Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Kernel Density Estimation for Text-Based Geolocation

Mans Hulden
University of Colorado Boulder
mans.hulden@colorado.edu

Miikka Silfverberg
University of Helsinki
miikka.silfverberg@helsinki.fi

Jerid Francom
Wake Forest University
francojc@wfu.edu

Abstract

Text-based geolocation classifiers often operate with a grid-based view of the world. Predicting document location of origin based on text content on a geodesic grid is computationally attractive since many standard methods for supervised document classification carry over unchanged to geolocation in the form of predicting a most probable grid cell for a document. However, the grid-based approach suffers from sparse data problems if one wants to improve classification accuracy by moving to smaller cell sizes. In this paper we investigate an enhancement of common methods for determining the geographic point of origin of a text document by kernel density estimation. For geolocation of tweets we obtain a improvements upon non-kernel methods on datasets of U.S. and global Twitter content.

Introduction

Text-based geolocation has received much attention recently. This is partly attributable to the availability of large geotagged datasets through Wikipedia and Twitter, which allows the evaluation of possibilities of geolocating a document solely through its text content.

The problem of geolocating a document has often been approached as a supervised classification problem where the task is to associate a document with a discrete cell of the earth's surface, given some previous training data for which the coordinates are known. Many standard classification approaches such as Naive Bayes can be directly adapted to address the problem of placing a document of unknown origin on a geodesic grid. The grid-based approach has also been shown to be quite competitive (Serdyukov, Murdock, and Van Zwol 2009; Wing and Baldridge 2011; Roller et al. 2012) with other more complex models, such as ones based on topic modeling (Eisenstein et al. 2010; Eisenstein, Ahmed, and Xing 2011) and more involved feature engineering (Han, Cook, and Baldwin 2012). However, discretizing the world into fixed-size bins entails a trade-off between accuracy and the amount of data available—the smaller and potentially more accurate the grid becomes, the more acute is the data sparsity problem for each cell. This has been partly addressed by work such as Roller et al. (2012), which uses a grid that adaptively shrinks in size so that each cell accommodates roughly the same number of training documents. In this work we address the data sparsity problem through smoothing out relevant features on a geodesic grid by kernel density estimation, while maintaining the grid model. We show that kernel density estimation offers consistent and robust results compared with completely discretized models. The amount of bookkeeping is also minimal compared with many other methods to handle grid data sparsity.

The paper is laid out as follows: first, we examine the basic geolocation of text documents using Naive Bayes and Kullback-Leibler divergence, and also introduce our adaptation of kernel density estimation to these methods. In the following sections we present the data and the details behind our experiments for geolocation using the U.S. GEOTEXT data set and a global WORLDTWEETS data set. This is followed by the results and discussion.

Geolocation on a geodesic grid

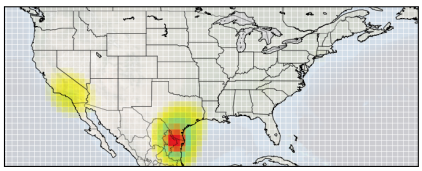
In the current work we discretize the earth's surface into square cells C that come in various sizes depending on the experiment; $10^\circ \times 10^\circ, 5^\circ \times 5^\circ, 2^\circ \times 2^\circ, 1^\circ \times 1^\circ$, and $0.5^\circ \times 0.5^\circ$.

Under this model, we treat geolocating a text document as a classification task where the object is to place an unknown text document in the most appropriate cell $\hat{c} \in C$. In our experiments, the features we use for classification are simply the individual words in a document. Although we only use words, there is in principle no reason why additional features, linguistic and non-linguistic (Eriksson et al. 2010; Youn, Mark, and Richards 2009), could not be integrated in the methods described below.

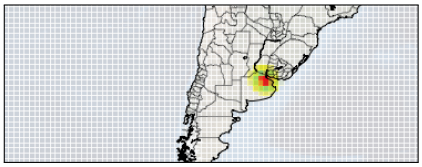
We briefly look at two popular approaches for classifying documents on a grid—Naive Bayes and Kullback-Leibler divergence—and discuss our adaptation of these to use kernel density estimates. Other classification methods can naturally be used. However, given the large number of classes in the task (e.g. 259,200 for $0.5^\circ \times 0.5^\circ$ granularity), simpler methods such as Naive Bayes are more practicable.

