University of Zurich[UZH]

Spatial Proximity as Similarity in Geographic Space: Using Topic Modeling to Detect Spatially Related Entities and Context

Thesis August 11, 2016

Taya Goubran
of Zurich ZH, Switzerland

Student-ID: 13-757-018
taya.goubran@uzh.ch

Advisor: **Marc Novel**

Prof. Abraham Bernstein, PhD
Institut für Informatik
Universität Zürich
http://www.ifi.uzh.ch/ddis

Acknowledgements

Foremost, I would like to thank my supervisor, Marc Novel, for his endless intellectual as well as personal support and patience. I would have not been able to achieve this work without him and I could have not asked for a better advisor. I would also like to thank Prof. Dr. Abraham Bernstein for giving me the opportunity to write my thesis under his supervision as I am highly interest in his research area and admire his work. Finally, I would like to thank my friends and especially my family for their boundless moral support and bearing with me through tough times.

Zusammenfassung

In einer Zeit mit unendlichen und leicht zugänglichen Daten, liegt oft wertvolle Information versteckt in unstrukturierte Text. Hier wird ein unüberwachtes Modell für Themenerkennung verwendet um geographische und räumliche Ähnlichkeiten aus online Benutzerbewertungen zu extrahieren. Durch die Verwendung von unterschiedlichen Datensätzen und Modellparametern sind mehrere Abstraktionsstufen bezüglich geographischen Ähnlichkeiten ermittelt worden. Hotels, die die gleiche Themen zugeteilt sind, weisen geographische Ähnlichkeiten auf. Die Lage der Hotels entsprechen deren zugeteilten Themen und das Themagewicht ist proportional zur ihrer geographischen und semantischen Bedeutung. Die Betrachtung der Thema-Wörter und deren Gewicht verschafft Einsicht in geographische kontextuelle Ähnlichkeiten.

Abstract

In a time with endless and easily accessible data, valuable information is hidden in the unstructured format of text. Here, an unsupervised topic model is used is detect geospatial proximity from online user text reviews. By tuning the model parameters and using different dataset, the generated topics have shown different degrees of abstraction in terms of geographical proximity. Hotels assigned to the same topics share geographical similarities. The location of the areas formed by those hotel corresponds to the topic keywords and its size is proportional to the topic weight. The combination of keywords and weight provides insight into contextual similarities.

Contents

1 Introduction **1**

2 Probabilistic Topic Modeling **5**
2.1 Latent Dirichlet Allocation 6
2.1.1 Posterior Inference 9
2.2 Hierarchial Topic Model 11

3 The Evaluation **13**
3.1 The Experiment 13
3.1.1 The Data 13
3.1.2 The Tools 14
3.1.3 The Process 15
3.1.4 Methodology 15
3.2 Results 16
3.2.1 Keywords 17
3.2.2 Distance 19
3.2.3 Density 19
3.2.4 Similarity 20
3.2.5 Results compared to Dataset 2 21
3.2.6 Hierarchical Topic Model Results 26

4 Discussion and Limitations **31**
4.1 Discussion 31
4.2 Limitations 35

5 Future Work **37**

6 Conclusions **39**

A Appendix **45**
A.1 Variational inference 45
A.2 Keywords 47
A.3 Interpolation - Nearest neighbours 47
A.4 HLDA Tree Output Results 47

1

Introduction

In an age where most communications, businesses and every day tasks are associated with the internet, now more than ever organizing and exploring digital data in an efficient manner has become vital. The field of Natural Language Processing (NLP) revolves around computers understanding the human natural language. One of the concerns of NLP is ambiguity. For example, the word "bank" can mean either a financial institution or refer to the slopes by the river [Suryawanshi et al., 2011], which based on the context can be identified as the former or the latter. While humans can easily identify the meaning of an ambiguous word based on its context, it is more challenging for computers. Therefore understanding the given context requires the computer to solve ambiguity. Since topics can may help identify the context, I find that topic modeling provides a solution by clustering words based on the topics of the document it occurs in. So the computer would be able to identify the meaning of a word based on the topic it belongs to. Besides ambiguity, we have to deal with context extraction in NLP. For example, saying "A is near B" depends on the context of the sentence. Nearness is relative since the sentence could be "Germany is near France" (a global level) or "The shop is near the station" (local level) [Denofsky, 1976]. Because countries fall into another topic than shopping or transportation, I believe that different topics would fall into into different contexts. There is often the tendency of describing the location of some place by mentioning the known nearby places to facilitate the description. Therefore, I believe that closer places are more often mentioned together than distant ones and as mentioned by Tobler "everything is related to everything else, but near things are more related than distant things"[Tobler, 1970] . If the context can be identified, the proximity relationship between the objects can be better understood. My research is concerned with whether I am able to extract geospatial proximity relations through thematic resemblance. I am interested in automatically extracting contextual information which may enable us to find proximity relations between objects mentioned in digital documents. I refer to contextual topics as topics, which are categorized according to their identified context. My approach consists of applying topic modeling on online user hotel reviews posted by the online community from the web platform Tripadvisor [Tripadvisor, 2016]. I am using topics to cluster the hotels into groups with the aim of inferring contextual information regarding nearness and to answer the following research questions:

RQ 1 Can we infer geospatial proximity relations between entities from thematic resemblance using topic models?

RQ 2 Can we infer the context of the proximity relation using topic models?

H1.1 The hotel is near an underground station.

H1.2 The hotel is near a certain POI.

H2.1 The hotel is in an urban or rural area.

H2.2 The type and/or size of the POI can be determined

For example, if a hotel is often mentioned together with a certain POI, then it is likely that they are close to each other and if another hotel is often mentioned with the same POI, then it is also possible that the two hotels are close to each other. This may not always be the case due to transitivity limitations mentioned below. Context is implicit in textual documents and is difficult to extract automatically without human interaction [Suryawanshi et al., 2011]. Identifying proximity from text can prove to be difficult due to its dependency on context. Using NL syntax parser and Named Entity Recognition as a method for identifying objects in a document, one can find key words such as "near", "close", "adjacent" among others to extract such relations in unstructured text. This method is a pure syntactical analysis with no ability to identify the given context. The context has to be explicitly given. However, topic modeling doesn't use the syntax of the document, but considers the documents as bag of words, where the words are the only observable. Consider the following example: In document A is written "The hotel X has a great view over the river and is a walking distance to Big Ben." and in document B is written "The hotel Y is directly by Big Ben". A human can intuitively infer using context from these two statements that both hotels are not only with close proximity to each other but also to the river Thames. Classic NLP methods would require a huge knowledge base, decision trees or rules to extract the required context [Xu and Klippel, 2012]. Using transitivity, one can imply that "A is near B", "B is near C" hence "A is near C". However, such statements are vague and it is hard to decide where to cut this transitivity [Minock and Mollevik, 2013]. There comes a point where the transitive closure does not apply. To avoid raising the question of where the transitivity relation ends, I use thematic resemblance of words instead to minimize this vagueness. Please note that I do not imply that objects that fall within the same topic are more likely to be near each other, but that documents that have the same topics may have the same context. For example reviews mentioning hotels that have the topic water or river may be near the river Thames. My question is whether I can automatically identify this proximity relation using topic modeling algorithms with minimum or no human interference.

Related Work

The notion of grouping documents by similarities has been long researched in the field of information retrieval using probabilistic models. Using the Dirichlet distribution as a key element to undergo topic models is a relatively new area first explored by Blei et al. in 2003 [Blei et al., 2003]. Further expansions of the model include supervised [Mcauliffe and Blei, 2008], dynamic [Blei and Lafferty, 2006], structural [Wang et al.,

2011b] and hierarchical topic models [Blei et al., 2004]. Topic model have since been used to undergo many experiments and test hypothesis. [Yin et al., 2011] have used GPS-associated documents along with the coordinates of their authors to capture the different interests that are consistent to certain locations and find common topics across several location. [Eisenstein et al., 2010] have predicted author location by exploiting words that are highly affiliated with certain regions and areas with geographically coherent linguistic features. Hong et al. [Hong and Davison, 2010] have expanded upon Eisenstein et al. by using micro-blogging sites such as Twitter to predict the origin of authors of new previously unseen tweets. [Adams and McKenzie, 2013] use travel blog entries to find similarity of locations, that share the same topics.

So far these papers have either used topic models to detect the location of authors or to differentiate the locations based on linguistic features. This thesis is closely related to [Adams and McKenzie, 2013], however with the addition of not only detecting semantic similarity but also geographical proximity. Using hotel reviews of a single city limits the immensely semantic diversity in terms of cultures and topics by confining the vocabulary to certain words, that are expected to be discussed describing traveling. The confinement to certain topics is intentional in order to concentrate mainly on geographical topics and to detect context. The aim is not to differentiate topics by obtaining one topic of geographical features, but to split the geographical features themselves into topics providing further insight into these features. The remainder of the thesis is structured in the following manner. First the intuitive idea behind topic models is shown to grasp how topics are created. Afterwards the formal model, Latent Dirichlet Allocation (LDA) is introduced and the inference of the posterior distribution using Gibbs sampling is explained. In the end of chapter 2, hierarchical models are introduced and shown has it derives from LDA. Chapter 3 describes the experiment setup, the dataset and the evaluation methods used to analyse test the hypotheses. Also the results are compared to the hierarchical topic model. Finally the results and limitations are discussed.

2

Probabilistic Topic Modeling

The basic idea behind topic modeling is to determine the latent topical structure of documents using a generative probabilistic model [Blei et al., 2003]. Every document consists of a set of topics, which in turn consist of a set of words. For istance, if we consider a document discussing the benefits of sports, it would probably have the topic of sports and health. Words like running, cardio and stretching belong to topic sports, while words such as life expectancy, disease and fit belong to health. The distribution of topics in a document and the distribution of the words in topics are however unknown. These model variables, among others, are called latent or hidden variables and to be learned by the model. The only observable variable by the model is the words in documents.

There are multiple known models that are able to model topics. They are not confined to textual models, but can also be applied to images [Blei and Jordan, 2003]. Among them, the most basic one Latent Dirichlet Allocation (LDA), which was first introduced by [Blei et al., 2003], and is used to infer topics from documents.

Using the basics of LDA, other topic models have emerged like:

- supervised LDA, where every docuemnt is paired with a response (like good, bad, average and not driven by the terms used in the corpus) and its aim is to infer topics predictive of a certain response [Mcauliffe and Blei, 2008]
- syntactic LDA, in which syntax and semantic are factored into the model [Boyd-Graber and Blei, 2010].
- labeled LDA, considering only topics provided in the labels and only inferring their distribution in the documents [Ramage et al., 2009].
- hierarchical topic models, which organizes documents into a topical hierarchy providing different abstraction levels [Blei et al., 2004].
- structural topic modeling incorporates meta-information into the model and see its effect on the created topics (topic prevalence)found and the words effecting the construction of that topic (topic content) [Wang et al., 2011b]

In this thesis we use LDA and compare it with hierarchical models to determine if the different level of geographical context in terms of proximity can be detected.

2.1 Latent Dirichlet Allocation

LDA is the most fundamental generative probabilistic model for topic extraction of its sort using the Dirichlet distribution family. Here, it was used here to test if topic models in general are able to capture geographical topics enabling other advanced model extensions to build upon the shown results. In the following, I explain how the model works, the intuition behind it as well as the method used for topic inference, Gibbs sampling.

LDA is better than other so far existing probabilistic models such as pLSA, because it accounts for more than one topic per document,avoids overfitting, and can better estimate the probability of a new unseen documents achieving better performance results. [Hofmann, 1999] We consider the documents as a mixture of topics and the topics themselves as a mixture of words. The words are from a fixed finite vocabulary.
Note that LDA relies on the assumption of exchangeability, where the order of words in a document as well as the order of documents in the corpus is irrelevant to the model, usually referred to as bag-of-words. While this assumption simplifies the use of inference techniques and the fast analysis of large corpora, it disregards useful multi-token words [Steyvers and Griffiths, 2007]. Since the order of words such as "New York" is ignored, the word is thereby considered two different tokens and could be classified into two different topics. This results in information loss and influences the model quality.

The intuitive creation of models can be shown best using an example building on [Chen, 2011, Underwood, 2012]. Let us assume that an author produces a document by following these three steps:

- choose the length of the document
- choose the topics of the document and their distribution (which topics are in the document and their proportion)
- for each word
    1. choose a topic according to the distribution chosen in the step before
    2. choose a word according to the probability of that word being in that topic

For example, we choose to create a document with 100 words (N= 100) and choose two topics (topic 1 = food with 70% and topic 2 animals = with 30% probability). For every 10 words: 7 are taken most likely from topic 1 and the rest from topic 2. Each word from topic food is selected according to the word distribution within the food topic ("salad" is more likely than "nutmeg"). Since the assumption is that the documents is generated in this manner, we try to backtrack the process in order to infer the underlying distribution.

To learn the topics, we assume the whole corpus has fixed k topics and the unknown variables are observed and correct, except for one word. Calculate the probabilities for this one word given the current state and readjust the model. If this procedure is

computed often enough the model converges towards the correct distribution [1]. It works as follows:

1. randomly assign for each word in the document a topic
2. for each word, go through each topic t and calculate:
   a) $P(topic|document)$, the proportion of words in that document belonging to that topic
   b) $P(word|topic)$, the proportion of this word type across all documents belonging to that topic
3. the multiplication of these terms is the probability of that word belonging to that topic

If we initially assign a word to the wrong topic ("salad" assigned to animals instead of food), the model checks how many other words in that document are assigned to topic food and how often the word salad in other documents has been assigned to the animal topic. After enough iterations the model should assign the word salad to the food topic.

For a more formal explanation, [Blei et al., 2003] has given the following notation:

- a *word*: is a discrete item of the vocabulary V. The *vth* word is annotated by an indexed V-dim unit basis vector with a single value at the *vth* position is set to one, while all others are zero. The *vth* word in a vocabulary, in short, is defined as $w^v = 1$.
- a *document*: is denoted by $\mathbf{w} = (w_1,w_2,...,w_N)$ for a document of the length N.
- a *corpus*: also called collection, is denoted by $D = (\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_M)$ for a corpus of the length M.
- k: is the number of topics, assumed known and fixed given as an input parameter.

Before LDA the closest model for finding topics in documents was pLSI (probabilistic latent semantic indexing) using a similar approach [Blei et al., 2003]. However, it was not specified how the topic weights are initially set and therefore not being able to generalize the model on new unseen documents. LDA has avoided this problem by introducing the Dirichlet distribution leading to its name *Latent Dirichlet Distribution* [Steyvers and Griffiths, 2007]. Dirichlet is the conjugate prior of the multinomial distribution and is best described as a "distribution over a distribution". In our case, the document consist of a distribution over topics, which are in turn a distribution over words. Since the true topic distribution is unknown, we assume a prior k-dimensional topic distribution vector α. This hyperparameter is used as a prior weight of the topics of a given document. It is assumed that each topic is present in the corpus with a certain proportion, which avoids the disadvantages of pLSI. It is used to calculate θ, which denotes the topic proportion of each document. The figure 2.1 shows the topic distributions in the k-1 simplex for

[1] This is the intuitive explanation of Gibbs sampling. A more formal approach follows

symmetric α in (a) and asymmetric α (b). Because ours is a continuous distribution, the triangle shows a probability density, where the parameter α gives the mean/variance of the distribution [Frigyik et al., 2010]. Since there is no need to initially assume that any topic is more prevalent in a document than others, it is often however set to all topics equally. The word distribution in topics given by β is fixed and to be learned by the model. Here the parameters are summerized in an overview:

- α: is a Dirichlet prior distribution. The k-vector hyperparameter gives a priori the assumed distribution of the topics in the entire corpus before having observed any data.
- θ: is a k-vector of the topic distribution of a single document and is calculated given α.
- β is a k×V matrix. A hyperparameter giving a priori the assumed distribution for each word to a topic, denoted by $\beta_{ij} = p(w^j = 1|z^i = 1)$, where z^i is the word assignment to topic i

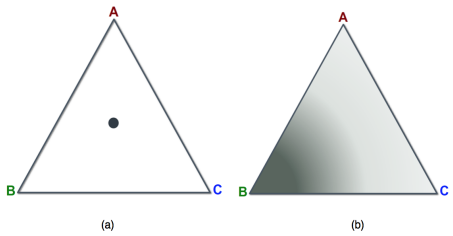
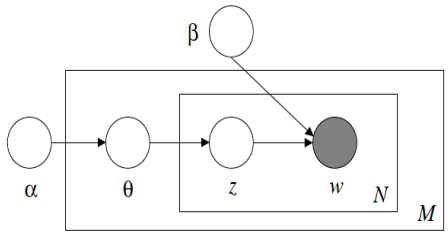


Figure 2.1: distribution of 3 topics in a 2-dim simplex (a) For k = 3, α is discrete point in a 2-simplex, where A=B=C $\frac{1}{3}$ and (b) continuous probability density, where α is higher in B than A and C [Paul, 2013]

LDA is called a generative probabilistic model based on the probabilistic process in which it generates the documents as explained in the example above. The mixture of topics can be inferred by inverting that process. Assuming the data is generated by the model, the aim is to find the most appropriate set of variables to explain the observed data [Steyvers and Griffiths, 2007]. Documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. Here is a formal approach of the algorithm explained above. For each document **w** in the corpus D [Blei et al., 2003]:

1. Choose $N \sim$ Poisson(ζ), any document according to any distribution
2. Choose $\theta \sim$ Dir(α), the topics proportion in a document according to a the topic distribution in the cor
3. For each w_n in N:



[h]

Figure 2.2: LDA graphical model

a) Choose a topic $z_n \sim$ Multinomial(θ)

b) Choose a word w_n from p($w_n \mid z_n, \beta$)

The model does not restrict a word to one topic enabling to infer polysemy, like the word *bank* can appear in both topics, topic river and topic money [Steyvers and Griffiths, 2007].

The graphical model shown in figure 2.2 illustrates the relationship between these variables. Arrows indicate dependencies and the plates are repetitive steps denoting the number of samples for each plate to generate the model. The word w is the only observable variable (shaded) in a document of length N that is a part of the collection D of size M. z is the assignment of a word w to a topic k. α is Dirichlet prior for topic distributions θ, and β is a word-topic matrix initialized with equal values to be adjusted by the model. Both are input parameters and can enhance the model quality. Given α and β, the joint distribution of the latent variables is as follows:

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta), \tag{2.1}$$

where θ is the topic mixture, z is the set of topics N, and $\mathbf{w}$ are the words in a document. This equation is intractable and the posterior distribution needs to be approximated using an inference algorithm. In LDA mostly Gibbs sampling or variational Bayes are used to approximate this distribution. In the following, Gibbs sampling is discussed.

2.1.1 Posterior Inference

The only given observable variable is the word in the document. The word-topic distribution β and the document-topic distribution θ have to be inferred. The are two common approaches for posterior inference; sampling and optimization [Hoffman et al., 2010]. Gibbs sampling is a famous sampling method based on Monte Carlo Markov Chain (MCMC), which takes a high-dimensional problem and samples the posterior variables from a low-dimensional subset of the problem [Steyvers and Griffiths, 2007]. The iterative sampling leads the model to converge. Gibbs sampling follows the intuition explained above. [Blei et al., 2003] uses variational Bayes, which tries to optimize a

parametric distribution that is close the real distribution of the posterior using Kullback-Leibler divergence . In the following, inference using Gibbs sampling is explained being the method used by the modeling tool, Mallet [McCallum, 2002]. An overview about variational Bayes is given in the appendix A.1 and a more detailed explanation can be found in [Hoffman et al., 2013, Jordan et al., 1999, Wainwright and Jordan, 2008].

