Text Mining Methods for Mapping Opinions from Georeferenced Documents

Duarte Choon Dias
dcd@ist.utl.pt
Instituto Superior Técnico
Av. Professor Cavaco Silva
2744-016 Porto Salvo,
Portugal

ABSTRACT

Most documents can be said to be related to some form of geographic context, although traditional text mining methods simply model documents as bags of tokens, ignoring other aspects of the encoded information. Recently, geographic information retrieval has captured the attention of many different researchers that work in fields related to text mining and data retrieval, envisioning the support for tasks such as map-based document indexing, retrieval and visualization. This work concerns with the study of automated methods, based on language model classifiers, for assigning documents to opinion classes and to the geospatial coordinates of latitude and longitude that best summarize their contents. Using this information, I then analyzed the possibility of building thematic maps portraying the incidence of particular classes of opinions, as extracted from documents, in different geographic areas. An extensive experimental validation has been carried out, using documents from Wikipedia and reviews from Yelp. The best performing method for geocoding textual documents combines character-based language models with a post-processing technique that uses the coordinates from the 5 most similar training documents, obtaining an average prediction error of 265 Kilometers, and a median prediction error of just 22 Kilometers. In what concerns opinion classification, the best performing method also used character-based language models, obtaining an accuracy of 0.80 when performing the classification on a two-point scale, and of 0.5 when considering a five-point opinion polarity scale. A technique known as Kernel Density Estimation was used for the construction of thematic maps, and an empirical analyzes has shown that the maps obtained through automatic extraction indeed correspond to an accurate representation for the geographic distribution of opinions.

Keywords

Text Processing, Geographic Information Retrieval, Document Geocoding, Language Model Classifier, Opinion Mining, Thematic Mapping

1. INTRODUCTION

Most textual documents can be said to be related to some form of geographic context and, recently, Geographical Information Retrieval (GIR) has captured the attention of many different researchers that work in fields related to retrieving and mining contents from large document collections. We have for instance that the task of resolving individual place references in textual documents has been addressed in several previous works, with the aim of supporting subsequent GIR processing tasks, such as document retrieval or cartographic visualization of text documents [16]. However, place reference resolution presents several non-trivial challenges [15, 17, 2], due to the inherent ambiguity of natural language discourse (e.g., place names often have other non geographic meanings, different places are often referred to by the same name, and the same places are often referred to by different names). Moreover, we have that there are many vocabulary terms, besides place names, that can frequently appear in documents related to specific geographic areas, and GIR applications can also benefit from this information. Instead of trying to correctly resolve the individual references to places that are made in textual documents, it may be interesting to instead study methods for assigning entire documents to geospatial locations [25, 1].

Geographical information can be used to make many types of interesting geographical analysis. In my Msc thesis I am interested in extracting the opinions of people towards a certain topic or service, and analyze these opinions in a geographical context, through the usage of thematic maps.

The extraction of opinions from textual documents is also a recent and growing area of research. There are many applications for this new technology, such as the classification of movie and product reviews. With the increasing number of opinionated documents available on the Internet, it becomes increasingly important to analyze the underlying opinions expressed in their contents. Opinion mining can be, for instance, done at many levels, namely at the document level, sentence level or phrase level. This work addressed the classification of opinions at a document level.

My thesis relates to exploring automated techniques to identify the geographical location that best describes the content of textual documents, with the objective of building a system that discovers and maps opinions towards certain themes, expressed in the context of particular locations. This system has three major modules, namely one for georeferencing textual documents, another for mining opinions expressed in textual documents, and finally a module for the construction of maps with the geographical distribution of opinions. In my MSc thesis, I studied and compared automated techniques for georeferencing textual documents, and for mining opinions expressed in them, using only the raw text as evidence. Both text mining problems were addressed using language model classifiers, where compact representations of both sets of documents were built using character-based or token-based language models capturing the document's main statistical properties.

In order to georeference textual documents, I relied on a discrete binned representation of the Earth's surface. The bins from this representation, corresponding to equally-distributed triangular areas of the Earth's surface, are initially associated to textual documents (i.e., all the documents from a training set that are known to refer to each particular bin were used). New documents are then assigned to the most similar bin(s). Documents are finally assigned to their respective coordinates of latitude and longitude, with basis on the centroid coordinates associated to the bin(s). I experimented with different post-processing techniques for assigning the geospatial coordinates in the final step, namely by using (i) the centroid coordinates of the most probable bin, (ii) a weighted average with the coordinates from the most probable bins, (iii) a weighted average with the geospatial coordinates from the neighbouring bins, and (iv) a weighted average with the coordinates from the *knn* most similar training documents contained within the most probable bin. Experiments with a collection of Wikipedia articles showed good results for the proposed general document geocoding approach. The best performing method combines hierarchical classifiers based on character-based language models, with the post-processing technique that uses the weighted average with the coordinates from the 5 most similar documents within the most probable bin, obtaining an average prediction error of 265 Kilometers, and a median prediction error of just 22 Kilometers.

The analysis of opinions expressed in textual documents was done using two different scales, namely a two-point, and a five-point opinion scale. In the case of the five-point scale, I also experimented with a meta-algorithm know as metric labeling, initially designed for ordinal regression problems [20]. The algorithm corrects the assigned class labels, in order to assure that similar data receives similar labels. Also in order to increase the computational performance of both the training and classifying process, I developed a hierarchical classification method. Experiments with a collection of reviews extracted from `yelp.com` showed that using language models to mine the opinions from textual documents presents promising results. The best classification method for both scenarios relied on character-based language models, obtaining an accuracy of 0.8 in the two-point scale. For the five-point scale case, the best method relied on a hierarchical classifier, obtaining an accuracy of 0.5.

Using the geographical information and the opinions extracted from documents, I studied techniques in order to represent the geographical distribution of opinions. A technique know as Kernel Density Estimation was used to estimate density surfaces, representing the distribution of different opinion classes.

The rest of this paper is organized as follows: Section 2 presents related work, while Section 3 details the proposed text mining methods for document geocoding and opinion mining, as well as the methods that are used for creating the thematic maps. Section 4 presents the experimental validation of the proposed methods, describing the considered datasets, the evaluation protocol, and the obtained results for different configurations of the text mining methods. Section 4 also presents several examples of thematic maps derived automatically from textual data. Finally, Section 5 presents the most important conclusions and points possible directions for future work.

