GEOLOCATION INFERENCING ON SOCIAL MEDIA USING GAUSSIAN MIXTURE MODEL

An Undergraduate Research Scholars Thesis

by

NAZIF ALI

Submitted to the Undergraduate Research Scholars program at
Texas A&M University
in partial fulfillment of the requirements for the designation as an

UNDERGRADUATE RESEARCH SCHOLAR

Approved by Research Advisor: Dr. James Caverlee

May 2017

Major: Computer Engineering

TABLE OF CONTENTS

ABSTRACT

Geolocation Inferencing on Social Media Using Gaussian Mixture Model

Nazif Ali
Department of Computer Science and Engineering
Texas A&M University

Research Advisor: Dr. James Caverlee
Department of Computer Science and Engineering
Texas A&M University

Modeling human behavior over social media can provide valuable insights into crowd behavior. It can be used as sensory data to understand and predict how crowds react to a certain local or international event. This can lead to applications that can predict elections, track flu and detect earthquakes. However, this analysis requires data that are geo-tagged, and most of the social media data has no location associated with it. Many models and algorithms have been proposed to find the location of a user based on his or her social media profile. Unfortunately, most methods are not scalable or robust enough to work perfectly in real world applications. In this research, I have tested and improved Gaussian Mixture Model (GMM) on tweets ranging from 325,875 to 2,332,305 to predict a Twitter user's location based purely on the tweet content. The experiments test different tokenization approaches, dataset sizes, temporal feature and languages in the dataset to conclude that GMM can indeed solve the location-sparsity issue in social media and pave way for location-based personalized information services.

CHAPTER I

INTRODUCTION

Social media provides a great platform for understanding human behavior relating to different languages, cultures and events. Associating data with a particular geographic location provides a powerful tool for modeling social behavior and trends like predicting elections or observing linguistic differences (Jurgens 2015). However, there is little social media data that are geolocation-annotated; several different techniques have been proposed to infer a user's physical location based on his or her social media profile data.

Some of the proposed techniques to infer location rely on evaluating the contents of a user's tweet. This technique requires classifying words in the user's tweets which correspond to a certain local geo-scope. The underlying assumption is that the content may contain location information or jargon that can be associated with a certain location (e.g. the use of 'howdy' might imply the user is from Texas) (Cheng 2010). Other techniques depend on mining the locations of the user's network of friends and followers on a particular social media site. One of the first such techniques was proposed by Backstrom, Sun and Marlow (2010) using Facebook as their social network site. They analyzed known locations of the user's friends to find a location that had the highest probability of being the user's true location. Over the years, multiple extensions to this model have been proposed over different social media sites (e.g. Twitter) (McGee 2013).

Unfortunately, different models work differently depending on the parameters, like the size of dataset and the temporal recency of data. Also, widespread difference among the data, conditions and metrics used to assess each model makes it difficult to compare models in real

applications. Hence, in real world scenarios, these models often fail to reach the productivity they were promised to achieve. My research will consider multiple factors in evaluating the worth of GMM in predicting locations based purely on tweet content.

Different tokenization assumptions will be tested to expose the location information hidden in the training corpus and to suggest improvements. Effects of the size of training dataset and temporal recency of GMM will also be tested, and finally, I will separate the dataset into different languages to learn that language does have an impact on the quality of results produced by GMM. From all these tests, we find that GMM has a great potential to make social media informational applications a reality by accurately estimating tweet locations.

CHAPTER II

METHODS

To conduct experiments, I utilized a publicly available GMM library developed by a group of commercial programmers. The library takes large number of geo-tagged tweets to train the GMM classifier which then predicts the estimated location of test tweets in terms of latitude and longitude values. The model will be evaluated on several factors over tweets in multiple languages collected from across the globe to find the limiting factors and suggest improvements to the model.

What is GMM?

