Methodology workflow proposed in Martin and Schuurman (2017) for social media GIS research. Each of these steps is facilitated within the Social Spatial tool.

4.4.1. Organization

Social-Spatial was designed around two specific goals. First, to implement the methodology advocated in Martin and Schuurman (2017), and to create a program that is modular, flexible, and does not require an understanding of computer science to operate. The program is presented as a proof-of-concept, and new methods will be incorporated over time. It has been designed so that as new elements are developed they will not interfere with already designed and implemented components. This is accomplished by using modular components that can be opened in and closed as they are needed.

4.4.2. Flexibility in Program flow

Social-Spatial has been designed into modules as an effort to provide a researcher with a high flexible tool that gives them the freedom to explore and

investigate in a non-linear fashion. While the function of each module may not new be new, they have not been organized into a program that creates the potential for an interplay between discovery, visualization, and reflection. By having the capacity to interact with data, refine theories, and develop are redevelop a set of primary data social-spatial is unique in its approach to geographic social media research.

4.4.3. Development Methods

In previous iterations of the software, many components of Social Spatial existed as command-line-interfaces (CLIs). CLIs provide rapid and easily reconfigurable tools that are perfect for creating working iteratively, but due to their reliance on user knowledge of command line execution environments are often not accessible to all users. Implementing a graphical user interface may reduce the configurability of software, but the advantages of usability and visual capability make it attractive – especially as the goal is to make big data analysis more accessible to a range of geographers.

We have developed this software as an open-source project. It sits in a public code repository called ‘github’ (https://github.com/mikedotonline/SocialSpatial) and can be accessed by anyone with a web browser. It is our hope that by making this software open source it will reach a larger audience, and encourage others to engage in the development of new modules, or incorporate our methods into other existing GIS environments, such as QGIS.

Social Spatial is developed in the Python programming language. Python is a well-used language and has an extensive list of pug-in modules that extend its functionality. Social Spatial utilizes many of these including the Psycopg2 module for database connections, Gensim (Řehůřek and Sojka, n.d.) for topic modelling, Cairo for graphics (Pycairo, n.d.), and PyQT (Summerfield and Mark, 2008) for the graphical user interface (GUI.

The code structure used in Social-Spatial follows the pattern of Model View Controller (MVC) (Freeman, 2015). The MVC architecture is helpful for the module nature of Social-Spatial because it is primarily focused on decoupling all elements of the data (model), interface (view), and logic (controller). The interface was designed using

Qt-Designer, an application that accompanies a PyQT installation. Qt-Designer application allows the GUI to be designed visually, enabling the interface and code to be decoupled. As a result, the data models and logic of the program can be used as a CLI as well as a GUI.

4.5. **Software Features**

The following section provides an in-depth look at each of the modular tools (referred to a modules) in Social-Spatial, depicted in figure 4.3. Descriptions of each module is summarized by research stage: data acquisition, exploration, analysis, and output. Module applicability to research stages has been summarized in figure 3.4.

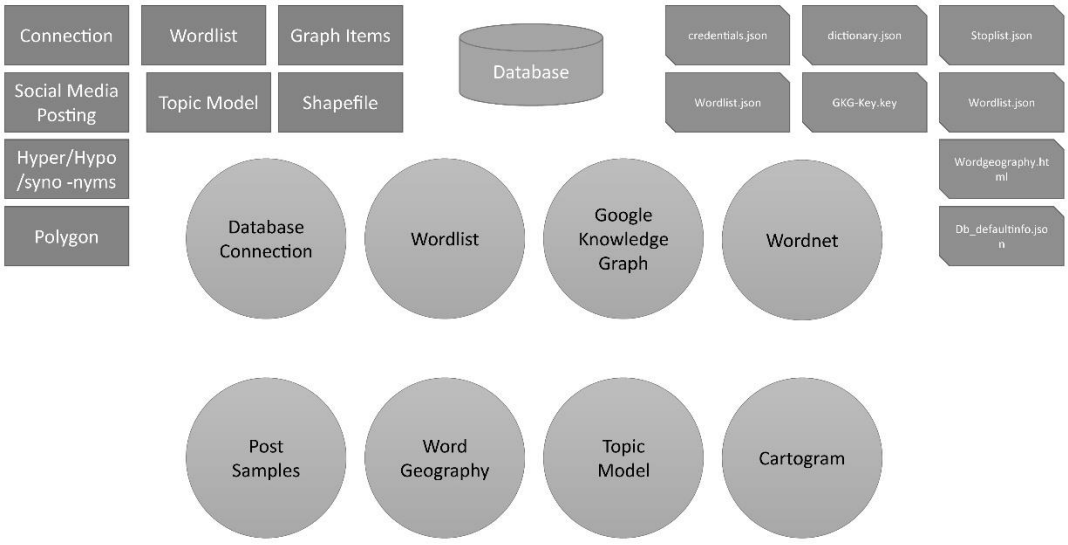


Figure 4.3. Component diagram of Social Spatial. Program Modules (green circles) form the core modules implemented. Data Models (blue squares) allow data to flow throughout the program from module to module. Data files (orange squares) allow for easy program configuration via outside text editors. A database (grey disk shape) with spatial extensions is used as a datastore for social media postings and various spatial data formats as necessary.

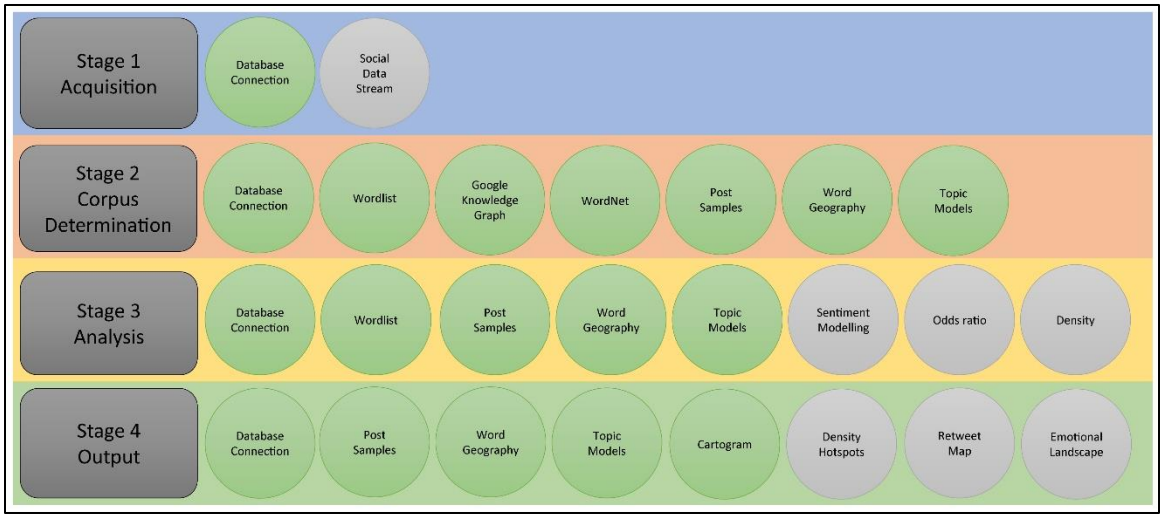


Figure 4.4. Module utility at each stage of research. Green circles represent modules currently developed. Grey circles are future modules not fully developed

4.5.1. Data Acquisition

The data acquisition stage incorporates both the gathering of social data, and the sorting of the information. Social Spatial currently assumes that the user has a dataset already implemented and stored in a PostGIS database, along with tables that delineate the area boundaries they intend to investigate. A module may be included at a later date facilitating the harvest of various social media data feeds, however due to the 'always-on' nature of that module it would not be appropriate to use Social-Spatial for this purpose.

4.5.2. Data Exploration

Data exploration is highly interactive, and to a large extent this is where Social Spatial excels. It is at this stage that the researcher builds an extensive list of words used to identify the social media data that are most relevant to their research. This is similar to traditional qualitative research methods when an investigator purposefully selects participants for a study. Social Spatial implements four modules to aid the researcher in this, The Wordlist, Google Knowledge Graph, Wordnet, and Topic Modelling.

The wordlist houses an import and export functionality to and from the JSON data format. In fact, all of the import and export functions of Social Spatial rely on the JSON (JavaScript Object Notation) format for an important reason. It is flexible (the schema can be augmented if new data structures become relevant), human-readable (it is stored as plain text), and it interoperable (many programs support this format). The wordlist provides the functions to build a list of words and associated tags that describe them. For example, using the case of studying obesity, a wordlist was made that explores healthy and unhealthy foods. Words such as Bacon, Fat, and McDonald's have been given the tag #unhealthy, while words such as Kale, Spinach, Brusslesprout have been given the tag 'healthy'. Tags in Social Spatial can be used as qualitative themes, and sorted for easy access.

The Google Knowledge Graph and Wordnet modules integrate semantic word discovery. This enables researchers to expand a wordlist beyond those they already know. The Google Knowledge Graph3 uses a linked data structure to hold words that are related to specific types of identities, defined by schema.org4. For example, by searching the word "Pirates" by type 'movie' may yield Pirates of the Caribbean while using the type "product" may yield "Lego Pirate Man". Wordnet is alternative that uses an ontological approach to word discovery. Wordnet identifies synonyms, hypoynyms and hypernyms (Miller et al., 1990). While synonyms are words semantically equivalent, hypernyms and hoponyms are words that sit above and below (respectively) in a graph structure of meaning. For example, using the word 'bar' a synonym is be 'rod' or 'speakeasy', a hypernym is 'candy', and a hyponym is 'snickers' or 'mars bar'. Using this method, all words in the wordlist can be checked for semantically related words.

The topic modelling module is useful during the exploration and analysis phase. In the exploration phase, it acts as a mechanism for discovery. The tool can search for all postings within a single geographical boundary (i.e. a city) or iterate through each sub-area (i.e. each city neighbourhood), building a topic model for each sub-area. Functionally the topic modelling module builds a set of terms according to the parameters set by the researcher. Depending on what default parameters are set in the model, the results change. The number of topics will change how generic, or large in

[3] https://developers.google.com/knowledge-graph/)

[4] http://schema.org

scope a set of topics will be. The number of words per topic will limit the number of words per topic. The ‘number of passes’ parameter will increase the probability of each topic and topic word, at the expense of computation time. The alpha coefficient is set at default for short messages, but can be changed for when longer input documents are used. Generating topics and topic words can identify new items to add to the wordlist that the researcher may be unaware of, be it because they are regionally, culturally, age, or otherwise specific beyond their knowledge.

The Post Samples and Word Geography modules aid researchers by providing a visual link between the wordlist and the source material. By selecting different words from the wordlist the user can use the post samples module to see the raw data that emerges from searching against them. Alternatively, users can use the word map to see how these posts are distributed across space. Additionally, the Post Samples module is able to talk to the word map, allowing the researcher to move between lexical and cartographic exploration in real time. The word map is a flexible visualization, the backend of the cartographic visualization is a connection to Leaflet (www.leaflet.com). Leaflet web framework for mapping, using html, CSS and Javascript. This visualization is written as a web document contained in the data directory of Social Spatial, and may be configured to the users desired, be it marker colour, functionality, or can even be changed to a different web framework, such as OpenLayers, another excellent cartographic platform.

4.5.3. Analysis

Many methods of social media analysis are currently available and can be organized into three groups: social science, GIScience, and computer science.

