Arcada Working Papers 3/2017
ISSN 2342-3064
ISBN 978-952-5260-80-9

Topic Modelling Analysis of Instagram Data for the Greater Helsinki Region*

Shuhua Liu[i] and Patrick Jansson[ii]

Abstract

In this study we explore an Instagram dataset that collected public Instagram posts and comments from the greater Helsinki region during a three months period[1]. We perform a variety of topic modelling analysis on the dataset, to grasp overall topic presence and prevalence on the social media platform. We focus on the analysis of English data in this paper, and will present the handling of Finnish and Swedish data in our next report.

Keywords: topic modelling, LDA, social media, instagram, urban development

1 INTRODUCTION

This study is our first step towards understanding our cities and region through social media content analysis. Over half of the world's population live in cities nowadays, and cities are composed of large complex systems with physical, cyber and social components. Many city authorities and city planners face various challenges in planning future developments, in deploying, maintaining and optimizing urban infrastructure. Understanding the urban dynamics, city systems and interactions has never been so important and crucial for smooth functioning of modern cities and regions.

Conventional ways for collecting data to support our understanding of cities are deemed more reliable but very labor intensive, expensive, slow, do not scale easily and often produce data that are sparse, with coarse location granularity and minimal context. On

* Funding from Helsinki Region Urban Research Program (http://www.helsinki.fi/kaupunkitutkimus/) and Arcada Foundation TUF (http://tuf.arcada.fi) are gratefully acknowledged.
[i] Arcada University of Applied Sciences, Dept. of Business Management and Analytics, [Shuhua.liu@arcada.fi]
[ii] Arcada University of Applied Sciences, Dept. of Business Management and Analytics, [Patrick.jansson@arcada.fi]
[1] Data collection by Digital Geography Lab, University of Helsinki, is gratefully acknowledged.

the other hand, modern day citizens generate and share large amounts of information about where they are and what they are doing on social media, leaving marks and notes of their interaction with the urban environment. Such social media data are sometimes biased and less reliable, but much cheaper, easy and faster to collect in massive amount as timely, geo-tagged data, with fine-grained location data, rich demographics and more context information (Tass and Hong, 2014).

It is our belief that social media presents a rich and timely source of community knowledge and information that could potentially be very valuable assets for enriching our understanding of topics in the regional development of the greater Helsinki. We propose that useful community knowledge may be extracted from social media data to complement information from conventional channels, thus enable innovative analytical approaches for understanding important issues in urban planning and regional development.

The latest breakthrough developments in AI, deep learning methods and tools have brought rapid developments in natural language technology and multimedia information processing, empowered automated tools for social media content analysis. This brings new opportunities for better making use of community knowledge in the social media channels. In this study we apply topic modelling methods and tools to explore an Instagram dataset that collected public Instagram posts and comments from the greater Helsinki region. We perform a variety of topic modelling analysis on the dataset, to grasp and reveal the presence and prevalence of different topics on the social media platform, and to investigate relevant topics related with cycling and public transportation.

2 DATA AND METHODS

2.1 Data

Data was collected from publicly accessible Instagram accounts in the Helsinki region during the summer time of 2016 (June-August), by the Digital Georgraphy Lab at University of Helsinki (https://www.helsinki.fi/en/researchgroups/digital-geography-lab). The database contains text content of the posts and comments as well as related metadata: Date, userID, number of likes, link to image content.

Our analysis so far focused on the text content extracted from the Posts and Comments fields. Language detection on the text corpus found content in 47 languages[2], the most frequent being Finnish (169,826) and English (111,157), followed by Estonian (17,415), Russian (8552) and Swedish (12,500). Posts/comments with no text or text that isn't in a specific language amounts to 22,891 entries, they are discarded from further analysis. Posts and comments are considered separate entries/documents, but can be merged ea-

[2] For language detection, we used the "langdetect" library ported to Python from Google's language detection tools. Language detection is applied to posts and comments separately, but we do not separating a single entry (post or comment) into smaller units to identify possible mixed use of multiple languages within the range of a post or comment, although we are aware that on social media people sometimes tend to mix in a word or two of some other language in their post which is otherwise in an original language.

sily at analysis. Hashtags are kept as effective content. An overview of the English, Finnish and Swedish data is given in Table 1.