GMM is a probability density function represented as the weighed sum of multiple Gaussian densities. It is represented as,

$$P(\boldsymbol{x}|w_i,\ \mu_i,\ C_i) = \sum_{i=1}^{M} w_i . g(\boldsymbol{x}|\mu_i, C_i) \quad \text{.....(Equation 1)}$$

where $\boldsymbol{x}$ is the data vector, i ranges from $1, \dots, M$, w_i are the weights of each component and $g(\boldsymbol{x}|\mu_i, C_i)$ are the component Gaussian densities in the mixture with μ_i and C_i being the corresponding mean and covariance matrix.

In the context of geo-inferencing, the geospatial locations of the appearance of each unique word in the tweet corpus (or training sample) is collected and fitted on a GMM. This creates a dictionary of different words and their corresponding geospatial distributions across the globe. To predict a tweet's location, the classifier combines the GMM of each word in the tweet to give a GMM for the tweet itself. The resulting GMM can be used to find the most likely location of the tweet and the probability of the tweet being within some radius. Figure 1 shows

the GMM of word ‘texas’. The benefit of using GMM is that it can identify different clusters/populations in the data and assign weights to them.

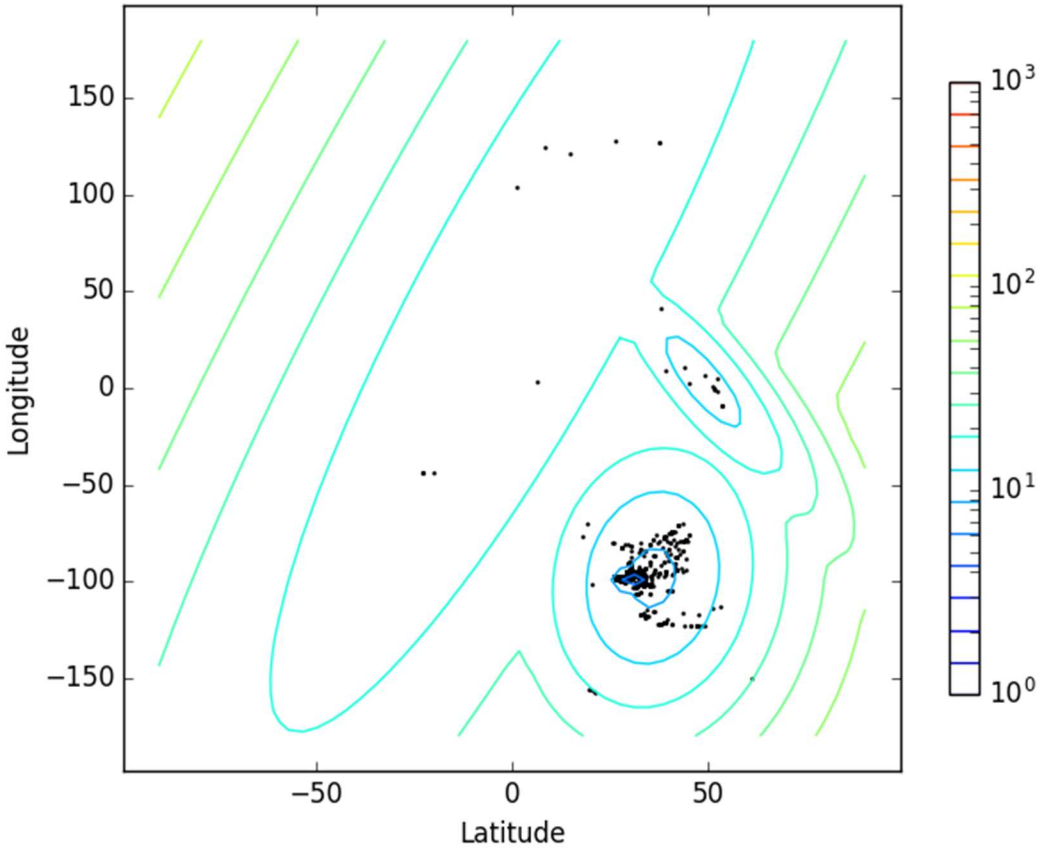


Figure 1. GMM of Word ‘texas’

Evaluating the model

Two metrics are used to evaluate the quality of GMM, namely Median Error and the Mean Error. An error is defined as the distance between the predicted location and the actual location of the tweet. The actual location is typically generated by the smartphones or laptops when users geo-tag their tweets; therefore, it can be considered as the ground truth. Median Error gives us the error of the 50^{th} percentile point in the sorted list of errors, while Mean Error gives us the average of all errors calculated in the test data.