Applications

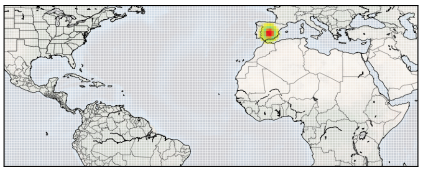
Since the Twitter platform started providing information about the geographic location of tweeters based on self-



lmao no **mames**



porque no te vas a cagar **boludo** despues queres que hablee



venga, va mañana me animo a ir con coleta

Figure 1: Classification output identifying Spanish dialects and regional usage (Mexican, Argentine, and Peninsular Spanish) in three tweets using the WORLDTWEETS corpus. The most geographically indicative words are boldfaced.

reports or built-in GPSs in devices, we have witnessed a flurry of creative applications that take advantage of this new resource. Most applications involve supervised learning algorithms that identify tweet locations from words. The topics of investigation have included the study of language variation along various dimensions: socioeconomic (Alis and Lim 2013), regional (Hong et al. 2012; Kondor et al. 2013), and social variables of language (Eisenstein 2013). Others have tapped into the opportunities of supporting language and cultural education through knowledge of the geographic origin of tweets (Grosseck and Holotescu 2008; Borau et al. 2009). The possibility of identifying disaster and crisis tweets has also been researched (Corvey et al. 2012). One of the intended applications of the current work is the ability to perform fine-grained dialect distinctions automatically, providing opportunities for automatic text classification on non-tweet material. Figure 1 shows an example output of our tweet location classifier tool GEOLOC, discussed in this paper, when trained on tweets from around the globe and provided with short Spanish-language tweets from Mexico, Argentina, and Spain.

Multinomial Naive Bayes

To geolocate a document with words $w_1, \ldots, w_n$ using Naive Bayes and words as features, we assume the standard document classification approach of estimating

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} \hat{P}(c) \prod_i \hat{P}(w_i|c) \tag{1}$$

The prior probability of a document originating from a cell $\hat{P}(c)$ is obtained by observing the number of documents emanating from a cell in the training data, divided by the total document count $|T|$.[1]

$$\hat{P}(c) = \frac{\#(t,c) + \alpha}{|T| + \alpha|C|} \tag{2}$$

Likewise, the conditional estimate $\hat{P}(w_i|c)$ is obtained from the counts of words found in training documents in a particular cell.

$$\hat{P}(w_i|c) = \frac{\#(w_i,c) + \beta}{\sum_{j \in V} \#(w_j,c) + \beta|V|} \tag{3}$$

[1]We use $\#()$ to denote counts.

Here, α and β are our cell and word priors, respectively, and V is the vocabulary seen during training.

Kullback-Leibler divergence

In classifying with Kullback-Leibler (KL) divergence, we try to find a cell whose word distribution matches the distribution of the document. Using the definition of KL-divergence

$$KL(P||Q) = \sum_i P(i) log \frac{P(i)}{Q(i)} \tag{4}$$

we assume P to be the document's distribution of words and Q a given earth cell's distribution of words. Then, for a current document D to be classified, we estimate the probability of the ith word P_{di} to be $\#(w_i \text{ in } D)/|D|$, and the divergence becomes

$$\sum_i P_{di} \log(\frac{P_{di} \sum_{j \in V} \#(w_j,c) + \beta|V|}{\#(w_i,c) + \beta}) \tag{5}$$

using the same quantities as in Naive Bayes for the distribution of a cell's words $\hat{P}(w_i|c)$ in the denominator. Classifying a document entails finding the cell with minimum KL-divergence.

Kernel density estimation

As mentioned above, when geolocating documents with the above methods, using smaller grid sizes leads to an immediate sparse data problem since very few features/words are observed in each cell. The idea behind kernel density estimation is to smooth out the counts of documents and words over a larger region, while simultaneously being able to maintain a small cell size for accurate location classification. Figure 3 illustrates this smoothing effect by showing the distribution of a feature as a single point (in red) and as a density resulting from kernel density estimation.