2. RELATED WORK

The relationship between language and geography has long been a topic of interest to linguists [12]. Many studies have, for instance, shown that geography has an impact on the relationship between vocabulary terms and semantic classes. For instance the term *football*, in the United States, refers to the particular sport of American football. However, in regions such as Europe, the term *football* is usually associated to different sports (e.g., soccer or, less frequently, rugby football). Terms such as *beach* or *snow* are also more likely to be associated to particular locations. In this study, we are interested in seeing if vocabulary terms, and textual contents in general, can be used to predict geographical locations.

Overell investigated the use of Wikipedia as a source of data for georeferencing textual articles, in addition to article classification by category, and individual place reference resolution [18]. Overell's main goal was to resolve place references in documents, for which global document georeoreferencing could serve as an input feature. For document georeferencing, Overell proposed a simple model that uses only the metadata available (e.g., article title, incoming and outgoing links, etc.) and not the actual text.

Anastácio et al. surveyed heuristic approaches to assign documents to geographic scopes, based on recognizing place references in the documents and afterwards combining the recognized references [3]. The authors specifically compared approaches based on (i) the occurrence frequency for the references, (ii) the spatial overlap between bounding boxes associated to the references, (iii) hierarchical containment between the references, using a taxonomy of administrative divisions, and (iv) graph-propagation methods using again a taxonomy of administrative divisions. Experiments with a collection of Web pages from the Open Directory Project showed that hierarchical containment achieved very good results. In this paper, we are also studying approaches for georeferencing the entire contents of textual documents, but using only the raw text as input, instead of relying on place references recognized in the texts.

Adams et al. studied the relationship between topics in textual documents and their geospatial distribution [1]. While most work in geographic information extraction and retrieval relies on specific keywords such as place names, Adams et al.

proposed an approach that uses only non-geographic expressions, seeing if ordinary textual terms are also good predictiors of geographic locations. The proposed technique uses Latent Dirichlet Allocation (LDA) to discover latent topics present in the collection of documents. LDA is essentially an unsupervised generative method for modeling documents as probabilistic mixtures of topics, which are in turn modeled as a probability distribution over a word vocabulary. After fitting the LDA model, the authors use Kernel Density Estimation (KDE) to interpolate a density surface, corresponding to a geospatial region, over each LDA topic. Noticing that each document can be seen as a mixture of topics, the authors use map algebra operations to combine the density surfaces from each topic, finally assigning documents to the location having the highest density.

Eisenstein et al. investigated the dialectal differences and the variations in regional interests over Twitter users, using a collection of georeferenced *tweets* and probabilistic models [9]. These authors tried to georeference USA-based Twitter users with basis on their *tweet* content, concatenating all the *tweets* for each single user and using Gaussian distributions to model the locations of the Twitter users. The approaches proposed here are instead based on a discrete representation for the Earth's surface, and on relatively simple probabilistic models built over this discrete representation.

Wing and Baldridge, in a very similar study to the one that is reported in this paper, compared different approaches for automatically georeferencing documents, also based on statistical models derived from a large corpus of already georeferenced documents, such as Wikipedia [25]. The Kullback-Leibler divergence between the language model for a test document, and language models for each cell in a discrete gridded representations for the Earth's surface, was used to predict the most likely grid cell for a document. A similar approach was proposed for the temporal resolution of documents, capable of determining the date of publication of a story, based on its text [14]. Again, the authors built histograms encoding the probability of different temporal periods for a document, later using the Kullback-Leibler divergence to make the predictions. The work reported in this paper is very similar to that of Wing and Baldridge, but I propose to use (i) a different scheme for partitioning the set of documents into bins of equal area, according to their geospatial locations, (ii) a different language modeling approach for classifying documents according to the most similar bins, (iii) a hierarchical decomposition approach for improving the computational performance of the classification method, and (iv) different post-processing techniques for assigning the geospatial coordinates with basis on the obtained classification scores.

Opinion Mining is different from most other traditional Information Extraction and document classification tasks, as it presents several novel challenges that require creative solutions. In a study with movie reviews, Pang et al. tested the difficulty that humans have in discerning good words to use for the classification of sentiments [20]. Suppose we want to classify a document by counting words that are associated with positive or negative opinions. Choosing one such list of words is not a trivial task, as words may have different interpretations according to context.

Pang et al. also tested whether the classification of opinions could be addressed as a special case of the classical topical classification of texts [20]. To this end, three classification models were tested, namely Naive Bayes, Maxium Entropy and Support Vector Machines to classify the polarity of opinions (i.e., positive or negative opinions). The authors achieved state of the art results, obtaining an accuracy of 0.82 for the Naive Bayes classifier, an accuracy of 0.81 for the Maxium Entropy model, and an accuracy of 0.83 with Support Vector Machines.

Instead of just classifying a document as having either a positive or negative sentiment, one can think that it would be interesting to measure the positivity of a document. Classifying opinions towards a multi-point scale presents even more challenges, since each category has now a more ambiguous vocabulary. Pang et al. addressed the rating-inference problem, attempting to classify documents using a three-point scale or a four-point scale [19]. For this purpose, they used a meta-algorithm known as metric-labeling. The basic idea of this method is to use an initial label preference function that gives an estimation of how to label the items. Using the initial labels, it is possible to compute the label of a new item according to the similarity towards other labeled items, in order to ensure that similar data receives similar labels. As the initial label function the authors used a One vs All SVMs or a SVM regression model, also testing the usage of these models alone. All the tests indicated that metric labeling presents better results than just using the baseline models, obtaining significantly better results with the three-point scale.

3. THE PROPOSED METHODS

In this section I describe the proposed methods for georeferencing textual documents, for extracting opinions, to create thematic maps representing the distribution of opinions. My system starts by assigning a pair of coordinates to each document of a given collection of textual documents, later using the studied techniques for opinion mining to extract the overall opinion expressed in each document, resorting to a two-point scale scheme (i.e., positive and negative opinion), or a five-point scale scheme. Finally, using the information extracted with the two previous methods, the system generates a density surface over a map, using a technique know as Kernel Density Estimation.

3.1 LingPipe Language Model Classifiers

The LingPipe language model classifiers perform joint probability based classification of textual documents into categories, based on either character-based or token-based language models (i.e., in the experiments that were performed, both these two different classification approaches were tested). The general idea is to build a language model $P(text|cat)$ for each category cat, afterwards building a multinomial distribution $P(cat)$ over the categories, and finally computing joint log probabilities for the classes according to Bayes's rule, yielding:

$$\log_2 P(cat, text) \propto \log_2 P(text|cat) + \log_2 P(cat) \qquad (1)$$

In the formula, $P(text|cat)$ is the probability of seeing a

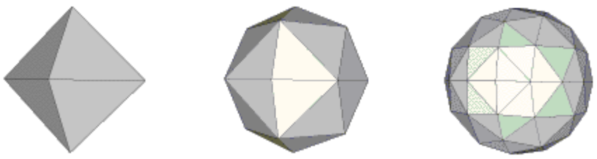


Figure 1: Decompositions of the Earth's surface for triangular meshes with resolutions of 0, 1 and 2.

given *text* in the language model for a category *cat*, and $P(cat)$ is the marginal probability assigned by the multinomial distribution over the categories.