Table 1: Data Overview

	Finnish	English	Swedish
# of posts	88877	75354	7221
# of comments	80949	35803	5279
Total word count	1359174	1340185	97007
Vocabulary size	260902	140275	27811
Longest post (words)	2087	1985	950
Average length posts	89	111	69
Longest comments (words)	1577	996	437
Average length comments	38	42	35

In this report, we will focus on the analysis of English data. The processing and analysis of the Finnish and Swedish data will be reported in a subsequent paper.

2.2 LDA Topic Modelling Method

Topic modeling offers a sophisticated treatment of the topic extraction problem with an unsupervised approach. Topic modelling has been widely used for tasks such as corpus exploration, document classification and information retrieval. It proves to be a powerful technique for finding hidden thematic structure in large text collections.

LDA (Latent Dirichlet Allocation) topic modeling and its variations represent the simplest and most popular methods for discovering topic structure and extracting topics from document collections. They are probabilistic models based on a hierarchical Bayesian analysis method (Blei et al, 2003; Blei, 2012). Topics are defined as a distribution over a fixed vocabulary of terms, documents are defined as a distribution over topics, with the distributions all automatically inferred from analysis of the text collection.

Given a document collection, assuming there are K topics $\beta = \beta 1$ to βK, each of which is a distribution over a fixed vocabulary of the corpus, LDA topic modelling will derive the posterior distribution (or maximum likelihood estimate) of the K topics in such a way that the language model most likely generated the documents in the collection. The K topics assume mixed memberships in each document and each document embraces multiple topics.

We applied the online LDA method by Hoffman, Blei and Bach (2010), which is implemented in Genism (https://radimrehurek.com/gensim/models/ldamodel.html). Online LDA fits topic models to massive data using an online variational Bayes (VB) algorithm for Latent Dirichlet Allocation. It can handily analyze massive document collections, including those arriving in a stream. In their study, Hoffman et al fitted a 100-topic topic model to 3.3M articles from Wikipedia in a single pass using online LDA. They

showed that online LDA can find topic models as good or better than those found with batch VB, and in a fraction of the time. We consider this a very suitable modelling tool for our purpose, and it prepares us for handling much larger dataset at a later stage.

3 TOPIC MODELLING ANALYSIS OF INSTAGRAM DATA IN ENGLISH

The English part of the Instagram data contains 75354 Posts, 35803 Comments; 1,340,185 words and a vocabulary size of 140,275, with longest posts 1985 words, on average 111 words per post; longest comments 996 words, on average 42 words per comment.

All data entries detected as in English are included in our analysis, cleaned and preprocessed. Preprocessing performed tokenization and removed stopwords[3]. Further cleaning removed unrecognised words, special symbols, as well as words with less than three characters. No stemming was done, but it's possible to switch between case sensitive or not.

Most frequent words: *thank/thanks, great, love, nice, good, like, beautiful, cool, gorgeous, shot, picture, awesome, amazing, wonderful, perfect, welcome, happy, hear, wish.* This clearly tells in general very positive tones of the English or bilingual Instagram community in the region. Negative words are only very sparsely found.

Most frequent hashtags: *#helsinki, #finland, #summer, #suomi, #vscocam, #visithelsinki, #vsco, #visitfinland, #instagood, #thisisfinland, #travel, #igscandinavia, #nature, #igfinland, #architecture, #ourfinland, #igersfinland, #photooftheday, #onnea, #sea.*

3.1 Topic Models and Topics Overviewof all English Data

We first apply online LDA modelling methods to the English corpus, with both *Posts and Comments* included (denoted as PostsComments), to obtain a general picture of all the potentially useful topics covered in the data collection. We then compare with results from LDA analysis of *Posts* data only. In order to understand the effcts of hashtags, we compare models that included *hashtags* with those that excluded hashtags.

One most important parameter in LDA topic modeling analysis is the number of topics. For our purose, we intially tested with different options: k = 2, 3, 4, 6, 8, 10, 12 , 16, 20, 26, 30, 35, 38, 40, 46, 50. When the number of topics becomes too big, there tend to be too much overlapping between many topics. We set 50 as the maximum topic number for this set of experiments.

[3] *Note:* NLTK stopwords list, with the possibility to add new stopwords; A comprehensive stopwords collection for different languages: https://github.com/stopwords-iso, https://github.com/stopwords-iso/stopwords-fi, has almost 10 times more stopwords than the NLTK list (1298 compared to 153 in NLTK for English) -processing takes much longer time.

