# Geographic Feature Type Topic Model (GFTTM): Grounding Topics in the Landscape

**Benjamin Adams**[1]

[1]**Centre for eResearch, Department of Computer Science, The University of Auckland, New Zealand, b.adams@auckland.ac.nz**



## ABSTRACT

Probabilistic topic models are a class of unsupervised machine learning models used for understanding the latent topics in a corpus of documents. A new method for combining geographic feature data with text from geo-referenced documents to create topic models that are grounded in the physical environment is proposed. The Geographic Feature Type Topic Model (GFTTM) models each document in a corpus as a mixture of feature type topics and abstract topics. Feature type topics are conditioned on additional observation data of the relative densities of geographic feature types co-located with the document's location referent, whereas abstract topics are trained independently of that information. The GFTTM is evaluated using geo-referenced Wikipedia articles and feature type data from volunteered geographic information sources. A technique for the measurement of semantic similarity of feature types and places based on the mixtures of topics associated with the types is also presented. The results of the evaluation demonstrate that GFTTM finds two distinct types of topics that can be used to disentangle how places are described in terms of its physical features and more abstract topics such as history and culture.

Keywords: Text mining; Topic modeling; Bayesian inference; Volunteered geographic information

## INTRODUCTION

The rendering of a place in natural language text will to some degree reflect the physical properties of that place. This is as true of narrative descriptions of first-person experiences as it is of highly stylized writings, such as encyclopedia entries and literary works. Consequently, in a large corpus of place descriptions the words used to describe environments that share specific types of features will exhibit statistical regularities. For example, documents describing travels in mountainous areas will tend to have more discussions of climbing and hiking activities than will similar documents describing densely populated urban areas. These regularities allow us to identify topics most associated with specific feature types.

With the ever increasing availability of geographic information online we have at our disposal not only many descriptions of places but also an unprecedented amount of information about the geographic features present there. In this paper a probabilistic topic model called the Geographic Feature Type Topic Model (GFTTM) is presented that uses these two different kinds of evidence to identify the latent topics that are associated with specific kinds of geographic feature types Blei (2012). As with other probabilistic topic models, each topic is represented as a distribution over words. This model provides us with a representation of feature types derived from observations of how people write about them rather than in terms of a fixed, formal top-down ontological definition Bennett et al. (2008). Furthermore, it provides us with a means to measure the semantic similarity of feature types with respect to the myriad physical characteristics, activities, and social constructs associated with those types as reflected in these writings. As such, this approach is compatible with the representation of types by the set of common affordances that they provide without artificially restricting the representation to a designed ontology Kuhn (2002); Janowicz and Raubal (2007); Sinha and Mark (2010).

A number of probabilistic topic models have been developed, which condition the topics on geographic labels or other kinds of spatio-temporal information Mei et al. (2006); Wang et al. (2007); Ramage et al. (2009); Eisenstein et al. (2010); Hao et al. (2010); Adams and Janowicz (2012); Hong et al. (2012). The GFTTM is distinct from these topic models, because it is the first topic model to directly condition topics on the structure of the geographic environment, in particular the features present in those environments. In addition, the generative model of GFTTM is unique in that it represents each document as a mixture of both feature type topics that are based on physical characteristics and abstract topics, which are not based on the physical characteristics.

This paper is organized as follows. The next section provides some motivational background for why we want to investigate the interplay between language and landscape. Section 3 presents background information on probabilistic topic modeling. Section 4 introduces the GFTTM generative model and describes a Gibbs sampling algorithm for doing inference on the model. Section 5 presents an evaluation of GFTTM using volunteered geographic information, and section 6 discusses the results of this evaluation. Section 7 concludes the paper and describes avenues for future research.

## 1 MOTIVATION

Understanding the relationship between language and geography is an ongoing multidisciplinary pursuit. Linguistic geography is a well-established field in linguistics where the variation and diffusion of linguistic elements are studied, e.g. studying differences in dialects and geographic diffusion of words Kurath (1949); Bailey et al. (1993); Labov (2007). The geographic perspective on this research investigates the social and geographical processes that give rise to these variations Trudgill (1974). The specific relationship between language and the physical geographic environment that people inhabit is the study of ethnophysiography Mark and Turk (2003). This research examines how cultural context can imbue strong relationships between the features of landscapes with terms in language that are tied to important concepts in the culture. These terms need not be restricted to physical description but can reflect, e.g., strong spiritual beliefs broadly held within a community about features and categories of features Mark et al. (2007, 2012). The cross-cultural aspects of how language use and landscape are related is also an important aspect of this work Burenhult and Levinson (2008). In related work, environmental psychologists have studied how people relate landscape values and place attachment Brown and Raymond (2007).

People write about places not locations and ethnophysiography is close related to the study of place and place attachment, described as *topophilia* by Tuan Tuan (1974). In his book *Place and Politics* John Agnew Agnew (1987) argues that one necessary component of a place is a locale, which is a combination of the physical setting and configuration of the place and the types of activities that people engage in there. The accessibility to large corpora of place descriptions gives us an unprecedented opportunity to investigate the research problems of linguistic geography, ethnophysiography, and place through discovery of statistical patterns in the data Gregory and Hardie (2011); Adams and McKenzie (2013). However, the duality of place as a physical location and social construction – while commonly understood in geography – has not been reflected in the topic models of language that have been previously developed.

## 2 TOPIC MODELING

Probabilistic topic modeling is a class of text mining statistical models designed to identify the latent topics (or themes) that exist in a corpus of documents, and how each these topics are represented in each individual document Steyvers and Griffiths (2007). Since topic models are a type of unsupervised learning it means that a structure of the corpus' content can be discovered without requiring that examples be tagged for different pre-selected topics by a human user prior to training. The Latent Dirichlet allocation (LDA) model is the most popular probabilistic topic model and describes each document as a mixture of topics, where each topic is itself a probability distribution over words Blei et al. (2003). LDA is a generative model in that it describes how the documents that exist in the corpus were generated through a random process of picking each word that goes into the document given the distributions of words and topics defined by the model.