The book by Carpenter and Baldwin has the complete details on the language models used for estimating $P(text|cat)$, and on the multinomial distribution $P(cat)$ over the categories [6]. The multinomial distribution $P(cat)$ is basically estimated using a maximum a posteriori probability (MAP) estimate with additive (i.e., Dirichlet) priors.

In terms of the character-based or token-based language models, they are essentially generative language models based on the chain rule, which smooth estimates through linear interpolation with the next lower-order context models, and where there is a probability of 1.0 to the sum of the probability of all sequences of a specified length.

3.2 Document Geocoding

The proposed approach for representing the Earth's surface is based on discretizing space into a set of bins, allowing us to predict locations with standard approaches for discrete outcomes. However, unlike previous authors such as Serdyukov et al. [21] or Wing and Baldridge [25], which used a grid of squared cells of equal degree, the method proposed in this section is based on a Hierarchical Triangular Mesh[1] [8, 23]. This strategy results in a triangular grid that roughly preserves an equal area for each bin, instead of variable-size regions that shrink latitudinally, becoming progressively smaller and elongated as they get closer towards the poles. Notice that this binned representation ignores all higher level semantic regions, such as states, countries or continents. Nonetheless, this is appropriate for the purpose of this work, since documents can be related to geographical regions that do not fit into an administrative division of the Earth's surface.

The Hierarchical Triangular Mesh (HTM) offers a multi-level recursive decomposition for a spherical approximation to the Earth's surface – see Figures 1 and 2, both adapted from original images at the HTM website. The decomposition starts at level zero with an octahedron and, as one projects the edges of the octahedron onto the sphere, it creates 8 spherical triangles, 4 on the Northern and 4 on the Southern hemispheres. Four of these triangles share a vertex at the pole, and the sides opposite to the pole form the Equator. Each of the 8 spherical triangles can be split into four smaller triangles by introducing new vertices at the midpoints of each side, and adding a great circle arc segment to connect the new vertices. This sub-division process can be repeated recursively, until we reach the desired level of resolution. The triangles in this mesh are the bins used in the representation of the Earth, and every triangle, at any resolution, is represented by a single numeric ID. For each location given by a pair of coordinates on the surface of the sphere, there is an ID representing the triangle, at a particular resolution, that contains the corresponding point.

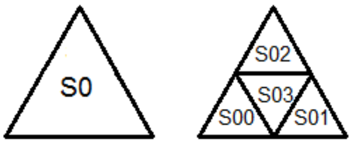


Figure 2: Recursive decomposition of the circular triangles used in the triangular mesh.

Notice that the proposed representation scheme contains a parameter k that controls the resolution, i.e. the area of the bins. Having course grained bins can lead to very rough estimates, but classification accuracy, with a thin-grained resolution, can also decrease substantially, due to insufficient data to build the language models associated to each bin. In the experiments that were conducted, this parameter ranged from 4 to 10, with 0 corresponding to the first-level division. Table 1 presents the maximum number of bins that would be generated at each of the considered levels of resolution. The number of bins n for a resolution k is given by $n = 8 * 4^k$. Table 1 also shows the area, in squared Kilometers, corresponding to each bin.

With the HTM-based discrete representation for the Earth's surface, the LingPipe[2] package was used to build character-based or token-based language models, afterwards using these models for associating, to each bin, the probability of being the best class for a given novel document. In the conducted experiments, the character-based language models were based on sequences of 8 characters. As for the token-based language models, sequences of tokens with a 2-gram model were captured, and the models considered white-spaces and unknown tokens separately.

After having probabilities assigned to each of the bins from the representation of the Earth, latitude and longitude coordinates are computed with basis on the centroid coordinates for the most probable bin(s). In this particular stage, experiments with four different post-processing techniques were conducted. These are as follows:

1. Assign geospatial coordinates with basis on the centroid of the most probable bin.

2. Assign geospatial coordinates according to a weighted average of the centroid coordinates for all the possible

[1] http://www.skyserver.org/htm/Old_default.aspx

[2] http://alias-i.com/lingpipe

Resolution	4	6	8	10
Total number of categories	2,048	32768	524,288	8,388,608
Average area of each bin (km^2)	28,774.215	17,157.570	1,041.710	67.031

Table 1: Number of bins in the triangular mesh and their corresponding geospatial area.

bins, where the weights come from the probabilities assigned to each of the bins by the classifier.

3. Assign geospatial coordinates according to a weighted average of the centroid coordinates for the most probable bin and for its adjacent neighbors in the hierarchical triangular mesh, again using weights corresponding to the probabilities assigned to each of the bins.

4. Assign geospatial coordinates according to a weighted average of the coordinates associated to the kn most similar documents in the training collection that are contained in the most probable bin.

Methods two and three, from the previous enumeration, require for the classifier to return calibrated probabilities, whereas the language modeling approach used in LingPipe is know to produce skewed and too extreme probability estimates. In the literature, there are many methods for calibrating classification probabilities by post-processing, but most of these methods are only defined for binary classification problems [4, 11]. In this particular multi-class problem, the values returned by the language model classifier were processed by using a sigmoid function of the form $(\sigma * score)/(\sigma - score + 1)$, where the σ parameter controlling the gradient of the curve was adjusted empirically.

In what regards method four, the similarity between documents was computed according to the cosine similarity between document feature vectors, built with basis on term bigrams. In the experiments that were conducted, the knn parameter ranged between 5 and 20.

Although language model classifiers can be used directly to assign documents to the most probable bins, they can be very inefficient in practice when considering a thin-grained resolution, due to the very large number of classes – see Table 1 – and due to the need for estimating, for each document, its probability of having been generated by the language model corresponding to each class. In this paper, I propose to use a hierarchical classification approach, where instead of a single classifier considering all bins from a detailed triangular mesh encoding the Earth's surface, a hierarchy of classifiers with two levels is used. The first level corresponds to a single classification model using bins from a coarse-grained division of the Earth, and the second level corresponds to different classifiers, one for each class from the first level, encoding different parts of the Earth with a thinner granularity. With this hierarchical scheme, classification can be made much more efficiently, as documents need to be evaluated with less language models.