To this end, when estimating $\hat{P}(c)$ and $\hat{P}(w_i|c)$ in the Naive Bayes and KL-divergence classification, instead of counts, we assign each document and feature a mass in a cell based on a Gaussian that is centered on the actual location where that feature or document was observed. We use a two-dimensional spherical (isotropic) Gaussian kernel to assign mass to each cell: the Gaussian has its means centered at the latitude and longitude of the observation, and each cell

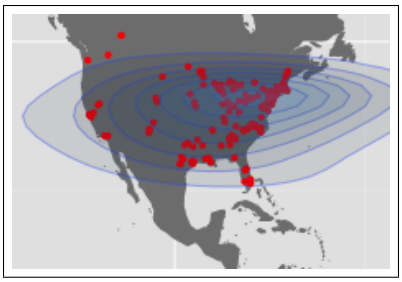


Figure 2: Aggregate density for an English word occurring throughout the United States.

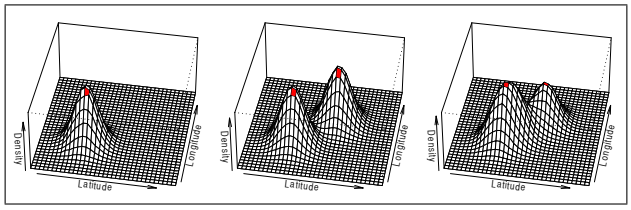


Figure 3: Illustration of areal smoothing effect for features with kernel density estimation, showing the how the contribution of a word feature is spread out over a large area.

receives the corresponding density measured at that cell's midpoint. We then replace our earlier count-based estimates with the respective kernel-based ones:

$$\hat{P}(c) = \frac{\hat{f}_H(t,c) + \alpha}{|T| + \alpha|C|} \quad (6)$$

and

$$\hat{P}(w_i|c) = \frac{\hat{f}_H(w_i,c) + \beta}{\sum_{j \in V} \hat{f}_H(w_j,c) + \beta|V|} \quad (7)$$

Here, the kernel function $\hat{f}_H$ is simply the sum of the relevant individual Gaussians at the midpoint (x, y) of cell c of the form

$$f(x,y) = \frac{1}{2\pi\sigma^2} e^{\frac{-\left((x-\mu_x)^2+(y-\mu_y)^2\right)}{2\sigma^2}} \quad (8)$$

where μ_x and μ_y are the observation coordinates.

We assume a spherical kernel function as a good preliminary evaluation entry point to the overall usefulness of kernel density methods as spherical Gaussians can be calculated quickly and only have one parameter to tune.[2]

Data

For our first experiments, we have used the GEOTEXT geotagged corpus. GEOTEXT is a corpus of 377,616 geotagged tweets originating within the United States by 9,475 users recorded from the Twitter API in March 2010, as documented in Eisenstein et al. (2010). A document in the dataset

[2]The only tunable parameter is a single standard deviation σ, expressing the width of the dispersion of mass of an observation.

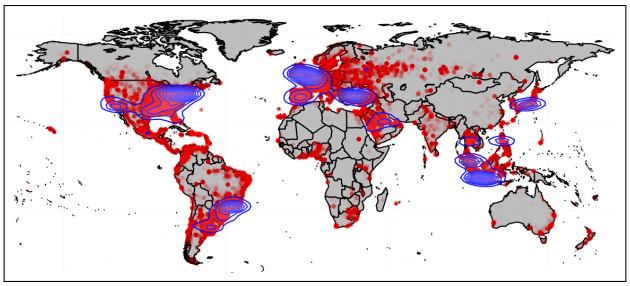


Figure 4: Density map of 1 million tweets sampled from the larger WORLDTWEETS dataset.

consists of all tweets from a single user, concatenated. We use the training/test/dev splits that come with the dataset and are used elsewhere, yielding 5,685 documents in the training set and 1,895 documents in the development and test sets. While the corpus is relatively small—the raw data occupying 54MB—it has the advantage of public availability.[3]

For the second experiment, we used a larger global set of geolocated tweets, WORLDTWEETS, collected during three weeks beginning at the end of January 2014. The set contains 4,870,032 randomly selected geotagged tweets from a wide variety of languages and locations around the globe (as seen in figure 4). We held out 10,000 tweets for development and 10,000 for testing. In this data set, each tweet was considered a separate document (unlike GEOTEXT where tweets from one user are combined into one document).