The idea behind Gibbs Sampling is assuming that the model is correct for all but one word w^i. If we sample often enough and update the model after each iteration, the model will converge to the true distribution. For every word the algorithm calculates the probability of belonging to each topic k conditioned on the remaining variables being constant. Here we refer to N_{imk} as the number of tokens of type w^i in document m and assigned to topic k, N^{-st}_{imk} is the same excluding the current token $w^{i,m}$ of document m and (.) is the sum of all other values on the index. For example, $N^{-st}_{i(\cdot)_k}$ is the number of words of type w^i in all the documents, that are assigned to topic k excluding the current observed token. The conditional distribution of assigning topic k to the word i in document m is calculated (non-normalized) as follows:

$$\begin{aligned} P(z_{i_m} = k|\mathbf{z}^{-i,m}, \mathbf{x}, d, \alpha, \beta) & \\ &= p(x_{i,m}|\phi^k)p(k|d_m) \\ &\propto \frac{N^{-i,m}_{i_{(\cdot)k}} + \beta}{N^{-i,m}_{(\cdot)(\cdot)k} + W\beta} \frac{N^{-i,m}_{(\cdot)mk} + \alpha}{N_{(\cdot)m(\cdot)} + T\alpha} \end{aligned} \tag{2.2}$$

The first term denotes the probability of word w under k, whereas the second term is the probability of topic k in document d. If enough words of the type w are assigned to topic k, then the probability of the current token w^i belonging to topic k increases (left term) and if many tokens in document m belong to topic k, then the probability of any word including the current token w^i belonging to topic k increases as well (right term). In other words the probability of one word belonging to a topic k depends on how often a that word type occurs in a topic k across the corpus and how intense the topic k is present in the current document m. The intuition can be shown by this example. If we have a document discussing health and sports, and we trying to assign the word "training" to a topic. The algorithm would first look at the topic distribution of the word "training" across the whole corpus and would find that maybe $\frac{1}{2}$ of words "training" are assigned to "sports" and $\frac{1}{4}$ assigned to "health", while the rest are scattered across the remaining topic. Then it would look at the topic distribution within that document and would see that 50% of the tokens are assigned to "heath" and 50% of the tokens are assigned to "sports". So the probability of that single word "training" in document m is 0.25 for the topic "sports" and 0.125 for the topic "health". Due to the random assignments of the beginning of the algorithm, the first few iterations are discarded.

2.2 Hierarchial Topic Model

LDA is confined by its number of topics. Estimating the right number of topics can be daunting and inaccurate. In the hierarchical topic model this assumption is relaxed and the number of topics are set by the model. The hierarchical structure of this model as shown by [Blei et al., 2010] enables the detection of relationships between topics as opposed to LDA. More abstract topics are near the root while the more concrete ones are towards the leaf. The hierarchical topic model extends the probabilistic model of the Chinese Restaurant Process (CRP), which simulates the seating arrangement of M customers in a restaurant with the an endless count of infinitely large tables. Each observation corresponds to a customer entering the restaurant and sitting at one of the tables. Let's say the first customer enters the restaurant and sits on any table. The *mth* customer chooses tables according to the following distribution:

$$\begin{aligned} p(\text{occupied table i} \mid \text{previous customers}) &= \frac{m_i}{\gamma + m - 1} \\ p(\text{next unoccupied table} \mid \text{previous customers}) &= \frac{\gamma}{\gamma + m - 1} \end{aligned} \tag{2.3}$$

The seating plan here maps the arrangement in one restaurant representing only one level of the hierarchy. The nested Chinese Restaurant Process (nCRP) is an extension representing the levels in the hierarchy. Imagine that there are an endless number of restaurants with an endless number of infinitely large tables in a city. The first restaurant, representing the root, has on each table a card with the name of another restaurant, which in turn has other restaurant names on their tables, while every restaurant is only mentioned once (no loops or iterations). The first day a customer goes to the first restaurant and sits on a table chosen by the probability mentioned in equation 2.3 and reads the card mentioning another restaurant. On the second day he goes to that mentioned restaurant and again sits on a table chosen according to equation 2.3. After L days the customer would have visited L restaurants representing a path of the length L in an infinite tree. If we consider M customers following that same approach, the collection of chosen paths represent a subtree in the infinite tree. nCRP is used to express uncertainty about the L-level subtrees.

If we imagine the restaurant as a document, the table as topics and the customers as words, the nested CRP can be applied to create a hierarchical topic model. The process in [Blei et al., 2004] is as follows:

1. Let c_1 be the root topic
2. For each of level l $\in$ {2,..,L}:
   a) Draw a topic from the topic c_{l-1} using equation 2.3 and set c_l as the new topic
   b) Draw an L-dimensional topic proportion vector from θ for Dir(α)
   c) For each word $n \in$ {1,..,N}:

i. Draw $z \in \{1,..,L\}$ from Mult(θ)

ii. Draw w_n from the topic associated with topic c_z

The basic approach is similar to LDA, but an additional step is done to add the depth on each step building the tree structure. The equation 2.3 is used to choose an existing or create a new child of the current node as a topic. The nCRP is thereby used to relax the assumption of a fixed number of topic. The posterior is sampled using Gibbs sampling as well. Details about the specifics of the algorithm, inference and parameter estimation is beyond the scope of this thesis and can be found in [Blei et al., 2004, Blei et al., 2010, **?**, Wang et al., 2011a]

The core statistical details are not discussed here, because the mathematical background is not the subject of the thesis. However, the topic models have been discussed in enough detail to understand the conceptual idea of how topics are generated and their distribution in documents. It is also sufficient to understand the output of the topic modeling tool.

3

The Evaluation

In this chapter, first the data extraction process as well as the tools used to evaluate the data are explained. Different topic model outputs are then analyzed and compared using different methods, such as the topic keywords, grouping hotels with the same topic as well as comparing two hotels by their topic distributions. Finally HLDA results are compared to LDA.

3.1 The Experiment

3.1.1 The Data

Tripadvisor showed 1063 hotels, which is the entire list returned by Tripadvisor when searched for hotels in London [Tripadvisor, 2016]. For every hotel, all reviews were crawled, that were written in the English language. Usually reviews talk more or less about the rooms of the hotel, cleanliness, service and food among other common topics concerning accommodation - which are not our topic of interest. Fortunately, reviews also talk about the location of the hotel with regards to city attractions and underground stations. In order to focus more on the geographical information and discard topics about features of the hotels themselves, words, which are present in every document are removed. These words neither help in the distinction of topics nor give any geographical information about the entity. Typos are also accounted for, so words appearing two times or less are filtered out of the corpus. Documents are then considered as a bag of words, in which the order of the word is indifferent. For each hotel, metadata such as the address, latitude and longitude were extracted to pinpoint the location of the hotels on the map for reference to proximity. We have preprocessed the data to different extents hoping to manipulate the algorithm and shift its focus on different words. The remaining 800 hotels are shown in figure 3.1(black). While most hotels lie around Hyde Park and Westminster Abbey, others are scattered beyond the city center[1].

POI Points of interest (POI) were extracted in order to determine if they are a clustering criterion. For example many hotels are located near *Hyde Park* and we wish to differentiate them from ones located by the airport for instance. First the POIs have been

[1] Here, I refer to City of London, City of Westminster and Kensington as the city center due to their central location and density of attractions

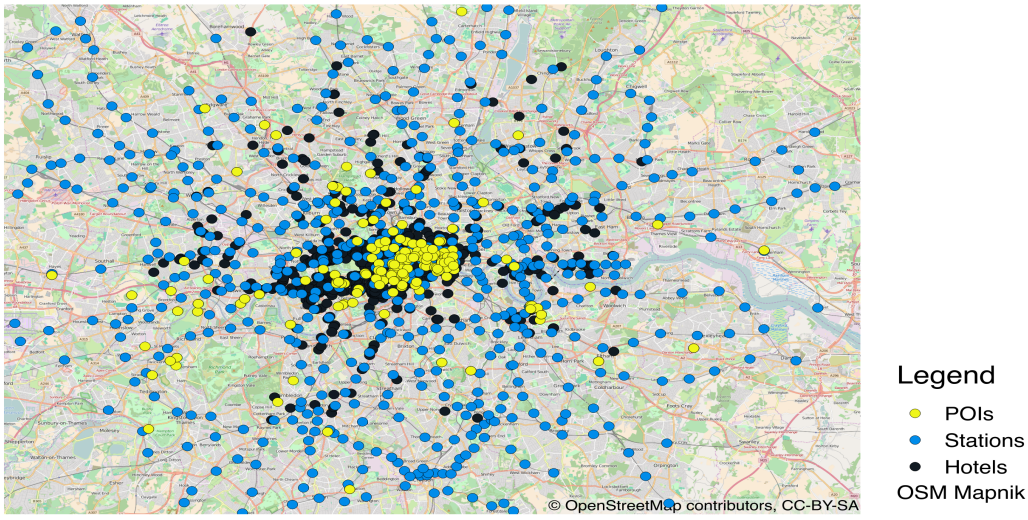


Figure 3.1: Hotels (black), POIs (yellow) and stations (blue) of London

extracted from [britainexpress.com, 2016] , along with their metadata such as address, coordinates and nearest station. However, only 181 POIs were listed by the website. Tripadvisor had a wider range of attractions under "Things To Do". Therefore the Tripadvisor corpus was favoured in order to capture more of the activities discussed by reviewers. 1207 "Things to do" were scraped, which were then fed to the Google API [Google, 2016] to get their coordinates. Using the data from Tripadvisor enables also the search for POIs that may be located a bit outside of London and does not confine the search to physical attractions points, but also incorporates experiences such as a lookout, a market or an antique road. In order to get only the relevant POIs, only the ones mentioned in the reviews were filtered out, resulting in 490 POIs.

Transportation A list of 640 stations of London together with longitude or latitude location and postal code were taken from [Bell, 2000]. Figure 3.1 shows hotels, POIs and stations layed out on an OpenStreetMap [OpenStreetMap, 2016] of London using QGIS.

3.1.2 The Tools

Scrapy The Tripadvsor data was extracted using a python library for crawling the web called Scrapy [Scrapy, 2016]. There is a Tripadvisor API, which could have been used, but it limits the reviews to 200 characters, which would have been insufficient in our case. Scrapy creates a spider that iterates over each page of the list of hotels, goes into each hotel page and extracts the reviews on all pages. Some hotel reviews are long and only a snippet of the review is shown, so by clicking on the first review on the first hotel page all reviews are automatically expanded enabling us to get the full review text. Note that

the clicked review is always the first review on each page. So starting from the second page the first review has to be ignored so that it appears only once in the dataset.

QGIS To test our hypothesis, a geographic information system was used to map the locations of the hotels, POIs and stations. It not only visualizes the proximity of two entities, but also enables us to analyze the resulting clusters from our approach. The extracted coordinates are in latitude and longitude coordinates, which were then changed to a UTM zone (30U for London) to get the distances in meters for later calculations [QGIS, 2016].

3.1.3 The Process

There are multiple libraries in different programming languages that perform topic modeling. *Gensim* is a python library - stands for generate similar [Řehůřek and Sojka, 2010]- is used to model topics according to the approach by Blei et al which is explained above [Blei et al., 2003] *Mallet* is a java library - stands for MAchine Learning for LanguagE Toolkit [McCallum, 2002]- which estimates the posterior using Gibbs Sampling according to the approach by [Steyvers and Griffiths, 2007] *topicmodels* is a R package for topic modeling, especially used for structural topic models [Bettina Grün, 2016]

3.1.4 Methodology

The documents in the dataset consist of all reviews of each hotel resulting in one document per hotel, where the reviews are the text of the documents. Since some hotels do not have many reviews, a threshold of 50 reviews has been set in order to be considered for the corpus. This limited our dataset to 800 hotels. We define a word as tokens seperated by either a space or a special character. A word count across all documents determined the frequency of the words in our vocabulary facilitating the search for very rare and very frequent words. Words that occurred in every document (occured 800 times) such as “and”,“the”,“London” etc. as well as words occuring once or twice are removed in order to account for typos. The latter also helps in cases where the reviewers deliberately write words that are not grammatically correct such as “soooo great” instead of “so great” or abbreviate like “amzg” for “amazing”.

Using the list of extracted London stations and POIs, other datasets were created. To eliminate confusion when comparing the datasets, they are enumerated and referred to as:

1. dataset 1, in which very frequent words and very rare words are discarded.
2. dataset 2, which only consists of words that are in the list of underground stations.
3. dataset 3, which only consists of words that are in the list of POIs.

Although not backed by any specific literature, I am thereby trying to shift the emphasis of the algorithm to stations and POIs respectively. Since most reviews mentions the properties of the hotel itself more in terms of service, cleanliness and facilities, and less in

terms of location, which is my quantity of interest. While the first dataset categorizes the hotels in topics such as "good", "bad", "fancy" and so forth, I was trying to emphasize the location of the hotel. For the remainder of this thesis, the datasets are treated equally following the same processes for topic modeling.

Both Gensim and Mallet have been used to perform LDA, while in this case one outperforms the other in different aspects. Gensim recognized the POIs and stations better than Mallet, but Mallet has shown better results in regards to clustering hotels according to their location, which helps verify the hypotheses. Other than finding the right parametrization for LDA, the correct number of topics is the keys for getting good topic models. We have found that the number of topic correlates with the granularity of the clustering.

Multiple models have been run with the three datasets using the default values for the hyperparameters ($\alpha = 50$ and $\beta = 0.01$). [Steyvers and Griffiths, 2007] for 3,5,10,20 and 30 topics. Mallet generates multiple files such as an evaluator and an inferencer files based on a trained model to evaluate the likelihood of a held-out testing set and infer their topics, respectively. Also a file of the topic distribution of the trainingset is stored detailing the document i, topic number and it's topic proportion θ_i. Another file lists the top keywords of every topic and their optimized Dirichlet parameter α as topic weight, which is proportional to the overall document topic proportion θ_i.

3.2 Results

The visualization for 10, 20, 30 topics needs to be on a larger scale and is shown in the appendix. For visibility reason, only the difference between 3 and 5 topic clusterings will be compared, which is sufficient to support the argumentation. The notation $x.y$ is used to refer to the yth topic of the model using x topics number. The results are first evaluated within the same dataset using different number of topics to determine the best model for clustering the hotels and afterwards compared across the different datasets with the same number of topics to highlight the difference between the datasets.

Before defining the meaning of a good cluster, we differ between the proximity of a location and the proximity of a feature. The former is a physical attraction, that is defined by coordinates in a map and the latter is defined by an abstract feature, such as "river", "area" or "building". A geographical feature is not defined by certain coordinates, but is a more general feature, that could be applied to many attractions. Therefore the meaning of a good cluster differs according to the semantics used by the model. A good cluster in terms of proximity to an attraction is one that has a high density (many hotels relative to the area) or a sub-cluster of another that provides a finer granularity. A good cluster regarding a geographical feature, however, is difficult to measure according to size or exclusivity of the cluster, but has to be evaluated according to the semantic of the assigned topic.

Increasing the number of topics often achieves finer geographical granularity. However, too many topics results in overfitting. Although the log likelihood is higher with increasing number of topics denoting better model performance, the quality of the result-

# Topics	3.00	5.00	10.00	20.00	30.00	50.00	100.00
dataset 1	-9684341.29	-8065596.17	-8003811.20	-7951543.12	-7923884.77	-7895893.90	(-7854866.80)
dataset 2	-13818.78	(-13786.03)	-13817.45	-13828.83	-13887.96	-13866.88	-13959.65
dataset 3	-18146.02	-18081.65	-17912.91	-17917.91	-17912.91	(-17884.58)	-17965.71

Table 3.1: Log likelihood overview with the highest value of each dataset in brackets

ing topics decreases in my opinion. It has been shown that log likelihood and perplexity as a measure for model performance does not necessarily correlate with human judgement [Chang et al., 2009]. In dataset 1 the topics shift the focus on more accommodation-specific topics. As for datasets 2 and 3,the highest likelihood is by 5 and 50 topics, respectively, as shown in table 3.2. The cluster hulls and nearest neighbour of the remaining models are shown in the appendix A.3.

3.2.1 Keywords

The top keywords represent the meaning of the topics. As shown in table 3.2.1 the words are grouped into 3 or 5 topics, upper and lower table, respectively. Keywords in topic 3.0 contain words that describe travelling or accommodation in general. The two other topics 3.1 and 3.2, however, mostly capture famous locations in London. Comparing the lower table, it is easy to see that topic 3.0 corresponds to topic 5.3, but upon a closer look, one can detect that topic 3.1 is roughly split into 5.0 and 5.1, and 3.2 is roughly split into topics 5.2 and 5.4.

Topic	α	top keywords
3.0	0.33224	concierge st lounge birthday square afternoon drinks dinner upgraded attentive oxford complimentary club upgrade spa delicious love superb garden
3.1	0.21339	inn premier bridge tower hilton parking buffet dlr river meal thames dinner st pm travelodge executive eye drinks pub
3.2	0.65886	paddington court hyde cross kensington heathrow st euston basement square kings museum victoria pm apartment buffet oxford pancras british

Topic	α	top keywords
0	0.15315	bridge tower st westminster eye river thames victoria waterloo ben bank trafalgar market buffet parliament paul tate liverpool hilton
1	0.21394	inn premier parking hilton dlr travelodge buffet meal wharf canary dinner pm greenwich usual excel airport kids pool westfield
2	0.23778	square st cross euston kings garden oxford pancras covent british russell theatre museum soho king leicester buffet pm eurostar
3	0.37521	concierge lounge afternoon birthday dinner drinks attentive upgraded club spa complimentary upgrade delicious love professional superb oxford appointed pool
4	0.7573	paddington court hyde kensington heathrow basement apartment earls gloucester fridge victoria pm earl bayswater dated cereal thin elevator hall

Table 3.2: Top Keywords per Topic for 3 and 5 topics

While the model is able to capture geographical locations in London, it is rather

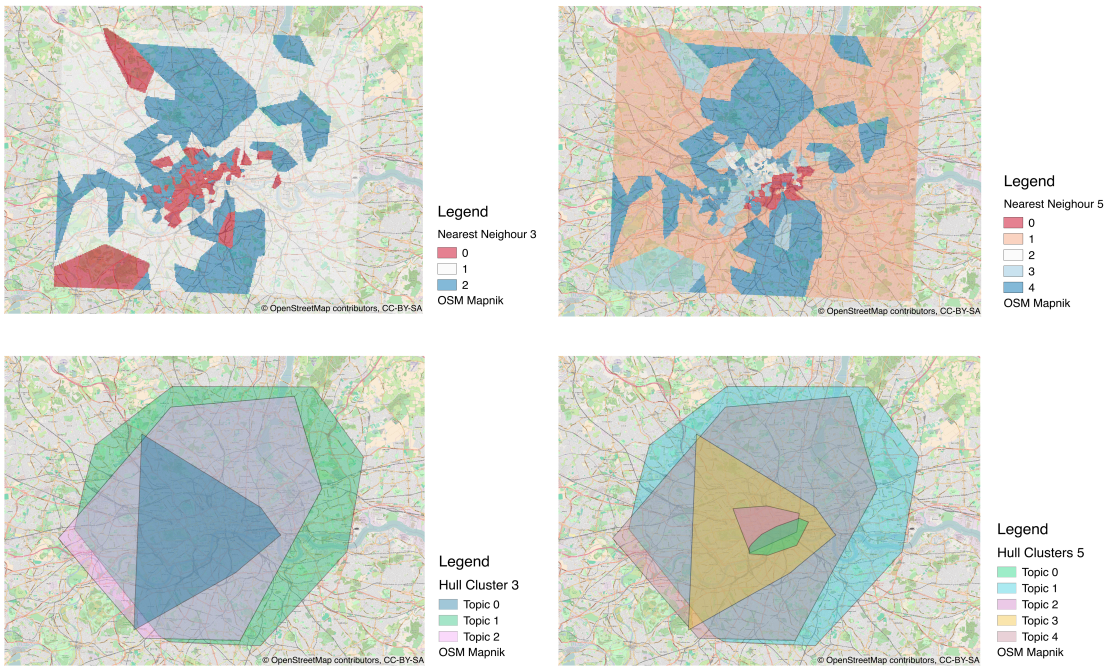


Figure 3.2: top: nearest neighbours of 3 (left) and 5 topics (right), bottom: hull clusters of 3 (left) and 5 topics (right)

difficult to access their proximity given only the keywords without being familiar with London attractions and their location. However, since that is often not the case, the location of the hotels and their distances to stations and POIs are visualized using QGIS. In order to analyze the group of hotels assigned to the same topic, they have been grouped and colour-coded according to their most dominant topic. This has been done for every topic number in each dataset. The results are 15 (3 datasets using 5 different topic numbers) different groupings of hotels have emerged.