I also propose a technique that takes advantage of the properties from the hierarchical triangular mesh in order to reduce the number of classes in each of the models from the second level of the classification hierarchy. If a given bin from the decomposition of the Earth does not contain any training documents assigned to it, and if only one of its neighbouring bins in the mesh contains documents, then a single class from the hierarchical triangular mesh of the immediately smaller resolution is used, in order to represent this region in the classification model.

In a related previous work, Wing and Baldridge [25] reported on very accurate results (i.e., a median prediction error of just 11.8 Kilometers, and a mean of 221 Kilometers) with a similar, but non-hierarchical classification approach, based on the Kullback-Leibler divergence between language models. However, these authors also claim that a full run with all their experiments (i.e., six different strategies) required about 4 months of computing time and about 10-16 GB of RAM when run on a 64-bit Intel Xeon E5540 CPU. The proposed hierarchical classification approach can substantially reduce the required computational effort.

3.3 Opinion Mining

This section presents the proposed approach for analyzing the opinions expressed in textual documents. Two approaches based on language model classifiers were considered, namely one using a two-point scale (i.e., polarity of opinion) and another using a multi-point scale, in this latter case considering a five-point scale. For both approaches, the LingPipe[3] package was used to build character-based and token-based language models, as described in Section 3.1. The approaches used were based on the sentiment analysis tutorial described in the LingPipe Website[4], with some minor modifications. In the conducted experiments, the character-based language model classifier was based on sequences of 8 characters, while the token-based language model classifier used sequences of tokens with a 2-gram model, and with the white-spaces and unknown tokens modeled separately.

One important difference between the two-point scale opinion analysis and the multi-point scale case is that, in the second case, we can leverage the fact that nearby classes should be more similar than far-away classes, and similar data should receive similar labels. In order to model this idea, a meta-algorithm know as metric-labeling, originally proposed for ordinal regression problems (i.e., problems where we have a natural ordering between the possible outcomes), was used to smooth the results of the language model classifier [19]. Using the assumption that similar data should receive similar labels, we can correct the results given by the classifiers by considering the real classes from the knn most similar documents of the training dataset, using the cosine similarity between documents. The language model classifier was used as the initial label preference function

[3] `http://alias-i.com/lingpipe`

[4] `http://alias-i.com/lingpipe/demos/tutorial/sentiment/read-me.html`

$LM(x,l)$, which gives an estimation on how to label the documents (i.e., give the classification scores for a document x and a label l). Also a metric distance between labels d was used, and a set of documents $knn(x)$ with the knn nearest examples in the training data for document x, according to the cosine similarity $cosine\_sim(x,y)$ between documents x and y. The problem of labeling items can be solved by choosing the labels that minimize the formula bellow, where α represents a trade-of and/or scaling parameter, adjusted empirically:

$$\sum_{x \epsilon test} \left[-LM(x,l) + \alpha \sum_{y \epsilon knn(x)} d(l_x, l_y) cosine\_sim(x,y) \right] \tag{2}$$

In order to speed up both the training and the classification, a hierarchical classification approach was also developed. This approach is similar to the one presented for the problem of georeferencing textual document, and is also composed of two levels of classification. In the first level we have a single classifier that distinguishes between positive and negative documents. In the second level we have two classifiers, namely one that distinguishes levels of negativity, and a second classifier that distinguishes levels of positivity. Category 3 was considered in both the positive and negative classifier, since this category is usually associated with a neutral opinion.

3.4 KDE and Thematic Maps

Having geocoded documents assigned also to an opinion class, we can then create thematic maps showing the geographic distribution of opinions. I specifically used a simple interpolation method known as Kernel Density Estimation (KDE) for producing density maps. The general approach that is used for the construction of the maps is composed of the following steps.

1. We start with point clouds corresponding to the latitude and longitude coordinates from the set of documents associated to a particular opinion class (i.e., we start from a point cloud for each opinion class);

2. The KDE method is used to produce, from the point clouds associated to each opinion class, density surfaces showing the incidence of each particular opinion class, at the different regions. The density estimation procedure uses a Gaussian kernel for interpolating from the available data that is located within a given bandwidth. The estimates are essentially produced according to the formula shown bellow, where K is a Kernel function which integrates to one, d_i is the geographical distance between data point i and location (x,y). and the bandwidth parameter is selected according to the method originally proposed by Sheather and Jones [22]:

$$f(x,y) = \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{d_i}{h}\right) \tag{3}$$

3. We combine the density surfaces from each opinion class through map algebra operations, essentially adding the surfaces in order to produce the final thematic maps that show the intensity and the polarity of the opinions associated to particular regions.

When using opinions in a two-point scale, we combine the two different surfaces, obtained through the KDE method, in order to form a continuous surface with values from -1.0 to 1.0, according to the formula $KDE(positive) - KDE(negative)$. Thus, regions where there are more negative or positive opinions will be associated, respectively, to a negative or a positive value. Regions where there are few opinions, or where we have an equal proportion of positive and negative opinions, will be assigned to a value that is close to zero. A similar procedure is used for the case when we have opinions in a five-point scale, but we instead normalize the KDE values for each opinion class into the interval $[0; 0.4]$, and we then map each opinion class to an interval within the limit values of -1.0 and 1.0. The opinions corresponding to a value of 3 in the three-point scale are assigned to the interval from -0.2 to 0.2, and since documents expressing a neutral opinion are often closer to expressing a negative evaluation, we perform the mapping in way such that regions where there's a higher density of neutral opinions, end up being associated to a value that is closer to -0.2.

4. EXPERIMENTAL EVALUATION

This section describes the experimental methodology used for comparing the proposed methods, afterwards discussing the obtained results.

4.1 Datasets and Methodology

The experiments concerning with document geocoding used a sample of the articles from the English Wikipedia dump from 2012. Included in this dump are a total of 4,080,270 articles, of which 430,032 were associated to latitude and longitude coordinates. Previous studies have already shown that Wikipedia articles are a well-suited source of textual contents for the purpose of georeferencing documents [18, 25].