Effects of tokenization

Each tweet in the training sample is broken down at white spaces to generate a list of tokens. Tokens can be words, punctuations, numbers, dates, emojis, @-mentions, hashtags, URLs and so on. The underlying assumption of this whole research approach is that some tokens are better indicators of geospatial location than others. For example, if a tweet uses the word 'rockets' then there is a good chance it was originated from Houston. To test the geolocation scope of each token, I conducted the following experiments.

No changes

A control experiment was done to compare the changes brought about by adding or removing different tokens in the following experiments. URLs and @-mentions were eliminated by default because usernames of other twitter users and names of websites and links does not have much geo-scope in them.

Removing punctuations

When initial analysis on GMM was conducted, it was hypothesized that punctuations and emojis do not contain any geolocation scope within themselves. It was also noticed that trailing punctuations were separating similar words. For example, 'again', 'again!', 'again,', 'again.', 'again..' and 'again?' were being treated as different tokens with their own GMMs when they are essentially the same word. Hence, the punctuations "!#$%&'()*+,-./:;<=>?[\]^_`{|}~ were removed which eliminated many emojis and any trailing, leading or within-word punctuations.

Adding bigrams

Phrases also have a certain geo-scope in them that single words do not. For example, appearance of the word 'Arlington' in a tweet can imply that the tweet is from any of the 21 towns in United States named Arlington. However, the phrase 'UT Arlington' can tell us that the

tweet is most probably form Arlington, Texas. Hence, I trained GMM on all the unique consecutive two-word phrases that appeared in the tweet corpus.

Removing stop words

Stop words are the list of most common words used in a language (e.g. 'the', 'is', 'at', 'which', and so on). Stop words have little location information in themselves and are commonly removed from textual analysis. Hence, 2,781 stop words from Danish, Dutch, English, Finnish, French, German, Hungarian, Italian, Kazakh, Norwegian, Portuguese, Russian, Spanish, Swedish and Turkish languages were removed using the Python's Natural Language Toolkit library.

Removing numbers

Since the punctuations were removed from tokens, all the dates, times, monetary amounts, temperature values, percentages, fractions and decimals were converted to simple integer numbers. Therefore, all these numeric tokens were removed from the dictionary to find any substantial changes in the results.

Size of training and test data

The hypothesis is that increasing the size of training data should improve the results, and the size of test data should have no impact on results. The bigger the training corpus, the more data points per word, which gives a more realistic fit on the GMM. Five different training samples were created with sizes ranging from 325,875 to 2,332,305 tweets. Five different test samples were created with sizes ranging from 117,614 to 588,033 tweets while the training sample was fixed to 651,754 tweets to reveal any trends in the median and mean errors.

Temporal recency of training data

Due to changing user base of a social media platform, the prediction capability of a training model decrease by half after every 4 months (Jurgens 2015). This introduces the

problem of retraining the model on more tweets every four months which is complicated and takes time and resources. To test if an old GMM model can be valid on latest tweets, two models were trained on tweets that were randomly selected from two seasons, Spring of 2014 and Fall of 2015, roughly 20 months apart. GMM from each sample was tested on the latest tweets from March 2017 which were downloaded from the Twitter API's tweet streaming facility.