The number of topics is proportional to the granularity of the topic semantics. With a small number of topics, some hotels are assigned to the most dominant topic despite their low distribution within the document. However, with a higher number of topics, the words are spread across more topics increasing the classification bins and thereby assigning the hotel documents to a more accurate topics.

The upper row of figure 3.2 shows an interpolated classification of 3 and 5 topics using the nearest neighbours method, classifies the unknown points by the value of their nearest neighbour. This provides a good overview of the data point, while visualizing distinct geographical areas on the map. Note that the topics are classified using a nominal scale, in which neither the numbers nor the colours have any numerical value. The bottom row visualizes a different interpretation of the data point using hulls to cluster each topic. The hull is built by connecting the outermost points of each topic. The resulting clusters visualizes the area size covered by each topic and to detect patterns. For instance,

the clusters show that 2 new, more area-defined subtopics (5.0 and 5.2) emerged from the previously existing 3 topics. Nearest neighbours show a more accurate hotel-topic classification by showing which areas of the overlapping hulls belong to which topic. The area classification also corresponds to the one given by the keywords mentioned above.

3.2.2 Distance

To measure the clusters in a more quantitative manner, the distances of hotels within a cluster are computed with the aim of detecting proximity. For each topic k, the distance between each hotel $i \in H$, with $i_k = j_k$ and $i < j$, is given by:

$$distance_k = \log \frac{\sum \text{dist}(i_k,j_k)}{\sum H_k}$$

where the *dist* value is given by a distance matrix from QGIS. The result is an accumulated distribution over distances of each cluster as shown in figure 3.3 (left). Topic 3.0 has the smallest average distance between two hotels within the same topic corresponding to its area. While topic 3.2 is roughly the same as 5.4, topics 3.1 and mostly 3.0 are spread out in topics 5.0, 5.1, 5.2 and 5.3. Topics 5.0 and 5.2 are small clusters resulting in a negative value after using the logarithm. While the area remains the same, the average distances differ from models with 3 or 5 topics, because the new clusters are sub-clusters of the existing ones and do not divide the big cluster into smaller ones, but rather take away from it's density.

Note that the average distance between 3.1 and 3.2 is quite similar (3.1 slightly larger) despite topic 3.2 having more than double the hotels in 3.1. The hotels in 3.1 are scattered across the area almost equally distributed resulting in a high total distance. To the contrary, most hotels in topic 3.2 are relatively close to each other with only a few further away resulting in large area size, but a small total distances, despite being double as much. This could be used to detect outliers in clusters. The average distance is good at detecting small clusters, but does not uncover the relation between size and quantity.

3.2.3 Density

While the distances between the hotels may be small, the density within a cluster might be low. 3 hotels are enough to form a a polygon area. If two clusters have the same area size, the one with the higher density is considered better. The density is used to abstract from the distances within a cluster and focus on the amount of hotels per area.

Talking in terms of the cluster area, polygons are built that cover the area of each topic by connecting the corresponding outer-most points in the map as shown in figure 3.2 (bottom). These hulls define an area, which is used to calculate the density of the hotels per cluster. The density of an area of topic k, is given by

$$density = \log \frac{\sum i_k}{areasize_k}$$

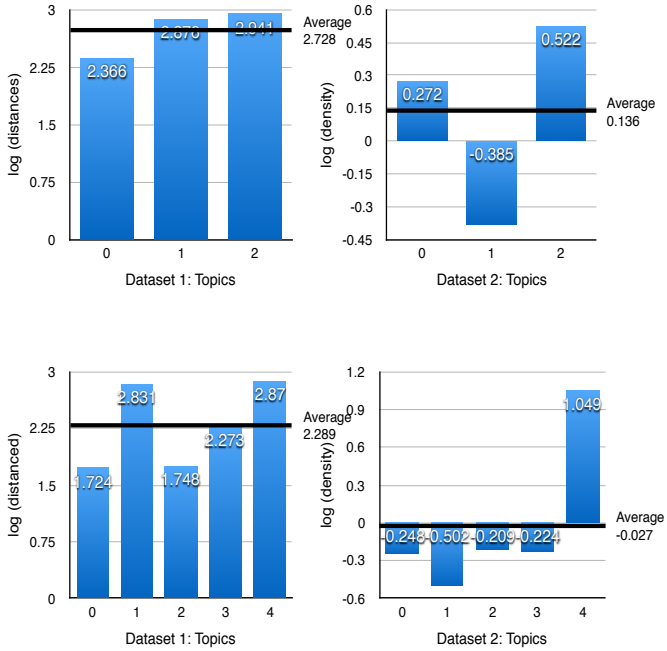


Figure 3.3: cluster distance and density of 3 (top) and 5 (bottom) topics respectively

As can be seen in figure 3.3 (right), the distance values of topics 3.0 and 3.2 differ widely, they appear to have close values in terms of density. This is due to topics 3.2 having more than double the hotels of 3.0 as well as more than double its area size. Topic 3.1, however, has the smallest amount of hotels, but occupies the largest area size with the largest distances, meaning that the hotels are widely distributed across the area with no clear agglomeration.

3.2.4 Similarity

In the field of information theory, entropy measures the uncertainty of a random variable. It describes the form of the distribution. If a distribution has extreme values, then it has low entropy value because the uncertainty of a probability outcome is low. For example, if the probability of an outcome is 100% likely to appear, then the informative value of that outcome occurring is zero. However, if the outcome is only 50% probable then the occurance of that outcome becomes more informative [Kruschke, 2010]. Relative entropy compares two probabilities by measuring their distribution distance. Relative entropy, also referred to as Kullback-Leibler divergence, is an asymmetric measure and computes the probability distribution P with respect to probability distribution Q as follows:

$$D_{KL}(P|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

with $D_{KL}(P|Q) \neq D_{KL}(Q|P)$. If the resulting distance value is small, then not

much information has been gained, because the distributions are similar [Adams and McKenzie, 2013].

Abstracting from the most dominant topic proportion of a document and looking at its total topic distribution, similarities between documents can be inferred using the KL-divergence divergence. The resulting distance could be an indicator of hotel as well as location similarities. If the distance between their distributions is small, then the difference between their topics is small, and thereby they are similar. In other words, when two hotels have the same topics with similar proportions they are more likely to be similar than those with different topic proportions. For each hotel, its 10 most similar hotels have been extracted and mapped in QGIS. While there are fewer hotels within a cluster, the area size along with the average density has decreased. KL-divergence achieved a finer cluster granularity by considering all the topic proportions of a document and not only the most dominant one as done so far. The similarity is based on the topic distribution and captures the semantic similarity of the documents. Often the hotels that are classified as similar are also close regarding their geographical location, however this is not always the case. Some similar hotels are not geographically close, but seem to share some feature, which may not be more or less obvious. For instance, a cluster of similar hotels entail the *Travelodge* and *Holiday Inn* hotels marking the hotels which are located outside the center of London, and are mostly near highways. Other examples are shown in figure 3.4:

- partitioning in east, west and center Greenwich, Dockland, Canary Wharf and London City Airport in the east (red), Wimbledon, Ealing, Wembley and Putney in the west (navy blue) and Paddington, Kensington, Westminster in the in the center (purple, light blue, green and orange)
- hotels along a highway connecting for example, north and south (dark blue)
- hotels along a railway (dark red)
- hotels along the river Thames (orange)
- hotels outside the city center (black)
- hotels in a certain district (white)

It is noticeable that similar hotels lie closer to each other, the higher the numbers of topics per model. The more topics are compared by their proportions, the more aspects (or topics) are compared by two hotels. If hotels are compared based on topics involving district, station and nearby attractions among others, then hotels sharing the same topic proportions are more likely to be close than others.

3.2.5 Results compared to Dataset 2

Dataset 1, which has been discussed so far consists of the total content of the reviews of hotels with the stopwords and very rare words removed. Here, another dataset, dataset

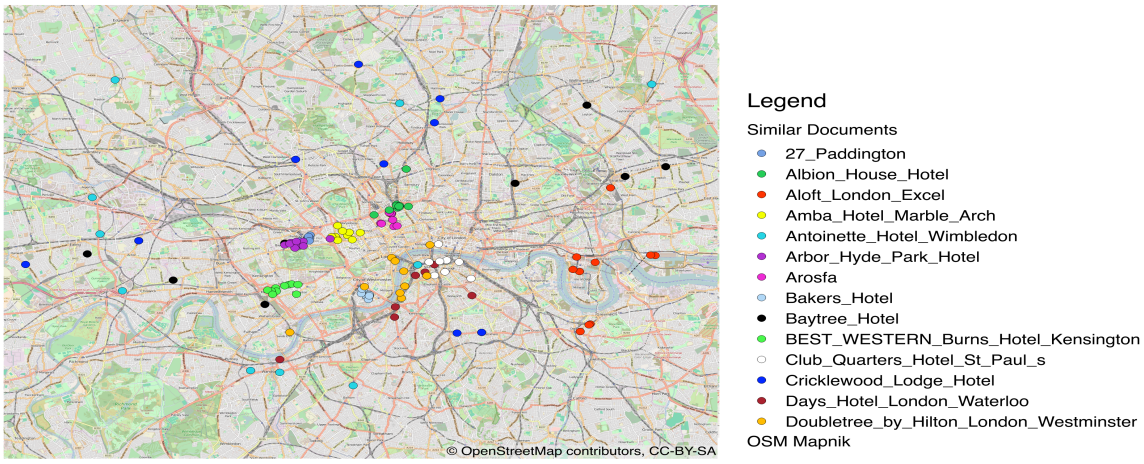


Figure 3.4: a sample of hotels that are considered similar

2, consisting of only the stations mentioned in the reviews is evaluated according the aforementioned measures and finally compared to dataset 1. Dataset 2 was created to emphasize location-related words – London stations in this case – and discard the hotel and accommodation specific vocabulary. The aim is to capture a relationship between the hotels and their neighbouring stations. It's reasonable to believe that reviewers would mention the closest station as a mean of reference to the connectivity to the rest of London. Stations are often named after an attraction close by, such as *Westminster*, *Tower Hill* or *Marble Arch* or its name entails a borough of London such as *South Kensington* or *Ealing Common*.

Without knowing the London Transportation system, the station or line names it is difficult to understand the top keywords given for every topic (see Appendix A.5). With a closer look, topic 3.2 can be considered a more general topic due to its abstract top words, such as *circus, square, street, park, station*. The hotel clusters based on their most dominant topic do not overlap as much as in dataset 1. The same pattern of a single topic spread across the map and the remaining topics form subclusters is also found here. The key difference, however, is that while dataset 1 formed clusters within each other, here the subclusters are almost mutually exclusive (they overlap slightly at the borders) as shown in figure 3.2 (bottom right).

As of the KL-divergence, the hotels that are considered similar seem to form a T-like or an arrowhead shape on the map as shown in figure 3.5. There could be many explanations, but I believe some of these patterns are clusters either along a certain railroad line. For instance:

- along the overground Southern Line (green)
- along the Docklands Light Railway (pink)
- or groups that lie along two lines with a common famous destination (blue dots

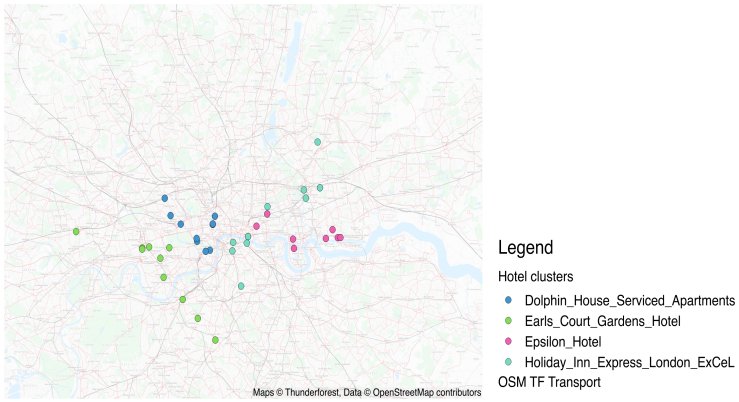


Figure 3.5: 4 clusters based on the distribution similarity of dataset 2

along Jubilee Line and Victoria Line, which intersect in Green Park)

- between two railroad lines, which probably indicate that both lines are fairly close (light blue dots between overground line and district line).

Although the idea of a main cluster and several subclusters is the same in both datasets, the way the topics are interpreted is different. Topics of dataset 1 seem to recognize the city center as a reference point and form surrounding topic clusters with varying radius. The emerged topics could be explained by the usage of new words in addition to the existing words. If one is visiting London and are accommodated in a hotel slightly outside of London, the tourists are likely to mention nearby locations in addition to the core must-visit locations in London.

However, when the dataset is limited to stations only, many city attractions – not all – and accommodation descriptive words are not considered and the vocabulary, thereby drastically decreases. Many tourists would mention the most important attractions close to the hotel as a descriptive feature. The reference of tourists is the usually the closest station and its connection to either neighbouring stations or popular centrally located stations. Although the railway lines reach from east to west and north to south, tourists are often interested in reaching the city center. The subcluster in figure 3.2 show that the stations from east and west do not overlap and the city center is either a cluster of it's own or is part of the main cluster. The interest in to the city center is not confined to the attraction sites, but also because of it's connectivity to the other destinations.

An overview of the accumulated distance and average density of both datasets in given in table 3.4. Note that the values are not logarithmized in order to get a better overview of the absolute differences between the models with differing number of topics and among the different datasets.

Dataset 3 is similar to dataset 2 in terms of preprocessing, however instead of stations, points of interest (POI) are used. It serves the purpose of trying to detect the proximity

of POIs to each other or their type/importance. Famous and important POIs would probably be mentioned in many if not all documents by at least a user or another. They may also be in topics differentiated by type, such as buildings, parks, etc. The top keywords are shown in table 3.2.5.

Topic	α	top keywords
0	0.20781	bridge london tower end garden greenwich market park covent lane wharf canary southwark museum view victoria thames grad stadium
1	0.48543	square theatre london street westminster covent garden park british end palace museum victoria tower west trafalgar leicester piccadilly national
2	0.73068	kensington park museum palace grad street victoria hyde ealing view end hill west garden notting square westminster covent station

Topic	α	top keywords
0	0.28174	street park grad square marylebone view regent west hampstead arch marble end baker paddington bond soho museum angel trafalgar
1	0.19179	bridge london tower market greenwich museum end southwark lane thames canary wharf street docklands brick bank east borough park
2	0.55248	kensington park museum street notting hill palace hyde gardens grad natural history view paddington marble arch ealing end victoria
3	1.08224	square covent london westminster garden park victoria palace end west bridge tower leicester buckingham big grad ben eye ealing
4	0.21831	theatre street national gallery square london museum british regent war shaftesbury soho bloomsbury royal avenue house tate bond piccadilly

Table 3.3: Top Keywords per Topic for 3 and 5 topics of dataset 3

While the topics give no apparent distinction between the types of the POIs, some words appear in every topic. This could indicate either the importance of these POIs or their central location. In the upper table of 3.2.5 words such as *garden park covent museum victoria* appear in each topic, which could be interpreted as generality or ambiguity.For instance, London is filled with many royal parks and gardens covering 19.75 square kilometers[Weston, 2002] and over 300 museums and galleries [Town, 2016] Ranging from Kensington in the west and Greenwich in the east, Covent Garden – which is not a garden but a district – is fairly centrally located. The keywords could either refer to the district or their more geographically specific stations.

The cluster areas formed by the most dominant topic of hotels give no clear distinction of a geographical property using few topics. The clusters are neither clusters within each other (like dataset 1) nor a main cluster with mutually exclusive subclusters (like dataset 2), but a mixture of both. The model with 3 topics seems to split London into east, west and center, but with quite a few outliers hindering a good visualization on the map. However, using a model with higher number of topics, several random clusters emerge as shown in figure 4.1. With increasing number of topics random smaller areas around the center of London are formed. The accumulated distances and density are shown in comparison to dataset 1 and dataset 2 in table 3.4.

	# Topic	Docs per Topics	Distance in km	Denstiy in $hotel/km^2$
dataset 1	3.0	159	232.49521984	1.87034845010646
	3.1	157	750.848225757	0.412033029164957
	3.2	404	872.076670359	3.32757632319388
	5.0	48	52.9078745281	0.564633494371763
	5.1	120	677.803952646	0.314929703820349
	5.2	75	56.0014182593	0.617743129305794
	5.3	128	187.3823102	0.597052279881816
	5.4	349	741.371724751	11.1911160094522
dataset 2	3.0	144	238.049291404	1.58145097253181
	3.1	48	126.5157763	0.445455280462926
	3.2	528	1272.70095918	1.07101221920089
	5.0	29	84.7709034732	0.34113273618294
	5.1	409	811.382983073	1.07338540718769
	5.2	63	175.201276609	0.518904228616867
	5.3	35	143.032026855	0.163256482780184
	5.4	184	180.410007875	5.9001872370751
dataset 3	0	129	409.723103918	0.365383884401454
	1	223	334.742467504	0.878422450385887
	2	368	730.678744784	0.774619661773524
	0	68	165.258315393	0.341096230971457
	1	97	246.693988064	0.627707143242634
	2	208	266.413054306	0.998420858079958
	3	310	840.857826381	0.588955055467435
	4	37	20.1734017164	7.46292155070935

Table 3.4: Comparing distance and density of dataset 1 dataset 2, dataset 3 (in absolute values)

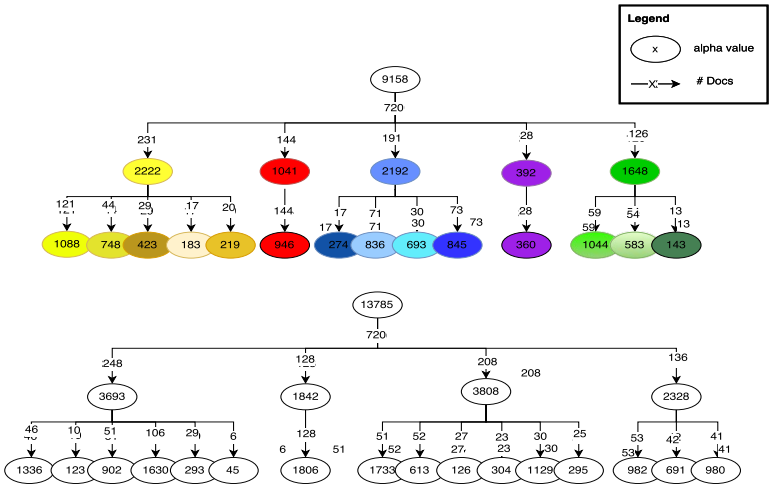


Figure 3.6: Hierarchical Topic Tree for dataset 2 (upper) and dataset 3 (lower)

3.2.6 Hierarchical Topic Model Results

In hierarchial LDA, the model creates a tree, in which each level represents a subtopic from the parent topic and each document is a path from root to leaf. The Mallet results of the hierarchical topic model differs from the one for LDA. The input is a mallet file containing the corpus and a single input parameter defining the tree depth, which is proportional to topic granularity. The higher the tree depth, the more subtopics are created. The output consists of a tree with each document as a subtree and all documents share the root node. The tree depth specifies how many topics are in a single document, while the tree breadth is determined by the algorithm itself and varies with each node. The number of children of a parent node specify in how many topics the parent node is split. For a better understanding, imagine the extreme case in which all documents are completely separate with absolutely no shared words (hence no shared topics), then the root would have as many children as there are documents. However similar documents share paths along the tree. In the following, the results of the hierarchical model of dataset 2 will be analyzed and discussed, because the size of the resulting tree is relatively small facilitating visualization. The remaining datasets would be compared afterwards.