The Wikipedia dump was processed in order to extract the raw text from the articles and for extracting the geospatial coordinates, using the software dmir-wikipedia-parser[5], which is based on manually-defined patterns to capture some of the multiple templates and multiple formats for expressing latitude and longitude in Wikipedia. Considering a random order for the georeferenced articles, about 91% of the georeferenced articles that could be processed were used for model training (i.e., a total of 390,032 articles) and the other 9% were used for model validation (i.e., a total of 40,000 articles). Table 2 presents a statistical characterization for the considered dataset, while Figure 3 illustrates the geospatial distribution of the locations associated to the Wikipedia documents. Notice that some geographic regions (e.g, North America or Europe) are considerable more dense in terms of document associations than others (e.g, Africa). Moreover, oceans and other large masses of water are scarce in associations to Wikipedia documents. This implies that the

[5] `http://code.google.com/p/dmir-wiki-parser/`

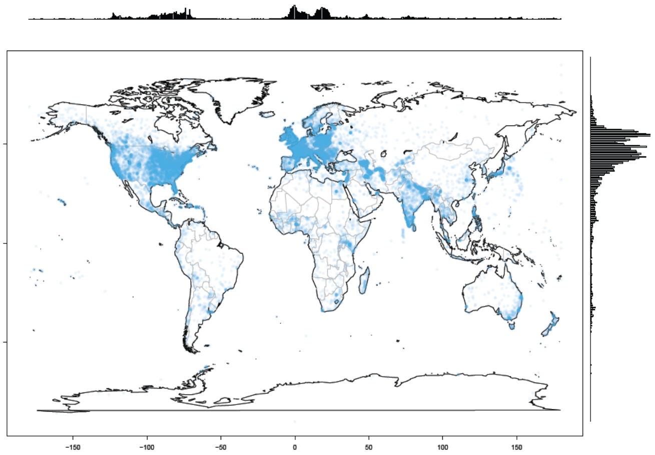


Figure 3: Geographic distribution for the Wikipedia documents.

Statistic	Train	Test
Number of Documents	390,032	40,000
Number of Words	160,508,876	16,696,639
Avg. Words per Doc.	411	417
St. Dev. of Words per Doc.	74.202	231.705

Table 2: Characterization of the Wikipedia dataset.

number of classes that has to be considered by the model is much smaller than the theoretical number of classes given in Table 1. In the Wikipedia dataset that was used, there are a total number of 1,123 bins containing associations to documents at resolution level 4, and a total of, 8,320, 42,331 and 144,693 bins, respectively at resolutions 6, 8 and 10.

In order to get some insights into the hypothesis which states that general textual terms can be indicative of geographic locations, documents were first filtered according to their containment of particular terms, and then the corresponding coordinates were plotted on a map. Figure 4 shows the geographic incidence of four different textual terms, namely *mountain* using green dots, *wine* using red dots, *snow* using blue dots, and *desert* using yellow dots. The figure shows that these particular terms are more associated to the regions that would be expected (i.e., terms like wine are more associated to regions such as France, and terms like desert are more associated to North Africa).

In the experiments reported for opinion mining, a sample of reviews retrieved from the website `yelp.com` was used (i.e., the data from the Yelp academic dataset[6]). Included in this collection are a total of 152,295 reviews. This collection includes reviews from clients of many business types, including restaurants, hotels and different kind of local shops. The reviews in the Yelp dataset include ratings given by the users, using a five-point scale. In a previous work, Cabral and Hortaçu observed that users do not see neutral reviews, as between positive and negative, but rather they tend to see them as negative [5]. In order to use this dataset in the polarity classification case, I considered that reviews rated with a 3 (*i.e.*, neutral) would be considered as negative reviews, since most of these reviews had a mix between negative and positive phrases.

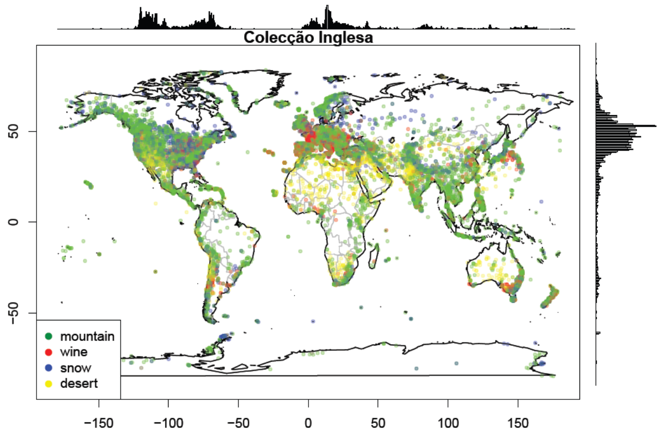


Figure 4: Geographic incidence of particular terms.

Statistic	Train	Test
Number of Reviews	112,295	40,000
Number of Words	17,038,778	6,041,821
Avg. Words per Review	151	151
St. Dev. Words per Review	125.094	124.762

Table 3: Characterization of the Yelp academic dataset.

Analysis type	Category	Collection Train	Collection Test
Polarity	positive	69,763	24,800
	negative	42,532	15,200
5 Star	1	8,599	3,022
	2	11,872	4,219
	3	22,061	7,959
	4	38,615	13,664
	5	31,148	11,136

Table 4: Number of reviews per category in the Yelp dataset.

Considering a random order for the articles, about 74% of the reviews that could be processed were used for model training (i.e., a total of 112,295 reviews) and the other 26% were used for model validation (i.e., a total of 40,000 articles). Table 3 presents a statistical characterization for the considered dataset.

Table 4 presents the number of reviews per category for the training subset and the test subset, where we can see that the number of negative cases is much smaller than the number of positive cases in both subsets.

4.2 Document Geocoding

Using the Wikipedia dataset, experiments were conducted with classification models relying on bin sizes of varying

[6] `http://www.yelp.com/academic_dataset`

					Geospatial Distance					
	Resolution		Classifier Accuracy		Centroid		All Bins		Neighbour Bins	
Method	k1	k2	1st Level	2nd Level	Average	Median	Average	Median	Average	Median
Character Models	0	4	**0.9609**	**0.8354**	405.214	240.017	438.762	228.829	386.271	219.379
	1	6	0.9411	0.6669	**254.846**	62.846	282.874	71.119	257.741	65.551
	2	8	0.9283	0.3989	268.480	**25.757**	283.761	48.039	269.493	28.569
	3	10	0.8912	0.1615	281.669	30.405	287.909	51.595	281.755	30.464
Token Models	0	4	0.9444	0.5209	693.764	240.0174	1544.543	228.8294	691.899	219.3789
	1	6	0.9103	0.4244	433.224	92.770	729.546	303.055	444.766	120.561
	2	8	0.8909	0.2747	455.025	44.621	572.406	191.522	457.858	50.349
	3	10	0.8297	0.1305	547.760	42.162	591.111	123.457	547.996	41.982

Table 5: The obtained results for document geocoding with different types of classifiers.

granularity. Table 5 presents the obtained results for some of the different methods that were under study (i.e., all types of classifiers and the three first post-processing strategies, which did not use the similarity of neighbouring documents), together with the error values for each bin size. The prediction errors shown in Table 5 correspond to the distance in Kilometers, computed through Vincenty's formulae[7] [24], from the predicted locations to the locations given at the gold standard. The accuracy values correspond to the relative number of times that it was possible to assign documents to the correct bin (i.e., the bin where the document's true geospatial coordinates of latitude and longitude are contained). The $k1$ and $k2$ values correspond to the resolution for the Earth's representation used at each level of the hierarchical classifier.