Let $\theta_i$ be the topic distribution for document $i$, $\phi_k$ be the word distribution for topic $k$, $z_{ij}$ be the topic for the $j^{th}$ word in document $i$, and $w_{ij}$ be that word. The LDA generative process is defined as follows Blei et al. (2003):

1. Choose $\theta_i \sim \mathrm{Dir}(\alpha)$, where $i \in \{1,\ldots,M\}$
2. Choose $\phi_k \sim \mathrm{Dir}(\beta)$, where $k \in \{1,\ldots,K\}$
3. For each word $w_{ij}$, where $j \in \{1,\ldots,N_i\}$

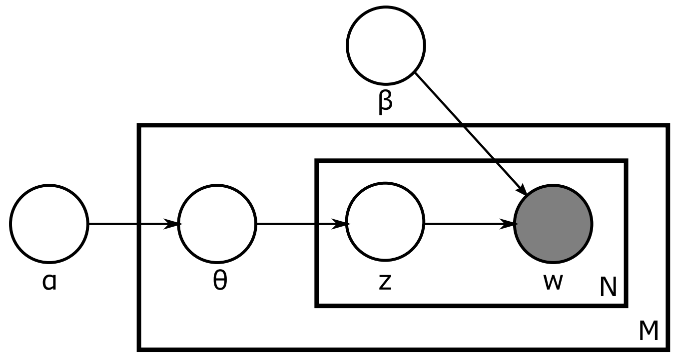


**Figure 1.** LDA plate notation

(a) Choose a topic $z_{i,j} \sim \text{Multinomial}(\theta_i)$.

(b) Choose a word $w_{i,j} \sim \text{Multinomial}(\phi_{z_{i,j}})$,

where in the previous $\text{Dir}(\cdot)$ is the Dirichlet distribution for the given parameter. Figure 1 shows the LDA model in plate notation.

Given the model we need to infer the most likely topics and words over topics to explain the existing corpus. Gibbs sampling Markov Chain Monte Carlo (MCMC) has been an effective approach toward implementing LDA inferencing Griffiths and Steyvers (2004). Gibbs sampling is a randomized algorithm to approximate the posterior distribution of a graphical model Bishop (2006). In MCMC a Markov chain is created with transitions that converge toward a stationary distribution Mackay (2005). In practice, the chain is followed for a fixed number of steps and a sample is drawn (this is the Monte Carlo part). The Gibbs sampling algorithm is a kind of MCMC developed by Geman and Geman (1984) to sample from the joint probability distribution of a set of random variables. Let $p(\mathbf{x}) = p(x_1, \ldots, x_n)$ be a joint probability distribution over $n$ variables. At the initial state of the MCMC the variables $x_1, \ldots, x_n$ are set to the values $x_1^{(0)}, \ldots, x_n^{(0)}$, respectively. The values sampled for the variables at step $\tau$ of the Markov chain are denoted $x_1^{(\tau)}, \ldots, x_n^{(\tau)}$. At each step, the value for each variable is updated in turn by sampling from the conditional probability, given that all other variables remain constant. The Gibbs sampling algorithm (from Bishop (2006)) is shown in Algorithm 1. Assuming there are not two variables that are perfectly

**Algorithm 1** Gibbs sampling algorithm (from Bishop (2006))

Initialize $\{x_i : i = 1, \ldots, n\}$
**for** $\tau = 1, \ldots, T$ **do**
  Sample $x_1^{\tau+1} \propto p(x_1 | x_2^{(\tau)}, x_3^{(\tau)}, \ldots, x_n^{(\tau)})$
  Sample $x_2^{\tau+1} \propto p(x_2 | x_1^{(\tau+1)}, x_3^{(\tau)}, \ldots, x_n^{(\tau)})$
  $\vdots$
  Sample $x_j^{\tau+1} \propto p(x_j | x_1^{(\tau+1)}, \ldots, x_{j-1}^{(\tau+1)}, x_{j+1}^{(\tau)}, \ldots, x_n^{(\tau)})$
  $\vdots$
  Sample $x_n^{\tau+1} \propto p(x_n | x_1^{(\tau+1)}, x_2^{(\tau+1)}, \ldots, x_{n-1}^{(\tau+1)})$
**end for**

correlated, Gibbs sampling will converge toward a steady state that is the desired distribution Gelman et al. (2004).

The LDA graphical model is easily extended and several variations of LDA have been developed. Of relevance to this work especially are the models that condition for topics based on additional geographic information. While these models consider geographic information in the form of place labels Mei et al. (2006) or location information Eisenstein et al. (2010); Adams and Janowicz (2012), none of these models consider the actual features present at the places being described. The GFTTM presented in the next