To get a general overview of topics covered in the Instagram English data, our first set of experiments perform topic modelling analysis of *Posts* together with *Comments*, with both the hastag sign and the tags included.

A 3-topic lda model can only bring up the topics formulated by most frequent terms. It seems that the hashtag sign has a single big effect, with one topic contains only words (topic 1, the most prevalent topic), one topic contains only hashtags (topic 2, slightly less prevalent), and the third topic a mixture of these two types of data (topic 3, least prevalent). We can also notice that, topic 1 is dominated by positive adjectives, topic 2 has more useful content topic terms, topic 3 more a random mixture of both. This gives us some perspective information, but is still far from a fair representation of the rich topics on Instagram.

> Topic1: love, great, cool, nice, like, good, very, really, beautiful, awesome, know, back, shot, amazing, best, hope, miss, lovely, come, photo, haha, finland, pretty
>
> Topic 2: #helsinki, #finland, #summer, #vscocam, #vsco, #suomi, #instagood, #visithelsinki, #visitfinland, #travel, #nature, #architecture, #igscandinavia, #igersfinland, #photooftheday, #thisisfinland, #ourfinland, #love, #vscofinland, #sea, #food, #instadaily, #igfinland, #vscorussia, #vscogood, #beautiful, #picoftheday, #typicalscandinavia, #scandinaviaclub, #nordics
>
> Topic 3: babe, tack, cheers, foto, greetings, cutest, Monday, color, flowers, beach, fuck, holy, gratis, #sweden, handsome, colours, perfection, pants, #topclasstattooing, #oldlines, #realtattoos, #brightandbold, gracias, #fashion, #tattooworkers, #whipshaded, #truetraditionaltattoos, #besttradtattoos, #realtraditional

Large amounts of modeling experiments were conducted. The 50 topics model seems to generate many overlapping topics. In Figure 1[4], we present a 30 topics LDA model. As we can see, the most prevalent topic is Topic 1, which mainly concerns ”travel” in Finland, Helsinki or Scandinavia. Some other interesting topics are revealed by the model:

- topic 5: VSCO, camera, sunset and flowers;
- topic 8: food, party, music, play, #scandinaviaclub, lunch;
- topic 11: beer, dinner, swimming, breakfest, land, market;
- topic 12: weekend, food and drink, beach, design, strawberries;
- topic 13: nature, naturelovers, forest, flowers, gardens, seagull, espoo, lonlyplanet;
- topic 14: music, concert, stage, airport, money;
- topic 21: pictures, shop, watch, game, photograph, denmark;
- topic 24: gallery, evening, album, artist, exhibition;
- topic 25: flight, coffee, shops, pizza, cafe, restaurant, arizona, california, africa;
- topic 26: island, boat, ship, ride, suomenlinna, journey, bike, games, retail, drummer;
- topic 30: interetsing mix of tatoo, fish, chocolate, pride, bear, stone and stylish ☺.

[4] LDA topics are visualized using LDAvis visualization tool (Sievert and Shirley, 2014).