The methods of social science are fluid, but for the purposes of Social Spatial they are limited to the process of developing thematic categories and coding responses (the data) according to themes. While it is not the objective to reproduce the excellent efforts of other programs that accomplish this same process, such as QSR NVivo or Altas Ti, some of the functionality is reproduced in the wordlist. Using the wordlist, a researcher can develop thematic tags, group these tags together, and then visualize how these tags are expressed textually and distributed spatially. Because of the short messages that Social Spatial was designed around, multi coding the same document

has not expressly been developed, but could easily be done accomplished by augmenting the SQL code used in the Post Samples module from using ‘OR’ commands to “AND” – in essence changing the selection from union to intersection. Social Spatial also provides access to easy exporting of the posts found using the Post Samples module, to JSON objects that can be imported to traditional environments such as NVIVO or others.

The analytical methods most relevant to GIScience are those used by Word Maps. There are many ways that GIScience methods can be used to analyse and represent social media postings (given the Cartesian metadata it is bundled with). These may include heat maps, odds ratios, geographically weighted regression, density functions or statistical models, such as spatial/temporal/multi kriging. While these functions are well developed into existing GISystems, they are not currently available in the version of Social Spatial that this paper is based upon. However, Social Spatial provides a tool to copy the SQL statement used to produce the dataset, making it easy to import these data to a GISystem of choice, such as QGIS or ArcGIS. The analytic capability of the current software is limited to point pattern and basic density evaluation, using the word map module. In future releases, it is hoped that more functionality will be integrated into Social Spatial or a tighter integration with a GISystem that users may already be familiar with.

Computer science also provides numerous methods that can be used to investigate text, and this release of Social Spatial has two integrated, with a third on the horizon. The first is topic modelling, where the user can identify latent topics present in the text and how they are distributed across any number of areal definitions. This is particularly useful if the user is interested in identifying patters that exist a different spatial configurations or scales, also known as sensitivity analysis, and described as the modifiable areal unit problem. The topic modelling module is also when used in conjunction with thematic coding, as it can identify new codes to investigate. Topic modelling however is computationally expensive, so it is important that this function be run using a computer with sufficient computing power, and a well configured BLAS (a math library for linear algebra explained in the software documentation).

While not yet fully implemented, Sentiment modelling is on the horizon for Social Spatial. While perhaps premature to include in this release, sentiment is an exciting

inclusion, provided by the NLTK software library. Sentiment modelling is a key method to bring to social spatial, as it will unlock a process for discovery of the emotions hidden in big data. This method will handle the volume and taxing nature of deciphering emotional meanings in short messages.

Using a mixture of social science, GIScience and computer science, Social Spatial attempts to bring together a set of key methods for social media researchers. In effect, these methods are gathered into a usable toolkit. In many cases, this software had been designed and illustrated here with short messages in mind, in particular Twitter. Although not tested for other mediums, there is nothing that currently limits Social Spatial to such methods nor short messages. We expect that, as new datasets are made available to researchers that contain geographic metadata, Social Spatial will be able to handle these as well.

4.5.4. Visualization and Output

Nearly all the modules of Social Spatial provide a method for output, be it visual, or data via JSON. The data objects include stopwords, wordlists, topic models, and social media messages. The visual outputs from social spatial are currently limited to the Word Cartography module, as provided by Leaflet, and the Word Cartograms module that produces PNG files. The cartograms it produces are expressions of spatial topic models that have been generated using topic modelling module. This module translates the textual output of the models into geographic space, printing the topics and words directly on the map canvas. Parameters are exposed in the module to configure the visual nature of the cartogram, such as font, color, and size of image produced. Drawing large cartograms is computationally taxing, due to the large size of the arrays used to composite multiple layers of text and geographic space. For simplicity, this module is best used in the WGS84/Latitude-Longitude projection/coordinate system. The cartograms produced using of this module can be made at any scale, and are particularly useful in exploring the effects of the modifiable areal unit problem.

Currently, expanding the graphic capabilities of Social Spatial beyond the current offerings is not a primary focus of development. However, as analytic functionality increases, new ways of expressing the cartographic data within them may be important to create. The use of web frameworks such as Leaflet or OpenLayers is the standard

method upon which all future cartographic output will be expressed, expect where additional processing is required (such as in the case of custom layer compositing in the Word Cartograms). Web framework are chosen to be the most interoperable, flexible and accessible method that is easily implementable for future developers.

4.6. Case Study using Obesity and Unhealthy Eating Tweets

To demonstrate the utility of Social Spatial, the following case study is given. In Canada, the PURE study is investigating the lived experiences of obesity in Vancouver, British Columbia (Gasevic et al., 2013; Walker et al., 2016). While the study has focused on a small cohort of participants (in number and spatial extent) in great detail, there is a need to understand the tapestry of food culture the participants exist within. One way to do so would be to investigate the way unhealthy foods are expressed in social media in the city.

4.6.1. Stage 1 - Data generation

During the months of January - October 2011, twitter data was acquired using a python script and stored in a PostGreSQL database with PostGIS extensions installed. Over the 10 month period, 1.5m tweets were collected.

4.6.2. Stage 2 - Keyword building

First, a list of words relating to the topic of unhealthy eating is compiled using the researcher's intuition:

```
Obesity obese fat fatty overeating overeat bloating bloated unhealthy inactive
sedentary sugar fructose lethargy lethargic sloth metabolism tired sad chest-
pain
```

Using these words as a starting point, the twitter corpus can be searched, both within a search area, or across the entire dataset. This is done through SQL "Like" commands, taking full advantage of lexical indexes and creating fast searches.

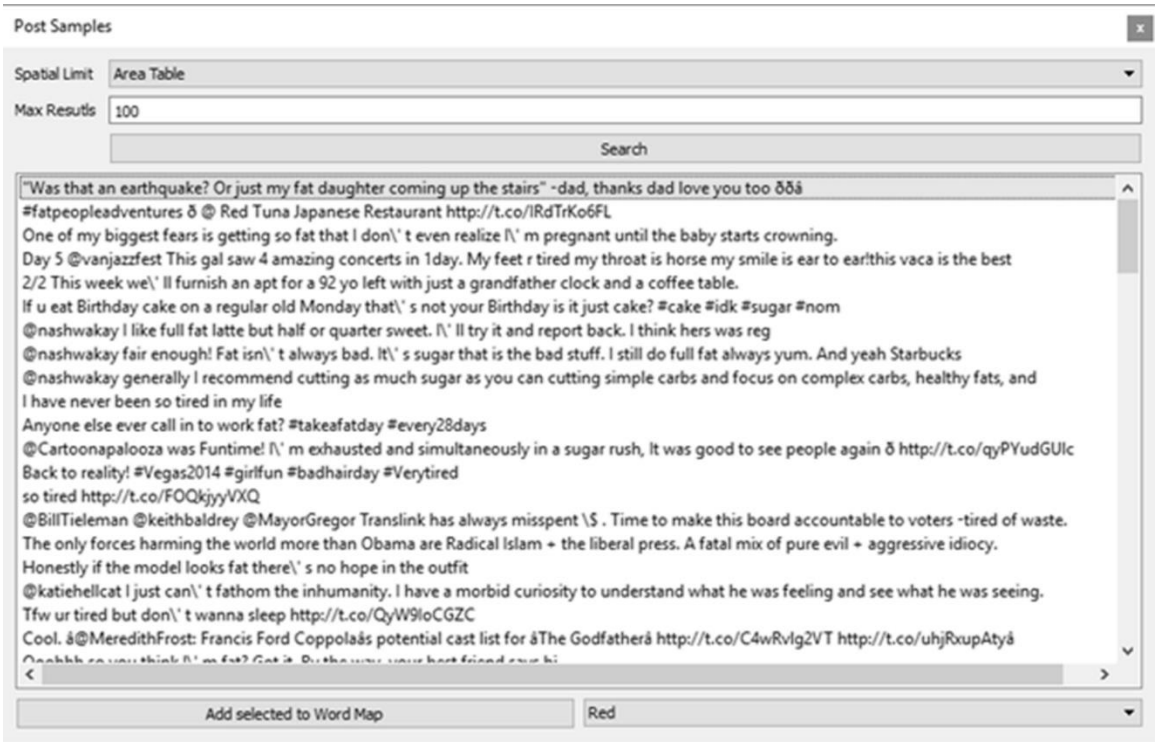


Figure 4.5. Using the Post Samples module to explore unhealthy eating habits based on keyword matching from the wordlist module

By looking at the tweets themselves, more keywords are determined and added to the growing list of words. For example, using figure 4.5 (above) the words ‘sugar’, ‘sweet’, ‘cake’ and the hashtag ‘fatpeopleadventures’ and ‘takeafatday’ are be added, while the word ‘tired’ is removed because it yields too many unrelated postings.

Using the post samples module, it was identified that products such as candy bars and cola useful in finding unhealthy eating tweets. To explore this further, the Google Knowledge Graph module was used to identify names of more unhealthy products. As an example, the words ‘Candy Bar’ and ‘soft-drink’ is searched using the category of words ‘products’ (figure 4.6). By using the Google Knowledge Graph, a key research output was identified: brand names and products of unhealthy foods are the most prolific and useful keywords in the twitter dataset, in terms of identifying tweets that can be used to build a comprehensive corpus

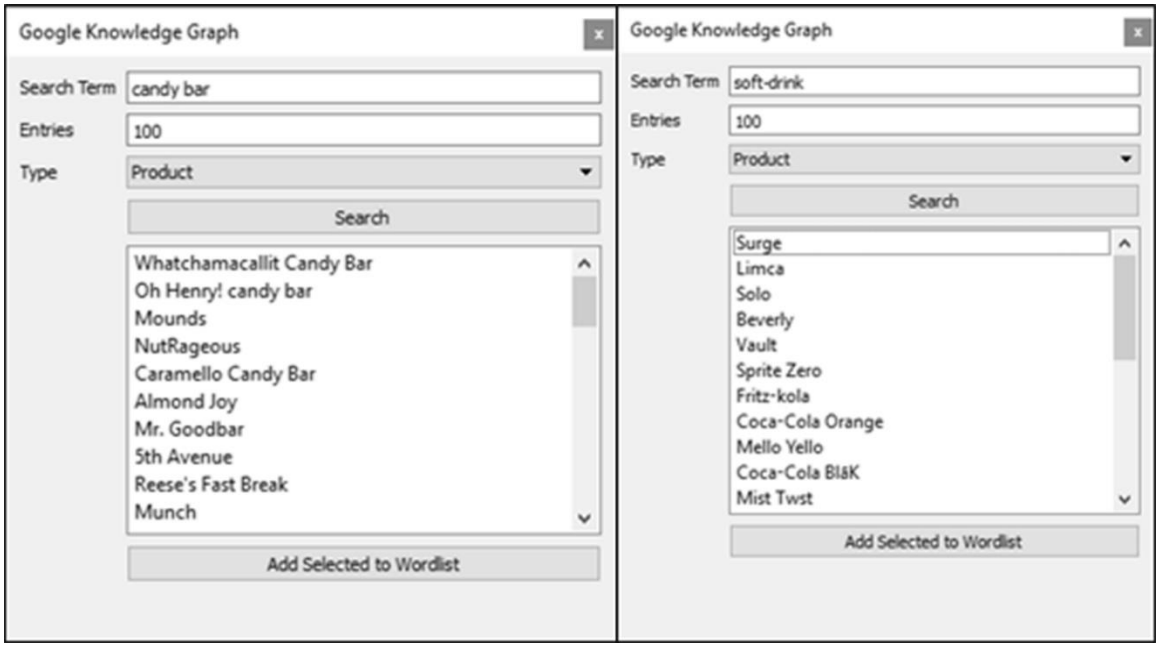


Figure 4.6. The Google Knowledge Graph module providing multiple product listings for candy bars and soft drinks

The wordnet module help to identify words that have many synonyms, for example, the words overeat. Searching for this words yielded ‘overindulge’ ‘gorge’ ‘binge’ ‘pig out’ and ‘glut’. Each of these words is then added to the wordlist, and a search of the tweet data is carried out using them to see what they yield. ‘Gorge’ proves to be unhelpful as it is matched to ‘gorgeous’ and is removed, or changed to ‘gorge ‘ as the space character stops the “Like” SQL command from identifying so called ‘Matryoshka words’, which can be useful at other times like pluralization, possession, or hashtags (figure 4.7).