| Symbol | Meaning | Data type representation |
|---|---|---|
| $D$ | Number of documents in corpus | scalar |
| $F$ | Number of feature types for corpus | scalar |
| $T^{feat}$ | Number of feature type topics | scalar |
| $T^{abst}$ | Number of abstract topics | scalar |
| $\mathbf{p_d}$ | Feature type distribution for document $d$ | $F$-dim vector |
| $N_d$ | Number of words in document $d$ | scalar |
| $W$ | Vocabulary size (number of unique words in corpus) | scalar |
| $\mathbf{W}$ | Corpus observation data | $D \times W$ sparse matrix |
| $\phi_t^{feat}$ | Probabilities of words given feature topic $t$ | $W$-dim vector |
| $\phi_t^{abst}$ | Probabilities of words given abstract topic $t$ | $W$-dim vector |
| $\varepsilon_f$ | Probabilities of feature type topics given feature type $f$ | $F$-dim vector |
| $\phi^{feat}$ | $W \times T^{feat}$ matrix | |
| $\phi^{abst}$ | $W \times T^{abst}$ matrix | |
| $\theta_d$ | Probabilities of abstract topics given document $d$ | $T^{abst}$-dim vector |
| $\pi_d$ | Binary switch probabilities on feature vs. abstract topic for document $d$ | 2-dim vector |
| $x_{di}$ | Binary switch assignment for word $i$, document $d$ | $\in \{abst, feat\}$ |
| $f_{di}$ | Feature type assignment for word $i$, document $d$ | element of F |
| $z_{di}$ | Topic assignment for word $i$, document $d$ | element of Tfeat $\bigcup$ Tabst |
| $w_{di}$ | Word assignment for word $i$, document $d$ | element of W |
| $\alpha$ | Dirichlet prior | |
| $\beta^{feat}$ | Dirichlet prior | |
| $\beta^{abst}$ | Dirichlet prior | |
| $\psi$ | Dirichlet prior | |
| $\gamma$ | Beta prior (parameters $\alpha$, $\beta$) | |

**Table 1.** Definitions for symbols used in this paper.

section is a generative model built on the assumption that certain physical features in the environment will generate some of the words in a place description.

# 3 GEOGRAPHIC FEATURE TYPE TOPIC MODEL (GFTTM)

The GFTTM is a generative model in the same vein as LDA and its many extensions. The observed data of a corpus of place descriptions is modeled as being randomly generated from a set of abstract topics and feature type topics, which are conditioned on a second set of observation data describing the features of that place. Like these other models the estimation of the parameters can be approximated using Gibbs sampling, the algorithm for which is described below. Table 1 is a reference listing of all symbols used subsequently along with their meanings.

## 3.1 Generative model

A document that describes a place is assumed to be a mixture of feature type topics and abstract topics. The mixture of feature type topics for a document is based on the relative densities of feature types within the spatial extent of the place described in the document. Let $D$ be the number of documents in a corpus, $F$ be the number of feature types, $T^{feat}$ be the number of feature type topics, $T^{abst}$ be the number of abstract topics. The generative process (shown in plate notation in Figure 2) for creating a corpus of place descriptions is defined as follows:

1. (a) For each feature type topic $t^{feat} \in \{1 \ldots T^{feat}\}$ choose a multinomial distribution over words $\phi_t^{feat} \propto Dirichlet\left(\beta^{feat}\right)$

   (b) For each abstract topic $t^{abst} \in \{1 \ldots T^{abst}\}$ choose a multinomial distribution over words $\phi_t^{abst} \propto Dirichlet\left(\beta^{abst}\right)$

   (c) For each feature type $f \in \{1 \ldots F\}$ choose a multinomial distribution over feature type topics $\varepsilon_f \propto Dirichlet\left(\psi\right)$



PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.816v1 | CC-BY 4.0 Open Access | rec: 5 Feb 2015, publ: 5 Feb 2015

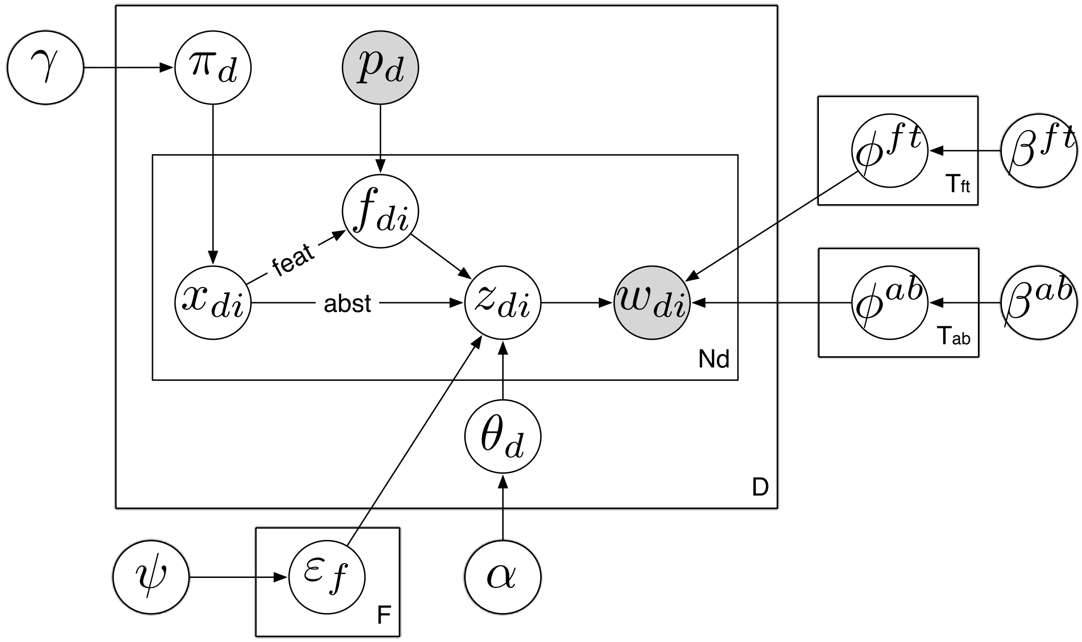


**Figure 2.** Geographic feature type topic model plate notation

2. For each document $d \in \{1 \ldots D\}$
   (a) Given a multinomial distribution over feature types in place described in document, $p_d$ (F-dimensional vector)
   (b) Choose a multinomial distribution over abstract topics $\theta_d \propto Dirichlet\,(\alpha)$
   (c) Choose a binomial distribution over feature type topics vs. abstract topics $\pi_d \propto Beta\left(\gamma^{feat}, \gamma^{abst}\right)$
   (d) For each word $w_{di}$ in document $d$
      i. Choose a binary switch $x_{di} \propto Binomial\,(\pi_d)$
      ii. If $x_{di} = abst$, choose an abstract topic $z_{di} \propto Multinomial\,(\theta_d)$
      Else if $x_{di} = feat$, choose a feature type $f_{di} \propto Multinomial\,(p_d)$ and then choose a feature type topic $z_{di} \propto Multinomial\left(\varepsilon_{f_{di}}\right)$
      iii. Choose a word $w_{di} \propto Multinomial\left(\phi_{z_{di}}^{x_{di}}\right)$

Given this model, the probability of a corpus, **W**, being generated is shown in Figure 3.