The results from Table 5 show that the method corresponding to the usage of character-based language models obtained the best results, with the best configuration having a prediction accuracy of approximately 0.40 in the task of finding the correct bin, while assigning documents to the correct geospatial coordinates had an error of 268 Kilometers on average. The documents that were assigned to the correct bin had an average distance towards the correct coordinates of 12 Kilometers. The results from Table 5 also show that both the second and third post-processing technique, in which the coordinates are adjusted with basis on a weighted average with all the bins or the neighboring bins, does not improve the results over the baseline method in which the centroid coordinates for the most probable bin is used. I believe that this is due to the fact that the language model classifiers that were used do not provide accurate and well-calibrated probability estimates. Even when using the described post-processing score calibration technique based on a sigmoid function, the method still produces too extreme estimates.

Table 6 presents the obtained results for language-model classifiers based on character n-grams (i.e., the best performing method from the previous experiment), when using the fourth post-processing method, in which the coordinates of latitude and longitude are assigned through a weighted average between the centroid coordinate of the most probable bin and the coordinates of the knn most similar training documents contained within that same bin. The first column of Table 6 indicates the number of considered nearest neighbors, while the $k1$ and $k2$ values correspond to the resolution for the Earth's representation used at each level of the hierarchical classifier based on character n-grams. The results show that using the 5 most similar documents, provided the best results, with an average distance error of just 265 kilometers and a median distance error of 22 kilometers.

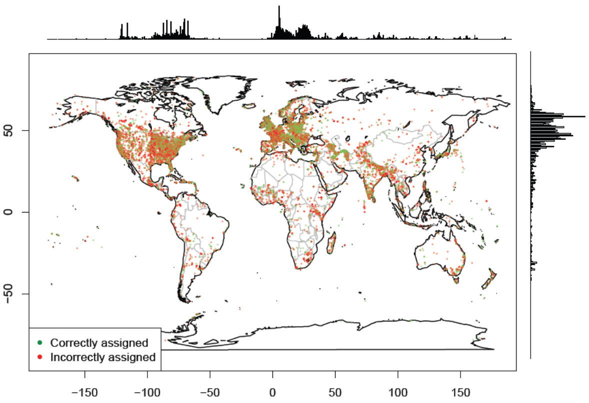


Figure 5: Estimated positions for the Wikipedia documents.

	Resolution		Geospatial Distance	
Nearest neighbours	k1	k2	Average	Median
5	0	4	289.750	90.515
	1	6	235.702	34.005
	2	8	265.734	**22.315**
	3	10	281.442	30.209
10	0	4	274.515	68.252
	1	6	233.982	32.092
	2	8	265.655	22.371
	3	10	281.460	30.208
15	0	4	271.045	64.008
	1	6	**233.928**	32.243
	2	8	265.744	22.480
	3	10	281.464	30.172
20	0	4	270.337	63.373
	1	6	234.197	32.687
	2	8	265.869	22.640
	3	10	281.466	30.170

Table 6: The obtained results for document geocoding with post-processing based on the k most similar documents.

A visualization of the results obtained with the best performing method can be seen in Figure 5, where the map represents the geospatial distribution for the predicted locations. The figure shows that errors are evenly distributed, and also that Europe and North America remain the regions

[7] http://en.wikipedia.org/wiki/Vincenty's_formulae

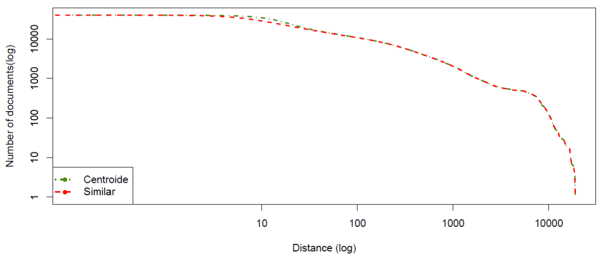


Figure 6: Distribution for the obtained errors, in terms of geospatial distance towards the correct coordinates.

		Multi-Scale		
Model	Polarity	Normal	Metric Labeling	Hierarchical
Characters	**0.8030**	0.4940	0.4947	**0.5008**
Tokens	0.7577	0.4458	0.4461	0.4488

Table 7: The obtained results in terms of accuracy for multi-scale sentiment analysis.

of highest density.

Figure 6 illustrates the distribution for the errors produced by the character-based classifiers, in terms of the distance between the estimated coordinates and the true geospatial coordinates, when using the baseline method that assigns the centroid coordinates for the most probable bin, and when using the post-processing method that uses the coordinates from the 5 most similar documents. This figure plots the number of documents whose error (i.e., the distance towards the correct result) is greater or equal than a given value, using doubly logarithmic axes. Figure 6 shows that the proposed post-processing method based on the analysis of the most similar documents assigns coordinates to the majority of examples with a small error in terms of distance, with 31634 documents having an error smaller than 100 Kilometers. Worse results are shown for the baseline method, with about 29903 documents having an error smaller than 100 Kilometers.

4.3 Opinion Mining

Table 7 presents the results for the two-point and five-point opinion mining experiments, for both the token based and the character based language models. We can see that character based language models achieved the best accuracy in all experiments. In the two-point scale case, an accuracy of 0.80 was achieved. The five-point scale case shows worse results, where the best method (i.e. using an hierarchical classification approach) only achieved an accuracy of 0.50. We can also see that the metric labeling approach only increased the accuracy by 0.07 points. In fact, of the 40,000 test cases, the metric labeling method only affected 323 cases. I believe this is due to the fact that the LingPipe language model classifiers do not provide well-calibrated probability estimates, and instead they only focus on finding the most probable category.

Table 8 shows the precision achieved for each opinion class by the three methods tested for the five-point scale case. We can see that, for the baseline and the metric labeling method, opinion classes 2 and 3 present worse results, as to be expected since these two categories have perhaps the most ambiguous language. Interestingly the hierarchical method increases the precision for category 3 by approximately 0.10 points. This may be due to the fact that both second level classifiers have been trained with examples of category 3. Even if the initial polarity classifier considers that a review that belongs to category 3 is positive, the second classifier can still classify it has belonging to category 3.