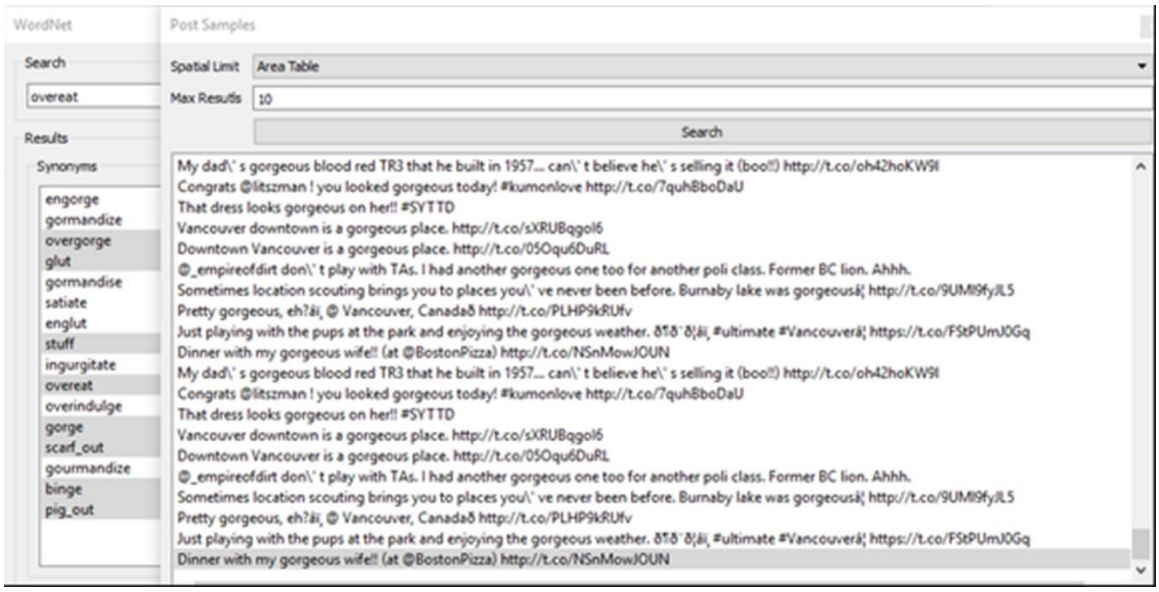


Figure 4.7. The Wordnet module provides synonyms to the word 'overeat' and the post samples generated from these synonyms

Finally, one more module is used to identify additional keywords. While the Google Knowledge Graph and Wordnet modules are useful in identifying words based on other known words, eventually there all of the known options for additional words will be exhausted. The Topic Modelling is used to perform a search of the tweet corpus in the area or areas being researched in order to identify words that are not otherwise known, and that are used by the twitter users themselves. Topic Modelling groups words together that are related to one another, in this case, topics related to unhealthy eating. When applied to the dataset used here it identifies the names of ice cream parlours “earnest ice cream” and “soft peaks”, a donut shop “Cartems”, a type of restaurant “poutinerie”, hashtags “#epicmeal” and “gastropost” (figure 4.8). These words are then searched against the corpus of tweets, which yields more words again.

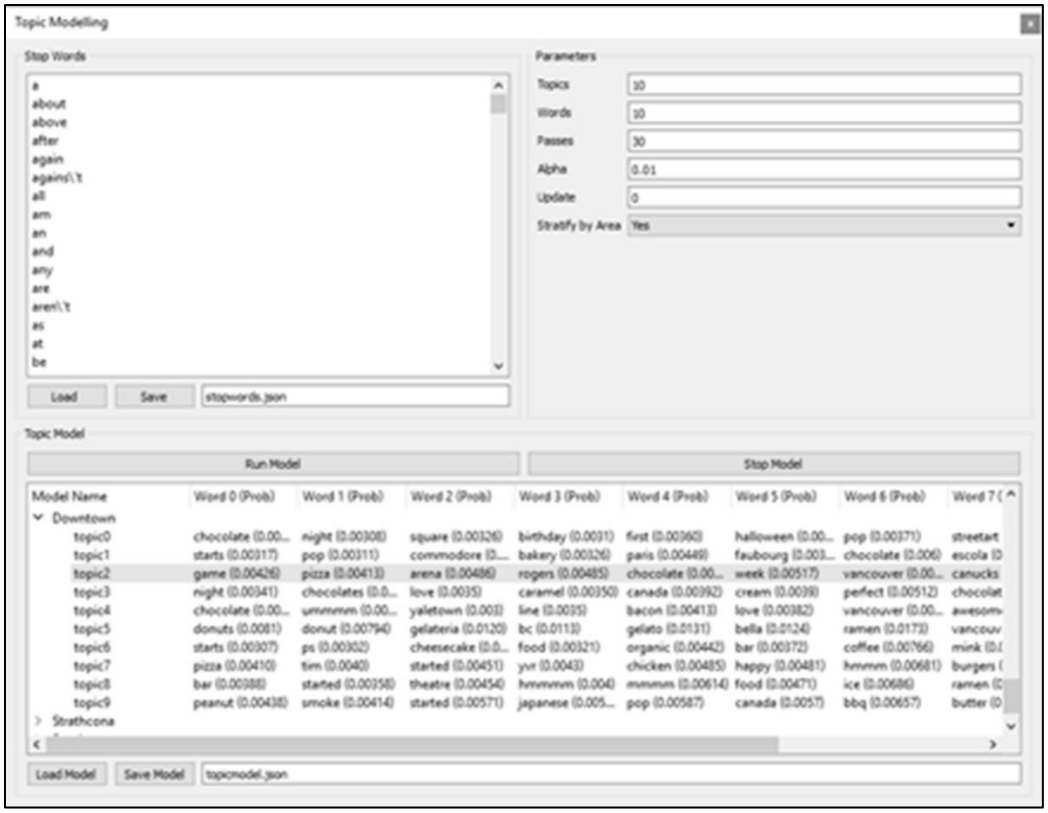


Figure 4.8. Topic Modelling results, calculated for each neighbourhood in Vancouver, BC

The process used above highlights the iterative nature of social media research. At all times, the researcher find new words, checks them against the data, searches for alternatives, and continues to develop the primary data they are using and identify primary themes that can be used later on in their analysis.

The process of keyword building yields two important results. 1) the subset of data that contains all relevant material to base analysis on; and 2) the themes that each keyword can be organized into. This allows for the analysis results to be organized around the concepts that are emergent. In the case of the unhealthy eating and the Pure study, the themes branding, fast food, and emotional eating have been identified.

4.6.3. Stage 3 - Analysis

Based on the three themes identified, two analyses are carried out: 1) thematic variations in tweet locations are visually inspected using the Word Geography module; and 2) topic models are created that explore how the themes are discussed across

different areas of the city. The two methods can be combined alongside one another, for example: topic modelling identifies that alcohol is a major topic of conversation in the tweets, and sports are as well. Using the Word Geography module, we can see that posts concerning alcohol and tweets are clustered in the downtown core of Vancouver, in particular around drinking establishments in the Stadium district. This search leads to a new thematic discovery in the study, that bars and sports may be significant contrubtors to the ways that unhealthy foods are discussed, and may later lead to policy development targetted to sports patrons.

Figure 4.9. Word Geography module, with unhealthy eating and alcohol realted posts in the stadium area of Vancouver, BC

Beyond the analytical capabilities of Social Spatial, the GIS literature offers several novel approaches using the tools already present in traditional GIS environments such as QGIS or ArcGIS. In an effort not to recreate their functionality, Social Spatial can produce a SQL query based on the wordlist used, that can be implemented in GIS software. In this study, this functionality was used to identify unhealthy tweet density, and normalized tweet density by population. An odds ratio was also performed in addition against the themes healthy and unhealthy.

4.6.4. Stage 4 - Output

The final step on the case study is preparing research output. Three methods are used from Social Spatial; 1) key quotes via the Post Samples module (figure 4.7); 2) Topic Model visualization is prepared via the Cartogram module, and; 3) figures of tweet occurrence (figure 4.9). While generating visually appealing representations of findings is not the main focus of Social Spatial, the program provides easy access to export the findings. During the study of unhealthy eating, the consumption of alcohol emerged as a theme. It visualizes where alcohol was being tweeted from, the researchers used the “copy to SQL” function of the Post Samples module and performed basic cartography in QGIS to generate figure 4.10, illustrating the high density of postings in the neighbourhood of Downtown and the low densities found near family residences. Connecting the alcohol and unhealthy tweets together, a union of the two layers is completed, using the “Copy to SQL” function again, and using an AND function between the two queries.

Further work would be required to disambiguate these tweets from the general popularity of Twitter downtown, however it would be a starting point for further geographic analysis.

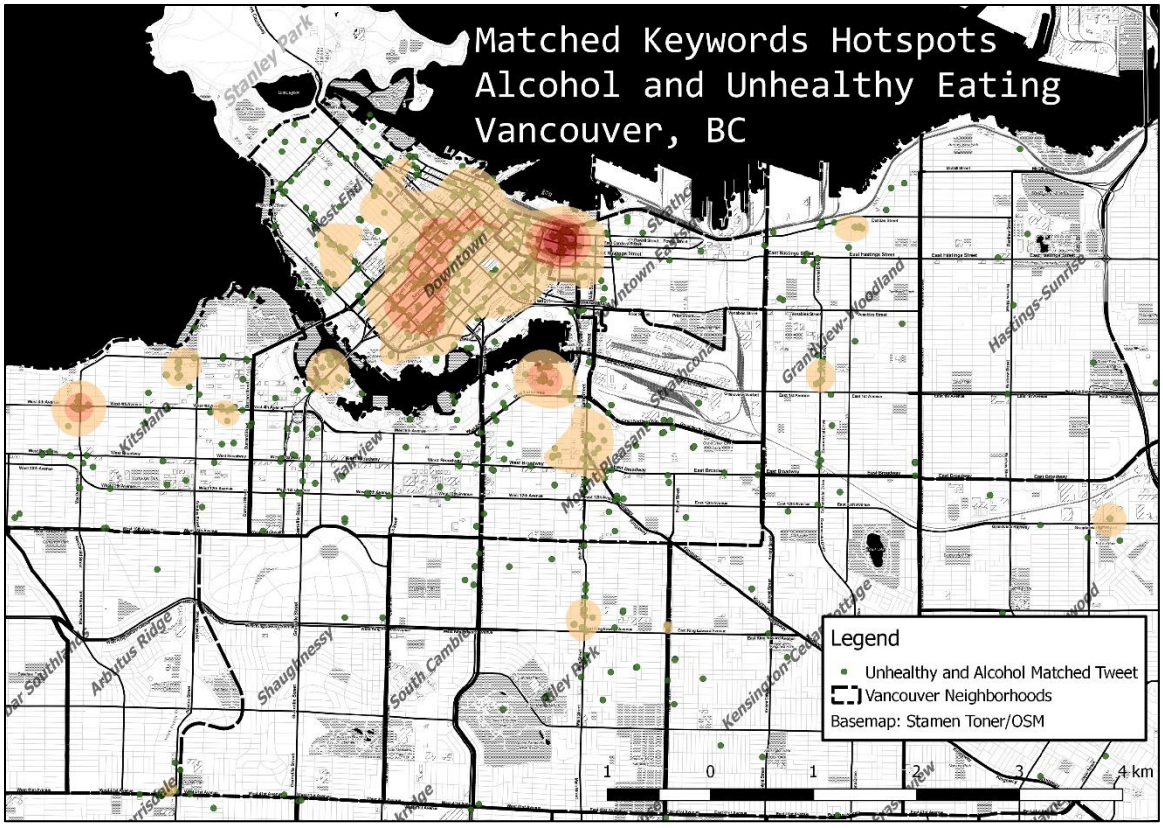


Figure 4.10. Matched keywords of both unhealthy eating and alcohol with tweets. Map generated through Social-Spatial and SQL copied to QGIS for cartography. Basemap courtesy Stamen Design (www.stamen.com) and OpenStreeMap (www.osm.org)

4.7. Discussion and Conclusion

Social Spatial is a progression from earlier efforts made by GIScience scholars to fuse qualitative methods with traditional GIS environments, such as those of Kwan and Ding (2008), and Jung (2007).The work of these scholars on ArcGIS extensions and CAQ-GIS were important steps that illustrated that it was indeed possible to integrate qualitative data while providing the tools required for analysis.