### 3.2 Gibbs sampling inference

In this section a method is described for using Gibbs sampling to estimate the parameters of the GFTTM given the observed variables: the corpus **W** and feature type distributions **P** and hyperparameters: $\beta^{feat}$, $\beta^{abst}$, $\psi$, $\alpha$, $\gamma$. In a Gibbs sampling simulation a series of iterations is run where the topic assignment for each word in the corpus is updated in sequence (see Griffiths and Steyvers (2004)). An update rule is applied to determine what topic a word should take on in the current iteration given the assignments of topics for all the other words in the corpus. The update rule defines the proportional weight for each possible assignment and then a random selection is made based on those weights and the word is updated. This calculation is done for each word in one iteration of the algorithm.

Let the following list of definitions hold for the update rules defined below.

- $n_{w,\sim di}^{abst\,z}$ : number of times word $w$ is assigned to abstract topic $z$.
- $n_{d,\sim di}^{abst\,z}$ : number of times a word in document $d$ is assigned to abstract topic $z$.

$$
\begin{aligned}
P(\mathbf{W}|\phi^{ft},\phi^{ab},\varepsilon) &= \prod_{d=1}^{D} P\left(\mathbf{w_d}|\phi^{ft},\phi^{ab},\varepsilon,\theta_d,\pi_d,p_d\right)\\
&= \prod_{d=1}^{D}\prod_{i=1}^{N_d} P\left(w_{di}\,|\,\phi^{ft},\phi^{ab},\varepsilon,\theta_d,\pi_d,p_d\right)\\
&= \prod_{d=1}^{D}\prod_{i=1}^{N_d}\Bigg[P\left(x_{di}=\text{abst}\,|\,\pi_d\right)\sum_{t=1}^{T^{ab}} P\left(w_{di}\,|\,z_{di}=t,\phi_t^{ab}\right)P\left(z_{di}=t\,|\,\theta_d\right)\\
&+ P\left(x_{di}=\text{feat}\,|\,\pi_d\right)\sum_{t=1}^{T^{ft}} P\left(w_{di}\,|\,z_{di}=t,\phi_t^{ft}\right)\sum_{f=1}^{F} P\left(z_{di}=t\,|\,f_{di}=f,\varepsilon_f\right)P\left(f_{di}=f\,|\,p_d\right)\Bigg]
\end{aligned}
$$

**Figure 3.** Probability of a corpus

$$
\begin{aligned}
&P\left(x_{di}=abst,z_{di}=z\,|\,w_{di}=w,\mathbf{x}_{\sim di},\mathbf{z}_{\sim di},\mathbf{w}_{\sim di},\alpha,\beta^{abst},\gamma\right)\\
&\propto \frac{n_{w,\sim di}^{abst\,z}+\beta^{abst}}{\sum_{w\prime} n_{w\prime,\sim di}^{abst\,z}+W\beta^{abst}}\cdot\frac{n_{d,\sim di}^{abst\,z}+\alpha}{n_{d,\sim di}^{abst}+T^{abst}\alpha}\cdot\left(n_{d,\sim di}^{abst}+\gamma^{abst}\right)
\end{aligned}
$$

**Figure 4.** Gibbs sampling update rule for abstract topics

- $n_{d,\sim di}^{abst}$ : number of times a word in document $d$ is assigned to an abstract topic.
- $n_{w,\sim di}^{feat\,z}$ : number of times the word $w$ is assigned to feature type topic $z$.
- $n_{z,\sim di}^{f}$ : number of times feature type topic $z$ is assigned to feature type $f$.
- $n_{z,\sim di}^{F}$ : number of times feature type topic $z$ is assigned to a feature type.
- $n_{d,\sim di}^{feat}$ : number of times a word in document $d$ is assigned to a feature type topic.
- $n_{d,\sim di}^{f}$ : number of times a word in document $d$ is assigned to feature type $f$.
- $\mathbf{p_d}$ : probability vector of features for document $d$.
- $p_{df}$ : probability of feature type $f$ for document $d$.

The Gibbs sampling update rule for abstract topics is shown in Figure 4. This update rule is similar to the update rule for Gibbs sampling LDA Griffiths and Steyvers (2004). The update rule for feature type topic is shown in Figure 5. Thus, the probability of a word being assigned a given feature type topic is proportional to the probability of the feature given the document times the probability of other topics assigned to that feature and the probability of other words assigned to that feature type topic.

$$
\begin{aligned}
&P\left(x_{di}=feat,f_{di}=f,z_{di}=z\,|\,w_{di}=w,\mathbf{x}_{\sim di},\mathbf{z}_{\sim di},\ \mathbf{w}_{\sim di},\mathbf{p}_d,\psi,\beta^{feat},\gamma\right)\\
&\propto \frac{n_{w,\sim di}^{feat\,z}+\beta^{feat}}{\sum_{w\prime} n_{w\prime,\sim di}^{feat\,z}+W\beta^{feat}}\cdot\frac{p_{df}}{n_{d,\sim di}^{f}/n_{d,\sim di}^{feat}}\cdot\frac{n_{z,\sim di}^{f}+\psi}{n_{z,\sim di}^{F}+F\psi}\cdot\left(n_{d,\sim di}^{feat}+\gamma^{feat}\right)
\end{aligned}
$$