Figure 3.6 shows the hierarchical tree for dataset 2 and dataset 3. The values in the nodes represent the alpha dirichlet parameter denoting the topic weights and the values along the edges represent the number of documents along that path. Let us consider in the upper tree with 3 levels - which consists of dataset 2 containing only the stations - the root (topic weighted the highest) represents word that appear in all the documents, which coincides with the results in LDA, where the main cluster area covers all hotels. 231 hotels share the same child topic (weighted with 2222) and 121 of those are considered similar, because they share the same path from root to leaf. These nodes have the highest topic weight of 1088 and entail the following top keywords: *gate lancaster bayswater arch marble queensway paddington edgware royal victoria* .

In terms of document similarities, the value on the edges leading to the leaves denote the number of similar documents that share the same topic distribution. I believe that the edge values could be considered as an indicator for topic distinction quality. If the value is too high, then the model could not accurately distinct topics and has assigned unclassifiable documents to a general topic. However, if the edge value is too low, then the model could not detect similarities between topics and considered - in the extreme case, where edge value is 1 - each document as a separate topic. A well-balanced tree would yield more stable results.

It is difficult to detect hotels geographical proximity based on document similarity since the topic distribution per document is not given by Mallet. However, in figure 3.8 the stations of every leaf topic are colour-coded to attempt to infer proximity. Note there are different degrees colour-intensities of each node group denoting the assignment to their parent node. These match those given in figure 3.6.

The stations within the city center are grouped by their proximity to each other. They may be connected or intersect in a certain station, but the points appear to have a circular form. As for stations outside the center of London, they seem to be along railroad. The purple points a lie along the *Northern* line, the red points connect the *Central* and *District* lines while the far right dark green points are along the *TfL Rail* and connects to the *DLR*. Points of the same colour palette do not seem to have a certain pattern, however the yellow points seem to be the dominating colour south of the Thames, which additionally provide riverboat services. Interestingly, this coincides with the fact that the keywords of the parent topics (leaf parent) seem to be fairly similar containing stations such as *station, victoria, westminster, paddington* and *bank*. The root topic lists stations such as *oxford circus, st pancras* and *green park*, which are words shared in all documents. I believe the root position of the former two stations is proportional to their importance; *St. Pancras*, also called *King's Cross*, has the highest number of underground lines passing through, followed by *Paddington* and *Bank*. *Oxford Circus* is at the intersection of *Oxford Street* and *Regent Street*, which are the main shopping streets in London and have a vivid nightlife [VisitLondon.com, 2016] As for the latter station, *Green Park*, it is difficult to assess whether its position in the root topic is due to it's importance of surrounding the *Buckingham Palace* or because the two words could be rather ambiguously interpreted by the model. As mentioned in the previous chapter 2, topic models consider the documents as bag-of-words, the order of the words is irrelevant. The terms *green* and *park* could be referring to the colour green, the green *District* line or any other station entailing the word "green" and due to the large number of parks in London, here *park* could refer to any of them. This provides a good example of the limitations of the bag-of-words model. N-grams could be used to rectify this undesired ambiguity [Wallach, 2006].

Hierarchical topic model determines the number of found topics by increasing the tree breadth and provides a hierarchy based on term occurrence in documents.

In terms of detecting geographical proximity, the hierarchical output tree of dataset 1 is too large to visualize on a map. However, the Mallet output tree structure is shown in the appendix A.4. The tree breadth is considerably larger than in dataset 2, which is plausible due to the larger diversity of terms. With 41 leaves, their edge values depicting

the document similarity are often below 10. Even for a hierarchy of depth 3, some topic keywords are often too specific, and their co-occurrence is likely to be limited to a single document. The root topics consist of random words that appear in every document descriptive of a hotel accommodation such as *buffet pm dinner drinks st complimentary attentive*. Despite the specificity of the leaf topics, some are able to detect proximity. For example, the leaf topic keywords which contain locations are found and their walking distance approximated using Google Maps [Google, 2016].

- *inn stratford parking westfield dlr olympic travelodge express east staybridge*: revolve around the district *Startford* in east London, which has a Westfield shopping mall (with Staybridge Suites in the same building), the Queen Elizabeth Olympic Park and the Travelodge Hotel, all within 2.25 kilometers (a 27-minute walk) as well as being connected to the Dockland Light Rail (DLR).

- *soho square garden covent leicester theatre oxford museum british radisson*: describes the attractions in the neighbouring districts *Soho* and *Covent Garden* with many theatres in the area (the route Soho Theatre, The Royal Theatre and British Museum is a 28-minute walk). The ambiguity of which theatre is probably due to the large number of theatres in this area.

- *club plaza blackfriars crowne lounge st lane paul bridge drinks* : a 10 minute walk from the Blackfiars Bridge to St. Paul's Cathedral with the Crowne Plaza on the way, which is famous for it's posh lounge.

- *paddington hyde apartment lancaster heathrow gate darlington express basement fridge*: an express connection, called "Heathrow Express" between paddington and Heathrow airport takes 15 minutes instead of 31 minutes.

- *chelsea fulham copthorne football broadway match club millennium stadium millenium*: the Millennium Copthorne Hotels at Chelsea Football Club are located in Stamford Bridge - a stadium and home of the Chelsea Football Club - in the area Fulham west of London. The specificity of the hotels mentioned is probably due to the very near proximity to the stadium.

These examples denote the locations mentioned within a topic are either considered near to each other or are located within each other. The latter case usually involves a district or an area, with it's famous attractions.

Compared to the parent nodes, the ambiguity of the mentioned locations rises. This is expected due to the topic abstraction intended by the hierarchical tree. The locations are often districts or popular stations. The examples below show the smallest walking distance along all stations and the shortest railway connection of the two furthest stations (geographically, not according to the railway connection):

- *court earls victoria kensington inn cross st earl premier kings*: a distance of 9.7 km and a walk of 2 hours and 3 minutes or 36 minutes via underground

```
 ---
106462/50 travelodge dlr ibis excel southwark waterloo parking airport pub aldgate
   157042/46 premier inn tower bridge meal inns parking kids eye dinner
   41750/3 st paul pool spa grange pauls cathedral holborn bridge club
   8783/1 chelsea fulham copthorne football broadway match club millennium stadium millenium
 ---
```

Figure 3.7: a subtree with 3 leaf nodes. The first number is it's alpha parameter followed by the number of documents with the same topic distribution, which is so far referred to as edge value of the tree

- *cross pool dated kings travelodge wharf canary euston inn gym*: a distance of 10.7 and a walk of 2 hours and 9 minutes or 40 minutes via underground or 37 minutes via underground
- *travelodge dlr ibis excel southwark waterloo parking airport pub aldgate*: a distance of 16 km and a 3 hour and 19 minute walk. Interestingly, all these areas have one or both of the Travelodge and Ibis chain hotels.

Let us consider the last example to analyze the relationship between the leaf topics and their parent nodes in figure 3.7. The first two leaf topics mention some city attractions in this area such as St. Pauls Cathedral, London Eye and Tower Bridge among other words. However, the last leaf topic (which was also mentioned in the examples above) mentions a location in the west of London. Geographically they are far apart, but they share a contextual factor, namely, the stadiums and arenas (Staybridge Bridge in the west and Olympic Stadium, Mile End Park Stadium and the O2 Arena in the east).

When given enough data, HLDA is thereby able the capture the increasing the degree of location specificity and proximity provided by the hierarchy as well as contextual similarity between leaf topics.

However, I believe for the provided datasets using a hierarchy level larger than 3 would result in overfitting. While some leaf topics are still informative, others are too specific to travel and accommodation details. The keywords for the datasets with hierarchy level 3 and 4 are given in the appendix A.4.

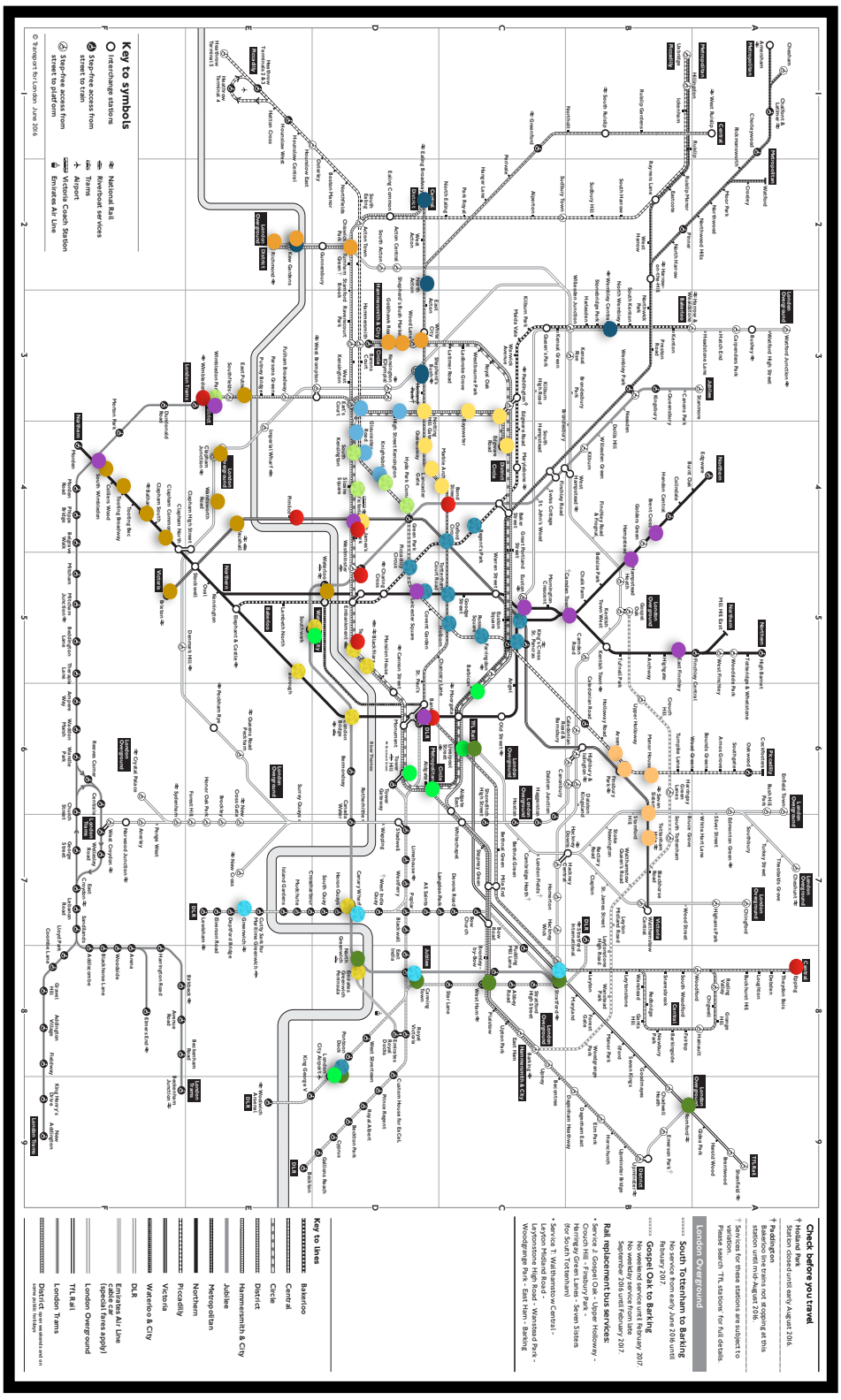


Figure 3.8: point colour represent stations of the same leaf topic and points of the similar colour-scale belong to the same parent topic (map taken from Transport for London[TfL,])

4

Discussion and Limitations

4.1 Discussion

It has been shown that topic modeling is able to capture geographical locations based on similarities of topics. Topics extract features that differ by topic number and dataset type. While the common feature is often abstract and not immediately apparent, the visualizations provide a mean to detect some similarities. After analyzing the data some general observations are found, which answer the research questions:

RQ 1 Can we infer geospatial proximity relations between entities from thematic resemblance using topic models?

RQ 2 Can we infer the context of the proximity relation using topic models?

- H1.1 The hotel is near an underground station.
- H1.2 The hotel is near a certain POI.
- H2.1 The hotel is in an urban or rural area.
- H2.2 The type and/or size of the POI can be determined

Topic numbers provide a good reference for specificity. In every model so far, at least one topic captures almost all hotels entailing common words that are likely to appear in every document, referred to here as the common topic. The common topic in dataset 1 mostly involves location irrelevant words, while in dataset 2 and 3, the common words are either abstract words such as “station”, “garden” and or important or central locations, such as “kensington” or “hyde”. For instance, there are many features associated with the word *kensington*; it’s name is part of a London Borough, a district, a garden, a palace and many stations among others. It’s position in the common topic is justified by its importance as well es it’s diversity in contextual meaning. In models with few topics, the most important locations in London are often to be found in the topic keywords and more often so in the common topic. However, there is no automatic way to differentiate the location-specific from non-location words. The word assignment to the same or different topics may differ by the topic number depending an topic granularity. The more topics in a model the more the keywords are split across topics and form smaller clusters with a defined geographical context. Some words appear in several topics due

to either ambiguity (kensington), generality (park), or importance (victoria – in dataset 2). In the common topic, I believe there is no automatic way to differentiate the words denoting importance from those that are random descriptive words without using a named entity recognition tool such as [NER, 2016]. However, the topic weights α are a good indicator for topic generality. [1]. If the topic has a low α value, then the keywords or their combination are not very common in documents, although common enough to be classified as keywords. Topics with lower α values define a smaller geographical cluster area. More topics lead to more classification options and the model can produce more narrow statements. For instance, some topic keywords with a low alpha value mention 3 destinations, which are in are 10 minute walk away from each other [2]. The common topic can often be identified by the highest α value depicting its generality as well as their wider geographical context. This generality is seen in the tree structure of HLDA as well. The leaves of HLDA represent the LDA topics with the lowest α values. I believe, this verifies the first research question, RQ 1, based on the existence of smaller clusters, that entail hotels falling within the same topics and the similarities identified using the KL-Divergence. The topics of dataset 1 form geographical clusters of different sizes with all centroids in more of less in the city center. While the content of the clusters revolves around the city center, its size is an indication of its the geographical limits. Dataset 2 creates clusters, which divide London into fairly distinct areas. Stations by hotels that are closer together are more likely to be mentioned in the same documents resulting in a fairly distinct topic classification - excluding the common topic. In dataset 3, topic models with a small topic number have shown no apparent cluster formation hindering a general statement. I believe this is due to high density of POIs in the city, which are likely to be visited by tourists accommodated in near as well as far located hotels. However, with larger topic numbers, smaller clusters denote geographical context. The larger the cluster, the more higher the level of abstraction with regards to geographical locations. While keywords of smaller clusters mention a few POIs that are fairly close to each other, keywords of larger ones are more spread out. To compare two entities with each other, KL-divergence is able to compare document similarities based on their topic distribution capturing geospatial proximity based on the shared topics. Here, the 10 most similar hotels were grouped together. Alternatively, a threshold can be set for similarity distance based on user requirements of context proximity. However, if the hotels are located fairly outside the city, similar hotels may not be as dense together as the ones closer to the city, which is justified by the lower hotel density outside London.

The second research question, RQ 2, can only be partially verified. In terms of geographical proximity, the topic keywords may be an indication of context. If the keywords are locations with large geographical distance, then the hotels being assigned to that topic may as well be spread out and if their distance is small - or nested within each other - the hotels are more likely to be more densely located. Using the combination between the topic keywords and document topic proportion α one can roughly approxi-

[1] These α weights are the hyperparameter values optimized during the model training. Their value is proportional to the overall document topic proportion θ_i

[2] This example is shown in the previous section 3.2.6

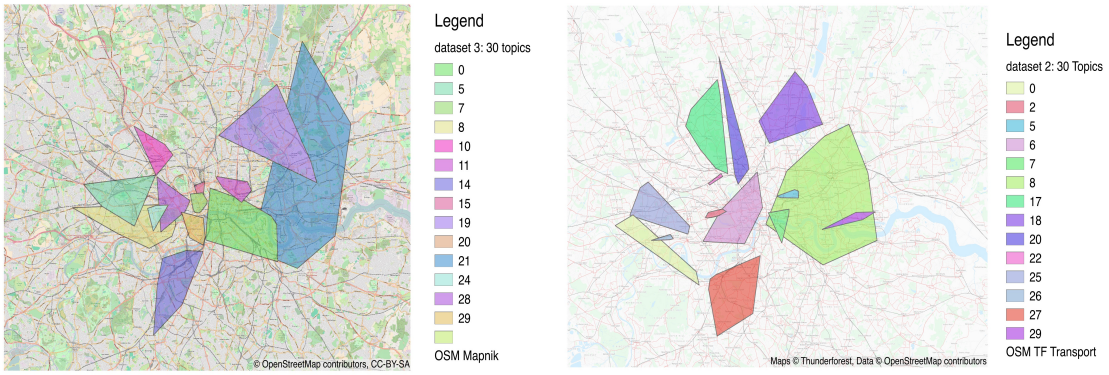


Figure 4.1: Smaller clusters from 30 topics of dataset 2 and 3

mate the location of the hotels [3]. In the following each hypothesis is discussed separately to show the degree of context information gained by topic models.