Preprocessing of both the GEOTEXT and WORLDTWEETS textual data included a basic cleanup and tokenization by simply replacing all non-alphanumeric symbols (except `#`, `@`, `'`) with single spaces, lowercasing all Latin characters, and segmenting on whitespace.

Replication

The program code and relevant instructions for running all experiments are available at our website.[4] We release the main program, GEOLOC, as a stand-alone utility for geolocating arbitrary documents using the methods described in this paper, and also the WORLDTWEETS dataset.

Details

In classifying tweet documents using the above methods we also consider some additional useful implementation details that other authors have taken advantage of.

When classifying a document in a cell $\hat{c}$, we will not assign the actual midpoint coordinates of the cell to the document. Rather, we will use the centroid of all documents in that cell seen during training time. Using a centroid instead of the midpoint of a most likely cell to estimate location has been noted to yield good results (Roller et al. 2012), which is confirmed by our experiments on the development set.

[3]Unfortunately, some larger data sets, such as UTGEO2011 (Roller et al. 2012), are not publicly available.

[4]http://geoloc-kde.googlecode.com

To include a word type in the model, we require a minimum number of occurrences of it during training. Such a threshold has been found to have a large impact by other authors (Wing and Baldridge 2011). Additionally for the GEOTEXT corpus experiments,, we use a list of 572 stopwords for English that we ignore during training. The list originates in the *Textgrounder* project (Roller et al. 2012). In all tests, unknown words during classification are simply discarded, as our preliminary tests on the development set showed consistently worse results with attempts to model unknown word distributions.

Tuning

We tune the following parameters for the density estimation method: (1) the standard deviation of the two-dimensional Gaussian: σ, (2) the vocabulary threshold h, (3) the prior β for words. The document/cell prior α is fixed at 1.

Methods

We tune the parameters on the development set separately for both the standard and density-based methods. In total, we report on four models: NAIVEBAYES—Standard Naive Bayes, KULLBACK-LEIBLER—KL-divergence, NAIVEBAYES$_{kde2d}$—Naive Bayes using kernel density estimation, and KULLBACK-LEIBLER$_{kde2d}$—KL-divergence using kernel density estimation, each described above. In all approaches, we tune for mean location error (rather than median). Table 2 shows the mean errors on the development sets in kilometers with varying vocabulary thresholds and grid sizes. As a baseline, we use the strategy of always choosing the cell with largest document count seen in training.

A coarse grid search over σ, β, and h (threshold) was used to fix σ, after which a finer-grained 3d grid search was used to tune β , h (0-20), and the grid size in degrees (0.5,1,2,5,10) (part of it is shown in table 2). The cell prior α was fixed from the beginning at 1.0 as it has very little effect on the GEOTEXT datasets where documents are long and the prior gets quickly overwhelmed by the data as one 'document' consists of many concatenated tweets from a single user.

Results

The main results of our experiments of the test set of GEOTEXT are given in table 1. Following the results from tuning on the development set, the grid sizes were set to 5 degrees for the non-kernel experiments and 1 degree for the kernel-based ones—for comparison, we also report the kernel-based version results on a 5-degree grid size. Overall, the effect of the kernel-based approach is reflected primarily in the mean error distance, while the median error is roughly similar in both types of approaches. Also, there is little significant difference between the Naive Bayes and KL-divergence approaches. We see an improvement to prior work on the same data set both as regards the mean error and the median error. The KL-divergence without kernel density estimation produces the smallest median error at 333.4km, while NAIVEBAYES$_{kde2d}$ yields the smallest mean error at 764.8km. The fact that KL-divergence slightly outperforms Naive Bayes in the non-kernel setting is a result also found in Wing and Baldridge (2011).