**Figure 5.** Gibbs sampling update rule for feature type topics



PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.816v1 | CC-BY 4.0 Open Access | rec: 5 Feb 2015, publ: 5 Feb 2015

### 3.3 Similarity of feature types

The similarity of feature types can be measured in terms of the relative entropy of feature type topic probabilities for any pair of feature types. For a symmetric measure it is a function of the Jensen-Shannon divergence between the multinomial distributions of topics for two feature types. Jensen-Shannon divergence is defined in Equation 1.

$$JS(P \parallel Q) = \frac{1}{2} D_{\mathrm{KL}}(P \parallel M) + \frac{1}{2} D_{\mathrm{KL}}(Q \parallel M), \tag{1}$$

where $P$ and $Q$ are two multinomial distributions, $D_{\mathrm{KL}}$ is the Kullback-Leibler divergence (defined in Equation 2), and $M = (P+Q)/2$.

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}. \tag{2}$$

In this context the similarity of feature types is based entirely on the probabilities of words used, conditioned on the presence of specific feature types. This similarity measure can be combined with other methods of measuring semantic similarity of feature types (e.g., based on ontologies) Janowicz et al. (2011).

### 3.4 Similarity of places

The decomposition of documents into feature type topics and abstract topics allows us also to compare the similarity of places in terms of these two different types of topics. We can, for example, use the documents associated places to compare two place either in terms of similar descriptions of feature types at those places, or in terms of abstract topics. This allows us to explore the ways in which places are similar or different in a more nuanced way.

## 4 EVALUATION

To evaluate the GFTTM a training set of georeferenced Wikipedia articles from the United States and geographic feature data for the United States from Geonames.org[1] were used. Geonames.org classifies each feature with one of 645 feature codes [2]. These feature codes are defined informally with short textual descriptions and are organized into a shallow taxonomy of 9 broad categories (A country, state, region; H stream, lake; L parks, area; P city, village; R road, railroad; S spot, building; T mountain, hill, rock; U undersea; V forest, heath). Figure 6 shows a sample of these two sources of data that are georeferenced in the vicinity of Yosemite valley in California.

After removing all articles with fewer than 200 words, the total training corpus consisted of 36,994 Wikipedia articles with 75,772 unique terms. To generate the appropriate feature type ratios for each document, the set of all geographic features within 5 km of the georeference location for the article were extracted from the Geonames database. Extremely rare or sparse feature codes were removed (i.e., those where there are never more than four within 5 km of the georeference location for any article). Following the removal of rare feature types there were 85 feature codes remaining. Because some features, such as building (S.BLDG) are much more common than others, such as mountain (T.MT), the feature counts are normalized based on the maximum number of a given feature type within any 5 km buffer zone (See Table 4 for maximum feature type counts within 5km radii of georeferenced Wikipedia articles). The rationale for this is that any feature type that is disproportionately more present than normal within a place is more likely to be described in a written account of that place. For example, although there are relatively few mountains compared to the number of buildings in the Los Angeles area, descriptions of L.A. will likely reference mountain related words more often than based on simple counts of features, because there are more mountains than an average American city.

Several experiments were run, varying the hyperparameter values. Picking the appropriate hyperparameter values for topic modeling is a bit of an art and will depend on the size and quality of the corpus as well as the number of topics that are being modeled. However, in our experiments we found that $\alpha = 0.5$,

[1] http://geonames.org
[2] http://www.geonames.org/export/codes.html



PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.816v1 | CC-BY 4.0 Open Access | rec: 5 Feb 2015, publ: 5 Feb 2015

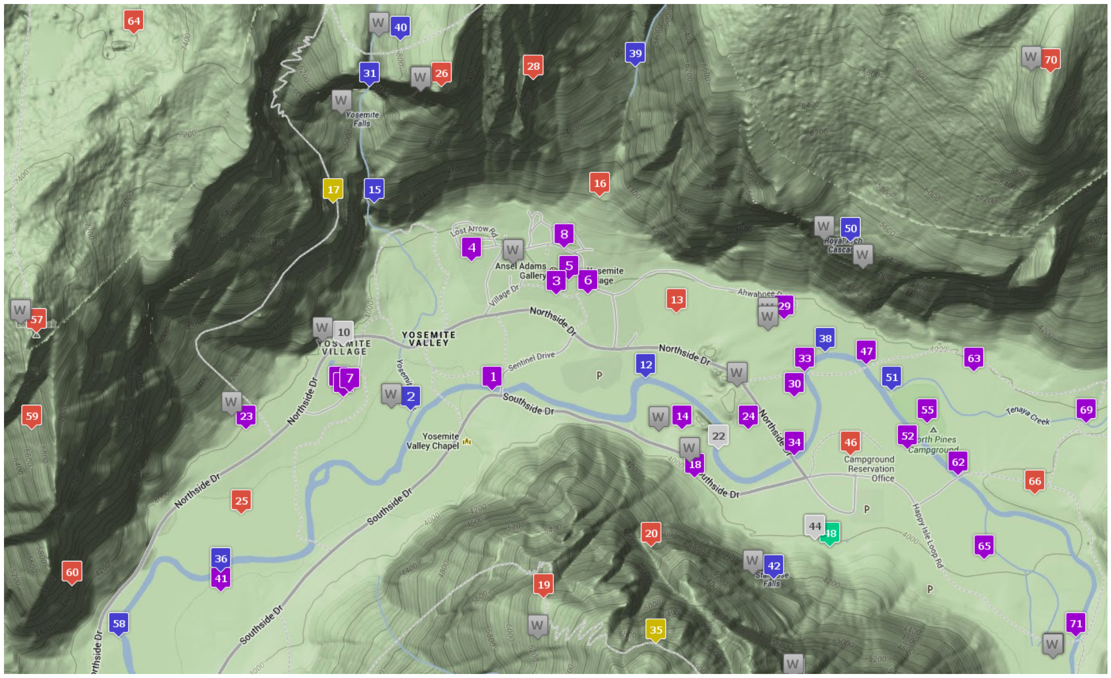


**Figure 6.** Sample of geonames.org features and georeferenced Wikipedia articles in the vicinity of Yosemite valley. The 'W' icons refer to Wikipedia articles and the numbered icons refer to geonames.org features. The colors correspond with the 9 broad feature type categories defined by geonames.org.