More than ten years later, qualitative GIS has begun to come of age. Outside of GIS, QSR's NVivo software has become a mainstay software for 1.5 million social science researchers engaged in qualitative data analysis5. As NVivo demonstrates, digital systems can be an essential tool. Natural Language Processing literature has

[5] https://www.qsrinternational.com/nvivo/who-uses-nvivo

proven to be a key set of tools for handling the massive number of data points big data integrates, and is a rapid growth area of computer science that other disciplines are taking note of, including geography(Frank et al., 2013; Ghosh and Guha, 2013; Gore et al., 2015). What has been lacking is a next generation qualitative GIS to implement these key technologies in a spatially aware interface.

Social Spatial is a first step towards a qualitative GIS for big social geospatial data. It embraces modern programming techniques, languages, data formats, and methods. It is modular and allows for researchers to act iteratively, testing new thematic hypothesis as they develop. It provides a pathway for researchers to publish the settings, keyword lists, and results of their research alongside research publications, increasing research transparency and trust. What this program cannot do, it readily hands off to the researcher so they may use programs that are better at those particular activities, for example in providing easy tools for data exfiltration to traditional GIS software. Many areas for growth and refinement clearly exist. The list of potential analytical techniques continues to grow and the list of potential refinements to usability and stability of the program is large, too. However, in the modern landscape of software development the authors have chosen to 'fail-early' and provide the 'minimal-viable-product' to its potential users to demonstrate the utility of the program, while they continue to create a better product, using a 'perpetual-beta' approach (Hennig, 2017). This approach is facilitated through the use of GIT, an online versioning and source control platform hosted by Github. We see this as an improvement to the way previous qualitative GIS tools have been created and shared. This approach not only makes the code used open access, but also encourages the participation of the wider community in not only validating the methods used and provides the opportunity for injecting new code in a collaborative way. We see this as a progression in the way that code is written in GIScience. As methods become increasingly complex and unsupervised, it is important that we ensure that we remain as transparent in our implementations as possible and continue to recognise the potential for all GIScientists to contribute to new tools that represent all of our needs.

Chapter 5.

Conclusion

This dissertation is the presentation of method, methodology, and software that together produce a qualitative GIS for big data and social media. Building a qualitative GIS is more than producing a software or set of algorithms that are capable of handling qualitative information in light of new data. It requires critical examination of how the code is embodied and ways in which the data are contextually sensitive.

In the first article, a natural language processing technique was implemented to produce topic models visually across areal units, and applied to Vancouver, BC. In this study the component parts of the algorithm were teased apart using a critical appraisal. This appraisal of the topic modelling algorithm revealed the ways that stop words, data selection, and parameter settings can alter the output of the model. The cartographic algorithm investigation examined how font, color, and placement can alter the way the output is produced – and how this affects representation. Using an area based approach to the topic modelling algorithm allowed for the geographical problems to be identified - such as the modifiable areal unit problem and the challenge of spatializing qualitative information.

The second article presented a methodology in four stages that can be used for undertaking qualitative GIS research using big data and social media. The four stages include 1) data acquisition; 2) corpus formulation; 3) analysis and; 4) representation. At each stage, the methods used by GIScience and computer science were critically examined. The approach taken in this review recognizes researcher positionality and suggests methods that work with the partial perspective of a researcher. In the discussion of the article, issues related to data representativeness and ethics, researcher positionality, the challenge of hybridity, and algorithmic complexity are explored.

The third and final article presents Social Spatial, a software that implements the method and methodology presented in the previous two articles. The purpose of writing this software was twofold, 1) to provide access to methods that might otherwise required specialized computer science skills and, 2) to expose as many of the parameters that go

into qualitative social media research methods. The software was released as open-source with thorough internal documentation to a collaborative code repository to encourage others to contribute and make the programming as transparent as possible. Extensive use of settings files was employed to expose parameters – word lists, model parameters, and stop words. An advantage of using files is that it encourages users to publish them alongside research findings. Publishing these exposes the choices made by researchers, creating transparency in qualitative social media research methods and the code/spaces they were enacted within without increasing the burden of research documentation.

The contribution of this dissertation is to not only produce a qualitative GIS, but at the same time to reveal its internal components. While other researchers have focused on examining the impact of big data and social media analysis, this work instead investigates the qualitative GIS itself using the tools that critical and qualitative GIS scholars have been building since its inception in the 1990s. In order to do so effectively, and as a contribution to the field itself, a qualitative GIS was built along the way. Thus the call for a qualitative GIS made by Elwood and Cope in 2009 was fulfilled.

This dissertation is limited in that it provides a method, methodology and software. It does not provide a specific implementation of these in an in-depth, peer-reviewed, multi-investigator qualitative research study from end-to-end. Doing so would provide a better opportunity to see how the dissertation can achieve its goal of creating a truly qualitative GIS for big data and social media. However, this is beyond the scope of this dissertation.

The implications of this research provide a potential pathway for future research and offer advice for future scholars wishing to engage in qualitative GIS. First, integrating human geography and qualitative GIS with the discipline of computer science and natural language processing is imperative. Nothing will stop the latter from progressing and developing ever more capable algorithms to study social phenomena, especially as they relate to geography. Should human geography and qualitative GIS be unable or unwilling to interact with these algorithms and the researchers that produce them, the rich history, theory and methods of human geography, critical GIS, and the inroads qualitative GIS has made will be undermined. Qualitative GIS and social media

big data studies need to advance the approaches they have cultivated in sync with changes to technology and data sources available.

It is my hope that future scholars continue to investigate the way in which context, situated knowledges and critical can be used to understand the relationships that exist between place and space. The ways that data and methods work on qualitative information must be placed within a framework that is sensitive to the context and situations they are produced within. If human geographers and GIScientists take up this challenge, future research will help to maintain a healthy skepticism of how closely algorithms are tied to social conventions and culture – thus avoiding black boxed technologies and supporting integration of qualitative data.

References

Agarwal A, Xie B, Vovsha I, et al. (2011) Sentiment analysis of Twitter data. *Proceedings of the Workshop on Languages in Social Media*, Association for Computational Linguistics: 30–38. Available from: http://dl.acm.org/citation.cfm?id=2021114 (accessed 11 March 2017).

Allen J (2003) Natural language processing. John Wiley and Sons Ltd.: 1218–1222. Available from: http://dl.acm.org/ft_gateway.cfm?id=1074630&type=html (accessed 22 June 2015).

Alonge A, Calzolari N, Vossen P, et al. (1998) The Linguistic Design of the EuroWordNet Database. In: *EuroWordNet: A multilingual database with lexical semantic networks*, Dordrecht: Springer Netherlands, pp. 19–43. Available from: http://link.springer.com/10.1007/978-94-017-1491-4_2 (accessed 17 October 2017).

Asghar MZ, Khan A, Ahmad S, et al. (2014) A Review of Feature Extraction in Sentiment Analysis. *Journal of Basic and Applied Research International* 4(3): 181–186. Available from: http://www.researchgate.net/publication/283318740_A_Review_of_Feature_Extraction_in_Sentiment_Analysis (accessed 7 December 2015).

Babbage C (1864) *Passages from the life of a philosopher*. London: Longman, Green, Longman, Roberts, & Green. Available from: https://www.worldcat.org/title/passages-from-the-life-of-a-philosopher/oclc/258982 (accessed 8 March 2017).

Backstrom L, Sun E and Marlow C (2010) Find me if you can. In: *Proceedings of the 19th international conference on World wide web - WWW '10*, New York, New York, USA: ACM Press, p. 61. Available from: http://portal.acm.org/citation.cfm?doid=1772690.1772698 (accessed 5 October 2017).

Balasuriya L, Wijeratne S, Doran D, et al. (2016) Finding street gang members on Twitter. In: *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, IEEE, pp. 685–692. Available from: http://ieeexplore.ieee.org/document/7752311/ (accessed 11 March 2017).

Bauer S, Noulas A, Seaghdha DO, et al. (2012) Talking Places: Modelling and Analysing Linguistic Content in Foursquare. In: *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, IEEE, pp. 348–357. Available from: http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=6406375 (accessed 22 June 2015).

Bazeley P and Jackson K (2013) *Qualitative data analysis with NVivo*. Sage Publications Limited. Available from: https://scholar.google.ca/scholar?cluster=10380075664962249095&hl=en&as_sdt=0,5&as_ylo=2011#0 (accessed 15 October 2015).

Becker H, Naaman M and Gravano L (2011) Beyond Trending Topics: Real-World Event Identification on Twitter. *Fifth International AAAI Conference on Weblogs and Social Media*. Available from: http://academiccommons.columbia.edu/catalog/ac:135415 (accessed 26 February 2016).

Bell N, Schuurman N, Oliver L, et al. (2007) Towards the construction of place-specific measures of deprivation: a case study from the Vancouver metropolitan area. *The Canadian Geographer/Le Géographe canadien*, Blackwell Publishing Inc 51(4): 444–461. Available from: http://doi.wiley.com/10.1111/j.1541-0064.2007.00191.x (accessed 15 September 2016).

Bello-Orgaz G, Jung J and Camacho D (2016) Social big data: Recent achievements and new challenges. *Information Fusion*, Elsevier 28: 45–59. Available from: http://www.sciencedirect.com/science/article/pii/S1566253515000780 (accessed 16 October 2017).

Bird S, Klien E and Loper E (2009) *Natural Language Processing with Python*. O'Reilly Media. Available from: http://shop.oreilly.com/product/9780596516499.do (accessed 9 March 2017).

Blei DM, Ng AY and Jordan MI (2003) Latent dirichlet allocation. *The Journal of Machine Learning Research*, JMLR.org 3: 993–1022. Available from: http://dl.acm.org.proxy.lib.sfu.ca/citation.cfm?id=944919.944937 (accessed 7 December 2015).

Bolelli L, Ertekin S and Giles CL (2009) *Advances in Information Retrieval*. Boughanem M, Berrut C, Mothe J, et al. (eds), Lecture Notes in Computer Science, Berlin, Heidelberg: Springer Berlin Heidelberg. Available from: http://www.springerlink.com/index/10.1007/978-3-642-00958-7 (accessed 13 January 2016).