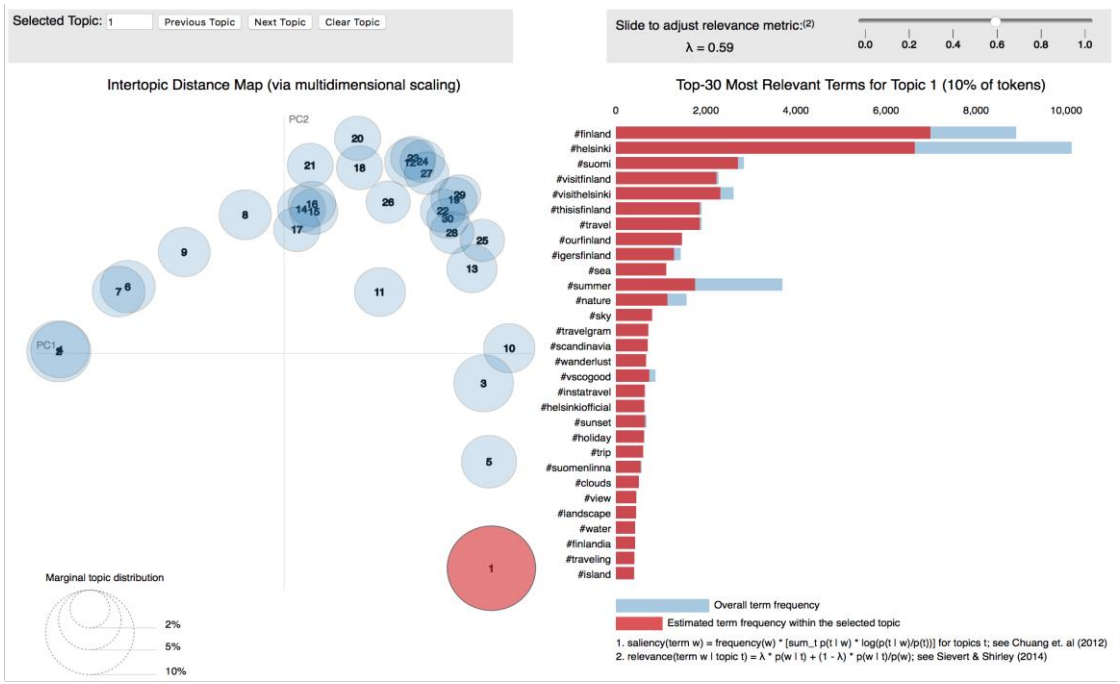

Figure 1. 30 topics lda model with LDAvis, English data, posts with comments, full hastags

3.2 Posts vs PostsComments

Next we perform topic modelling analysis using only content of Posts with hashtags, but excluding Comments related with the each post. A example 30-topics lda model is shown in Fingure 2.

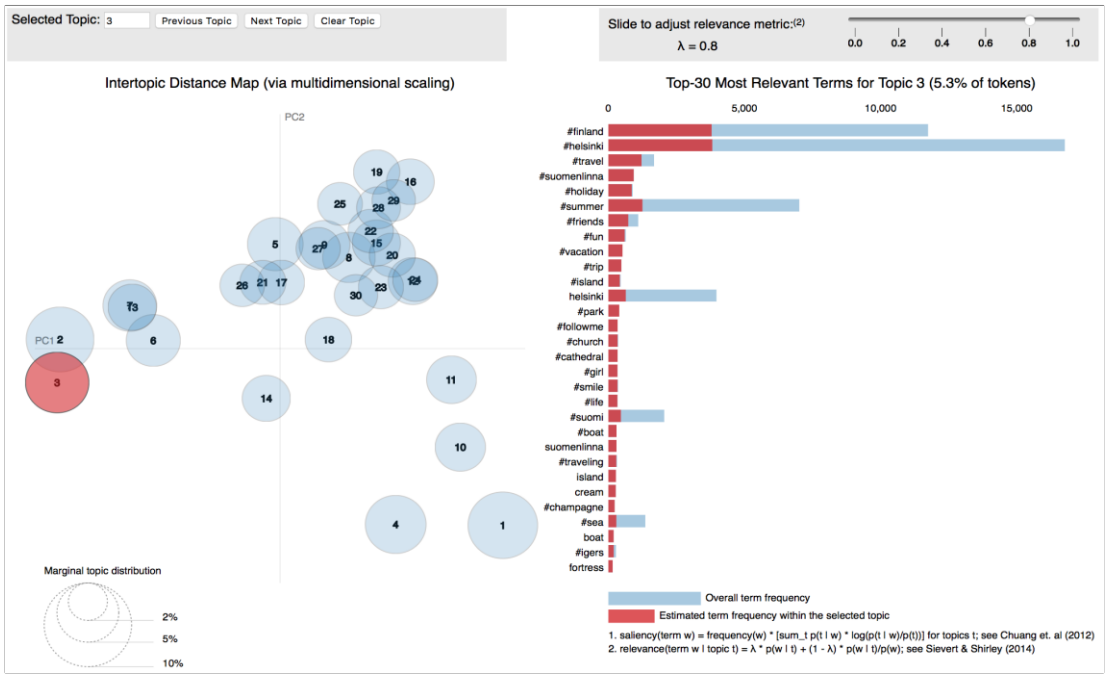

Figure 2. 30 topics lda model with LDAvis, English data, posts only, no comments, full hastags

Some topics revealed in the 30 topics model include:

- topic 2: summer, nature, sun, sky, sea, sunset, clouds, finland, helsinki, espoo, city, landscape, forest, tuska, linnamäki, trees, street, love, terrace, harbour;
- topic 3: travel, suomenlinna, friends, vacation, trip, park, island, cathedral, church, boat;
- topic 6: visithelsinki, visitfinland, ourfinland, discoverfinland, thisisfinland, weareinfinland, urban, city, beauty, bar, tour, cityscape, road, citylife, seaside, bridge, people, midnight, tram;
- topic 8: food, foodpom, lunch, instafood, fresh, restaurant, beer, vagan, salad, foodie, punavuori;
- topic 11: dinner, freind, training, pride, workour, throwback, summernight, office, oldbuildings;
- topic 14: europé, utrafinland, 1dhelsinki, wedding, otrahelsinki, tourist, elore, traveller, tourism;
- topic 16: park, fitness, gym, photographer, rsanature, vacation, cycling, bodybuilding, workout, korkeasaari, zoo;
- topic 23: work, sunshine, july, walk, cathedral, market, company, restanrant, swimming;
- topic 26: interior, hietsu, dance, hotel, töölö, capital, handmade, decor, backpacker, artist;
- topic 27: party, icecream, lauttasaari, business, design, rainyday, nikon, finnishgrils, mattolaituri, kruunuvuori, silence, baltic, urbandecay, kahvi, soul, chapel, minimalism, diesel, dogs;
- topic 28: sunday, wine, vintage, cake, ride, play, holidays, ourdoor, usa, beachlife, seashore;
- topic 30: church, rock, kids, drink, kaivopuisto, midnightsun, book, wood, olympiastadion.

We can notice that many topics of this Posts-only model seem to contain terms only loosely related to a coherent context comparing to when modelling using PostsComments. This is a bit surpriseing as we consider removing Comments should not have had much effect on the scope of topics as the majority of Comments on instagram often simply offer congratulations or other kinds of compliments, rather than bringing new topics. The modeling result could probably be explained by the reason that when removing Comments from content input it caused a considerable change to the amount of data as well as the proportion of words and hashtags in the content.

On the other hand, the Posts only models do bring up more topic terms. In addition, the best choice of topic number would change when the input text content changes.

3.3 Hashtags vs No Hashtags

Next we continue topic modeling analysis on the PostsComments data, comparing the effects of including and excluding hashtags. We have two options: (1) remove both the hash sign and the tags, so no hashtag related infomation at all in the content; (2) remove the hash sign, but keep the tags. We consider the 2nd alternative a better choice as removing all hashtag information could remove much useful topic information. In addition, it's a rather common practice that hashtags are inserted in the middle of a Post text. Removing them would mean losing information that is an important part of a Post text as well.

Our large amount of experiments show that, with the first option we indeed lose good information in understanding the topics. Keeping the hashtags as they are does bring value into the LDA model. When hashtags are included in analysis, removing the hash sign is a good alternative to keeping it. With hash sign removed, a 30-topic lda model shown in Figure 3 is more similar to Figure 1 than Figure 2.