Separating languages

Twitter is a platform where users from across the world engage in live, public conversations. Many businesses want to leverage this global platform by connecting with users that have diverse backgrounds and interests. They want to talk in the language that their audience understand, so they can target relevant ads in their language. Therefore, I separated the training dataset into four different languages to see how GMM would work on each of them. English was chosen because it is the most popular language on Twitter. Spanish, Japanese and Arabic were chosen because they are spoken in geographically and culturally very diverse regions and also because they are written in morphologically different scripts.

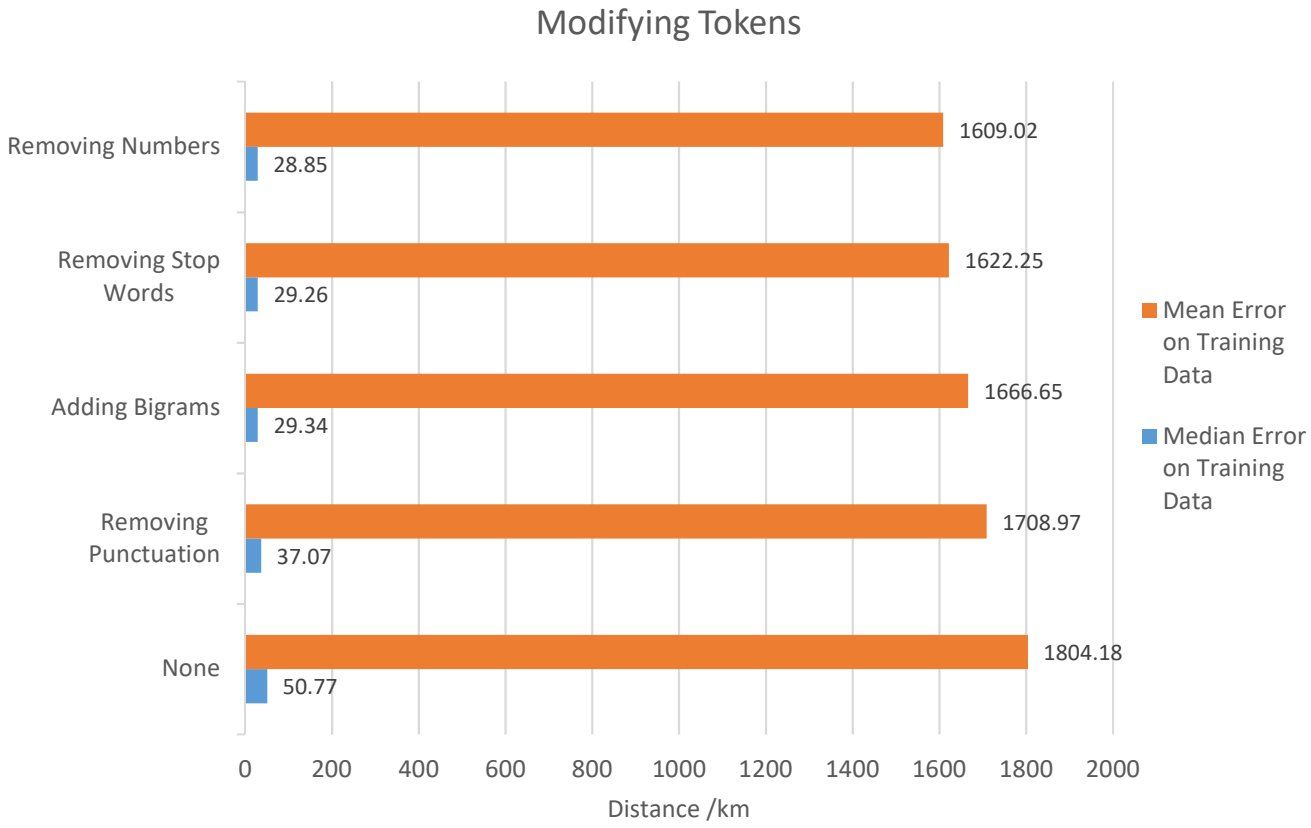
CHAPTER III

RESULTS

The results of each experiment described in the Methods section are discussed below.