Bonevski B, Regan T, Paul C, et al. (2014) Associations between alcohol, smoking, socioeconomic status and comorbidities: Evidence from the 45 and Up Study. *Drug and Alcohol Review* 33(2): 169–176. Available from: http://doi.wiley.com/10.1111/dar.12104 (accessed 15 September 2016).

Boyd D and Crawford K (2012) Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*. Available from: http://www.tandfonline.com/doi/abs/10.1080/1369118X.2012.678878 (accessed 15 June 2017).

Brown M and Knopp L (2008) Queering the Map: The Productive Tensions of Colliding Epistemologies. *Annals of the Association of American Geographers*, Taylor & Francis Group 98(1): 40–58. Available from: http://www.tandfonline.com/doi/abs/10.1080/00045600701734042 (accessed 19 September 2016).

Burrows S, Auger N, Gamache P, et al. (2012) Individual and area socioeconomic inequalities in cause-specific unintentional injury mortality: 11-Year follow-up study of 2.7 million Canadians. *Accident Analysis & Prevention* 45: 99–106.

Caldarola EG and Rinaldi AM (2016) Improving the Visualization of WordNet Large Lexical Database through Semantic Tag Clouds. In: *2016 IEEE International Congress on Big Data (BigData Congress)*, IEEE, pp. 34–41. Available from: http://ieeexplore.ieee.org/document/7584918/ (accessed 9 March 2017).

Causer T and Wallace V (2012) Building a volunteer community: results and findings from Transcribe Bentham. *Digital Humanities Quarterly , 6 (2012)* . Available from: http://discovery.ucl.ac.uk/1362050/ (accessed 5 May 2017).

Chambers R (1994) Participatory rural appraisal (PRA): Analysis of experience. *World Development* 22(9): 1253–1268. Available from: http://www.sciencedirect.com/science/article/pii/0305750X94900035 (accessed 22 August 2015).

Chi M, Plaza A, Benediktsson JA, et al. (2016) Big Data for Remote Sensing: Challenges and Opportunities. *Proceedings of the IEEE* 104(11): 2207–2219. Available from: http://ieeexplore.ieee.org/document/7565634/ (accessed 11 October 2017).

Chikersal P, Poria S, Cambria E, et al. (2015) Modelling Public Sentiment in Twitter: Using Linguistic Patterns to Enhance Supervised Learning. In: Springer, Cham, pp. 49–65. Available from: http://link.springer.com/10.1007/978-3-319-18117-2_4 (accessed 11 March 2017).

Cody EM, Reagan AJ, Mitchell L, et al. (2015) Climate change sentiment on Twitter: An unsolicited public opinion poll. Physics and Society; Computers and Society: 11. Available from: http://arxiv.org/abs/1505.03804 (accessed 16 September 2015).

Crampton JW (2009) Cartography: maps 2.0. *Progress in Human Geography* 33(1): 91–100. Available from: http://phg.sagepub.com/content/33/1/91.refs (accessed 13 August 2015).

Crampton JW, Graham M, Poorthuis A, et al. (2013) Beyond the geotag: situating 'big data' and leveraging the potential of the geoweb. *Cartography and Geographic Information Science* 40(2): 130–139. Available from: http://www.tandfonline.com/doi/abs/10.1080/15230406.2013.777137 (accessed 4 November 2016).

Crane G, Fuchs B and Smith A (2000) The symbiosis between content and technology in the Perseus Digital Library. *Cultivate Interactive* (2). Available from: http://pubman.mpdl.mpg.de/pubman/faces/viewItemOverviewPage.jsp?itemId=escidoc:2277435 (accessed 5 May 2017).

Crawford K, Miltner K and Gray M (2014) Critiquing Big Data: Politics, Ethics, Epistemology. *International Journal of Communication* 8: 1663–1672. Available from: http://web.a.ebscohost.com/ehost/detail/detail?vid=0&sid=eb8f1d70-313b-4365-808f-63f3267c3593%40sessionmgr4007&bdata=JnNpdGU9ZWhvc3QtbGl2ZQ%3D%3D#AN=97249722&db=ufh (accessed 17 October 2017).

Crooks A, Croitoru A, Stefanidis A, et al. (2013) #Earthquake: Twitter as a Distributed

Sensor System. *Transactions in GIS* 17(1): 124–147. Available from: http://doi.wiley.com/10.1111/j.1467-9671.2012.01359.x (accessed 16 June 2015).

Curry M (1994) Image, practice and the hidden impacts of geographic information systems. *Progress in Human Geography*. Available from: http://phg.sagepub.com/content/18/4/441.short (accessed 15 October 2015).

Dalton CM and Thatcher J (2015) Inflated granularity: Spatial 'Big Data' and geodemographics. *Big Data & Society*, SAGE Publications 2(2): 2053951715601144. Available from: http://bds.sagepub.com.proxy.lib.sfu.ca/content/2/2/2053951715601144.abstract (accessed 12 November 2015).

Daniel R, Hall D, Nallapati R, et al. (2009) Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, STROUDSBURG, PA: ACL, pp. 248–256.

Dixon LJ, Correa T, Straubhaar J, et al. (2014) Gendered Space: The Digital Divide between Male and Female Users in Internet Public Access Sites. *Journal of Computer-Mediated Communication*, Blackwell Publishing Ltd 19(4): 991–1009. Available from: http://doi.wiley.com/10.1111/jcc4.12088 (accessed 15 June 2017).

Duggan M (2015) *Mobile Messaging and Social Media 2015 | Pew Research Center*. Available from: http://www.pewinternet.org/2015/08/19/mobile-messaging-and-social-media-2015/ (accessed 27 September 2016).

Elwood, Sarah; Cope M (2009) *Qualitative GIS: A Mixed Methods Approach*. SAGE Publications. Available from: https://books.google.com/books?hl=en&lr=&id=vnAbJ8spyeYC&pgis=1 (accessed 22 June 2015).

Elwood S (2006) Beyond Cooptation or Resistance: Urban Spatial Politics, Community Organizations, and GIS-Based Spatial Narratives. *Annals of the Association of American Geographers*, Taylor & Francis Group 96(2): 323–341. Available from: http://www.tandfonline.com/doi/abs/10.1111/j.1467-8306.2006.00480.x (accessed 19 September 2016).

Elwood S (2008) Volunteered geographic information: future research directions motivated by critical, participatory, and feminist GIS. *GeoJournal* 72(3–4): 173–183. Available from: http://link.springer.com/10.1007/s10708-008-9186-0 (accessed 16 June 2015).

Elwood S (2009) Mixed methods: Thinking, doing, and asking in multiple ways. *The SAGE Handbook of qualitative geography*. Available from: https://books.google.ca/books?hl=en&lr=&id=pvAqzyZhQ24C&oi=fnd&pg=PA94&ots=XXwlKpVBw6&sig=LpR-9-pur_xqMBAKLyb5pZTtClo (accessed 15 October 2015).

Elwood S and DeLyser D (2010) *Mixed methods: Thinking, doing, and asking in multiple ways*. 1st ed. Thousand Oaks, CA: SAGE Publications. Available from:

https://books.google.ca/books?hl=en&lr=&id=pvAqzyZhQ24C&oi=fnd&pg=PA94&dq=Mixed+Methods:+Thinking,+Doing,+and+Asking+in+Multiple+Ways&ots=XXBnQq0In3&sig=F1nOOp4SIptLC_VWvhXDl7aMAlY (accessed 15 June 2017).

Elwood S and Ghose R (2001) PPGIS in Community Development Planning: Framing the Organizational Context. *Cartographica: The International Journal for Geographic Information and Geovisualization*, University of Toronto Press 38(3–4): 19–33. Available from: http://utpjournals.press/doi/10.3138/R411-50G8-1777-2120 (accessed 13 October 2017).

Elwood S and Leszczynski A (2011) Privacy, reconsidered: New representations, data practices, and the geoweb. *Geoforum* 42(1): 6–15. Available from: http://www.sciencedirect.com/science/article/pii/S001671851000093X (accessed 5 May 2017).

Elwood S, Goodchild MF and Sui D (2013) Prospects for VGI Research and the Emerging Fourth Paradigm. In: *Crowdsourcing Geographic Knowledge*, Dordrecht: Springer Netherlands, pp. 361–375. Available from: http://www.springerlink.com/index/10.1007/978-94-007-4587-2_20 (accessed 19 September 2016).

England KVL (1994) Getting Personal: Reflexivity, Positionality, and Feminist Research∗ . *The Professional Geographer*, Taylor & Francis Group 46(1): 80–89. Available from: http://www.tandfonline.com/doi/abs/10.1111/j.0033-0124.1994.00080.x (accessed 9 March 2017).

Ferrara E, Varol O, Menczer F, et al. (2013) Traveling trends: social butterflies or frequent fliers. In: *Proceedings of the first ACM conference on Online social networks - COSN '13*, New York, New York, USA: ACM Press, pp. 213–222. Available from: http://dl.acm.org/citation.cfm?doid=2512938.2512956 (accessed 15 March 2017).

Frank MR, Mitchell L, Dodds PS, et al. (2013) Happiness and the Patterns of Life: A Study of Geolocated Tweets. Available from: http://arxiv.org/abs/1304.1296 (accessed 15 March 2017).

Freeman A (2015) The Model/View/Controller Pattern. In: *Pro Design Patterns in Swift*, Berkeley, CA: Apress, pp. 527–552. Available from: http://link.springer.com/10.1007/978-1-4842-0394-1_27 (accessed 5 October 2017).

Frias-Martinez V, Soto V, Hohwald H, et al. (2012) Characterizing Urban Landscapes Using Geolocated Tweets. In: *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, IEEE, pp. 239–248. Available from: http://ieeexplore.ieee.org/document/6406289/ (accessed 15 March 2017).

Gao S, Janowicz K and Couclelis H (2017) Extracting urban functional regions from points of interest and human activities on location-based social networks. *Transactions in GIS* 21(3): 446–467. Available from: http://doi.wiley.com/10.1111/tgis.12289 (accessed 5 October 2017).

Gasevic D, Yew A, Teo K, et al. (2013) Built Environment, Physical Activity and Adiposity in the Prospective Urban Rural Epidemiology (PURE) Study Vancouver Cohort. *Canadian Journal of Diabetes*, Elsevier 37: S275. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1499267113004723 (accessed 2 May 2017).

Ghosh DD and Guha R (2013) What are we 'tweeting' about obesity? Mapping tweets with Topic Modeling and Geographic Information System. *Cartography and geographic information science*, Taylor & Francis 40(2): 90–102. Available from: http://www.tandfonline.com/doi/abs/10.1080/15230406.2013.776210 (accessed 30 March 2015).

Goodchild MF (2007) Citizens as sensors: the world of volunteered geography. *GeoJournal* 69(4): 211–221. Available from: http://link.springer.com/10.1007/s10708-007-9111-y (accessed 14 July 2014).

Gore RJ, Diallo S and Padilla J (2015) You Are What You Tweet: Connecting the Geographic Variation in America's Obesity Rate to Twitter Content. *PloS one*, Public Library of Science 10(9): e0133505. Available from: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0133505 (accessed 16 September 2015).

Greenwood S, Perrin A and Duggan M (2016) *Social Media Update 2016*. Available from: http://www.pewinternet.org/2016/11/11/social-media-update-2016/.

Haklay M, Singleton A and Parker C (2008) Web Mapping 2.0: The Neogeography of the GeoWeb. *Geography Compass* 2(6): 2011–2039. Available from: http://doi.wiley.com/10.1111/j.1749-8198.2008.00167.x (accessed 4 July 2015).