H1.1 *The hotel is near an underground station*: As before, similar hotels are closer to each other, the more centrally located they are, being likely to share the same station. The cluster of similar hotels often form a line or a circle. In the case of a small circle, the similarity is often based on a common station, destination or a railways line. Although the hotels do have common stations, which are considered either close by or easily accessible, it can not be automatically determined exactly which one. Especially in dataset 2, the common context of the hotels is considered either a station or a railway line. The higher the distribution of a topic in the document, the more reliable the similarity values become in terms of proximity. If topics are distributed equally in a document, then its informative value for distinction decreases. A higher probability of a topic denotes a higher affiliation of the document to the given topic. If hotels are grouped by their most dominant topic instead of similar probability distributions, then smaller clusters show either a slim rectangular or circular form as well. The size of the cluster could be identified by the topic weight, α. On average half the topics are considered small clusters. Smaller clusters are more location specific and a common station (or stations) can be inferred from the topic keywords. For a model using 30 topics - 16 of which considered small - roughly 45% of the hotels are classified 4.1 (left).

H1.2 *The hotel is near a certain POI.* The keywords in dataset 3 provide a good indicator for a geographical context. The resulting cluster sizes vary in size denoting different generality. Once the larger clusters have been excluded, smaller size cluster are fairly accurate of their closeness to POIs. The topic keywords denote the level

[3] Let us assume document d roughly consists of 70% topic 1, 20% topic 2 and 10% distributed among other topics. If the keywords of topic 1 and topic 2 are fairly close, then the location of document d is located in destinations both topics have in common. However if the locations are far apart, it is likely that document d is located midway.

of specificity.The document geographical context corresponds to the one of the keywords of its most dominant topic. For a model using 30 topics, 14 topics are considered small clustering 40% of the hotels 4.1 (right). Even large-sized cluster provide information about the generality of some POIs depending on the context derived from the topic keywords. [4].

H2.1 *The hotel is in an urban or rural area.* A topic distinction between pure urban and pure rural has could not achieved by the models, because there were barely any rural hotels in the dataset. However, it has been shown that some topics capture a district or an area located outside of the city center. The hulls of the most dominant topic in dataset 2 divides the city in east, west, north and south, in which the former 3 are located fairly outside the center and the latter less so. It is thereby fair to assume, that documents exhibiting such topics are accordingly located. This is also detectable using in dataset 1 using document similarities. Due to the circular cluster pattern a group of hotels located outside the city center are considered alike. However this is not an urban-rural classification of the hotels in term of topics.

H2.2 *The type and/or size of the POI can be determined.* The model did not create any noticeable topics, that show any common characteristic in term of type or size. I believe this is plausible, since , for instance, a distinction between gardens and buildings may be merely impossible considering most important attractions in London consist of a famous or royal building surrounded by a garden or park. The model seems to. However, I believe the POI size in terms of importance or popularity could be identified from topics by their occurrence and frequency. Less famous attractions are less likely to appear in the top keywords and even less so a multiple times. As mentioned in the point before, if the random topics are to be removed the remaining mostly topics cover the important POIs in London and often several times in different topics. However there is no automatic way to differentiate POIs from other words. For that a NER tool is required.

Some of the centrally located hotels belong to topics of large clusters, which is justified by their central location, accessibility and proximity to many other location. Note that larger-sized clusters entail the more general and important words and denotes a larger geographical context.

If not interested in the hotel location themselves, but in the location and proximity of London districts, stations or attraction, one can consider only the keywords and their topic weight α as reference for groupings and context. However, the need for identifying and specifying the location of the mentioned keywords still remains an obstacle. If the context is known, then one can rely on the documents sharing the same geographical context. The topic themselves do not distinct only railway roads, districts or closely

[4]Some topics are only dominant in one or two document, resulting in a cluster hull not being built. These are usually the topics with the lowest α weight.

located attractions as separate topic, but the context of the topic is applicable on the hotels regarding their context. While in some cases the relation could be "nearby", in other cases it could be "along the same railway line" or "outside city center" among others possible options given by the keywords context.

The advantage of using HLDA, is filtering all the random descriptive words from the dataset in the root topic. The amount of the location-irrelevant words reaching the leaf nodes is negligible and could be considered descriptive of the location it appeared with. The edge values are proportional to their given topic proportion, which could be used as a indicator for topic importance similar to α.

4.2 Limitations

Using topic modeling to infer geospatial proximity has been proven useful. Hotels have been grouped according to some common feature denoted by a shared topic. The general importance of that topic is given by its α parameter and its specific importance to the document i is given by the document topic proportion θ_i. The values are proportional as α is the Dirichlet prior of θ, meaning that if the general topic weight is low across the corpus then it is less prevalent in each document. The dataset used consists of user reviews of 800 hotels across London, while it seemed sufficient at first, I believe a larger corpus in terms of hotels as well as vocabulary would have proven useful. 8 hotels make 1% of the dataset and classifying 800 hotels into 30 topics, for instance, leaves on average 26.6 hotels per cluster. With such small numbers it is difficult to differentiate between outliers and a regular data point. While it could be shown in a small dataset, I am fairly certain that this statement is applicable of a larger scale.

For models with an large number of topics, overfitting was feared. Using a larger dataset of hotels would have facilitated the classification into urban and rural areas. A different datset source may have been more useful. Although extracting reviews from the hotels has achieved fair results, much of the vocabulary about the hotel itself, in terms of service and accommodation, was irrelevant and limiting the diversity of the other topics. I believe reviews of London attractions would have been more informative. The resulting topics could have been more diverse. While topic models were able to assign topics to documents in a manner that has shown a certain degree of classification, the keywords are often ambiguous. As mentioned before, it is unclear if the word *Kensington* refers to the borough, district, park, palace or station. Using n-grams would have proven useful in this case [Wallach, 2006]. However, despite the exact location being ambigious, the word is still bound by a geographic area. For instance, if the distinction between Kensingtion station and Kensington Palace cannot be made, then we are still certain that the hotel is somewhere in the Borough of Kensington and Chelsea. While a classification between stations and attraction may have been achieved by topic models, its exact context recognition by the keywords is not possible unless the other part of the n-gram was also in the keywords. In dataset 2, ambiguity in the case of station names like *Angel* or *Bank* was detected due to the preprocessing step. While extracting all station from documents, the differentiation between a station and a banking institution was

not possible. However, in dataset 1, the stations seldom appeared in the keywords and often so when the topic revolved around that area. Unfortunately, there is no certain way to verify that statement. HLDA has shown document similarities in terms of leaf topic keywords. However mapping the topic distribution to a specific document d was not given, hindering a graphical visualization of their proximity. The edge values are proportional to the topic weights may serve the same purpose as the topic weight α in LDA. However, the validity cannot be checked.

5

Future Work

LDA, as the most basic probabilistic topics model, has been proven to detect proximity between geographical entities by inferring topics from raw text. While the topic keywords and weights are an indicator for geographical size and context, the semantic identification of the generated topics is still required. Using more advanced models such as supervised models can help the data annotation process. Combining hierarchical topic models with n-grams can aid the word ambiguity along with detecting the hierarchy of administrative regions. Detecting proximity of location and grouping them in topics using raw text provides a way for recommendation systems to use the vast data provided by online users to give a prediction based on geographical proximity and context. The information is then not limited by distance or topic, but are able to use the combination of both. So far the results is shown with regards to hotels, but they may be applicable on other topics, such as POIs or cities. While only 800 hotels in the city of London were considered in this thesis. The results shown here could be scaled to larger areas with more topic diversity in terms of vocabulary and hierarchical administrative levels, such as cities or countries. I believe, it would achieve more accurate results in terms of ambiguous words and context distinction. Grouping hotels in clusters in a defined area of 623 square kilometers, overlapping is bound to happen due to word similarities and fewer abstraction level. Nevertheless, it has been shown that even in smaller more ambiguity prone areas geographical context can be inferred.

6

Conclusions

The research done in this thesis shows that topic models are able to detect spatially related entities. Using reviews of hotels, topic models are able to capture locations mentioned by the users. Depending on the preprocessing methods the resulting topics detect different contextual topics. Here, three datasets were created to manipulate the focus of the topic models. Dataset 1 consists of all the words used by users excluding stopwords and very frequent words, dataset 2 extracted only the mentioned stations of London and dataset 3 used only London's points of interest. Each document is represented by all the reviews of a hotel written in the English language. These documents were run through a topic modelling tool with varying topic numbers. By grouping the documents by their most dominant topic, clusters have emerged denoting geographical similarities. This was verified by mapping hotels with shared topics on the London map using the tools provided by a geographical information system.

The resulting topics showed different degrees of specificity regarding locations. The provided topic weights α indicate topic frequency and a high topic proportion in documents. A high topic weight is given to topics consisting of either general words or words, which are important and mentioned in many documents. The topic weight has been shown to be proportional to the cluster size. Lower topic weights describe topic keywords, which are less frequently used together but often enough to create a topic of their own. These words are more likely to define a smaller area, which is geographically limited by the topic keywords. If the location of the topic keywords are geographically wide-spread, then the size of the resulting hotel cluster along with its geographical context is spread accordingly. Smaller clusters define hotels with close proximity to each other and a common contextual feature given by the topic keywords. Common features could be "along a railway road", "stadium" or "near" any of the given topic keywords. Two hotels are compared by their similarities using KL-Divergence by measuring the distance between their topic distributions. The keywords as well as the topic weights of their given topics is an indicator for their contextual similarity.

While topic models are capable to detecting spatial proximity, the distinction of the geographical locations is not provided. The model is not able to differentiate between location-relevant and location-irrelevant words. For that, an NER tool is required.

Bibliography

[Adams and McKenzie, 2013] Adams, B. and McKenzie, G. (2013). *Inferring Thematic Places from Spatially Referenced Natural Language Descriptions*, pages 201–221. Springer Netherlands, Dordrecht.

[Bell, 2000] Bell, C. (2000). doogal.co.uk - postcodes, maps and code. *https://www.doogal.co.uk/london_stations.php*. Last accessed: 10.06.2016.

[Bettina Grün, 2016] Bettina Grün, K. H. (2016). R package for topic modeling – topicmodels. *https://cran.r-project.org/web/packages/topicmodels/topicmodels.pdf*. Last accessed: 02.06.2016.

[Blei et al., 2010] Blei, D. M., Griffiths, T. L., and Jordan, M. I. (2010). The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *J. ACM*, 57(2):7:1–7:30.

[Blei et al., 2004] Blei, D. M., Griffiths, T. L., Jordan, M. I., and Tenenbaum, J. B. (2004). Hierarchical topic models and the nested chinese restaurant process. In *Advances in Neural Information Processing Systems*, page 2003. MIT Press.

[Blei and Jordan, 2003] Blei, D. M. and Jordan, M. I. (2003). Modeling annotated data. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, pages 127–134, New York, NY, USA. ACM.

[Blei and Lafferty, 2006] Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 113–120, New York, NY, USA. ACM.

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

[Boyd-Graber and Blei, 2010] Boyd-Graber, J. L. and Blei, D. M. (2010). Syntactic topic models. *CoRR*, abs/1002.4665.

[britainexpress.com, 2016] britainexpress.com (2016). britainexpress.com – london attractions. *http://www.britainexpress.com/attraction-county.htm?County=Greater+London*. Last accessed: 05.06.2016.

[Chang et al., 2009] Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. L., and Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., and Culotta, A., editors, *Advances in Neural Information Processing Systems 22*, pages 288–296.

[Chen, 2011] Chen, E. (2011). Introduction to latent dirichlet allocation. *http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/*. Last accessed: 01.05.2016.

[Denofsky, 1976] Denofsky, M. E. (1976). How near is near ? A near specialist. Technical Report AI-M-344, Massachusetts Institute of Technology (Cambridge, MA US).

[Do and Batzoglou, 2008] Do, C. B. and Batzoglou, S. (2008). What is the expectation maximization algorithm? *Nature Biotechnology*, 26(8):897–899.

[Eisenstein et al., 2010] Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. (2010). A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1277–1287, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Frigyik et al., 2010] Frigyik, B. A., Kapila, A., and Gupta, M. R. (2010). Introduction to the Dirichlet Distribution and Related Processes. Technical Report 206.

[Google, 2016] Google (2016). Google – places api web service. *https://developers.google.com/places/web-service/*. Last accessed: 18.06.2016.

[Hoffman et al., 2010] Hoffman, M., Bach, F. R., and Blei, D. M. (2010). Online learning for latent dirichlet allocation. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A., editors, *Advances in Neural Information Processing Systems 23*, pages 856–864. Curran Associates, Inc.

[Hoffman et al., 2013] Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *J. Mach. Learn. Res.*, 14(1):1303–1347.

[Hofmann, 1999] Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 50–57, New York, NY, USA. ACM.

[Hong and Davison, 2010] Hong, L. and Davison, B. D. (2010). Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 80–88, New York, NY, USA. ACM.

[Jordan et al., 1999] Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Mach. Learn.*, 37(2):183–233.

[Kruschke, 2010] Kruschke, J. K. (2010). *Doing Bayesian Data Analysis: A Tutorial with R and BUGS*. Academic Press, 1st edition.

[Mcauliffe and Blei, 2008] Mcauliffe, J. D. and Blei, D. M. (2008). Supervised topic models. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T., editors, *Advances in Neural Information Processing Systems 20*, pages 121–128. Curran Associates, Inc.

[McCallum, 2002] McCallum, A. K. (2002). Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

[Minock and Mollevik, 2013] Minock, M. and Mollevik, J. (2013). Context-dependent âneârâ and âfârâ in spatial databases via supervaluation. *Data Knowledge Engineering*, 86(Complete):295–305.

[NER, 2016] NER, S. (2016). Stanford named entity recognizer. *http://nlp.stanford.edu/software/CRF-NER.shtml*. Last accessed: 04.08.2016.

[OpenStreetMap, 2016] OpenStreetMap (2016). Openstreetmap. *http://www.openstreetmap.org/*. Last accessed: 16.07.2016.

[Paul, 2013] Paul, M. (2013). An introduction to topic models. *https://www.cs.jhu.edu/~jason/465/PowerPoint/lect-topicmodels-mpaul.pdf*. Last accessed: 01.05.2016.

[QGIS, 2016] QGIS (2016). Qgis – quantum geographic information system. *http://www.qgis.org/en/site/*. Last accessed: 01.08.2016.

[Ramage et al., 2009] Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. (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*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Řehůřek and Sojka, 2010] Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. *http://is.muni.cz/publication/884893/en*.

[Russell and Norvig, 2003] Russell, S. J. and Norvig, P. (2003). *Artificial Intelligence: A Modern Approach*. Pearson Education, 2 edition.

[Scrapy, 2016] Scrapy (2016). Scrapy – an open source and collaborative framework for extracting the data you need from websites. *http://scrapy.org*. Last accessed: 10.05.2016.

[Steyvers and Griffiths, 2007] Steyvers, M. and Griffiths, T. (2007). *Latent Semantic Analysis: A Road to Meaning*, chapter Probabilistic topic models. Laurence Erlbaum.

[Suryawanshi et al., 2011] Suryawanshi, R. S., Thakore, D., and Raval, K. S. (2011). Context based word sense extraction in text: Design approach. *International Journal of Computer Science and Information Security*, 9(5):95.

[TfL,] TfL. Transport for London - tube. *https://tfl.gov.uk/maps/track/tube*. Last accessed: 01.08.2016.

[Tobler, 1970] Tobler, W. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46(2):234–240.

[Town, 2016] Town, L. (2016). Londontown.com – london museums and galleries. *Londontown.com*. Last accessed: 05.08.2016.

[Tripadvisor, 2016] Tripadvisor (2016). Tripadvisor. *https://www.tripadvisor.com*. Last accessed: 10.06.2016.

[Underwood, 2012] Underwood, T. (2012). The stone and the shell – topic modeling made just simple enough. *http://https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough*. Last accessed: 04.05.2016.

[VisitLondon.com, 2016] VisitLondon.com (2016). Visitlondon.com – top shopping destinations in london. *http://www.visitlondon.com/things-to-do/shopping/top-shopping-destinations#rb82VCYlk8YkDR2O.97*. Last accessed: 05.08.2016.

[Wainwright and Jordan, 2008] Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1-2):1–305.

[Wallach, 2006] Wallach, H. M. (2006). Topic modeling: Beyond bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 977–984, New York, NY, USA. ACM.

[Wang et al., 2011a] Wang, C., Paisley, J. W., and Blei, D. M. (2011a). Online variational inference for the hierarchical dirichlet process. In Gordon, G. J., Dunson, D. B., and DudÃk, M., editors, *AISTATS*, volume 15 of *JMLR Proceedings*, pages 752–760. JMLR.org.

[Wang et al., 2011b] Wang, H., Zhang, D., and Zhai, C. (2011b). Structural topic model for latent topical structure analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1526–1535, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Weston, 2002] Weston, W. (2002). Uk parliament website – common debates. *http://www.publications.parliament.uk/pa/cm200102/cmhansrd/vo020207/text/20207w18.htm*. Last accessed: 05.08.2016.

[Xu and Klippel, 2012] Xu, S. and Klippel, A. (2012). Developing nearness models from geocoding spatial entities in a news corpus. *GIScience 2012, extended abstracts*.

[Yin et al., 2011] Yin, Z., Cao, L., Han, J., Zhai, C., and Huang, T. (2011). Geographical topic discovery and comparison. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 247–256, New York, NY, USA. ACM.