Effects of tokenization

Results showed that removing punctuation and adding two-word phrases decreased errors significantly, while removing stop words and numeric tokens had little impact on the errors. This can be seen in Graph 1.



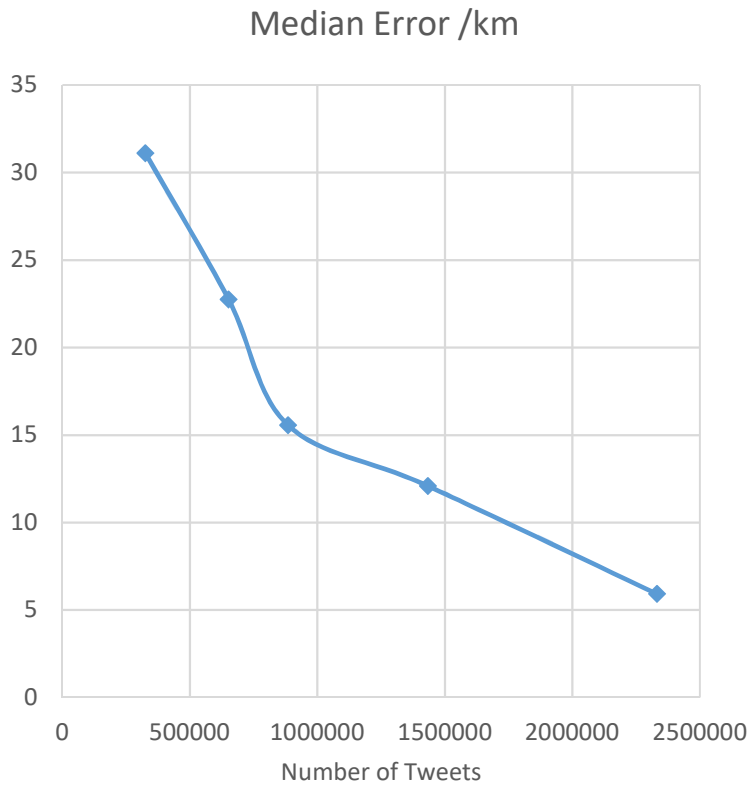
Graph 1. Modifying Tokens.

Removing punctuation reduced the median error by 27% and mean error by 5.3%. This solidifies the hypothesis made in the previous section that punctuations and emojis do not

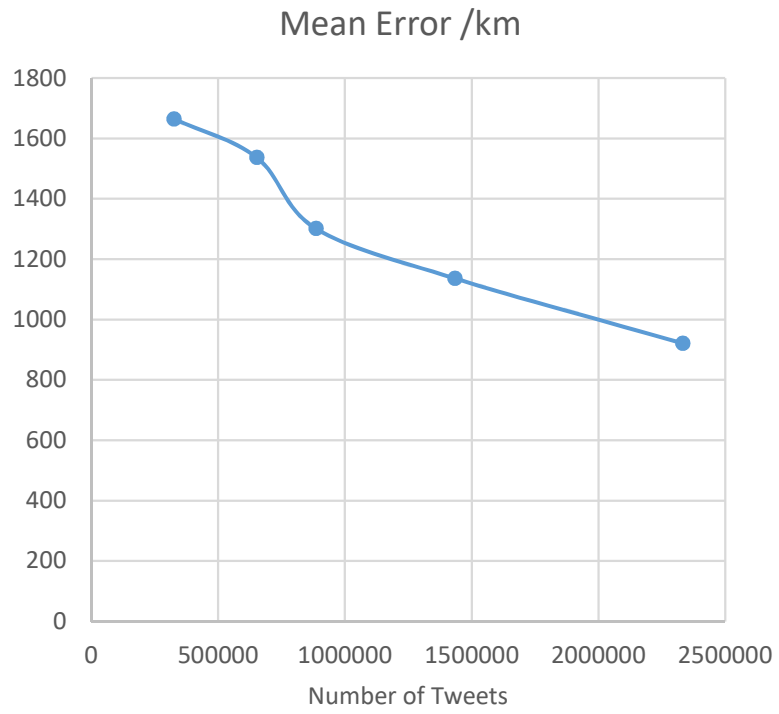
contain much geo-scope. Adding bigrams further reduced median error by 29% and mean error by 2.5%. While removing stop words had little impact on the median error, it reduced mean error by a substantial 2.7% margin.