| Topics | Top words |
|---|---|
| Feat. type Topic 1 | school student team community high city college area develop year athletic district |
| Feat. type Topic 2 | design build hall student study library art plan artist collect east park |
| Feat. type Topic 3 | trail area forest camp mountain park locate rock day dam rang nation |
| Feat. type Topic 4 | air force base fort army command train war squadron wing military |
| Feat. type Topic 5 | school house event day make open built east renovate summer annual surround |
| Abstract Topic 1 | apache territory mexico indian seminole mexican florida spanish reserve navajo santa |
| Abstract Topic 2 | british force command army battle attack burgoyne hooper arnold air advance general |
| Abstract Topic 3 | damage tornado tree destroy path unknown length coord utc source comment sustain |
| Abstract Topic 4 | flight crash aircraft airline accid pilot atr plane crew faa ntsb cargo |
| Abstract Topic 5 | limit speed mph ishi road truck yahi interstate rural statute highway divide |

**Table 2.** Sample topics found for Wikipedia and Geonames data (trained for 50 feature type topics and 100 abstract topics).



PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.816v1 | CC-BY 4.0 Open Access | rec: 5 Feb 2015, publ: 5 Feb 2015

| Feature type topic most-similar | Abstract topic most-similar |
|---|---|
| 1. Grand Canyon | 1. History of the Grand Canyon |
| 2. Dogpatch USA | 2. 1935 Labor Day hurricane |
| 3. Shawangunk Ridge | 3. Egg Rock |
| 4. Longmire, Washington | 4. Tantiusques |
| 5. Old Man of the Mountain | 5. Metacomet Trail |
| 6. Bandera County, Texas | 6. Weymouth Back River |
| 7. Panther Mountain (New York) | 7. Seven Falls |
| 8. Kalalau Valley | 8. Staten Island Greenbelt |
| 9. Hays County, Texas | 9. Eleven Jones Cave |
| 10. Rapidan Camp | 10. Stone Mountain |

**Table 3.** Top-10 most similar georeferenced Wikipedia articles to *Yosemite Valley* in terms of feature type topics and abstract topics (excluding all places within 100 km).

$\beta^{feat} = 0.1$, $\beta^{abst} = 0.1$, $\psi = 0.1$, $\gamma^{feat} = 0.01$, and $\gamma^{abst} = 2.0$ worked well to generate subjectively meaningful feature type and abstract topics.

To demonstrate the difference between the feature type and abstract topic mixtures for a place Table 3 shows the most similar place articles to the English Wikipedia `Yosemite_Valley` article in terms of feature type and abstract topics. The JS divergence as described in Equation 1 was used to find the top-10 most similar place articles.

## 5 RESULTS AND DISCUSSION

Table 2 shows sample results for the top words for both feature type topics and abstract topics from the training data. For these results we trained for 50 feature type topics and 100 abstract topics. The results shown in Table 2 are promising. For example, the terms *force* and *army* are assigned both to feature type topic 4 and abstract topic 2. In the abstract topic these terms are associated with other terms that are found in articles about historical battles. For example, there is reference to John *Burgoyne* and Benedict *Arnold*, famous generals from the American revolutionary war. In comparison for the feature type topic they are associated with other terms that will also be found in documents co-located with the S.INSM (military installation) feature type. An examination of the feature type topic mixture for S.INSM confirmed this as this topic was the primary topic, mixed with small amounts of feature type topics associated with the terms *island, bay, harbor* and *street, building, city*.

At first glance the similar places shown in Table 3 might not show a clear distinction between places that are feature type-similar and those that are abstract-similar. Both are heavily dominated by important natural features and places that have low human population and few manmade structures. However, upon closer examination of the content of the articles in question it is clear that the place descriptions in the abstract column nearly all contain significant historical sections variously describing Native American populations, national park history, and European settlement – all topics found in the Yosemite Valley article. The articles found in the feature type topic column do not contain these topics to such a degree.

One interesting result from this model is that the abstract topics seem to be more specific than in the traditional LDA model. In particular, words that are generally of lower probability in LDA topics, show up higher in GFTTM abstract topics. A possible explanation is that many of the more common terms in LDA topics are assigned to feature type topics allowing more rare words to move up in the abstract topics (which correspond to traditional LDA topics in the model). Further investigation into this phenomenon is merited.

## 6 RELATED WORK

Understanding the relationships between topics and geographic locations has been a very active research area in recent years. In this section we present relevant related work on spatial and geographic topic modeling.