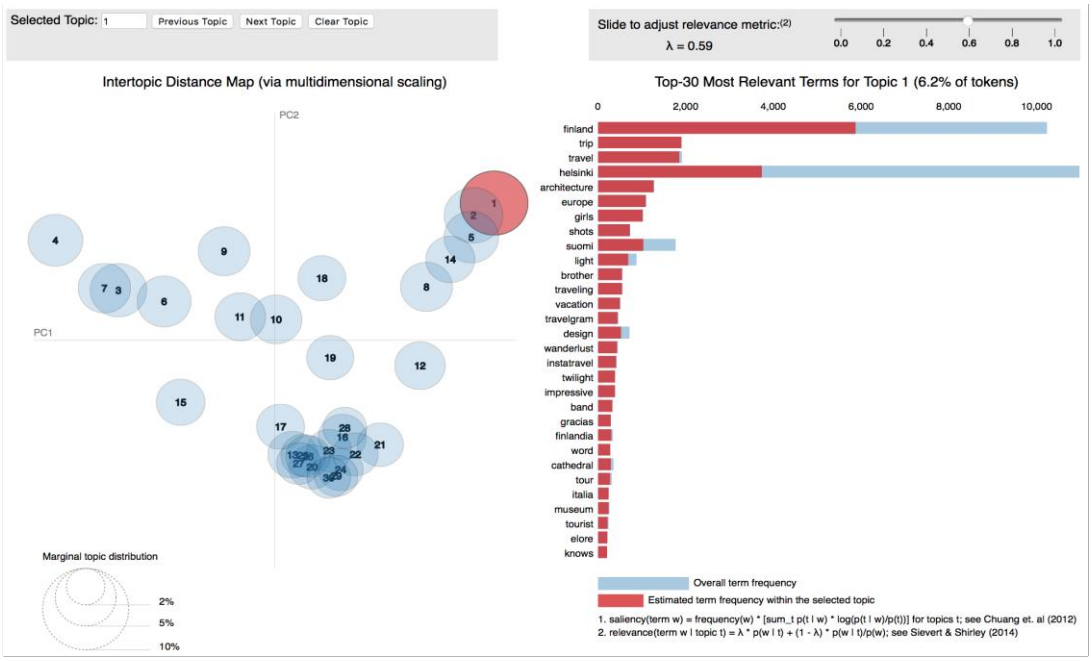

Figure 3. 30 topics lda model with LDAvis, English data, PostsComments, no hash sign, tags retained

In the 30-topic lda model, the most prevalent topic is Topic 1, which mainly concerns "trip, travel, vacation" in finland, helsinki or europe, with architecture, girls, design, band, finlandia, cathedral, museum, band, midnight, moomins, sauna, travelgram, instatravel as highly relevant terms. Some other interesting topics are revealed by the model:

- topic 5 : lovely, finnish, water, island, suomenlinna, rock, stockholm, boat, july, hate, pizza, cafe, seagual, square
- topic 12: vsco, vscocam, igersfinland, scandinaviaclub, typicalscandinavia, vscoelore, elortheearth, visitfinland, ourfinland, justgoshoot, livefolk, paint, picnic, painting
- topic 13: gallery, food, moment, dinner, artist, eloringtheearth, cake, foodporn, sushi, wine, chocholate, strawberries;
- topic 14: sunset, family, beach, vscogood, flight, nordics, scandinavia, evening;
- topic 15: hair, working, hard, office, understand, book, tickets, money, body, health;
- topic 16: work, shop, airport, inspiration, hell, weired, lake, swim, sale, solinor;
- topic 20: party, birthday, onnea, town, brand, iphone, handsome, land, buddy, quality;
- topic 21: rain, cream, flower, vegan, lights, metal, bread, icecream, garden, sports, aalto;
- topic 22: style, midsummer, outfit, eerience, topclasstatooing, realtatoos, ootd, fashion;
- topic 26: show, festival, dress, model, dreams, instamood, performance, premiere;
- topic 30: norway, norwegian, latergram, tallinn, hotel, bird, lifeofadvanture, album.

4 SUMMARY

In this study we explored the Instagram data that collected public Instagram posts and comments from the greater Helsinki region. We performed LDA topic modelling analysis on the English data to understand overall topic presence and prevalence in the data. In general, Instagram data contains large amounts of terms to convey compliments

(congratulations, thanks), excitements or other positive tone and sentiments. Removing such highly frequent non-topical terms would helpt to bring up more novel topics, esepciall when analyzing PostsComments content.

When we are mainly concerned with topics presence and prevalence on instagram, it's helpful to include both Posts and Comments for analysis, to remove the hash sign but retain the hashtags. Although we can't say for certain that Posts only approach has only adversal effect on the topic models, it would be safer to include Comments for analysis. Removing hash sign not only makes the content more coherent but also eliminates redundance in topic terms.

We can assume that the topics discovered in English data mostly represents impressions and concerns from tourists' perspective or an international perspective. The more local perspectives on our city would reside more in the data in Finnish. We should also be aware that the dataset is still limited in size and time span, and we would need to explore some larger complmentary data sources.

REFERENCES

Blei D, A. Ng and M. I. Jordan, Latent Dirichlet Allocation. *Advances in Neural Information Processing Systems.* 601-608, 2003

Blei D. Probabilistic Topic Models. *Communications of the ACM,* 55(4):77-84, 2012

Hoffman M, D. Blei and F. Bach, Online learning for Latent Dirichlet Allocation. *Advances in Neural Information Processing Systems* 23, 856-864, 2010

Sievert C. and K. Shirley, LDAVis: A Method for Visualizing and Interpreting Topics. *ACL Workshop on Interactive Language Learning, Visualization and Interfaces,* Baltimore, June 27, 2014.

Tasse Dan and Jason I. Hong, Using Social Media Data to Understand Cities. Proceedings of NSF Workshop on Big Data and Urban Informatics, 2014.