A

Appendix

A.1 Variational inference

In order to understand the variational inference, we first have to explain the background of Bayesian statistics [Russell and Norvig, 2003], which is used here for conditional probabilistic inference. According to Bayes' rule the posterior distribution $P(h_i|d)$ is given as follows:

$$P(h_i|d) = \alpha P(d|h_i)P(h_i)$$

where h_i are the different hypotheses (in our case the latent topic and word distributions) and d is the observed data (the words). $P(h_i)$ is called the prior, which gives the probability of a certain hypotheses without (or prior) having observed any data and. The term $P(d|h_i)$ is called the likelihood, which gives the distribution of the data given a certain hypothesis. In the real world this is easier to compute than the posterior distribution. For example, we want to know the probability of having cavity (hypothesis) given a toothache (data). Counting the cases of people who have cavity given having a toothache is easier than determining the cases in which toothache leads to cavity [Russell and Norvig, 2003]

LDA is a three-level hierarchical Bayesian model consisting of documents as a distribution over topics, which in turn are a distribution over words. The hierarchy and the dependencies are best seen in the LDA graphical model and expressed using Bayes' rule, the joint distribution for each document is as follows :

[h]

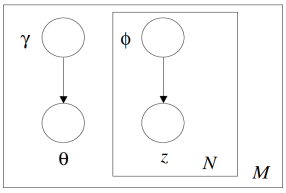


Figure A.1: variational distribution

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta), \tag{A.1}$$

where the first term is the hypothesis prior given by the Dirichlet distributed α and the second term is the likelihood. After marginalizing over the latent variables, we get

$$p(\mathbf{w}|\alpha, \beta) = \int p(\theta|\alpha)(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_n^j} \theta_d \tag{A.2}$$

Due to the coupling of β and θ in the likelihood term this equation is intractable to compute. Here Blei et. al. [Blei et al., 2003] used Jensen's inequality to get a tight lower bound on the likelihood. By removing and adjusting the edges in the graphical model, we lift the dependency between β and θ (figure A.1), simplifying the model using the free variational Dirichlet distributed parameter γ, multinomial distributed parameter ϕ.

$$q(\theta, z|\gamma, \phi) = q(\theta|\gamma) \prod_{n=1}^{N} q(z_n|\phi_n)$$

The aim is to optimize these parameter to find the tightest lower bound of the log likelihood. This is done by minimizing the Kullback-Leibler divergence between the variational distribution and the true posterior. Setting the derivative of the Kullback-Leibler divergence to zero, we get the following equations

$$\begin{aligned} \phi_{ni} &\propto \beta_{iw_n} exp E_q[log(\theta_i)|\gamma] \\ \gamma_i &= \alpha_i + \sum_{n=1}^{N} \phi_{ni} \end{aligned} \tag{A.3}$$

The algorithm is initialized with random ϕ and γ and updated using the variational EM (expectation maximization) procedure to maximize the lower bound [Do and Batzoglou, 2008]. The EM procedure computes the expected values of the variables given the current state of the model (E-step), then assumes these values are correct and updates the parameters (M-step). This procedure is repeated until convergence. An example is given by [Paul, 2013] as follows

- E-step

$$P(topic = 1|w^i = "apple", \theta_d, \beta_1) = \frac{P(w^i = "apple", topic = 1|\theta_d, \beta_1)}{\sum_k P(w^i = "apple", topic = k|\theta_d, \beta_k)}$$

- M-step

$$\text{new } \theta_{d1} = \frac{\sum_i P(topic_i = 1|w^i, \theta_d, \beta_1)}{\sum_k \sum_i P(topic_i = k|w^i, \theta_d, \beta_k)}$$

$$\text{new } \beta_{1w} = \frac{\sum_i I(w^i = word) P(topic_i = 1|w^i = word, \theta_d, \beta_1)}{\sum_v \sum_i I(w^i = v) P(topic_i = k|w^i, \theta_d, \beta_k)}$$

Gibbs sampling does not directly estimate the word-topic distribution and topic-document distribution, θ and ϕ respectively, but are inferred from the terms above, where:

$$\begin{aligned} \theta_k^{n(d)} &= \frac{N_{i_{(\cdot)k}}^{-i,m} + \beta}{N_{(\cdot)(\cdot)k}^{-i,m} + W\beta} \\ \phi_i^{n(k)} &= \frac{N_{(\cdot)mk}^{-i,m} + \alpha}{N_{(\cdot)m(\cdot)} + T\alpha} \end{aligned} \tag{A.4}$$

A better understanding and more details about the derivation of the upper equations is given by [Blei et al., 2003] in their introductory paper of LDA and [**?**] in their chapter about probabilistic topic models as well as in [Hoffman et al., 2013, Jordan et al., 1999, Wainwright and Jordan, 2008]

A.2 Keywords

A.3 Interpolation - Nearest neighbours

A.4 HLDA Tree Output Results

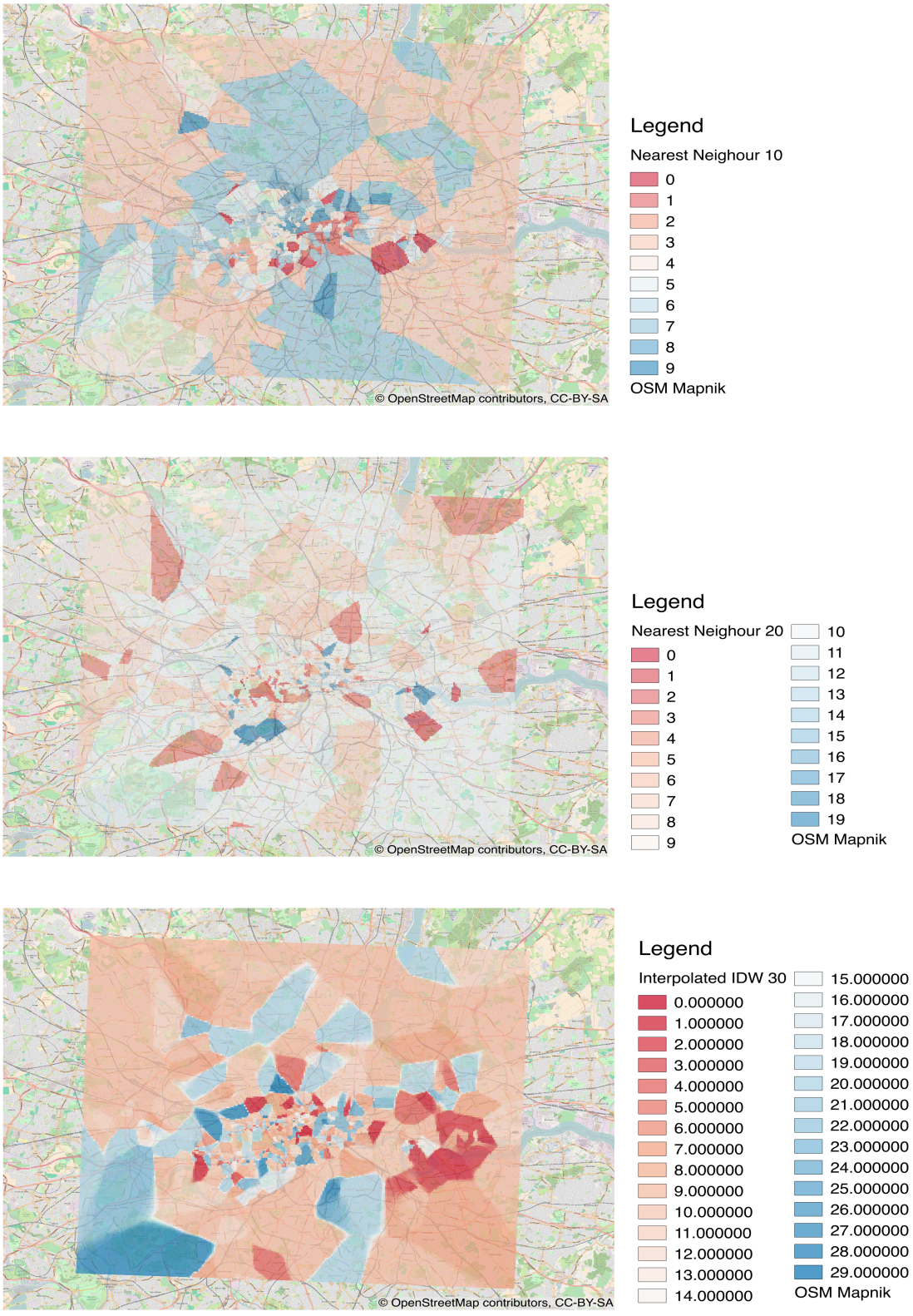


Figure A.2: Nearest neighbours of 10, 20 and 30 topics of dataset 1

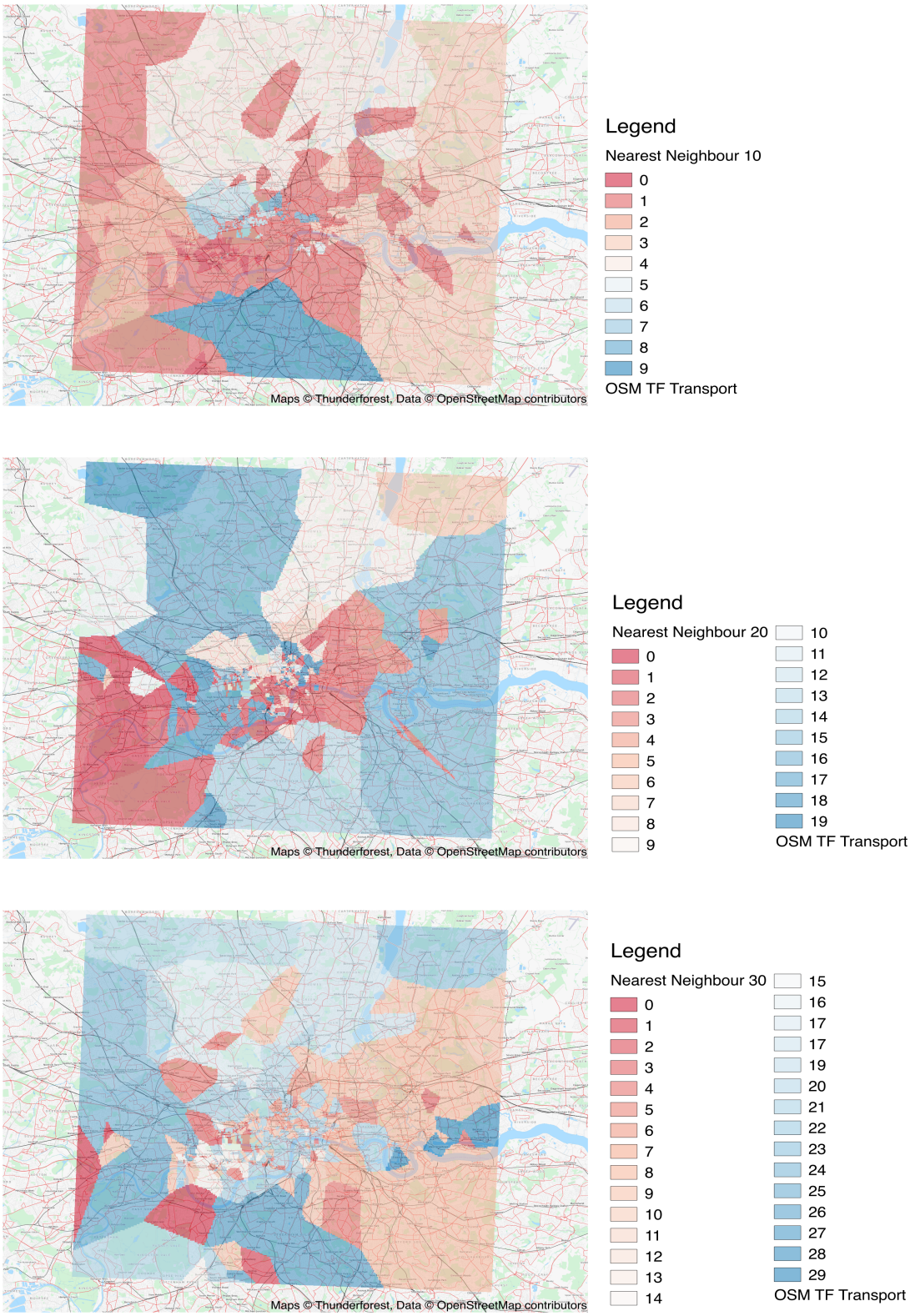


Figure A.3: Nearest neighbours of 10, 20 and 30 topics of dataset 2

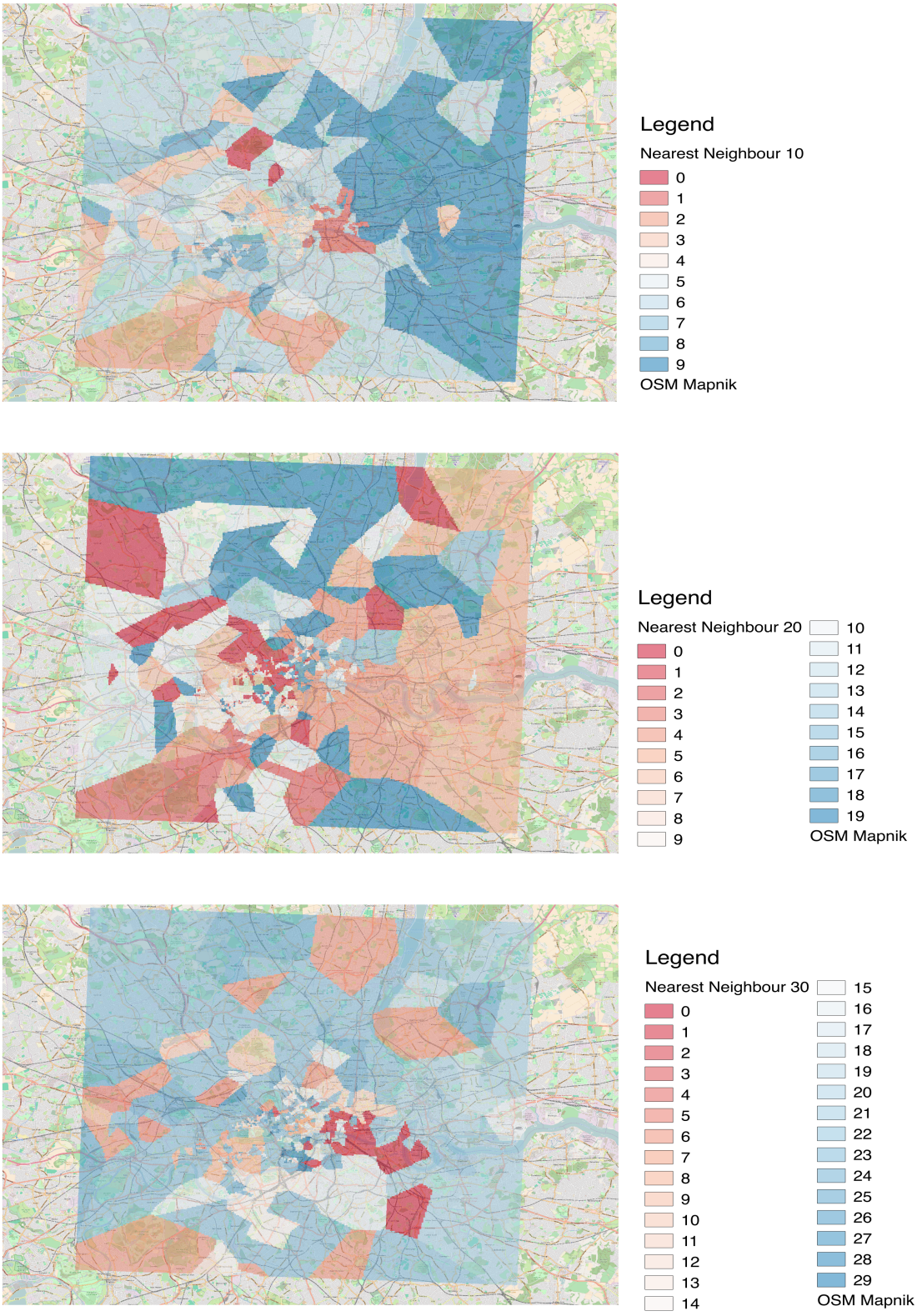


Figure A.4: Nearest neighbours of 10, 20 and 30 topics of dataset 3

Dataset 1: tree level = 3

4925805/721 buffet pm dinner drinks st west complimentary meal attentive usual
195920/325 hyde euston hilton bayswater cross st kensington queensway travelodge parking
567130/274 paddington victoria basement court cereal earls filthy awful thin disgusting
251189/27 spa afternoon oxford concierge club birthday arch pool marylebone hyde
53460/11 dlr excel airport pool docklands canary wharf parking cable aloft
24796/5 zetter soho townhouse honesty quirky rookery dean hazlitt boutique love
9002/4 camden lock inn market pub waterloo canal wellington markets woolwich
13118/3 court indigo earls mini complimentary drinks earl hide hendon parking
27399/1 wharf canary dlr dated pool britannia docklands views international meal
266390/121 square garden covent trafalgar concierge soho birthday leicester theatre piccadilly
93820/29 hyde paddington lancaster gate bayswater heathrow kensington queensway oxford gardens
64531/22 oxford green hilton mayfair bond radisson baker st concierge marble
144840/35 st kensington james south westminster palace apex ermin sofitel concierge
30275/5 wine cheese square hub compact soho covent leicester garden premier
23965/3 marriott lounge marble arch oxford executive hyde concierge lane pool
175689/18 afternoon chesterfield lounge palace mayfair buckingham concierge milestone champagne birthday
35051/3 melia lounge club piccadilly meridien le starwood spanish concierge portland
18118/3 marriott cottage swiss pool lounge regents executive concierge kilburn vale
9759/3 finchley colonnade inn express parking warwick venice avenue cat paddington
142861/144 club hilton concierge victoria lounge square executive kensington palace trafalgar
186408/64 gloucester kensington court apartment earls museum museums albert garden earl
114245/26 bridge paddington hilton heathrow tower executive lounge express market borough
121720/9 russell square british museum st oxford euston covent garden concierge
48254/13 travelodge lane sheraton green cross wembley kings parking dated starwood
11525/2 ace hoxton shoreditch cool holborn trendy bag vibe atmosphere hip
39846/15 british museum square arosfa garden russell euston st ridgemount bloomsbury
43897/6 tower hill bridge apex complimentary aldgate bank upgraded novotel indigo
36172/8 greenwich dlr parking cutty sark arena concert novotel ibis cafe
26888/1 citizenm tablet lighting citizen cool movies bridge mood design tate
123914/81 inn premier parking westfield hammersmith stratford dlr olympic mall meal
60261/36 oxford garden british museum covent grange bloomsbury holborn radisson square
51365/7 spa river shuttle thames clapham pool chelsea battersea junction rafayel
168587/33 premier inn tower bridge waterloo court eye travelodge meal inns
18485/2 apartment pool hall dolphin apartments pimlico green bethnal gym victoria
5728/2 victoria cheese wine complimentary glass pm palace buckingham sky queen
9574/1 river plaza apartment thames ben views balcony vauxhall parliament eye
18277/17 soho nadler victoria kitchenette belgrave windermere vauxhall oxford machine heart
28217/3 court stay base earl earls kitchenette nadler kensington microwave fridge
29370/4 tune paddington lambeth north liverpool waterloo westminster eye extras towel
41658/6 westminster pimlico parliament inn lounge executive hilton computer imac ben
6619/3 crown pool moran pub wembley parking kilburn cricklewood irish swimming
24047/1 marriott eye ben river thames hall county westminster lounge pool
75298/33 cross kings inn premier st euston pancras king eurostar jesmond
50220/13 euston st pancras cross kings eurostar pullman thistle king pantry
33526/2 st paul pauls cathedral spa pool grange bridge club atrium
6511/2 malmaison barbican square mal club cocktails charterhouse cocktail stylish dinner
40785/2 eye horseguards royal waterloo river thames embankment views square trafalgar
26681/3 hyde bayswater grand club queensway royale shaftesbury hill notting kensington
44129/6 wharf canary dlr hilton lounge executive marriott quay views seasons
8917/2 chelsea fulham copthorne football broadway match club millennium stadium millenium
5330/1 cricket st wood danubius lords regents baker johns john parking
6605/1 spa tech pool montcalm lights birthday shoreditch technology controlled tablets
19899/1 andaz liverpool drinks complimentary hyatt snacks mini soft wine st