Halevy A, Norvig P and Pereira F (2009) The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems*, IEEE 24(2): 8–12. Available from: http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=4804817 (accessed 7 May 2015).

Hao M, Rohrdantz C, Janetzko H, et al. (2011) Visual sentiment analysis on twitter data streams. In: *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, IEEE, pp. 277–278. Available from: http://ieeexplore.ieee.org/document/6102472/ (accessed 11 March 2017).

Hao Q, Cai R, Wang C, et al. (2010) Equip tourists with knowledge mined from travelogues. In: *Proceedings of the 19th international conference on World wide web - WWW '10*, New York, New York, USA: ACM Press, p. 401. Available from: http://dl.acm.org.proxy.lib.sfu.ca/citation.cfm?id=1772690.1772732 (accessed 26 February 2016).

Haraway D (1988) Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies* 14(3): 575. Available from: http://www.jstor.org/stable/3178066?origin=crossref (accessed 8 March 2017).

Haraway DJ (1988) Simians, cyborgs, and women : the reinvention of nature. In: *Simians, cyborgs, and women : the reinvention of nature*, Routlage, p. 287.

Available from: https://books.google.ca/books?id=23ff1vCaLPIC&dq=simians+cyborgs+and+women+situated+knowledges&lr=&source=gbs_navlinks_s (accessed 8 March 2017).

Hargittai E (2010) Digital Na(t)ives? Variation in Internet Skills and Uses among Members of the ?Net Generation? *Sociological Inquiry*, Blackwell Publishing Ltd 80(1): 92–113. Available from: http://doi.wiley.com/10.1111/j.1475-682X.2009.00317.x (accessed 15 June 2017).

Harley JB (1989) Deconstructing the Map. *Cartographica: The International Journal for Geographic Information and Geovisualization*, University of Toronto Press 26(2): 1–20. Available from: http://www.utpjournals.press/doi/abs/10.3138/E635-7827-1757-9T53 (accessed 16 June 2015).

Harris T and Weiner D (1998) Empowerment, Marginalization, and "Community-integrated" GIS. *Cartography and Geographic Information Science* 25(2): 67–76. Available from: http://openurl.ingenta.com/content/xref?genre=article&issn=1523-0406&volume=25&issue=2&spage=67 (accessed 5 October 2016).

Harvey F and Chrisman N (1998) Boundary objects and the social construction of GIS technology. *Environment and Planning A*, SAGE Publications 30(9): 1683–1694. Available from: http://epn.sagepub.com/lookup/doi/10.1068/a301683 (accessed 5 October 2016).

Harvey F, Kwan M-P and Pavlovskaya M (2006) Introduction: Critical GIS. *Cartographica: The International Journal for Geographic Information and Geovisualization*, International Cartographic Association/Association Cartographique internationale. Available from: http://www.utpjournals.press/doi/abs/10.3138/04L6-2314-6068-43V6?journalCode=cart (accessed 15 October 2015).

Haughton D, McLaughlin M-D, Mentzer K, et al. (2015) Can We Predict Oscars from Twitter and Movie Review Data? In: Springer, Cham, pp. 41–54. Available from: http://link.springer.com/10.1007/978-3-319-09426-7_6 (accessed 5 October 2017).

Hecht B, Hong L, Suh B, et al. (2011) Tweets from Justin Bieber's heart. In: *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*, New York, New York, USA: ACM Press, p. 237. Available from: http://dl.acm.org.proxy.lib.sfu.ca/citation.cfm?id=1978942.1978976 (accessed 28 December 2015).

Hennig N (2017) *Keeping up with emerging technologies : best practices for information professionals*. Santa Barbara: Libraries Unltd Inc.

Herfort B, de Albuquerque JP, Schelhorn S-J, et al. (2014) Exploring the Geographical Relations Between Social Media and Flood Phenomena to Improve Situational Awareness. In: Springer International Publishing, pp. 55–71. Available from: http://link.springer.com/10.1007/978-3-319-03611-3_4 (accessed 15 March 2017).

Hong L, Ahmed A, Gurumurthy S, et al. (2012) Discovering geographical topics in the

twitter stream. In: *Proceedings of the 21st international conference on World Wide Web - WWW '12*, New York, New York, USA: ACM Press, p. 769. Available from: http://dl.acm.org.proxy.lib.sfu.ca/citation.cfm?id=2187836.2187940 (accessed 13 April 2016).

Huang H, Cao Y, Huang X, et al. (2014) Collective Tweet Wikification based on Semi-supervised Graph Regularization. *In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* 1: 380–389. Available from: http://www.aclweb.org/website/old_anthology/P/P14/P14-1036.pdf (accessed 9 March 2017).

Jayakody JRKC (2016) Natural language processing framework: WordNet based sentimental analyzer. *Proceedings of the International Research Symposium on Pure and Applied Sciences*, Faculty of Science, University of Kelaniya, Sri Lanka: 71. Available from: http://repository.kln.ac.lk/handle/123456789/15725 (accessed 9 March 2017).

Jockers M (2013) *Macroanalysis: Digital Methods and Literary History*. Urbana, Chicago, and Springfield: University of Illinois Press.

Johnson PA, Corbett JM, Gore C, et al. (2015) A Web of Expectations: Evolving Relationships in Community Participatory Geoweb Projects. *ACME: An International Journal for Critical Geographies*, [Okanagan University College, Dept. of Geography] 14(3): 827–848. Available from: http://142.207.145.31/index.php/acme/article/view/1235 (accessed 21 August 2017).

Jongeling R, Datta S and Serebrenik A (2015) Choosing your weapons: On sentiment analysis tools for software engineering research. In: *2015 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, IEEE, pp. 531–535. Available from: http://ieeexplore.ieee.org/document/7332508/ (accessed 11 March 2017).

Jung J-K (2007) Computer-aided qualitative gis (caq-gis) for critical researchers: an integration of quantitative and qualitative research in the geography of communities. State University of New York at Buffalo. Available from: http://dl.acm.org/citation.cfm?id=1329707 (accessed 22 June 2015).

Jung J-K (2015) Code clouds: Qualitative geovisualization of geotweets. *The Canadian Geographer / Le Géographe canadien* 59(1): 52–68. Available from: http://doi.wiley.com/10.1111/cag.12133 (accessed 22 June 2015).

Kamath KY, Caverlee J, Lee K, et al. (2013) Spatio-temporal dynamics of online memes. In: *Proceedings of the 22nd international conference on World Wide Web - WWW '13*, New York, New York, USA: ACM Press, pp. 667–678. Available from: http://dl.acm.org/citation.cfm?doid=2488388.2488447 (accessed 15 March 2017).

Kent JD and Capello HT (2013) Spatial patterns and demographic indicators of effective social media content during theHorsethief Canyon fire of 2012. *Cartography and Geographic Information Science*, Taylor & Francis 40(2): 78–89. Available from: http://www.tandfonline.com/doi/abs/10.1080/15230406.2013.776727 (accessed 13

March 2017).

Kim MC and Chen C (2015) A scientometric review of emerging trends and new developments in recommendation systems. *Scientometrics* 104(1): 239–263. Available from: http://link.springer.com/10.1007/s11192-015-1595-5 (accessed 7 December 2015).

Kitchin R (2013) Big data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography*, SAGE PublicationsSage UK: London, England 3(3): 262–267. Available from: http://dhg.sagepub.com.proxy.lib.sfu.ca/content/3/3/262.short (accessed 24 February 2016).

Kitchin R (2014) Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, SAGE PublicationsSage UK: London, England 1(1): 205395171452848. Available from: http://journals.sagepub.com/doi/10.1177/2053951714528481 (accessed 17 October 2017).

Kitchin R and Dodge M (2011) *Code/space: Software and everyday life*. Available from: https://books.google.ca/books?hl=en&lr=&id=ZHez2BXgIeQC&oi=fnd&pg=PP1&dq=kitchin+code+space&ots=oQ9hN-69Qr&sig=Vzdk56XWAr4ZBBdrjznIy-5WV40 (accessed 14 June 2017).

Kling F and Pozdnoukhov A (2012) When a city tells a story. In: *Proceedings of the 20th International Conference on Advances in Geographic Information Systems - SIGSPATIAL '12*, New York, New York, USA: ACM Press, p. 482. Available from: http://dl.acm.org/citation.cfm?doid=2424321.2424395 (accessed 22 November 2016).

Knigge L and Cope M (2006) Grounded visualization: integrating the analysis of qualitative and quantitative data through grounded theory and visualization. *Environment and Planning A*, Pion Ltd, London 38(11): 2021–2037. Available from: http://ideas.repec.org/a/pio/envira/v38y2006i11p2021-2037.html (accessed 13 August 2015).

Ko MN, Cheek GP, Shehab M, et al. (2010) Social-Networks Connect Services. *Computer* 43(8): 37–43. Available from: http://ieeexplore.ieee.org/document/5551044/ (accessed 2 February 2017).

Korayem M, Aljadda K and Crandall D (2016) Sentiment/subjectivity analysis survey for languages other than English. *Social Network Analysis and Mining*, Springer Vienna 6(1): 75. Available from: http://link.springer.com/10.1007/s13278-016-0381-6 (accessed 5 October 2017).

Krzywicki A, Wobcke W, Bain M, et al. (2016) Data mining for building knowledge bases: techniques, architectures and applications. *The Knowledge Engineering Review* 31(2): 97–123. Available from: http://www.journals.cambridge.org/abstract_S0269888916000047 (accessed 9 March 2017).

Kwan M-P (2016) Algorithmic Geographies: Big Data, Algorithmic Uncertainty, and the

Production of Geographic Knowledge. *Annals of the American Association of Geographers*, Routledge. Available from: http://www.tandfonline.com.proxy.lib.sfu.ca/doi/full/10.1080/00045608.2015.1117937 (accessed 24 February 2016).

Kwan M-P and Ding G (2008) Geo-Narrative: Extending Geographic Information Systems for Narrative Analysis in Qualitative and Mixed-Method Research∗ . *The Professional Geographer*, Taylor & Francis Group 60(4): 443–465. Available from: http://www.tandfonline.com/doi/abs/10.1080/00330120802211752 (accessed 22 June 2015).

Landauer TK, Landauer and K T (2006) Latent Semantic Analysis. In: *Encyclopedia of Cognitive Science*, Chichester: John Wiley & Sons, Ltd. Available from: http://doi.wiley.com/10.1002/0470018860.s00561 (accessed 5 October 2017).

Lane R (2017) *The big humanities: Digital humanities/digital laboratories*. 1st ed. New York: Routlege. Available from: https://books.google.ca/books?hl=en&lr=&id=dAecDQAAQBAJ&oi=fnd&pg=PT6&dq=the+big+humanities+digital+humanities+/+digital+libraries+richard+lane&ots=aRGo4OuS4l&sig=rksMN4Gpxrj6e4GVIaYCVoGpe_A (accessed 5 May 2017).

Larsen ME, Boonstra TW, Batterham PJ, et al. (2015) We Feel: Mapping Emotion on Twitter. *IEEE journal of biomedical and health informatics* 19(4): 1246–52. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25700477 (accessed 16 September 2015).

Latour B (1987) *Science in action: How to follow scientists and engineers through society*. Available from: https://books.google.ca/books?hl=en&lr=&id=sC4bk4DZXTQC&oi=fnd&pg=PA19&dq=science+in+action+latour&ots=WamKCqbdUz&sig=IPDzigbxd-_AfLQ7OD6kwpWZqUk (accessed 14 June 2017).

Lee J-G and Kang M (2015) Geospatial Big Data: Challenges and Opportunities. *Big Data Research* 2(2): 74–81.

