Kernel density estimation for text-based geolocation

University of Colorado **Boulder**

Mans Hulden *mans.hulden@colorado.edu*
Miikka Silfverberg *miikka.silfverberg@helsinki.fi*
Jerid Francom *francojc@wfu.edu*

UNIVERSITY OF HELSINKI FACULTY OF ARTS

WAKE FOREST UNIVERSITY

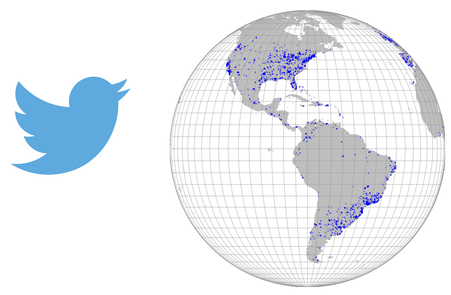
Abstract Text-based geolocation classifiers often operate with a grid-based view of the world. Predicting document location of origin based on text content on a geodesic grid is computationally attractive since many standard methods for supervised document classification carry over unchanged to geolocation 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.

Available resources

GeoLoc, a stand-alone utility for geolocating arbitrary documents, instructions for running all experiments, and the WorldTweets dataset are available at http://geoloc-kde.googlecode.com.

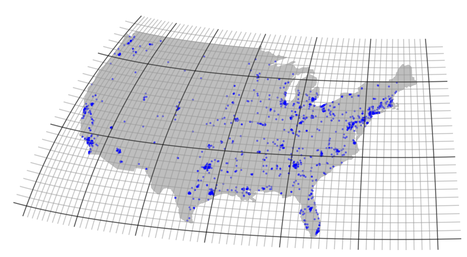
Geolocation on a geodesic grid

- **Text-based geolocation** often treated as a classification task where the object is to place an unknown text document in the most appropriate cell on a geodesic grid.



- **Grids** are simple to implement with standard methods for supervised document classification and competitive with other more complex models (Serdyukov et al., 2009; Wing and Baldridge, 2011).

- **Data sparsity problem** more acute for each cell the smaller and potentially more accurate the grid becomes (Roller et al., 2012).



Standard classifiers

Naive Bayes

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

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

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

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

Where $\hat{P}(c)$ the cell prior and $\hat{P}(w_i|c)$ as a conditional estimate for words found in a particular cell.

Kullback-Leibler divergence

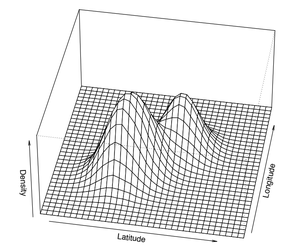
$$KL(P||Q) = \sum_i P(i) log \frac{P(i)}{Q(i)} \qquad (4)$$

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

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

Kernel density estimation

- **Kernel density estimation addresses the data sparsity problem** by smoothing out the counts of documents and features over a larger region.



- **Each document and feature is assigned a mass** in a cell based on a Gaussian that is centered on the actual location where that feature or document was observed when estimating $\hat{P}(c)$ and $\hat{P}(w_i|c)$ in Naive Bayes and KL-divergence classification.

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

$$f(x,y) = \frac{1}{2\pi\sigma^2}\ e^{\frac{-\left((x-\mu_x)^2+(y-\mu_y)^2\right)}{2\sigma^2}} \qquad (6)$$

Where μ_x and μ_y are the observation coordinates.

NaiveBayes$_{kde2d}$

$$\hat{P}(c) = \frac{\hat{f}_H(t,c) + \alpha}{|T| + \alpha|C|} \qquad (7)$$

$$\hat{P}(w_i|c) = \frac{\hat{f}_H(w_i,c) + \beta}{\sum_{j \in V} \hat{f}_H(w_j,c) + \beta|V|} \qquad (8)$$

Experiments

- **Data**: First, the publicly available GeoText corpus (Eisenstein et al., 2010) containing 377,616 geotagged tweets originating within the United States by 9,475 users and second, WorldTweets, a global set of 4,870,032 randomly selected geotagged tweets collected Jan/Feb 2014. 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.

- **Models**: Naive Bayes and Kullback-Leibler divergence classifiers with and without kernel density estimation

- **Grids**: $10^\circ \times 10^\circ, 5^\circ \times 5^\circ, 2^\circ \times 2^\circ, 1^\circ \times 1^\circ$, and $0.5^\circ \times 0.5^\circ$ cells.

- ***kde2d* tuning** for (1) the standard deviation of the two-dimensional Gaussian: σ, (2) the vocabulary threshold h, (3) the prior β for words.

GeoText Results

Method	Mean error(km)	Median error(km)	Grid size
Most frequent cell	1157.4	756.5	5°
NaiveBayes	855.0	352.3	5°
Kullback-Leibler	802.0	**333.4**	5°
NaiveBayes$_{kde2d}$	767.0	397.1	5°
Kullback-Leibler$_{kde2d}$	767.3	400.0	5°
NaiveBayes$_{kde2d}$	**764.8**	357.2	1°
Kullback-Leibler$_{kde2d}$	781.2	380.0	1°

WorldTweets Results

Method	Mean error(km)	Median error (km)	Grid size
Most frequent cell	10929.8	11818.9	1°
NaiveBayes	2678.9	637.0	1°
Kullback-Leibler	2777.6	681.2	1°
NaiveBayes$_{kde2d}$	**2429.0**	**531.7**	1°
Kullback-Leibler$_{kde2d}$	2691.0	578.0	1°

Upshot

- **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.

- **Can be extended to include combined sources of knowledge** to yield a location prediction (e.g IP address information, discussion topic information, census data, *inter alia*)

References

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*, pages 1277–1287. Association for Computational Linguistics.

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*, pages 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*, pages 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*, pages 955–964, Portland, Oregon, USA. Association for Computational Linguistics.