Dataset 1: tree level = 4

4628172/721 buffet pm dinner drinks st attentive professional complimentary meal upgraded
641971/464 court basement paddington hyde cereal kensington earls bayswater narrow elevator
65902/56 travelodge premier inn westfield stratford waterloo chelsea olympic parking stadium
84491/52 parking inn hammersmith express dlr greenwich wembley novotel north pub
32798/2 st paul pauls cathedral spa pool grange bridge club atrium
9423/2 hyatt churchill club oxford regency lounge arch marble concierge square
67673/24 gloucester kensington club concierge museum inn museums history albert natural
54537/24 kensington court albert museum museums rembrandt hall south earls harrods
213391/322 square russell cross euston st british kings museum hilton garden
74368/58 victoria bridge waterloo pimlico parking borough pub cheese wine market
319355/222 filthy disgusting awful travelodge victoria refund stained horrible stains carpets
230763/26 spa afternoon concierge birthday dinner champagne landmark dorchester hyde milestone
9653/5 baker oxford sherlock holmes marylebone plaza montagu blandford st boutique
38693/2 langham club lounge intercontinental oxford hyde afternoon spa lane concierge
10489/2 pantry cross kings pancras st cake eurostar corridor northern cakes
3683/3 honesty church chocolate camberwell maddox tapas sauce quirky dvd complimentary
8501/2 westfield staybridge stratford olympic suites mall views inn kitchenette microwave
29891/1 lounge champagne palace executive buckingham pantry snacks birthday victoria attention
5162/1 colonnade warwick venice avenue cat paddington afternoon canal charming bakerloo
60372/59 bayswater hyde soho queensway grand wine nadler cheese club shaftesbury
50806/40 paddington hyde heathrow lancaster gate express darlington oxford tune rhodes
84703/12 square trafalgar leicester covent garden st hilton theatre concierge lane
37773/7 zetter cool malmaison townhouse shoreditch hoxton ace soho trendy square
33834/3 westminster pimlico inn parliament lounge hilton imac computer doubletree executive
7226/3 sumner arch marble oxford hyde lounge boutique basement blocks king
268704/184 soho afternoon concierge attentive birthday boutique garden beautifully delicious love
93973/98 square oxford museum inn garden covent british bloomsbury theatre russell
60478/51 oxford arch marble square bond sloane concierge hyde st radisson
27811/8 arosfa victoria ridgemount british museum owners lounge garden court tessa
118497/16 cross euston kings premier st pancras inn eurostar king pullman
15568/2 hub square premier covent garden inn compact leicester control tech
30859/3 eye marriott ben river thames pool westminster hall county waterloo
12738/2 marylebone oxford westbury bond pool gym mayfair concierge st polo
37201/12 hyde lancaster paddington gate kensington gardens royal thistle views oxford
2263/2 jumeirah lowndes knightsbridge harrods carlton admiral sister hyde spa harvey
26729/2 citizenm tablet lighting citizen cool movies bridge mood design tate
75815/56 bridge eye market waterloo ibis southwark borough shard bank aldgate
59451/8 tower apex hill bridge temple complimentary court st upgraded upgrade
26277/3 hilton kensington bush executive lounge olympia westfield holland concierge shepherds
111091/27 premier inn tower bridge parking meal inns dlr kids dinner
50680/5 court earl earls stay base kensington kitchenette nadler garden heathrow
42477/6 marriott lounge executive marble arch concierge pool oxford hyde cottage
33465/4 chesterfield mayfair green afternoon athenaeum flemings palace buckingham piccadilly concierge
36392/3 paddington heathrow hilton express indigo executive lounge airport novotel hyde
32173/16 dlr excel travelodge airport ibis parking novotel docklands stratford railway
22690/14 greenwich dlr parking cutty sark cafe rouge ibis blackheath museum
7084/2 liverpool tune balham market extras st spitalfields pub towel lane
45794/14 victoria horseguards royal eye river thames grosvenor palace embankment executive
63727/14 st james buckingham palace westminster ermin sofitel abbey concierge ermins
7658/16 westfield olympic inn stratford stadium wembley mall casino parking views
51832/16 pool river excel spa thames dlr shuttle clapham rafayel junction
19675/5 spa chelsea battersea pool pestana square sloane bridge harbour victoria
8816/3 cross hilton angel brent shuttle inn doubletree cookie islington chelsea
31834/7 wharf canary dlr dated pool docklands britannia views international buffet
7499/1 courthouse oxford carnaby afternoon hilton champagne circus regent sandwiches pool
13509/33 apartment garden holborn apartments covent green kitchenette studio citadines curzon
30902/15 andaz liverpool hall drinks complimentary snacks hyatt pool mini baglioni

Dataset 2: tree level = 3

9158/720 park street circus court st pancras hampton green covent oxford
2222/231 station lee bank victoria westminster paddington euston epping wimbledon circus
1088/121 gate lancaster bayswater arch marble queensway paddington edgware royal victoria
748/44 london greenwich bridge southwark borough embankment arena wharf canary blackfriars
423/29 clapham putney common tooting wimbledon brixton waterloo junction wandsworth vauxhall
183/17 palace finsbury alexandra lane arsenal park sisters tottenham hale manor
210/20 shepherds bush chiswick city market goldhawk white richmond lane kew
2162/191 lee station victoria bank paddington westminster street square wimbledon epping
274/17 ealing acton wembley broadway stadium royal lane kew shepherds gardens
836/71 kensington royal west south earls high gloucester albert knightsbridge court
693/30 london greenwich west stratford wharf tower city airport town canning
845/73 square euston russell holborn circus tottenham pancras leicester garden regents
1648/126 station lee victoria bank paddington square westminster circus epping euston
1044/59 street london tower liverpool aldgate city barbican airport southwark hill
583/54 park royal kensington albert hyde corner knightsbridge south sloane wimbledon
143/13 stratford london arena liverpool greenwich romford ham west airport city
1041/144 square paddington circus leicester garden covent street angel piccadilly euston
946/144 lee station victoria westminster bank epping wimbledon pimlico bond embankment
392/28 hampstead lee road euston wembley station paddington hendon leicester cottage
360/28 finchley bank town park stadium victoria brent station west wimbledon

Dataset 3: tree level = 3

13785/720 london street museum westminster tower kensington piccadilly park circus big
3693/248 garden covent end grad victoria west view square ealing park
1336/46 market bridge lane london brick tower east bank centre southwark
123/10 house hampstead kenwood heath road keats abbey fenton kilburn roundhouse
902/51 square british theatre bloomsbury london russell holborn museum library covent
1630/106 kensington gardens hill notting bayswater park palace station arch hyde
293/29 kensington apollo hammersmith notting hill eventim hyde chiswick carnival westfield
45/6 broadway hackney east wharf canary shopping mare empire mall festival
3808/208 garden covent square victoria park end west view ealing grad
1733/51 theatre street royal national square bridge parliament british palace house
613/52 street park marble hyde marylebone baker bond grad arch paddington
126/27 grad stadium finsbury emirates west view victoria wireless ealing olympic
304/23 british covent hampstead canal library marylebone road garden market regents
1129/30 bridge southwark theatre tower tate war vic hall parliament national
295/25 clapham bridge battersea common brixton chelsea wimbledon grad gate crystal
1842/128 park garden covent hyde end ealing square view grad wimbledon
1806/128 kensington museum victoria chelsea palace history natural gardens science west
2328/136 garden covent park end grad west square victoria view ealing
982/53 greenwich london bridge wharf docklands canary olympic tower thames stadium
691/42 theatre westminster victoria palace apollo pimlico parliament square ben abbey
980/41 theatre street square national gallery soho regent shaftesbury avenue piccadilly

Dataset 3: tree level = 4

10943/720 london westminster street tower park piccadilly kensington ben bridge eye
4261/303 square covent garden west end view ealing leicester victoria angel
3000/162 london museum street park bridge end theatre south royal holborn
2090/93 theatre national palace street square shaftesbury bloomsbury avenue gallery chinatown
964/45 market lane bridge brick tower east southwark guildhall petticoat borough
114/20 finsbury grad stadium emirates festival wireless victoria palace crystal gate
19/4 mare hackney broadway market sutton conception immaculate athletics valley farm
184/25 kensington theatre apollo eventim grad museum south science chelsea gardens
177/22 kensington hammersmith apollo palace chiswick park hill high history end
7/2 fc united ham ground boleyn olympic stadium beavertown morris william
3/1 lodge hunting queen beavertown morris william sutton music stone walk
1007/116 palace victoria grad park kensington westminster museum station ealing view
542/21 bridge parliament war tate lambeth houses waterloo national hall gallery
544/63 kensington gardens paddington marble arch bayswater hyde notting hill station
307/32 clapham bridge battersea common chelsea sloane brixton thames greenwich parliament
738/72 garden covent end view west ealing leicester square london hyde
388/39 park kensington victoria chelsea grad palace history natural gallery harrods
212/23 sloane harvey hyde nichols palace square mayfair harrods royal knightsbridge
134/16 hill notting hammersmith carnival london chiswick westfield apollo centre road
247/33 museum library holborn garden covent trafalgar british university kensington palace
258/32 square british bloomsbury victoria russell grad canal hammersmith baker paddington
0/1 beavertown morris william sutton music stone walk cheyne john order
2710/230 covent garden square street ealing view park angel museum victoria
1606/156 park end west grad victoria london kensington museum view hammersmith
919/54 greenwich london bridge wharf docklands canary olympic thames tower stadium
1246/81 kensington gardens palace notting hill bayswater arch science whiteleys station
234/21 hampstead road house abbey heath primrose kenwood studios regents canal
698/74 kensington park gardens science end grad knightsbridge hammersmith sloane south
708/74 kensington history museum hyde natural high harrods palace west chelsea
980/96 square garden covent soho leicester view ealing angel trafalgar theatre
920/96 park victoria west grad hyde end palace buckingham westminster marble
511/46 street marylebone bond baker paddington regent museum collection wallace holmes
484/41 theatre palace westminster pimlico apollo victoria parliament big abbey houses
128/8 market design centre camden passage stadium holborn clerkenwell emirates business
0/1 beavertown morris william sutton music stone walk cheyne john order
349/19 garden covent square park view victoria grad end west palace
353/19 museum london tower westminster street royal cathedral bond borough leicester
405/13 bridge southwark theatre tate hall river thames vic blackfriars hms
54/6 arcade burlington jermyn savile wimbledon row arts academy avenue shaftesbury

Topic	α	top keywords
0	0.13199	victoria apartment westminster eye st river palace thames buckingham ben trafalgar apartments square parliament abbey pimlico james equipped houses
1	0.08529	bridge tower st waterloo market bank novotel paul thames south liverpool river borough apex eye shard southwark tate hill
2	0.1433	court kensington earls gloucester earl museum museums south albert heathrow hall harrods history natural neighborhood garden buffet chelsea lines
3	0.19412	inn premier parking dlr travelodge wharf canary meal greenwich excel buffet dinner westfield airport usual kids inns pm stratford
4	0.26938	afternoon birthday dinner attentive concierge delicious champagne superb mayfair treat spa green beautifully moment fabulous love professional appointed attention
5	0.3901	lounge executive hilton concierge club pool buffet upgrade upgraded drinks gym pm birthday marriott views dinner spa complimentary mini
6	0.22365	paddington hyde oxford heathrow bayswater arch express queensway lancaster marble kensington gate gardens marylebone neighborhood baker hill blocks airport
7	1.15119	basement pm thin cereal awful dated bread dont paper cheese carpets complained horrible shabby narrow worn ceiling towel refund
8	0.14769	cross euston st kings square pancras british russell museum garden covent king oxford theatre eurostar buffet holborn bloomsbury meal
9	0.31567	square soho design complimentary cool wine leicester garden stylish drinks lighting covent love designed boutique mini trendy heart cheese

Table A.1: dataset 1 with 10 topics

Topic	α	top keywords
0	0.15231	paddington hyde bayswater heathrow queensway lancaster express gate buses hill kensington gardens oxford basement notting airport elevator royal neighborhood
1	0.05158	dlr wharf canary greenwich excel stratford airport parking westfield docklands olympic arena views buffet concert jubilee river ibis railway
2	0.16223	afternoon concierge champagne hyde birthday dinner lounge mayfair moment attentive palace treat doorman superb elegant outstanding class langham love
3	0.17442	inn premier meal inns dinner kids usual parking ate pm buffet menu attentive putney chain meals superb adults lifts
4	0.10194	victoria westminster eye st waterloo ben palace buckingham river pub thames parliament abbey pimlico james houses marriott ermin views
5	0.18876	liverpool design cool wine cheese lighting designed tune compact andaz cafe trendy shoreditch lights fun drinks tablet love market
6	0.08886	kensington gloucester museum museums south albert harrods hall history natural knightsbridge hyde royal victoria neighborhood heathrow sloane gardens concierge
7	1.06955	basement dated awful thin cereal carpets bread shabby paper worn horrible dont ceiling disgusting refund narrow paint stained filthy
8	0.19954	travelodge parking hammersmith pub express north wembley buses bush west lodge westfield ride center scrambled chelsea takes stadium broadway
9	0.1862	hilton executive lounge club concierge upgrade drinks buffet upgraded mini gym dinner complimentary heathrow wine views pm king snacks
10	2.13932	pm slow buffet st con iron usual items works definately company group smile temperature information disappointing board earlier stars
11	0.62706	complimentary birthday drinks dinner attentive upgraded delicious menu mini appointed upgrade boutique professional toiletries deluxe superb love beautifully atmosphere
12	0.09539	court earls earl kensington nadler heathrow garden camden olympia fridge neighborhood stay base kitchenette block piccadilly district airport lines
13	0.10251	square garden russell museum british covent oxford theatre holborn bloomsbury buffet concierge st tottenham west meal theatres court royal
14	0.08734	cross euston kings st pancras king eurostar angel british paris islington pullman trains rail thistle library proximity jesmond refurbished
15	0.0737	bridge tower st market bank borough paul apex thames river southwark shard hill tate novotel south pauls cathedral ibis
16	0.0821	oxford arch marriott marble marylebone st bond baker regents lounge regent portland melia hyde concierge sumner blocks selfridges swiss
17	0.09218	square soho trafalgar leicester piccadilly st covent garden circus theatre heart lane theatres sofitel west james picadilly regent oxford
18	0.09068	apartment green apartments mayfair kitchenette chesterfield equipped studio living microwave fridge machine buckingham washing sofa suites dryer citadines athenaeum
19	0.0643	pool spa gym swimming sauna river thames steam views jacuzzi corinthia shuttle birthday chelsea dinner grange relaxing battersea montcalm

Table A.2: dataset 1 with 20 topics

Topic	α	top keywords
0	0.10833	square soho trafalgar leicester covent garden theatre st theatres heart piccadilly circus west club gallery lane oxford national district
1	0.80749	buffet dinner meal slow usual ate drinks menu iron items express sofa west fridge board housekeeping lifts chain conference
2	0.04249	dlr wharf canary greenwich excel airport docklands parking arena views river concert ibis thames jubilee railway quay cutty sark
3	0.04666	club chelsea plaza harrods knightsbridge sloane crowne millennium concierge lounge hyde fulham square millenium football kings harvey capital mandarin
4	0.65998	basement cereal cheese cereals triple narrow bread ensuite elevator shared thin soap croissants beans orange suitcase bacon ham fridge
5	0.06913	cross euston kings st pancras king eurostar british paris pullman thistle arosfa rail trains library jesmond dene alhambra pantry
6	0.09147	hilton lounge executive marriott concierge drinks property snacks upgrade upgraded gold club gym exec king wine pm members diamond
7	1.75668	pm party running clear customers apparently disappointing advised happened noticed explained impression glass corridor attitude contact live company lost
8	0.18468	afternoon concierge champagne dinner birthday moment attentive treat superb attention luxurious fabulous delicious outstanding love detail exceptional lounge class
9	0.18562	travelodge parking pub wembley north lodge express buses ealing putney wimbledon ride trains stadium takes refurbished cafe broadway bacon
10	1.8189	facing airport block sightseeing nicely tourist offers heathrow based conveniently maintained st july plan directions larger june center son
11	0.13751	inn premier meal inns kids dinner usual ate parking pm attentive chain meals menu superb adults costa professional breakfasts
12	0.0635	victoria st palace buckingham james westminster pimlico ermin abbey soffitel grosvenor ben pub parliament coach ermins belgrave theatre eye
13	0.067	paddington hyde heathrow lancaster express gate oxford indigo royal airport gardens darlington trains pub sussex rhodes rail kensington mercure
14	0.08019	apartment apartments kitchenette microwave fridge nadler studio equipped living machine stay base sofa washing citadines dryer washer serviced dishwasher
15	0.05863	square russell museum british garden covent oxford holborn theatre bloomsbury st court tottenham concierge grange theatres euston west russel
16	0.05611	pool spa swimming gym sauna steam river thames jacuzzi landmark views corinthia shuttle montcalm battersea relaxing clapham atrium swim
17	0.05835	bridge tower st paul market borough southwark shard tate thames river bank south pauls cathedral blackfriars globe pub theatre
18	0.09505	court kensington earls gloucester earl museums museum south albert hall history heathrow natural piccadilly royal harrods garden neighborhood hyde
19	0.0587	westfield stratford olympic bush mall olympia shepherds parking staybridge holland stadium shepherd views center suites overground international links west
20	0.95642	dated awful carpets shabby worn stained disgusting paint refund filthy stains horrible ceiling thin complained plug pictures paper damp
21	0.06513	hyde bayswater queensway kensington gardens hill notting grand club paddington gate oxford shaftesbury thistle palace royale deluxe basement upgraded
22	0.03255	tower liverpool apex hill east aldgate andaz tune st district bridge temple ibis lane market brick shoreditch bethnal barbican
23	0.48524	complimentary birthday upgraded drinks upgrade attentive appointed mini professional toiletries deluxe delicious superb dinner equipped king soft menu superior
24	0.224	design cool lighting designed wine trendy fun stylish lights love funky glass compact control cafe tablet drinks quirky atmosphere
25	0.05432	green mayfair piccadilly lane hyde palace chesterfield buckingham circus concierge club picadilly oxford bond sheraton starwood afternoon athenaeum intercontinental
26	0.05599	eye westminster waterloo river thames ben parliament views embankment bridge horseguards houses lambeth royal abbey north hall south trafalgar
27	0.07261	oxford arch marble marylebone st langham club baker bond regent hyde circus angel sumner radisson selfridges concierge lounge hyatt
28	0.40768	garden charming boutique atmosphere hall lounge delicious cosy charm character love gem beautifully cozy dinner neighborhood furnished traditional afternoon
29	0.04575	novotel hammersmith camden regents melia portland parking apollo cottage swiss kids lock kilburn st executive market regent canal spanish

Table A.3: dataset 1 with of 30 topics

Topic	α	top keywords
0	0.92393	royal kensington albert lee victoria station south earls gate court paddington bank knightsbridge park bayswater arch wimbledon lancaster high
1	0.35765	london greenwich city tower street wharf airport stratford bridgecanary hill southwark waterloo liverpool arena embankment blackfriars aldgate north
2	2.53119	circus square street park station lee bank garden covent euston victoria st leicester westminster pancras piccadilly epping paddington angel

Table A.4: dataset 2 with 3 topics

Topic	α	top keywords
0	0.12489	clapham putney broadway twickenham brixton wimbledon bridge kew court ickenham westminster waterloo junction common ealing south richmond hampton lee
1	1.27243	square circus street garden covent st victoria station park leicester euston pancras westminster bank lee angel paddington piccadilly hampton
2	0.2076	london street greenwich city airport tower wharf bridge stratford canary arena southwark liverpool aldgate hill waterloo bank lee west
3	0.1236	park hampstead road town wembley station acton lee finchley camden lane stadium palace bank green central heath alexandra west
4	0.60011	kensington lee royal station albert park paddington victoria bank gate south knightsbridge earls arch bayswater court marble circus street

Table A.5: dataset 2 with 5 topics

Topic	α	top keywords
0	0.47412	kensington south court royal park earls albert road knightsbridge high gloucester lee station street borough hyde sloane square victoria
1	1.58647	lee bank station victoria westminster circus garden square covent paddington angel epping leicester st wimbledon pancras euston knightsbridge green
2	0.14107	bush shepherds kew gardens acton wembley ealing stadium west richmond broadway royal white chiswick city lane junction ickenham putney
3	0.24242	london city greenwich street airport stratford wharf tower bridge canary hill liverpool arena aldgate west southwark east arsenal waterloo
4	0.1197	park hampstead road town finchley palace camden alexandra heath station green central kilburn kentish finsbury wembley stadium lee brent
5	0.21303	twickenham cross ickenham bridge charing london hampton greenwich waterloo hill southwark euston pancras blackfriars tower embankment st russell court
6	0.38011	street park arch marble marylebone bond baker piccadilly regents circus canary wharf wembley oxford edgware stadium covent corner angel
7	0.25389	gate paddington marble bayswater arch lancaster lee queensway station victoria hill notting circus royal oxford edgware bank albert earls
8	0.27569	square court road tottenham street euston russell pancras holborn charing st leicester circus cross garden camden town covent barbican
9	0.06028	clapham tooting putney brixton common junction broadway bridge wimbledon vauxhall balham fulham wandsworth battersea town london richmond waterloo oval

