Geolocation of Microbloggers

[Team Members]

**Geolocation of Microblogging Users**

**Introduction**

The average Twitter user has tweeted 307 times. Today, there is a massive amount of data available on social media platforms. However, this comes with a lot of noise. If processed correctly, this valuable information may be used for multiple business purposes including marketing and even location specific advertising.

In this project, we aim to mine and analyze Twitter datasets to predict the user’s location.

From a single tweet, the location can be identified from various sources including location coordinates, hashtags of events and the actual shared content.

**Preprocessing**

This initial step is done in order to

1. Filter out spammers – if the microblogs are less than a certain number of characters or if the user’s followers are far less in number than his friends
2. Word segmentation and stop word removal

We define the following two-fold approach in identifying the most probable home location of the user –

1. Location Inference from a Knowledge Database

2. Social Closeness model

**Location Inference from a Knowledge Database:**

We use a semantic knowledge base like Yago[[1]](#footnote-1) to compare the results of the identified location entities from the microblogs with the entities in the preexisting base. Accordingly, weights are assigned to the close matches of the particular location.

*Exact Matching Entities*

The location entities from the microblogs are extracted and classified into pre-defined locations which match perfectly, using a Named Entity Recognition (NER) tool.

*Fuzzy Matching Entities:*

Further, we employ fuzzy matching algorithms - N-gram or Q-gram-based Algorithms - to eliminate duplicate records which may arise due to typological errors, in order to find non-exact matches.

**Social Tightness Model:**

This model is based on the assumption that diﬀerent friends have diﬀerence importance to a user. Therefore, in order to accurately estimate a user’s location, we also take into consideration social closeness.

**Refinement**

Based on the aforementioned models, probable locations for each user are given cumulative weights. The highest ranking location in the set is predicted to be the user’s home location.

**Implementation Overview**

**Raw Data from Microblogs :**

Social media such as twitter is cluster of vague data; transforming such raw, free form real time text into meaningful information is a challenging task. We should clean such a raw data such as tweet by removing stop words and correcting misspelling, classifying into various domains/areas, locating the source of tweets and convert it into a more structured model like a histogram using various models such as FOAF and SIOC.(by maintaining privacy of the user.

**Preprocessing (Noise Removal):**

 Supervised learning methods to ( a trained dataset) obtain domain knowledge were initially used which contained both noisy and clean text. A more generalized method would be applying localized linguistic techniques to extract opinion expression from noisy texts. It is however very difficult to obtain reasonable size of domain data for the semi supervised learning of domain knowledge. A  java spell checker can named suggester along with a weighted function based on domain frequency of a word to suggest the correct spelling of a possibly misspelled word.

**Location Inference:**

Real-world applications depend on quantified uncertainty for  estimation of accuracy, precision and calibration. A content based approach can used to estimate the location of tweets using variants of gaussian mixture models.Applications in public health politics disaster management and other domains are increasingly turning to social Internet data to inform policy and intervention strategies. However, the value of these data is limited because the geographic origin of content is frequently unknown. Thus,here is growing interest in the task of location inference. Models such as Guassian Mixture Models(GMM) are based on trained on geotagged tweets, i.e., messages with user profile and geographic true origin points.For each unique n-gram, we fit a two-dimensional GMM to model its geographic distribution. To infer the origin of a new tweet, we combine previously trained GMMs for the n-grams it contains, using weights inferred from data; Figure shows an example estimate. This approach is simple, scalable and competitive with more complex approaches given an item, estimate its geographic

true origin.

**Conclusion**

In this project we attempt to accurately detect the location of microbloggers by (1) referring to a preexisting global knowledge base and (2) implementing the social closeness model. We estimate the efficiency of our prediction by comparing it to the actual user locations (validation set) and calculating the possible error distance. We propose to showcase the final output through an interactive web interface. The scope of this project is limited to microblogs tweeted in English. We are not considering multi-lingual tweets.

**References**

[1] Chuanyang Li, Xiuqin Lin, Bin Wu, Chuan Shi Location Inference Using Microblog Text and Friendships. ASONAM 2014

[2] Guoliang Li, Jun Hu, Kian-lee Tan, Zhifeng Bao, Jianhua Feng Effective Location Identiﬁcation from Microblogs. ICDE 2014

[3] Longbo Kong, Zhi Liu, Yan Huang SPOT: Locating Social Media Users Based on Social Network Context. VLDB 2014

[4] Cheng, Z., Caverlee, J., and Lee, K. 2013. A Content-driven Framework for Geolocating Microblog Users. ACM 2013

[5] Microblogging: A Semantic and Distributed Approach Alexandre Passant Tuukka Hastrup Uldis Bojars John Breslin LaLIC, Universite Paris-Sorbonne, 28 rue Serpente, 75006 Paris, France [firstname.lastname@paris4.sorbonne.fr](mailto:firstname.lastname@paris4.sorbonne.fr) DERI, National University Of Ireland, Galway, Ireland [firstname.lastname@deri.org](mailto:firstname.lastname@deri.org)

[6]Topic Time Series Analysis of Microblogs Eric Lai(ellai@uci.edu)Daniel Moyerxy moyerd@usc.edu Baichuan Yuanz ybcmath@zju.edu.cn Eric Fox x eric.fox@stat.ucla.edu Blake Hunter x{ blake.hunter@claremontmckenna.edu Andrea L.Bertozzi xbertozzi@math.ucla.edu Je  
rey Brantingham x [branting@ucla.e](mailto:branting@ucla.e)

[7] A LEXICON BASED ALGORITHM FOR NOISY TEXT NORMALIZATION AS PREPROCESSING FOR SENTIMENT ANALYSIS Sudipta Roy Sourish Dhar Saprativa Bhattacharjee Anirban Das Department of IT, Triguna Sen School of Technology, Assam University, Silchar Department of IT, Triguna Sen School of Technology, Assam University, Silchar Department of IT, Triguna Sen School of Technology, Assam University, Silchar Department of IT, Triguna Sen School of Technology, Assam University, Silcha

[8]Inferring the Origin Locations of Tweets with Quantitative Confidence Reid Priedhorsky,Aron Culotta, Sara Y. Del Valle Los Alamos National Laboratory Los Alamos, NM {reidpr,sdelvall}@lanl.gov Illinois Institute of Technology Chicago, IL aculotta@iit.edu

1. <http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/> [↑](#footnote-ref-1)