| Feat. code | Max. | Name | Feat. code | Max. | Name |
|---|---|---|---|---|---|
| A.ADMD | 32 | administrative division | S.HSP | 63 | hospital |
| H.BAY | 34 | bay | S.HTL | 280 | hotel |
| H.CHN | 19 | channel | S.INSM | 22 | military installation |
| H.CNL | 34 | canal | S.LIBR | 89 | library |
| H.GLCR | 14 | glacier(s) | S.LTHSE | 6 | lighthouse |
| H.HBR | 43 | harbor(s) | S.MALL | 38 | mall |
| H.INLT | 20 | inlet | S.MAR | 21 | marina |
| H.LK | 45 | lake | S.MN | 466 | mine(s) |
| H.OVF | 23 | overfalls | S.MNMT | 23 | monument |
| H.RPDS | 10 | rapids | S.MNQ | 18 | abandoned mine |
| H.RSV | 25 | reservoir(s) | S.MUS | 24 | museum |
| H.RSVT | 25 | water tank | S.OBPT | 9 | observation point |
| H.SPNG | 91 | spring(s) | S.PKLT | 13 | parking lot |
| H.STM | 39 | stream | S.PO | 30 | post office |
| H.STMB | 7 | stream bend | S.RECG | 9 | golf course |
| H.SWMP | 31 | swamp | S.REST | 105 | restaurant |
| H.WLL | 132 | well | S.RSRT | 14 | resort |
| L.AREA | 9 | area | S.RSTN | 24 | railroad station |
| L.INDS | 26 | industrial area | S.RSTNQ | 21 | aband. railroad station |
| L.LCTY | 11 | locality | S.SCH | 436 | school |
| L.OILF | 7 | oilfield | S.SQR | 8 | square |
| L.PRK | 267 | park | S.STDM | 11 | stadium |
| L.RESW | 5 | wildlife reserve | S.SWT | 5 | sewage treatment plant |
| P.PPL | 398 | populated place | S.THTR | 19 | theater |
| P.PPLQ | 17 | aband. populated place | S.TOWR | 35 | tower |
| P.PPLX | 193 | sect. of populated place | S.WHRF | 30 | wharf(-ves) |
| R.RDJCT | 42 | road junction | T.BAR | 25 | bar |
| R.TNL | 8 | tunnel | T.BCH | 14 | beach |
| R.TRL | 106 | trail | T.BNCH | 5 | bench |
| S. | 134 | spot | T.CAPE | 27 | cape |
| S.AIRH | 31 | heliport | T.CLF | 11 | cliff(s) |
| S.AIRP | 7 | airport | T.CRTR | 10 | crater(s) |
| S.ARCH | 17 | arch | T.DPR | 13 | depression(s) |
| S.BDG | 28 | bridge | T.GAP | 20 | gap |
| S.BLDG | 1140 | building(s) | T.ISL | 47 | island |
| S.CH | 344 | church | T.LAVA | 7 | lava area |
| S.CMP | 32 | camp(s) | T.LEV | 7 | levee |
| S.CMPQ | 7 | abandoned camp | T.MT | 34 | mountain |
| S.CMTY | 47 | cemetery | T.PLN | 8 | plain(s) |
| S.DAM | 21 | dam | T.RDGE | 16 | ridge(s) |
| S.FRM | 22 | farm | T.VAL | 27 | valley |
| S.FRMQ | 21 | abandoned farm | V.FRST | 17 | forest(s) |
| S.HSE | 8 | house(s) | | | |

**Table 4.** Feature codes and maximum counts within 5 km radius of any georeferenced Wikipedia article

### 6.1 Geographic topic models

One of the original probabilistic text mining approaches for finding themes associated with spatial regions was developed by Mei et al. (2006) and was evaluated for analyzing thematic change in blog entries. It uses a simpler unigram model than LDA Blei et al. (2003) for topics but mines for spatiotemporal patterns in the topics. The location associated with a document is a place label, not a spatially referenced location in longitude and latitude. The Location Aware topic model Wang et al. (2007) alters the LDA model to generate not only words from topics but also a location, where a "location" is a place name label. Each topic is characterized as a multinomial distribution over locations, just like words. The Geographic Topic Model (GTM) is a truly spatial topic model that generates spatial regions associated with topics Eisenstein et al. (2010). GTM is a cascading topic model that selects words conditioned by base topics that are further conditioned on spatial regions. The Location Topic model uses a binary switch variable similar to the one utilized in the GFTTM to differentiate between *local* and *global* topics to provide recommendations based on travelogue topics Hao et al. (2010). Hong et al. (2012) developed a model to find geographical topics in Twitter[3] and similar microblogging services, which conditions on geographic location as well as user information based on the assumption that individual users of Twitter will tend to be geographically localized.

### 6.2 Other approaches using topic models

In addition to the above models, which explicitly model geographic or spatial topics, more general-purpose topic models have been utilized to identify geographic topics. The relational topic model (RTM) models not only the documents in the corpus but also the network of links between the documents, and the RTM has been used to train for regional topics by linking documents that share a spatial relationship, e.g., they are tagged with the same geographic region Chang et al. (2010). The Supervised Latent Dirichlet allocation (SLDA) model associates an outcome variable with each topic and in their evaluation of GTM, Eisenstein et al. (2010) used a version of SLDA that models Gaussian distributions over the longitude and latitude trained from points associated with documents Blei and McAuliffe (2007). Adams and Janowicz (2012) presented a method for finding topics associated with places by doing a post-hoc analysis of the mean probabilities of basic LDA topics associated with geo-referenced documents.

## 7 CONCLUSION

In this paper a novel topic model, GFTTM, was proposed. GFTTM conditions some topics on the presence of feature types while other topics are treated as normal LDA topics. The model was evaluated using volunteered geographic information from Wikipedia and the Geonames.org website. The results of this evaluation demonstrated that the abstract topics and feature type topics trained using GFTTM form two distinct types of topics. These topics can be used to disentangle how places are described in terms of its physical features and more abstract topics such as history and culture.

GFTTM relies on a mapping between documents and feature data points based some degree of co-location. Therefore, the results of GFTTM must always be evaluated with due consideration of issues of scale and accuracy of the geo-location in both sets of source data. Furthermore, although the GFTTM is a more sophisticated generative model for how place descriptions are written, because it is trained on two sources of evidence rather than one it has two degrees of freedom for mismatches between the geocoded location and the actual place being described in the text. In addition, being a more complex model than LDA, inference is significantly slower than for LDA. Further investigation using different data sources will be needed to evaluate its practical usefulness for specific application domains.