4.4 Thematic Maps

In order to evaluate the generation of thematic maps portraying the geographic distribution of opinions, I again relied on the Yelp academic dataset. Different experiments were made with the Yelp dataset, separately accessing the issues related to the portrayal of opinions, from the issues associated to geospatial positional accuracy. Different maps were thus generated, namely:

1. Maps using the ground-truth data available from the Yelp dataset, namely the review ratings in a scale from 1 to 5, and the geospatial coordinates for the local business being reviewed. These maps correspond to the types of representations that one would produce when using perfectly accurate text mining components, capable of exactly discovering the geospatial coordinates of documents and their expressed opinions.

2. Maps using the ground truth geospatial coordinates available from the Yelp dataset, and using the opinion polarity information derived from the proposed opinion mining method, either when considering a two-point opinion scale or a five-point opinion scale.

3. Maps using the ground truth opinion polarity information from the Yelp dataset, and using the geospatial coordinates of latitude and longitude assigned by the proposed document geocoding method, in its most accurate configuration.

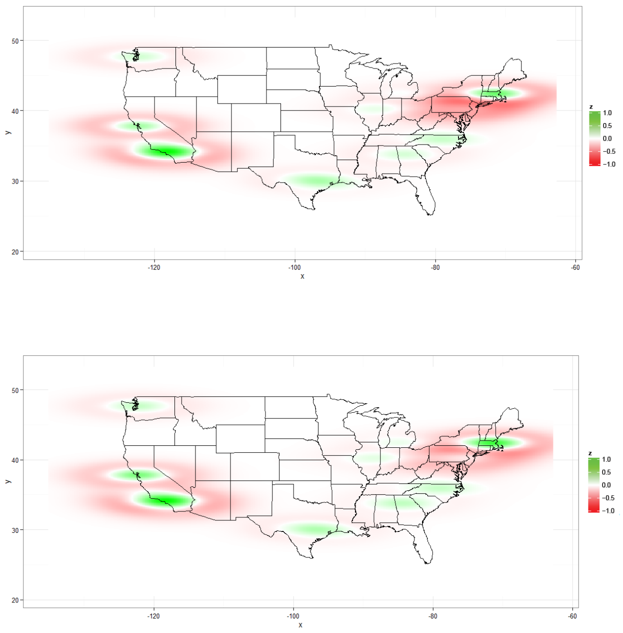
4. Maps using only information derived automatically, when processing the Yelp collection with the proposed text mining methods, in their most accurate configurations.

The maps produced from methods 2, 3 and 4, from the previous enumeration, were compared against the map produced by method one.

Category	Normal	Metric Labeling	Hierarchical
1	0.4173	0.4146	0.4302
2	0.2076	0.2079	0.2495
3	0.3118	0.3126	0.4194
4	0.6570	0.6603	0.6101
5	0.5530	0.5518	0.5392

Table 8: The obtained precision for each category for character based language models.

Figure 7: Thematic maps portraying the real geographic distribution of opinions



1st Accuracy	2nd Accuracy	Avg Distance	Median
0.3163	0.0404	4154.017	3657.000

Table 9: The obtained results for the estimated positions of the Yelp dataset.

Maps produced by method 1 and 2 from the previous enumeration are shown in figure 7, where the upper map represents the geographical distribution of opinions, using both real coordinates and opinions for each document. The bottom map represents the geographical distribution of opinions when estimating opinions, using the hierarchical classification approach for the five-point scale classification. We can see that the map with estimated opinions, closely resembles the real map, which proves that the proposed method for opinion mining suits the problem.

Figure 8: Thematic maps portraying the estimated geographic distribution of opinions

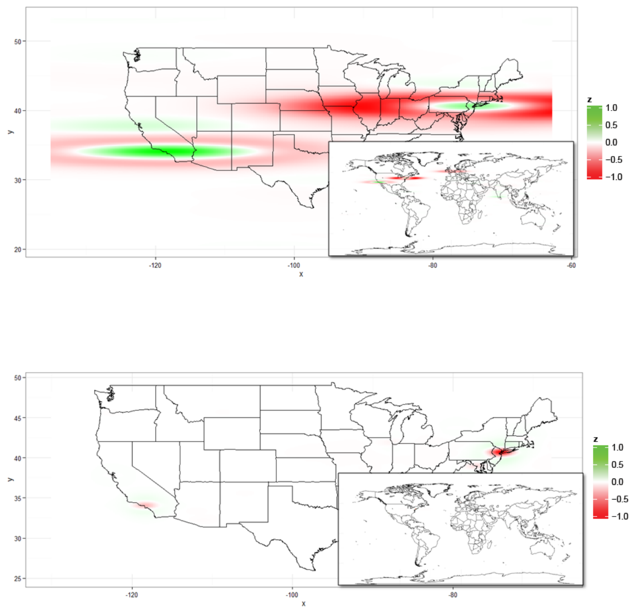


Figure 8 presents the estimated geographical distribution of opinions, where the upper map has the real opinion classifications, whereas the bottom map has both estimated coordinates and estimated opinion classes. We can see in the bottom figure that there are little density areas, clearly the estimated opinions nullify each other, as explained in Section 3.4. We can also notice in the mini world map that the georeference method classified some documents as belonging to the United Kingdom, probably because of some relation between equal city names in both the USA and the UK (e.g., like Cambridge or Oxford). In order to measure the quality of the estimated documents distribution of the Yelp collection, I analyzed the average distance error between the estimated position and the actual position of each review. Table 9 presents the results obtained when evaluating the performance of the document georeferencing module, trained with Wikipedia data, over the Yelp dataset. We can see that the results are not very accurate, with an average distance error of 4154 Kilometers and an accuracy of just 0.04. This is to be expected, since the dataset was georeferenced using a model trained with the Wikipedia collection. The Wikipedia collection contains very descriptive documents with an average of 411 words per document, while the Yelp collection is composed by short reviews with an average of 151 words per document. Furthermore these reviews contain many opinion based words, while the Wikipedia documents nearly has none.

5. CONCLUSIONS AND FUTURE WORK

This paper described automated methods to assign documents to geospatial coordinates of latitude and longitude that best summarizes their contents. The paper has also discussed one possible application for the georeferenced information that can be extracted this way, related to finding and representing the geographical distribution of opinions. I studied different methods to analyze the overall opinion expressed in textual documents. Using this information, I studied techniques to build thematic maps portraying the incidence of certain classes of opinions, in different geographic areas.