	Grid size				
Thr.	0.5°	1°	2°	5°	10°
0	895.0	889.1	900.8	914.0	1029.8
2	800.2	797.5	798.6	832.7	1029.0
3	753.5	748.9	752.1	795.6	1030.0
5	760.7	**747.7**	756.2	782.1	1021.6
10	789.5	788.91	785.8	783.6	1002.9
20	894.4	887.1	890.7	873.0	1014.1

Table 2: Mean error (km) on the development set for the Naive Bayes classifier with kernel density estimation using different vocabulary thresholds and grid resolutions.

We note, however, that the GEOTEXT data set appears to be too limited and non-standardized to be used for very reliable comparison between algorithms and methods, especially that of different authors. We found large fluctuations of performance on the development set based on text preparation factors such as text normalization, whether to retain user name and hashtag references that sometimes occur within tweets, punctuation removal strategy, and lowercasing. Indeed, these factors were often more important than the specific choice of algorithm, which motivates repetition of the test on the larger WORLDTWEETS. As mentioned above, in our final experiments, the tweets were used as is, with only lowercasing, punctuation removal and tokenization.

In the results for the larger global data set (table 3), WORLDTWEETS, we see a much larger, consistent improvement with the kernel density method, with the Naive Bayes kernel density classifier producing the smallest mean and median error.

The kernel-based methods require a much larger model. This is because for each word, the density for that word in each cell for the whole grid needs to be stored. Naturally, these matrices containing the densities are relatively sparse and have significant weight only near the means of the observation and hence occupy little space individually. Despite the larger model size, the kernel-based methods are not significantly slower at classification time. If classification speed is less of an issue compared with memory, the kernels can also be calculated on-demand. Doing so slows down classification time to roughly 0.5s per document from 0.04s when pre-calculated (at $1^\circ \times 1^\circ$) on GEOTEXT.

Discussion

The general areal smoothing approach presented here could also be extended to include frameworks where several individually inaccurate sources of knowledge are combined to yield a location prediction. This could include IP address information, discussion topic information, census data, and similar material. Under such scenarios, the density measure could be of arbitrary shape—e.g. the precise known area of an IP address pool, boundaries of a city, boundaries of a time

Method	Mean error(km)	Median error(km)	Grid size	Threshold
Most frequent cell	1157.4	756.5	5°	N/A
NAIVEBAYES	855.0	352.3	5°	5
KULLBACK-LEIBLER	802.0	**333.4**	5°	5
NAIVEBAYES$_{kde2d}$	**764.8**	357.2	1°	5
KULLBACK-LEIBLER$_{kde2d}$	781.2	380.0	1°	5
NAIVEBAYES$_{kde2d}$	767.0	397.1	5°	5
KULLBACK-LEIBLER$_{kde2d}$	767.3	400.0	5°	5

Table 1: Performance of different methods on the test set of GEOTEXT.

Method	Mean(km)	Median(km)
Most frequent cell	10929.8	11818.9
NAIVEBAYES	2678.9	637.0
KULLBACK-LEIBLER	2777.6	681.2
NAIVEBAYES$_{kde2d}$	**2429.0**	**531.7**
KULLBACK-LEIBLER$_{kde2d}$	2691.0	578.0

Table 3: Performance of different methods on the test set of WORLDTWEETS. Grid size is 1° and word threshold 5 throughout.

zone—and still integrated into a probabilistic model where Gaussians are used to model the uncertainty of geographic origin of an individual word.

An area of further investigation is also to evaluate the individual contribution of components that other authors have found to enhance accuracy; these include integration of topic models into the task (Eisenstein et al. 2010), k-d trees splitting of grid cells (Roller et al. 2012), n-gram information (Priedhorsky, Culotta, and Del Valle 2014) as well as exploiting tweet metadata such as user profile information (Han, Cook, and Baldwin 2013; 2014) and IP information (Backstrom, Sun, and Marlow 2010) in various ways.

In the current work, no effort has been made to constrain feature dispersion to known land masses. Including such information could also provide gains in accuracy, especially for fine-grained grid sizes. Likewise, more general linear (and other discriminative) classifiers—though more costly to train with very large amounts of data—may profit from the areal smoothing presented in this paper.