## REFERENCES

Adams, B. and Janowicz, K. (2012). On the geo-indicativeness of non-georeferenced text. In Breslin, J. G., Ellison, N. B., Shanahan, J. G., and Tufekci, Z., editors, *ICWSM*, pages 375–378. The AAAI Press.

Adams, B. and McKenzie, G. (2013). Inferring thematic places from spatially referenced natural language descriptions. In Sui, D., Elwood, S., and Goodchild, M., editors, *Crowdsourcing Geographic Knowledge*, pages 201–221. Springer Netherlands.

Agnew, J. (1987). *Place and politics: the geographical mediation of state and society*. Allen and Unwin.

[3] https://twitter.com/

Bailey, G., Wikle, T., Tillery, J., and Sand, L. (1993). Some patterns of linguistic diffusion. *Language variation and change*, 5(03):359–390.

Bennett, B., Mallenby, D., and Third, A. (2008). An ontology for grounding vague geographic terms. In *Formal Ontology in Information Systems - Proceedings of the Fifth International Conference (FOIS 2008)*, volume 183, pages 280–293. IOS Press.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York.

Blei, D. and McAuliffe, J. (2007). Supervised topic models. In *Advances in Neural Information Processing Systems*.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Brown, G. and Raymond, C. (2007). The relationship between place attachment and landscape values: Toward mapping place attachment. *Applied geography*, 27(2):89–111.

Burenhult, N. and Levinson, S. C. (2008). Language and landscape: a cross-linguistic perspective. *Language Sciences*, 30(2):135–150.

Chang, J., Blei, D. M., et al. (2010). Hierarchical relational models for document networks. *The Annals of Applied Statistics*, 4(1):124–150.

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.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian data analysis*. Texts in Statistical Science Series. Chapman & Hall/CRC, Boca Raton, FL.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.

Gregory, I. N. and Hardie, A. (2011). Visual GISting: bringing together corpus linguistics and geographical information systems. *Literary and linguistic computing*, 26(3):297–314.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235.

Hao, Q., Cai, R., Wang, C., Xiao, R., Yang, J.-M., Pang, Y., and 0001, L. Z. (2010). Equip tourists with knowledge mined from travelogues. In Rappa, M., Jones, P., Freire, J., and Chakrabarti, S., editors, *WWW*, pages 401–410. ACM.

Hong, L., Ahmed, A., Gurumurthy, S., Smola, A. J., and Tsioutsiouliklis, K. (2012). Discovering geographical topics in the twitter stream. In Mille, A., Gandon, F. L., Misselis, J., Rabinovich, M., and Staab, S., editors, *WWW*, pages 769–778. ACM.

Janowicz, K. and Raubal, M. (2007). Affordance-based similarity measurement for entity types. In Winter, S., Duckham, M., Kulik, L., and Kuipers, B., editors, *COSIT*, volume 4736 of *Lecture Notes in Computer Science*, pages 133–151. Springer.

Janowicz, K., Raubal, M., and Kuhn, W. (2011). The semantics of similarity in geographic information retrieval. *Journal of Spatial Information Science*, (2):29–57.

Kuhn, W. (2002). Modeling the semantics of geographic categories through conceptual integration. In Egenhofer, M. J. and Mark, D. M., editors, *GIScience*, volume 2478 of *Lecture Notes in Computer Science*, pages 108–118. Springer.

Kurath, H. (1949). *A word geography of the Eastern United States*. University of Michigan Press.

Labov, W. (2007). Transmission and diffusion. *Language*, 83(2):344–387.

Mackay, D. J. C. (2005). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, Cambridge, UK, 7.2 edition.

Mark, D. M. and Turk, A. G. (2003). Landscape categories in yindjibarndi: Ontology, environment, and language. In Kuhn, W., Worboys, M., and Timpf, S., editors, *Spatial Information Theory. Foundations of Geographic Information Science*, volume 2825 of *Lecture Notes in Computer Science*, pages 28–45. Springer Berlin Heidelberg.

Mark, D. M., Turk, A. G., and Stea, D. (2007). Progress on yindjibarndi ethnophysiography. In Winter, S., Duckham, M., Kulik, L., and Kuipers, B., editors, *COSIT*, volume 4736 of *Lecture Notes in Computer Science*, pages 1–19. Springer.

Mark, D. M., Turk, A. G., and Stea, D. (2012). Ethnophysiography of arid lands. *Landscape Ethnoecology:*





PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.816v1 | CC-BY 4.0 Open Access | rec: 5 Feb 2015, publ: 5 Feb 2015

*Concepts of Biotic and Physical Space*, 9:27–45.

Mei, Q., Liu, C., Su, H., and Zhai, C. (2006). A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In Carr, L., Roure, D. D., Iyengar, A., Goble, C. A., and Dahlin, M., editors, *WWW*, pages 533–542. ACM.

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 *EMNLP*, pages 248–256. ACL.

Sinha, G. and Mark, D. (2010). Toward a foundational ontology of the landscape. *Extended Abstracts of GIScience*, 2010.

Steyvers, M. and Griffiths, T. (2007). Probabilistic topic models. In Landauer, T., Mcnamara, D., Dennis, S., and Kintsch, W., editors, *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates.

Trudgill, P. (1974). Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography. *Language in Society*, 3(2):215–246.

Tuan, Y.-F. (1974). *Topophilia: A study of environmental perception, attitudes, and values*. Columbia University Press.

Wang, C., Wang, J., Xie, X., and Ma, W.-Y. (2007). Mining geographic knowledge using location aware topic model. In Purves, R. and Jones, C., editors, *GIR*, pages 65–70. ACM.





PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.816v1 | CC-BY 4.0 Open Access | rec: 5 Feb 2015, publ: 5 Feb 2015