Li L, Goodchild MF and Xu B (2013) Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science*, Taylor & Francis 40(2): 61–77. Available from: http://www.tandfonline.com/doi/abs/10.1080/15230406.2013.777139 (accessed 13 March 2017).

Li M, Zang S, Zhang B, et al. (2014) A Review of Remote Sensing Image Classification Techniques: the Role of Spatio-contextual Information. *European Journal of Remote Sensing*, Taylor & Francis 47(1): 389–411. Available from: https://www.tandfonline.com/doi/full/10.5721/EuJRS20144723 (accessed 16 October 2017).

Liben-Nowell D and Kleinberg J (2007) The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, Wiley Subscription Services, Inc., A Wiley Company 58(7): 1019–1031. Available from: http://doi.wiley.com/10.1002/asi.20591 (accessed 5 October 2017).

Lichman M and Smyth P (2014) Modeling human location data with mixtures of kernel densities. In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*, New York, New York, USA: ACM Press, pp. 35–44. Available from: http://dl.acm.org/citation.cfm?doid=2623330.2623681 (accessed 15 March 2017).

Lim KW, Chen C and Buntine W (2016) Twitter-Network Topic Model: A Full Bayesian Treatment for Social Network and Text Modeling. Available from: http://arxiv.org/abs/1609.06791 (accessed 11 March 2017).

Liu X, Li Y, Wu H, et al. (2013) Entity Linking for Tweets. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* 1(1): 1304–1311. Available from: http://www.aclweb.org/old_anthology/P/P13/P13-1128.pdf (accessed 9 March 2017).

Liu Y, Ester M, Hu B, et al. (2015) Spatio-Temporal Topic Models for Check-in Data. In: *2015 IEEE International Conference on Data Mining*, IEEE, pp. 889–894. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7373407 (accessed 13 April 2016).

Lu Y, Dong R and Smyth B (2016) Context-Aware Sentiment Detection from Ratings. In: *Research and Development in Intelligent Systems XXXIII*, Cham: Springer International Publishing, pp. 87–101. Available from: http://link.springer.com/10.1007/978-3-319-47175-4_6 (accessed 5 October 2017).

Marres N and Weltevrede E (2013) Scraping the Social? *Journal of Cultural Economy*, Taylor & Francis Group 6(3): 313–335. Available from: http://www.tandfonline.com/doi/abs/10.1080/17530350.2013.772070 (accessed 17 October 2017).

Martin ME and Schuurman N (2017) Area-Based Topic Modeling and Visualization of Social Media for Qualitative GIS. *Annals of the American Association of Geographers*: 1–12. Available from: https://www.tandfonline.com/doi/full/10.1080/24694452.2017.1293499 (accessed 2 May 2017).

McCallum A (2013) MALLET: A Machine Learning for Language Toolkit. Available from: http://mallet.cs.umass.edu (accessed 2 October 2017).

McNutt M (2016) Due process in the Twitter age. *Science* 352(6284). Available from: http://science.sciencemag.org/content/352/6284/387/tab-pdf (accessed 1 May 2017).

Mei Q, Cai D, Zhang D, et al. (2008) Topic modeling with network regularization. In: *Proceeding of the 17th international conference on World Wide Web - WWW '08*, New York, New York, USA: ACM Press, p. 101. Available from: http://dl.acm.org/citation.cfm?id=1367497.1367512 (accessed 7 May 2015).

Miller CC (2006) A Beast in the Field: The Google Maps Mashup as GIS/2. *Cartographica: The International Journal for Geographic Information and Geovisualization*, International Cartographic Association/Association

Cartographique internationale. Available from: http://www.utpjournals.press/doi/full/10.3138/J0L0-5301-2262-N779 (accessed 13 August 2015).

Miller GA, Beckwith R, Fellbaum C, et al. (1990) Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, Oxford University Press 3(4): 235–244. Available from: https://academic.oup.com/ijl/article-lookup/doi/10.1093/ijl/3.4.235 (accessed 9 March 2017).

Miller HJ (2010) The data avalanche is here. Shouldn't we be digging? *Journal of Regional Science*, Blackwell Publishing Inc 50(1): 181–201. Available from: http://doi.wiley.com/10.1111/j.1467-9787.2009.00641.x (accessed 5 July 2017).

Mitchell L, Frank MR, Harris KD, et al. (2013) The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place. Sánchez A (ed.), *PLoS ONE*, Public Library of Science 8(5): e64417. Available from: http://dx.plos.org/10.1371/journal.pone.0064417 (accessed 11 March 2017).

Monmonier M (1996) *How to Lie with Maps*. 2nd ed. University of Chicago Press. Available from: https://books.google.com/books?hl=en&lr=&id=7pHeBQAAQBAJ&pgis=1 (accessed 26 February 2016).

Morphet CS (1993) The mapping of small-area census data -- a consideration of the role of enumeration district boundaries. *Environment and Planning A*, SAGE Publications 25(9): 1267–1277. Available from: http://epn.sagepub.com.proxy.lib.sfu.ca/content/25/9/1267.abstract (accessed 23 February 2016).

Morse JM (1995) The Significance of Saturation. *Qualitative Health Research* 5(2): 147–149. Available from: http://qhr.sagepub.com/cgi/doi/10.1177/104973239500500201 (accessed 8 March 2017).

Neff G and Nagy P (2016) Automation, Algorithms, and Politics| Talking to Bots: Symbiotic Agency and the Case of Tay. *International Journal of Communication*, University of Southern California's Annenberg Center for Communication 10(0): 17. Available from: http://ijoc.org/index.php/ijoc/article/view/6277 (accessed 11 March 2017).

NumPy (n.d.). Available from: http://www.numpy.org/ (accessed 25 February 2016).

O'Reilly T and Battelle J (2009) Web Squared: Web 2.0 Five Years On. In: *Web 2.0 Summit*, San Franciscio: O'Reilly Media Inc., pp. 1–15. Available from: https://assets.conferences.oreilly.com/1/event/28/web2009_websquared-whitepaper.pdf (accessed 11 October 2017).

Ohmura M, Kakusho K and Okadome T (2014) Tweet Sentiment Analysis with Latent Dirichlet Allocation. *International Journal of Information Retrieval Research*, IGI Global 4(3): 66–79. Available from: http://www.igi-global.com.proxy.lib.sfu.ca/article/tweet-sentiment-analysis-with-latent-dirichlet-

allocation/127002 (accessed 7 December 2015).

Okolloh O (2009) Ushahidi or 'testimony': Web 2.0 tools for crowdsourcing crisis information. *Participatory Learning and Action* (59: Change at hand: Web 2.0 for development): 65–68.

Padmanabhan A, Wang S, Cao G, et al. (2014) FluMapper: A cyberGIS application for interactive analysis of massive location-based social media. *Concurrency and Computation: Practice and Experience* 26(13): 2253–2265. Available from: http://doi.wiley.com/10.1002/cpe.3287 (accessed 1 November 2016).

Páez A and Scott DM (2005) Spatial statistics for urban analysis: A review of techniques with examples. *GeoJournal* 61(1): 53–67. Available from: http://link.springer.com/10.1007/s10708-005-0877-5 (accessed 28 December 2015).

Pang B and Lee L (2008) Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, Now Publishers, Inc. 2(1–2): 1–135. Available from: http://www.nowpublishers.com/article/Details/INR-011 (accessed 5 October 2017).

Pavlovskaya M (2009) Non-quantitative GIS. *Qualitative GIS: A mixed methods approach*. Available from: https://books.google.ca/books?hl=en&lr=&id=vnAbJ8spyeYC&oi=fnd&pg=PA13&dq=related:PzvHw4maVZwJ:scholar.google.com/&ots=gQLctEzBKU&sig=chxuEwnLgm19IEmDsXtGXXhFy8M (accessed 29 November 2016).

Perrin A (2015) *Social Media Usage: 2005-2015*. Available from: http://www.pewinternet.org/2015/10/08/social-networking-usage-2005-2015/.

Pickles J (1995) *Ground truth: The social implications of geographic information systems*. Available from: https://books.google.ca/books?hl=en&lr=&id=8ER-jC1VB90C&oi=fnd&pg=PA1&dq=pickles++ground+truth&ots=rXUClIt7os&sig=nkBP6Y63l0Ffrk-Rw5q0rCkrTEM (accessed 5 October 2016).

Pickles John (1995) *Ground Truth: The Social Implications of Geographic Information Systems*. Guilford Press. Available from: https://books.google.com/books?hl=en&lr=&id=8ER-jC1VB90C&pgis=1 (accessed 16 June 2015).

Poorthuis A and Zook M (2015) Small Stories in Big Data: Gaining Insights From Large Spatial Point Pattern Datasets. *Cityscape: A Journal of Policy Development and Research @BULLET* 17(1).

PostGIS (n.d.). Available from: http://postgis.net/ (accessed 25 February 2016).

PostgreSQL (n.d.). Available from: http://www.postgresql.org/ (accessed 25 February 2016).

Pycairo (n.d.). Available from: http://cairographics.org/pycairo/ (accessed 25 February 2016).

Rebele T, Suchanek F, Hoffart J, et al. (2016) YAGO: A Multilingual Knowledge Base from Wikipedia, Wordnet, and Geonames. In: Springer, Cham, pp. 177–185. Available from: http://link.springer.com/10.1007/978-3-319-46547-0_19 (accessed 9 March 2017).

Redonnet B, Chollet A, Fombonne E, et al. (2012) Tobacco, alcohol, cannabis and other illegal drug use among young adults: The socioeconomic context. *Drug and Alcohol Dependence* 121(3): 231–239.

Řehůřek R and Sojka P (n.d.) Software Framework for Topic Modelling with Large Corpora. University of Malta. Available from: http://is.muni.cz/publication/884893/en (accessed 25 February 2016).

Richards L (1999) *Using NVIVO in Qualitative Research*. SAGE Publications. Available from: https://books.google.com/books?hl=en&lr=&id=foc2sdg0wa8C&pgis=1 (accessed 13 August 2015).

Ricker B (2017) Reflexivity, Positionality and Rigor in the Context of Big Data Research. In: Eckert J (ed.), *Thinking Big Data in Geography: New Regimes, New Research*, University of Iowa Press, pp. 96–118. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2911652 (accessed 16 October 2017).

Robertson C, McLeman R and Lawrence H (2015) Winters too warm to skate? Citizen-science reported variability in availability of outdoor skating in Canada. *The Canadian Geographer / Le Géographe canadien* 59(4): 383–390. Available from: http://doi.wiley.com/10.1111/cag.12225 (accessed 24 February 2016).

Rougier N (2009) Wordle.py. Available from: http://www.labri.fr/perso/nrougier/downloads/wordle.py (accessed 25 February 2016).

Rybarczyk G and Melis G (2017) Linking Human Health and Neighborhood Using Big Data: A Case Study in Torino, Italy. *Journal of Transport & Health*, Elsevier 5: S74–S75. Available from: http://www.sciencedirect.com/science/article/pii/S2214140517303833 (accessed 5 October 2017).

Sakaki T, Okazaki M and Matsuo Y (2010) Earthquake shakes Twitter users. In: *Proceedings of the 19th international conference on World wide web - WWW '10*, New York, New York, USA: ACM Press, p. 851. Available from: http://portal.acm.org/citation.cfm?doid=1772690.1772777 (accessed 15 March 2017).

Schuurman N (2000) Trouble in the heartland: GIS and its critics in the 1990s. *Progress in Human Geography* 24(4): 569–590. Available from: http://phg.sagepub.com/content/24/4/569.short (accessed 27 May 2015).