Table A.6: dataset 2 with 10 topics

Topic	α	top keywords
0	0.10268	bush shepherds ealing acton broadway white chiswick city royal lane kew gardens west piccadilly ruislip uxbridge hanger market goldhawk
1	0.78635	park green circus hampton piccadilly square westminster garden leicester welling wimbledon knightsbridge angel covent hyde borough corner victoria station
2	0.31542	street london tower southwark bridge hill liverpool city greenwich wharf airport aldgate embankment waterloo blackfriars canary borough monument arena
3	0.02558	stratford green end mile liverpool bethnal hackney newington stoke islington highbury heath high hoxton wharf fields pinner canary cambridge
4	0.1322	twickenham ickenham putney bridge stratford hampton court gardens westminster richmond barnes ham eltham brentford blackfriars morden station southgate queenstown
5	0.03156	heathrow terminal hounslow fulham city terminals richmond broadway liverpool king maryland knightsbridge upminster george belvedere southfields canning orpington greenford
6	0.42509	circus station lee bank victoria paddington oxford westminster epping piccadilly wimbledon angel arena bond cyprus ilford acton finchley welling
7	0.42265	cross square charing st pancras euston garden covent waterloo leicester angel town russell arsenal hampton holborn embankment camden london
8	0.51582	street circus arch marble park bond marylebone baker regents oxford st london canary piccadilly wharf angel euston holborn edgware
9	0.06925	palace alexandra finsbury park lane manor tottenham crystal arsenal liverpool green hart white hale house sisters central golders kentish
10	0.13786	wembley stadium park arena acton watford vale maida town oval north archway arsenal finsbury epping junction stanmore green kew
11	0.22771	court tottenham road square circus euston holborn russell street barbican oxford piccadilly regents paddington goodge leicester lane chancery warren
12	0.08132	street arch marble baker bond marylebone knightsbridge victoria lee george edgware pinner bexley kingsbury bexleyheath strawberry preston fields brent
13	0.36966	lee victoria station westminster bank epping pimlico oval paddington vauxhall brixton angel embankment wimbledon waterloo hoxton kennington east denmark
14	0.06944	clapham junction common tooting wimbledon brixton broadway putney balham wandsworth vauxhall kingston battersea wood fulham south stockwell colliers embankment
15	0.27154	gate marble bayswater lancaster paddington arch queensway notting lee station hill edgware albert royal victoria oxford earls oak queensbury
16	0.1275	greenwich london stratford airport city wharf tower canary west bridge ham arena arsenal north india hill town canning quay
17	0.08885	hampstead road town camden park finchley wembley green heath stadium kilburn cottage swiss brent hendon cross station central west
18	0.57133	kensington south royal albert court earls knightsbridge road high lee bank gloucester street west wimbledon kew sloane holland borough
19	0.78613	station bank lee covent victoria euston pancras garden st square leicester westminster epping paddington angel wimbledon borough waterloo oval

Table A.7: dataset 2 with 20 topics

Topic	α	top keywords
0	0.0459	putney twickenham ickenham bridge broadway fulham southfields ealing brentford richmond lane eltham barnes uxbridge downs imperial tooting mortlake syon
1	0.00872	tooting broadway bec orpington grove mitcham balham elephant boston hendon stockwell
2	0.28232	gate bayswater lancaster queensway marble arch notting hill paddington albert royal edgware oak queensbury queens white colliers romford
3	0.33276	lee station bank victoria garden stratford covent borough westminster euston fulwell pancras oakwood hampstead road borehamwood drayton belmont wimbledon
4	0.63566	station lee bank victoria paddington westminster wimbledon angel pimlico epping earls brixton oval vauxhall gloucester hampton belvedere westbourne denmark
5	0.08009	street hoxton barbican lane farringdon chancery liverpool holborn st moorgate bethnal shoreditch high whitechapel green hackney hampstead newington stoke
6	0.5022	square court road tottenham leicester st covent pancras euston russell garden holborn goodge victoria abbey warren welling finsbury birkbeck
7	0.17914	street tower hill aldgate liverpool church embankment borough barbican southwark monument blackfriars london fenchurch east whitechapel bermondsey arena gateway
8	0.46948	london wharf greenwich canary bridge city street airport stratford arena waterloo arsenal liverpool southwark barking embankment pinner north cannon
9	1.30852	circus station lee victoria bank oxford piccadilly paddington westminster street epping euston angel wimbledon marylebone waterloo bayswater stratford cyprus
10	0.53841	street marble circus arch baker piccadilly bond regents marylebone oxford park edgware euston paddington bank george james cyprus saints
11	0.16278	park corner hyde green arch bayswater battersea arsenal sutton acton barking royal city paddington pudding newbury blackheath
12	0.06294	west heathrow terminal fulham brompton broadway ham hounslow king terminals parsons imperial fields peckham manor warren maryland dartford gardens
13	0.13103	twickenham ickenham southwark blackfriars elephant embankment castle kennington north vauxhall lambeth barbican eltham canary hampton norbury upminster holland moorgate
14	0.63542	kensington south royal albert court earls knightsbridge road park high gloucester borough street holland sloane wimbledon clapton brompton barons
15	0.0497	house manor mansion dartford cannon brixton square bow victoria belvedere osterley devons forest boston hanwell wood camden
16	0.01135	portland great stratford highbury sutton
17	0.073	hampstead finchley green heath cross brent central hendon belsize east golders farm chalk palace alexandra wood west highgate cricklewood
18	0.05858	palace alexandra park crystal lane finsbury white hale hart arsenal green tottenham sisters islington highbury hayes golders cyprus ponders
19	0.05382	arch marble lee paddington queens cockfosters epping park centrale greenwich stratford arsenal leicester whyteleafe kensal elstree teddington woodmansterne kingston
20	0.114	town camden park road regents kentish arsenal arena holloway crescent mornington tufnell caledonian belsize whetstone oakleigh neasden woodgrange southgate
21	1.4349	square garden covent hampton st pancras leicester park knightsbridge angel green oval euston westminster borough welling wimbledon epping bond
22	0.07108	road park wembley kilburn vale maida stadium cottage swiss abbey johns town camden avenue warwick high edgware willesden waterloo
23	0.01958	liverpool ilford romford leytonstone walthamstow stratford central snaresbrook chingford newbury buckhurst wanstead chigwell shortlands bickley gants hiqhams oak loughton
24	0.11384	wembley stadium arena watford angel park archway stanmore oxford acton bond earls cyprus sudbury kingsbury junction mottingham canons greenford
25	0.119	bush shepherds acton ealing white city royal broadway lane town junction north wood uxbridge hanger market goldhawk clapham golders
26	0.12144	kew gardens west park richmond chiswick green wembley stadium westminster barnes turnham victoria ruislip ravenscourt brook stamford grange kenton
27	0.06539	clapham junction common wimbledon brixton tooting vauxhall battersea wandsworth waterloo balham bridge kingston putney wood stockwell east town streatham
28	0.24248	cross charing embankment bridge london waterloo greenwich watford archway epping stratford arsenal blackfriars kingston blackheath thameslink kennington feltham catford
29	0.11	tower west ham hill india town gateway quay canning north east beckton royal house limehouse custom international deptford poplar

Table A.8: dataset 2 with 30 topics

Topic	α	top keywords
0	0.20781	bridge london tower end garden greenwich market park covent lane wharf canary southwark museum view victoria thames grad stadium
1	0.48543	square theatre london street westminster covent garden park british palace museum victoria tower west end trafalgar leicester piccadilly national
2	0.73068	kensington park museum palace grad street victoria hyde ealing view end hill west garden notting square westminster covent station

Table A.9: dataset 3 with 3 topics

Topic	α	top keywords
0	0.28174	street park grad square marylebone view regent west hampstead arch marble end baker paddington bond soho museum angel trafalgar
1	0.19179	bridge london tower market greenwich museum end southwark lane thames canary wharf street docklands brick bank east borough park
2	0.55248	kensington park museum street notting hill palace hyde gardens grad natural history view paddington marble arch ealing end victoria
3	1.08224	square covent london westminster garden park victoria palace end west bridge tower leicester buckingham big grad ben eye ealing
4	0.21831	theatre street national gallery square london museum british regent shaftesbury soho bloomsbury royal avenue war house tate bond piccadilly

Table A.10: dataset 3 with 5 topics

Topic	α	top keywords
0	0.05543	hampstead road market house abbey heath canal camden regents south lock studios hill strand roundhouse primrose kenwood tossed venice
1	0.14032	bridge market london museum tower southwark street tate lane bank westminster modern greenwich borough thames gallery national south river
2	0.08913	clapham battersea apollo bridge wimbledon hammersmith common square chelsea west grad river power brixton eventim thames theatre centre chiswick
3	0.49865	street park arch regent marble kensington covent bond garden paddington square westminster soho station baker mayfair west circus marylebone
4	0.18131	theatre street national gallery london royal square tate westminster chinatown house war museum british parliament shaftesbury portrait piccadilly avenue
5	0.28783	kensington notting hill park gardens station bayswater paddington hyde marble arch palace grad ealing west end victoria view carnival
6	1.03088	square london garden westminster covent park victoria palace ben big tower eye buckingham end ealing trafalgar view leicester grad
7	0.20458	square covent garden museum british bloomsbury end russell london west holborn leicester soho library trafalgar view grad shaftesbury angel
8	0.3549	kensington museum history natural park street science chelsea palace south knightsbridge high victoria view grad gardens hyde harrods sloane
9	0.14521	london end bridge greenwich stadium tower wharf docklands olympic canary grad west park brick east lane market thames victoria

Table A.11: dataset 3 with 10 topics

Topic	α	top keywords
0	0.94442	street park arch westminster mayfair piccadilly buckingham marble regent bond palace circus cave lane station angel hyde harrods knightsbridge
1	0.05957	nichols harvey sloane kensington gallery memorial royal saatchi knightsbridge oratory wellington command bomber brompton canary south cadogan opera albert
2	0.08168	clapham bridge chelsea wimbledon battersea common brixton sloane club power court exhibition football cottage craven theatre center earls cemetery
3	0.09736	british square library shaftesbury theatre museum avenue russell holborn bloomsbury strand university stadium emirates economics school virgin trains marylebone
4	0.65091	london bridge tower westminster square south bank garden eye tate museum wharf national gallery modern canary river greenwich thames
5	0.10479	greenwich london olympic stadium canary wharf docklands bridge tower excel east park sark thames cutty city westfield queen stratford
6	0.30921	kensington notting hill paddington gardens bayswater palace station marble arch hyde road carnival portobello whiteleys grove westbourne winter market
7	0.17668	theatre westminster palace apollo victoria pimlico parliament square tate houses abbey station britain ben green bridge mall cathedral buckingham
8	0.05859	street museum collection baker wallace holmes sherlock hall soho wigmore marylebone air open british palladium carnaby bond place greenwich
9	0.44816	kensington museum natural history science park chelsea harrods knightsbridge south high palace gardens hill street notting wimbledon hammersmith holland
10	0.95788	garden covent london square victoria westminster end ben big leicester grad ealing view west eye abbey trafalgar tower piccadilly
11	0.32306	grad park view ealing victoria hammersmith west hyde end station paddington gate bayswater angel chiswick docklands westfield regent marylebone
12	0.08654	bridge southwark hall theatre museum hms royal belfast festival street waterloo millennium bankside market borough house war thames imperial
13	0.06957	war parliament churchill rooms imperial downing museum houses lambeth whitehall battersea britain southbank power hall cenotaph jubilee admiralty wellington
14	0.07261	hampstead road abbey house hill heath canal camden regents primrose lock studios roundhouse kilburn venice kenwood market museum keats
15	0.08879	market lane brick centre end east clerkenwell petticoat design road flower wall columbia stadium guildhall barbican emirates church street
16	0.0366	gallery national house royal jermyn arts arcade burlington academy row savile opera apsley theater angel clarence whitehall green casino
17	0.37051	covent square garden british museum bloomsbury leicester holborn view end west russell trafalgar victoria soho angel library ealing london
18	0.23345	theatre square street national gallery soho chinatown shaftesbury british royal avenue bloomsbury house road strand portrait piccadilly cross charing
19	0.8162	park grad victoria view kensington hyde west palace end ealing buckingham square angel cave south wimbledon strand mayfair harrods

Table A.12: dataset 3 with 20 topics

Topic	α	top keywords
0	0.28847	london bridge tower museum market tate modern south river thames borough westminster southwark bank gallery wharf national belfast hms
1	0.07827	museum baker collection wallace hall holmes sherlock street marylebone wigmore greenwich air palladium primrose open paddington buckingham fleet place
2	0.02558	exhibition cemetery court center earls brompton chelsea gogh hammersmith van stamford pimlico westfield kensal abney nightingale apple finborough hm
3	0.02552	arts jermyn arcade burlington academy row savile memorial command bomber theater chinatown casino criterion south beach palm monument battle
4	0.00602	guildhall threadneedle change millennium amphitheatre roman art stone england centre lane
5	0.13797	hill notting carnival hammersmith ealing chiswick studios shopping walpole manor pitzhanger bkd loftus lumiere architects treasury company lyric shard
6	0.58086	end victoria west grad ealing view park hammersmith gate strand hyde angel tossed storeys east chiswick highgate rich cenotaph
7	0.27231	theatre street gallery national shaftesbury avenue chinatown bloomsbury leicester bond british soho road modern tate cross charing piccadilly regent
8	0.56275	kensington museum natural history street south science high knightsbridge harrods chelsea park palace gardens hyde wimbledon holland british sloane
9	0.0404	apollo hammersmith theatre eventim cottage craven chiswick fulham westfield ravenscourt clapham cru television bbc bush lumiere baggage excess courtauld
10	0.04599	hampstead road heath house abbey kilburn kenwood studios keats primrose cinch roundhouse fenton cemetery frend lane highgate hunterian tricycle
11	0.09022	lane market brick centre east road end design petticoat clerkenwell flower columbia barbican wall spitalfields tower broadway guildhall hammersmith
12	0.50548	kensington station marble arch paddington hyde bayswater victoria gardens grad park palace view marylebone venice hammersmith west buckingham italian
13	0.07737	theatre house royal greenwich somerset opera holborn lyceum fleet whitehall hall aldwych adelphi end parliament courts festival justice embankment
14	0.09333	clapham bridge battersea chelsea wimbledon common power brixton sloane club football physic saatchi oval kia lane apsley vauxhall albert
15	0.16508	british square museum library russell bloomsbury holborn university london trains virgin trafalgar school economics shaftesbury tavistock marylebone charles wellcome
16	0.95277	london westminster tower bridge square eye abbey museum big piccadilly ben south circus buckingham bank trafalgar strand british street
17	0.24498	wonderland winter london kensington end eye crystal grad strand station tossed circus mayfair piccadilly farm show wellington roundhouse selhurst
18	0.06043	nichols harvey sloane gallery lane saatchi brompton oratory kensington court kyriance battersea parade zealand tower soho hyde
19	0.09771	stadium emirates festival wireless finsbury olympic hampstead docklands grosvenor hammersmith soho vic ground boleyn chissold tossed wood haringey merton
20	0.08478	bridge southwark hall vic imperial waterloo war parliament festival museum southbank centre city golden houses national millennium lambeth bankside
21	0.12453	london greenwich wharf canary docklands bridge olympic tower stadium sark cutty thames excel city westfield barrier east stratford queen
22	0.04921	market canal camden regents lock primrose borough hill roundhouse wharf venice canary whitehall zoo zsl chapel bankside parish beavertown
23	1.66443	park victoria palace ealing square westminster hyde view grad west buckingham end green wimbledon angel trafalgar cave leicester harrods
24	0.19368	notting hill kensington palace gardens bayswater road paddington market portobello whiteleys grove westbourne station memorial albert carnival arch marble
25	1.04145	garden covent square soho end leicester west park grad angel circus cave view piccadilly holborn bloomsbury trafalgar strand east
26	0.10549	national gallery street house royal portrait opera downing whitehall lane mall jermyn garrick embankment knightsbridge war clarence coliseum coward
27	0.04075	garden british covent museum library centre thames design hyde bridge athletics valley lee portcullis aquarium life sea bateaux zsl
28	0.47816	street regent bond arch marble station kensington paddington baker mayfair marylebone lane soho circus canary piccadilly wharf grosvenor chinatown
29	0.26619	theatre parliament palace westminster houses victoria war abbey apollo tate britain square pimlico ben churchill rooms cathedral imperial whitehall

Table A.13: dataset 3 with 30 topics

List of Figures

2.1 distribution of 3 topics in a 2-dim simplex (a) For k = 3, α is discrete point in a 2-simplex, where A=B=C $\frac{1}{3}$ and (b) continuous probability density, where α is higher in B than A and C [Paul, 2013] 8
2.2 LDA graphical model 9

3.1 Hotels (black), POIs (yellow) and stations (blue) of London 14
3.2 top: nearest neighbours of 3 (left) and 5 topics (right), bottom: hull clusters of 3 (left) and 5 topics (right) 18
3.3 cluster distance and density of 3 (top) and 5 (bottom) topics respectively 20
3.4 a sample of hotels that are considered similar 22
3.5 4 clusters based on the distribution similarity of dataset 2 23
3.6 Hierarchical Topic Tree for dataset 2 (upper) and dataset 3 (lower) . . . 26
3.7 a subtree with 3 leaf nodes. The first number is it's alpha parameter followed by the number of documents with the same topic distribution, which is so far referred to as edge value of the tree 29
3.8 point colour represent stations of the same leaf topic and points of the similar colour-scale belong to the same parent topic (map taken from Transport for London[TfL,]) 30

4.1 Smaller clusters from 30 topics of dataset 2 and 3 33

A.1 variational distribution 45
A.2 Nearest neighbours of 10, 20 and 30 topics of dataset 1 48
A.3 Nearest neighbours of 10, 20 and 30 topics of dataset 2 49
A.4 Nearest neighbours of 10, 20 and 30 topics of dataset 3 50

List of Tables

3.1 Log likelihood overview with the highest value of each dataset in brackets 17
3.2 Top Keywords per Topic for 3 and 5 topics 17
3.3 Top Keywords per Topic for 3 and 5 topics of dataset 3 24
3.4 Comparing distance and density of dataset 1 dataset 2, dataset 3 (in absolute values) 25

A.1 dataset 1 with 10 topics 55
A.2 dataset 1 with 20 topics 55
A.3 dataset 1 with of 30 topics 56
A.4 dataset 2 with 3 topics 57
A.5 dataset 2 with 5 topics 57
A.6 dataset 2 with 10 topics 57
A.7 dataset 2 with 20 topics 58
A.8 dataset 2 with 30 topics 59
A.9 dataset 3 with 3 topics 60
A.10 dataset 3 with 5 topics 60
A.11 dataset 3 with 10 topics 60
A.12 dataset 3 with 20 topics 61
A.13 dataset 3 with 30 topics 62