Size of training and test data

As hypothesized, GMM results can be improved by simply training it on large enough dataset. We see a steep decline in errors from a slight increase in training dataset size. The median error reduces by ½ when dataset size is increased by just 2.7 times. We can expect to see further reduction in median and mean errors with further increase in dataset size. The median error was reduced to as low as 6 km by simply increasing the training data size. This can be seen in Graph 2 and Graph 3.

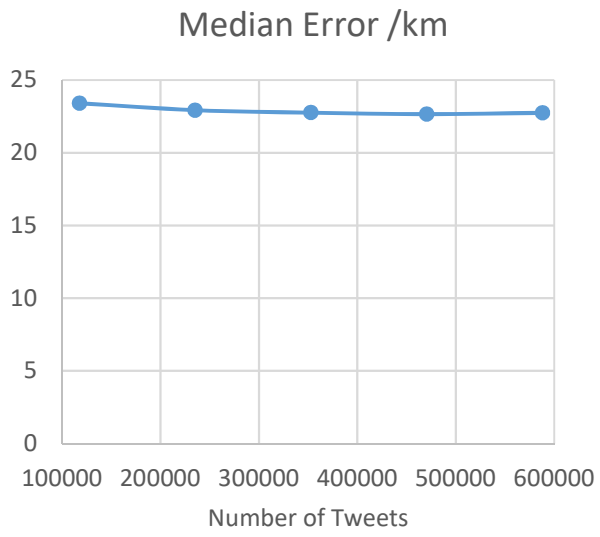


Graph 2. Median Error for Training Data.

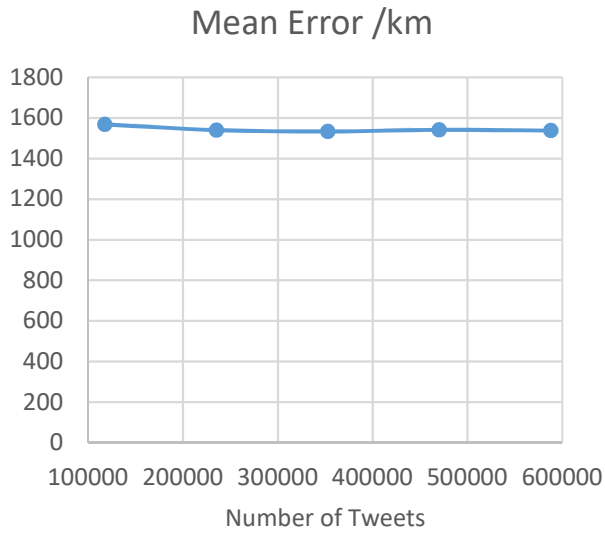


Graph 3. Mean Error for Training Data.

Again, as hypothesized, varying the size of test data had no effect on median or mean error. We do not need to train the model on larger or smaller dataset if the test dataset is of a different size. This can be seen as a straight horizontal line in Graph 4 and Graph 5.



Graph 4. Median Error for Test Data.



Graph 5. Mean Error for Test Data.