Schuurman N and Leszczynski A (2006) Ontology-Based Metadata. *Transactions in GIS*, Blackwell Publishing Ltd 10(5): 709–726. Available from: http://doi.wiley.com/10.1111/j.1467-9671.2006.01024.x (accessed 22 June 2015).

Shah H, Warwick K, Vallverdú J, et al. (2016) Can machines talk? Comparison of Eliza with modern dialogue systems. *Computers in Human Behavior*, Elsevier Science Publishers B. V. 58(C): 278–295. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0747563216300048 (accessed 9 March 2017).

Shelton T, Poorthuis A and Zook M (2015) Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. *Landscape and Urban Planning* 142: 198–211. Available from: http://www.sciencedirect.com/science/article/pii/S0169204615000523 (accessed 11 May 2015).

Sieber R (2004) Rewiring for a GIS/2. *Cartographica: The International Journal for Geographic Information and Geovisualization*, University of Toronto Press 39(1): 25–39. Available from: http://www.utpjournals.press/doi/abs/10.3138/T6U8-171M-452W-516R (accessed 16 June 2015).

Sieber R and Johnson P (2013) Situating the Adoption of VGI by Government. In: Sui D, Elwood S, and Goodchild M (eds), *Crowdsourcing Geographic Knowledge*, Dordrecht: Springer Netherlands. Available from: http://www.springerlink.com/index/10.1007/978-94-007-4587-2 (accessed 3 September 2014).

Sieber RE and Haklay M (2015) The epistemology(s) of volunteered geographic information: a critique. *Geo: Geography and Environment* 2(2): 122–136. Available from: http://doi.wiley.com/10.1002/geo2.10 (accessed 11 October 2017).

Slingsby A, Dykes J, Wood J, et al. (2007) Interactive Tag Maps and Tag Clouds for the Multiscale Exploration of Large Spatio-temporal Datasets. In: *2007 11th International Conference Information Visualization (IV '07)*, IEEE, pp. 497–504. Available from: http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=4272027 (accessed 22 June 2015).

Smith B and Fellbaum C (2004) Medical WordNet. In: *Proceedings of the 20th international conference on Computational Linguistics - COLING '04*, Morristown, NJ, USA: Association for Computational Linguistics, p. 371–es. Available from: http://portal.acm.org/citation.cfm?doid=1220355.1220409 (accessed 17 October 2017).

Smith N (1992) History and philosophy of geography: real wars, theory wars. *Progress in Human Geography*, Sage PublicationsSage CA: Thousand Oaks, CA 16(2): 257–271. Available from: http://journals.sagepub.com/doi/10.1177/030913259201600208 (accessed 13 October 2017).

Socher R, Perelygin A, Wu JY, et al. (2013) Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of the conference on empirical methods in natural language processing (EMNLP)* 1631: 1642. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.383.1327&rep=rep1&type=pdf (accessed 11 March 2017).

Steiger E, de Albuquerque JP and Zipf A (2015) An Advanced Systematic Literature Review on Spatiotemporal Analyses of Twitter Data. *Transactions in GIS*: n/a-n/a. Available from: http://doi.wiley.com/10.1111/tgis.12132 (accessed 16 September 2015).

Steiger E, Westerholt R, Resch B, et al. (2015) Twitter as an indicator for whereabouts of people? Correlating Twitter with UK census data. *Computers, Environment and Urban Systems* 54: 255–265. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0198971515300181 (accessed 15 March 2017).

Stephens M (2013a) Gender and the GeoWeb: divisions in the production of user-generated cartographic information. *GeoJournal*, Springer Netherlands 78(6): 981–996. Available from: http://link.springer.com/10.1007/s10708-013-9492-z (accessed 14 September 2016).

Stephens M (2013b) The Geography of Hate. *Floating Sheep*. Available from: http://www.floatingsheep.org/2013/05/hatemap.html (accessed 2 October 2017).

Stephens M and Poorthuis A (2015) Follow thy neighbor: Connecting the social and the spatial networks on Twitter. *Computers, Environment and Urban Systems* 53: 87–95. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0198971514000726 (accessed 15 March 2017).

Suchanek F and Weikum G (2013a) Knowledge harvesting from text and Web sources. In: *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, IEEE, pp. 1250–1253. Available from: http://ieeexplore.ieee.org/document/6544916/ (accessed 9 March 2017).

Suchanek F and Weikum G (2013b) Knowledge harvesting in the big-data era. In: *Proceedings of the 2013 international conference on Management of data - SIGMOD '13*, New York, New York, USA: ACM Press, p. 933. Available from: http://dl.acm.org/citation.cfm?doid=2463676.2463724 (accessed 8 March 2017).

Sui D and Goodchild M (2011) The convergence of GIS and social media: challenges for GIScience. *International Journal of Geographical Information Science*, Taylor & Francis 25(11): 1737–1748. Available from: http://www.tandfonline.com/doi/abs/10.1080/13658816.2011.604636 (accessed 27 June 2017).

Sukhija N, Tatineni M, Brown N, et al. (2016) Topic Modeling and Visualization for Big Data in Social Sciences. In: *2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld)*, IEEE, pp. 1198–1205. Available from: http://ieeexplore.ieee.org/document/7816979/ (accessed 11 March 2017).

Summerfield M and Mark (2008) *Rapid GUI programming with Python and Qt : the definitive guide to PyQt programming*. Prentice Hall. Available from: https://dl.acm.org/citation.cfm?id=1407353 (accessed 5 October 2017).

Teh Y and Jordan M (2010) Hierarchical Bayesian nonparametric models with applications. *Bayesian nonparametrics*. Available from: https://books.google.ca/books?hl=en&lr=&id=0GUzMF59AsgC&oi=fnd&pg=PA158&dq=Hierarchical+Bayesian+nonparametric+models+with+applications&ots=SVvTNMLFVW&sig=e0JOQgVzAs8_Kc_JR2PLAOfRJ6w (accessed 11 March 2017).

Tomlinson RF and Boyle AR (1981) The State Of Development Of Systems For Handling Natural Resources Inventory Data. *Cartographica: The International Journal for Geographic Information and Geovisualization*, University of Toronto Press 18(4): 65–95. Available from: http://utpjournals.press/doi/10.3138/7262-N455-7101-5347 (accessed 2 February 2017).

Tsou M-H, Yang J-A, Lusher D, et al. (2013) Mapping social activities and concepts with social media (Twitter) and web search engines (Yahoo and Bing): a case study in 2012 US Presidential Election. *Cartography and Geographic Information Science*, Taylor & Francis 40(4): 337–348. Available from: http://www.tandfonline.com/doi/abs/10.1080/15230406.2013.799738 (accessed 15 March 2017).

Twitter (2016) *Annual Report 2016*. San Fransiscio. Available from: http://files.shareholder.com/downloads/AMDA-2F526X/5372398874x0x935049/05E6E71E-D609-4A17-A8BD-B621324A950D/TWTR_2016_Annual_Report.pdf.

van Oers JA, Bongers IM, van de Goor LA, et al. (1999) Alcohol consumption, alcohol-related problems, problem drinking, and socioeconomic status. *Alcohol and Alcoholism* 34(1).

Varrazzo D (2010) PostgreSQL + Python | Psycopg. Available from: http://initd.org/psycopg/ (accessed 25 February 2016).

Vincent J (2016) Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day. *The Verge*, 24th March. Available from: http://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist.

Walker BB, Schuurman N and Hameed SM (2014) A GIS-based spatiotemporal analysis of violent trauma hotspots in Vancouver, Canada: identification, contextualisation and intervention. *BMJ open*, British Medical Journal Publishing Group 4(2): e003642. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24556240 (accessed 15 September 2016).

Walker BB, Schuurman N, Gašević D, et al. (2016) PT018 A Spatial Epidemiology Approach to Environmental CVD Risk: Case Studies of Food Environments and Obesity in Three Canadian Cities. *Global Heart* 11(2): e129–e130. Available from: http://linkinghub.elsevier.com/retrieve/pii/S2211816016304501 (accessed 5 July 2017).

Wang C, Wang J, Xie X, et al. (2007) Mining geographic knowledge using location aware topic model. In: *Proceedings of the 4th ACM workshop on Geographical information retrieval - GIR '07*, New York, New York, USA: ACM Press, p. 65. Available from: http://dl.acm.org/citation.cfm?id=1316948.1316967 (accessed 22

June 2015).

Wang N, Xu H and Grossklags J (2011) Third-party apps on Facebook. In: *Proceedings of the 5th ACM Symposium on Computer Human Interaction for Management of Information Technology - CHIMIT '11*, New York, New York, USA: ACM Press, pp. 1–10. Available from: http://dl.acm.org/citation.cfm?doid=2076444.2076448 (accessed 2 February 2017).

Wang X, Brown DE and Gerber MS (2012) Spatio-temporal modeling of criminal incidents using geographic, demographic, and twitter-derived information. In: *2012 IEEE International Conference on Intelligence and Security Informatics*, IEEE, pp. 36–41. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6284088 (accessed 16 September 2015).

Wangberg SC, Andreassen HK, Prokosch H-U, et al. (2008) Relations between Internet use, socio-economic status (SES), social support and subjective health. *Health Promotion International*, Centre for Comparative Social Surveys, City University, London 23(1): 70–77. Available from: https://academic.oup.com/heapro/article-lookup/doi/10.1093/heapro/dam039 (accessed 15 June 2017).

Widener MJ and Li W (2014) Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US. *Applied Geography* 54: 189–197. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0143622814001775 (accessed 11 March 2017).

Woods M, Paulus T, Atkins DP, et al. (2015) Advancing Qualitative Research Using Qualitative Data Analysis Software (QDAS)? Reviewing Potential Versus Practice in Published Studies using ATLAS.ti and NVivo, 1994-2013. *Social Science Computer Review*: 0894439315596311-. Available from: http://ssc.sagepub.com/content/early/2015/08/24/0894439315596311.abstract (accessed 15 October 2015).

Yeager CD and Steiger T (2013) Applied geography in a digital age: The case for mixed methods. *Applied Geography* 39: 1–4. Available from: http://www.sciencedirect.com/science/article/pii/S0143622812001634 (accessed 15 October 2015).

Yin J, Soliman A, Yin D, et al. (2017) Depicting urban boundaries from a mobility network of spatial interactions: a case study of Great Britain with geo-located Twitter data. *International Journal of Geographical Information Science* 31(7): 1293–1313. Available from: https://www.tandfonline.com/doi/full/10.1080/13658816.2017.1282615 (accessed 1 May 2017).

Yin Z, Cao L, Han J, et al. (2011) Geographical topic discovery and comparison. In: *Proceedings of the 20th international conference on World wide web - WWW '11*, New York, New York, USA: ACM Press, p. 247. Available from: http://dl.acm.org.proxy.lib.sfu.ca/citation.cfm?id=1963405.1963443 (accessed 13 April 2016).

Zook M and Poorthuis A (2014) Offline Brews and Online Views: Exploring the Geography of Beer Tweets. In: *The Geography of Beer*, Dordrecht: Springer Netherlands, pp. 201–209. Available from: http://link.springer.com/10.1007/978-94-007-7787-3_17 (accessed 9 March 2017).

Zook M, Graham M, Shelton T, et al. (2010) Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake. *SSRN Electronic Journal*. Available from: http://papers.ssrn.com/abstract=2216649 (accessed 22 June 2015).

Zook M, Kraak M-J and Ahas R (2015) Geographies of mobility: applications of location-based data. *International Journal of Geographical Information Science* 29(11): 1935–1940. Available from: http://www.tandfonline.com/doi/full/10.1080/13658816.2015.1061667 (accessed 4 November 2016).