In what concerns georeferencing textual documents, I studied techniques that rely on language model classifiers for assigning documents to corresponding geospatial regions, and then post-processing the classification results in order to assign the most likely geospatial coordinates of latitude and longitude. We have seen that the automatic identification of the geospatial location of a document, based only on its text, can be performed with high accuracy using simple supervised methods, and using a discrete binned representation of the Earth's surface based on a hierarchical triangular mesh. The proposed method is simple to implement, and both training and testing can be easily parallelized. The most effective georeferencing strategy uses language models based on character 8-grams, and assigns coordinates of latitude and longitude through the centroid coordinates of the k most similar documents contained within the most probable region for each document.

Using an approach based on language model classifiers, we have seen that the automatic analysis of opinions can be performed in two types of scales, namely a two-point scale and a five-point scale. Both classification cases achieved better results when using language models based on character 8-grams, where the two-point scale case achieved an accuracy of 0.80. We have also seen two types of techniques to analyze the opinions in a five-point scale. The best performing technique was based on a hierarchical classification scheme, which increases computational performance, and achieved an accuracy 0.5.

Using the information extracted from the previous methods, I also studied techniques to represent the geographical distribution of opinions using a technique know as Kernel Density Estimation in order to build density maps. We have seen that the representations created using density maps correspond to accurate representations for the geographic distribution of opinions. Although the technique used for adding density surfaces of each opinion class, proved to be effective, it can fail to convey information if the density surfaces from the opinion classes nullify each other. In future work, it would be interesting to study other techniques to represent several different density surfaces. I would also like to explore other types of thematic maps, such as choropleth maps, which I also think would fairly represent the sort of geographical information presented in my work.

It should be noticed that the proposed classification approach, based on language models, does not provide accurate and well-calibrated probability estimates for the different classes involved in the problem, instead focusing only on the simpler task of predicting which class is the most likely. For future work, instead of just experimenting with an heuristic score calibration method based on post-processing, it would be interesting to experiment with other classification approaches for assigning documents to the most likely bin(s), for instance through maximum entropy models. It would also be interesting to experiment with maximum entropy models using either expectation constraints specifying affinities between words and labels [7], or with posterior regularization [10], leveraging the fact that the presence of the words corresponding to place names is a strong indicator for the document belonging to a particular class.

Also regarding place name references in text, it is important to notice that although identifying a single location for an entire document can provide a convenient way for connecting texts with locations, useful for many different applications, many other applications could benefit from the complete resolution of place references in textual documents [15]. The probability distributions over bins, provided by the proposed method for document georeferencing, can for instance be used to define a document-level prior for the resolution of individual place names.

Finally, in terms of future work, it would also be interesting to experiment with document expansion techniques, particularly for the case of small documents, that could build pseudo-documents by constructing a neighborhood around the original ones (e.g., using linkage information between hypermedia documents, although we should note that we are primarily interested in georeferencing documents using only the text because there are a great many situations in which linkage information is unavailable, a particular example being historical documents in digital libraries). Such expansion techniques could be particularly useful for the case of georeferencing short texts posted in social media services [13].

6. REFERENCES

[1] B. Adams and K. Janowicz. On the geo-indicativeness of non-georeferenced text. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*, 2012.

[2] Einat Amitay, Nadav Har'El, Ron Sivan, and Aya Soffer. Web-a-where: geotagging web content. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2004.

[3] I. Anastácio, B Martins, and P. Calado. A comparison of different approaches for assigning geographic scopes to documents. In *Proceedings of the 1st Simpósio de Informática*, 2010.

[4] Antonio Bella, Cèsar Ferri, José Hernández-Orallo, and Marïa Ramírez-Quintana. Similarity-binning averaging: A generalisation of binning calibration. In *Intelligent Data Engineering and Automated Learning*. 2009.

[5] Luís Cabral and Ali Hortaçsu. The dynamics of seller reputation: Theory and evidence from eBay. Working paper, downloaded version revised in March, 2006.

[6] Bob Carpenter and Breck Baldwin. *Natural language processing with LingPipe 4*. LingPipe Publishing, 2011.

[7] Gregory Druck, Gideon Mann, and Andrew McCallum. Learning from labeled features using generalized expectation criteria. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in Information Retrieval*, 2008.

[8] G. Dutton. Encoding and handling geospatial data with hierarchical triangular meshes. In *Advances in GIS Research II*. Taylor and Francis, 1996.

[9] Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, 2010.

[10] Kuzman Ganchev, João Graça, Jennifer Gillenwater, and Ben Taskar. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 2010.

[11] Martin Gebel and Claus Weihs. Calibrating classifier scores into probabilities. In *Advances in Data Analysis*. 2007.

[12] B. Johnstone. Language and place. In *Cambridge*

Handbook of Sociolinguistics. Cambridge University Press, 2010.

[13] Sheila Kinsella, Vanessa Murdock, and Neil O'Hare. I'm eating a sandwich in glasgow: Modeling locations with tweets. In *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*, 2011.

[14] Abhimanu Kumar, Matthew Lease, and Jason Baldridge. Supervised language modeling for temporal resolution of texts. In *Proceeding of the 20th ACM conference on Information and Knowledge Management*, 2011.

[15] Jochen Lothar Leidner. *Toponym resolution in text: Annotation, evaluation and applications of spatial grounding of place names.* PhD thesis, University of Edinburgh, 2007.

[16] Michael D. Lieberman and Hanan Samet. Multifaceted toponym recognition for streaming news. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information*, 2011.

[17] Bruno Martins, Ivo Anastácio, and Pável Calado. A machine learning approach for resolving place references in text. In *Proceedings of the 13th AGILE International Conference on Geographic Information Science.* 2010.

[18] Simon Overell. *Geographic Information Retrieval: Classification, disambiguation and modeling.* PhD thesis, Imperial College London, 2009.

[19] Bo Pang and Lillian Lee. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 2005.

[20] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL Conference on Empirical Methods in Natural Language processing*, 2002.

[21] Pavel Serdyukov, Vanessa Murdock, and Roelof van Zwol. Placing flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in Information Retrieval*, 2009.

[22] Simon Sheather and Chris Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1991.

[23] Alexander S. Szalay, Jim Gray, George Fekete, Peter Z. Kunszt, Peter Kukol, and Ani Thakar. Indexing the sphere with the hierarchical triangular mesh. Technical report, Microsoft, 2005.

[24] Thadeus Vincenty. Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. *Survey Review*, 1975.

[25] Benjamin Wing and Jason Baldridge. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011.