Conclusion

We have shown that a kernel-based method alleviates some of the sparse data problems associated with geolocating documents on a discretized surface modeled as a geodesic grid and allows for the use of much smaller grids with less data. The kernel estimation can also be postponed until classification time, avoiding the storage of large models, at the cost of slightly slower classification. In such a case, the resulting model sizes are roughly comparable with those produced by strictly grid-based methods.

Using a kernel-based method significantly improves the mean error on the WORLDTWEETS data set, even when combined with a relatively simple Naive Bayes or KL-divergence based classifier.

References

Alis, C. M., and Lim, M. T. 2013. Spatio-temporal variation of conversational utterances on Twitter. *PloS one* 8(10):e77793.

Backstrom, L.; Sun, E.; and Marlow, C. 2010. Find me if you can: Improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, 61–70. New York, NY, USA: ACM.

Borau, K.; Ullrich, C.; Feng, J.; and Shen, R. 2009. Microblogging for language learning: Using Twitter to train communicative and cultural competence. In *Advances in Web Based Learning–ICWL 2009*. Springer. 78–87.

Corvey, W. J.; Verma, S.; Vieweg, S.; Palmer, M.; and Martin, J. H. 2012. Foundations of a multilayer annotation framework for Twitter communications during crisis events. In *LREC*, 3801–3805.

Eisenstein, J.; Ahmed, A.; and Xing, E. P. 2011. Sparse additive generative models of text. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 1041–1048.

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*, 1277–1287. Cambridge, MA: Association for Computational Linguistics.

Eisenstein, J. 2013. What to do about bad language on the Internet. In *HLT-NAACL*, 359–369.

Eriksson, B.; Barford, P.; Sommers, J.; and Nowak, R. 2010. A Learning-based Approach for IP Geolocation. In *Passive and Active Measurement*, 171–180.

Grosseck, G., and Holotescu, C. 2008. Can we use Twitter for educational activities. In *4th international scientific conference, eLearning and software for education, Bucharest, Romania*.

Han, B.; Cook, P.; and Baldwin, T. 2012. Geolocation prediction in social media data by finding location indicative words. *Proceedings of COLING 2012: Technical Papers* 1045–1062.

Han, B.; Cook, P.; and Baldwin, T. 2013. A stacking-based approach to Twitter user geolocation prediction. In *ACL (Conference System Demonstrations)*, 7–12.

Han, B.; Cook, P.; and Baldwin, T. 2014. Text-based Twitter user geolocation prediction. *J. Artif. Intell. Res.(JAIR)* 49:451–500.

Hong, L.; Ahmed, A.; Gurumurthy, S.; Smola, A. J.; and Tsioutsiouliklis, K. 2012. Discovering geographical topics in the Twitter stream. In *Proceedings of the 21st international conference on World Wide Web*, 769–778. ACM.

Kondor, D.; Csabai, I.; Dobos, L.; Szule, J.; Barankai, N.; Hanyecz, T.; Sebok, T.; Kallus, Z.; and Vattay, G. 2013. Using robust pca to estimate regional characteristics of language use from geo-tagged Twitter messages. In *Cognitive Infocommunications (CogInfoCom), 2013 IEEE 4th International Conference on*, 393–398. IEEE.

Priedhorsky, R.; Culotta, A.; and Del Valle, S. Y. 2014. Inferring the origin locations of tweets with quantitative confidence. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, 1523–1536. ACM.

Roller, S.; Speriosu, M.; Rallapalli, S.; Wing, B.; and Baldridge, J. 2012. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 1500–1510. Jeju Island, Korea: Association for Computational Linguistics.

Serdyukov, P.; Murdock, V.; and Van Zwol, R. 2009. Placing flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 484–491. ACM.

Wing, B., and Baldridge, J. 2011. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 955–964. Portland, Oregon, USA: Association for Computational Linguistics.

Youn, I.; Mark, B.; and Richards, D. 2009. Statistical Geolocation of Internet Hosts. In *Computer Communications and Networks*, 1–6.