Temporal recency of training data

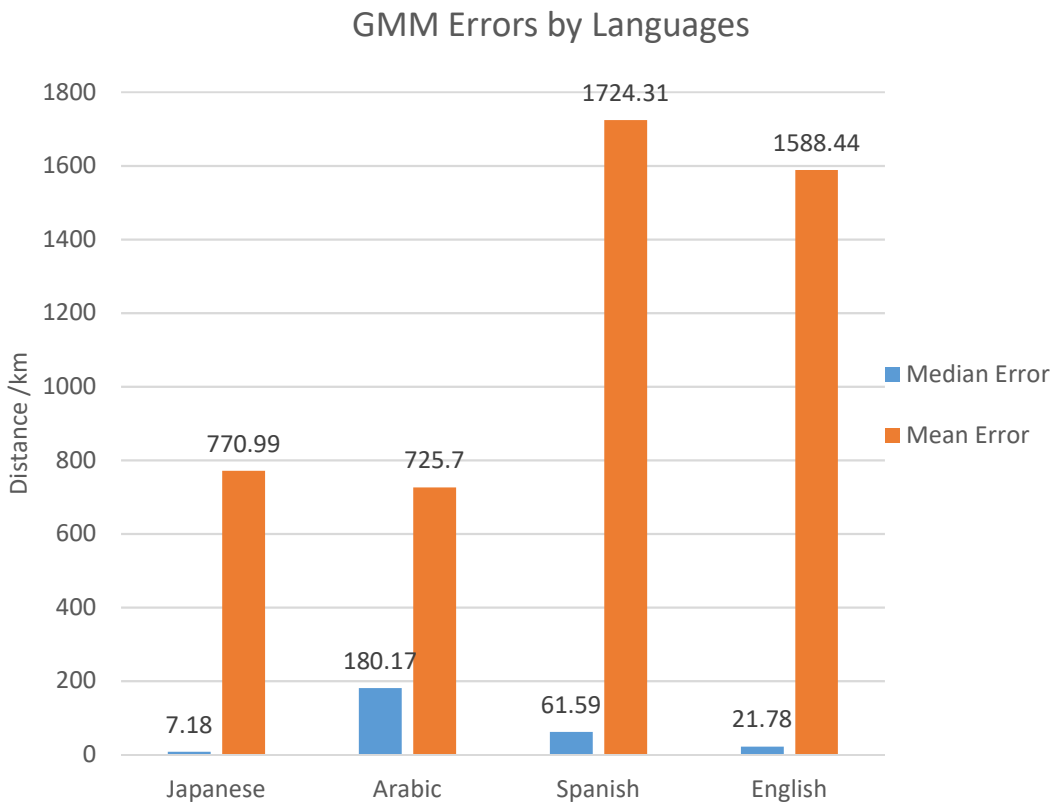
As Table 1 shows, GMM trained on old data can still give accurate results on new test data. The difference in median errors from both experiments is just 0.46 km which is very small. Businesses can use their resources to train GMM on a huge dataset just once and do not have to worry about training it again for a long time. Since the size of training sample can affect the accuracy of results, I kept it similar for the datasets from each season.

Table 1. Effect of the Age of GMM.

Training Sample	No. of Tweets in training sample	Median Error on 2017 tweets /km	Mean Error on 2017 tweets /km
Fall 2015	885425	15.56	1301.41
Spring 2014	858847	15.10	1234.48

Separating languages

Training GMM on different languages gave different quality of results as seen in Graph 6. Japanese was predicted with a very low median error of just 7.18 km, while Arabic had the highest median error of 180.17 km. Both, Japanese and Arabic tweets, had a low mean error compared to Spanish and English. This could be explained by the geographical spread of each language. Spanish and English are spoken around the world, increasing the geographical region under test, and hence, the mean errors. On the other hand, Japanese and Arabic are spoken in smaller geographic regions, keeping the mean errors low.



Graph 6. GMM Errors by Language.

Different quality in median errors could be explained by the fact that each language is morphologically different. The tokenization function applied in the GMM was geared towards English and other European languages. The same tokenization principal cannot be applied to other languages, like Arabic, where the words need to be broken further to find the smallest semantic unit with meaning. In short, further work could be done in tokenizing morphologically different languages and tweaking GMM to improve language-specific results.

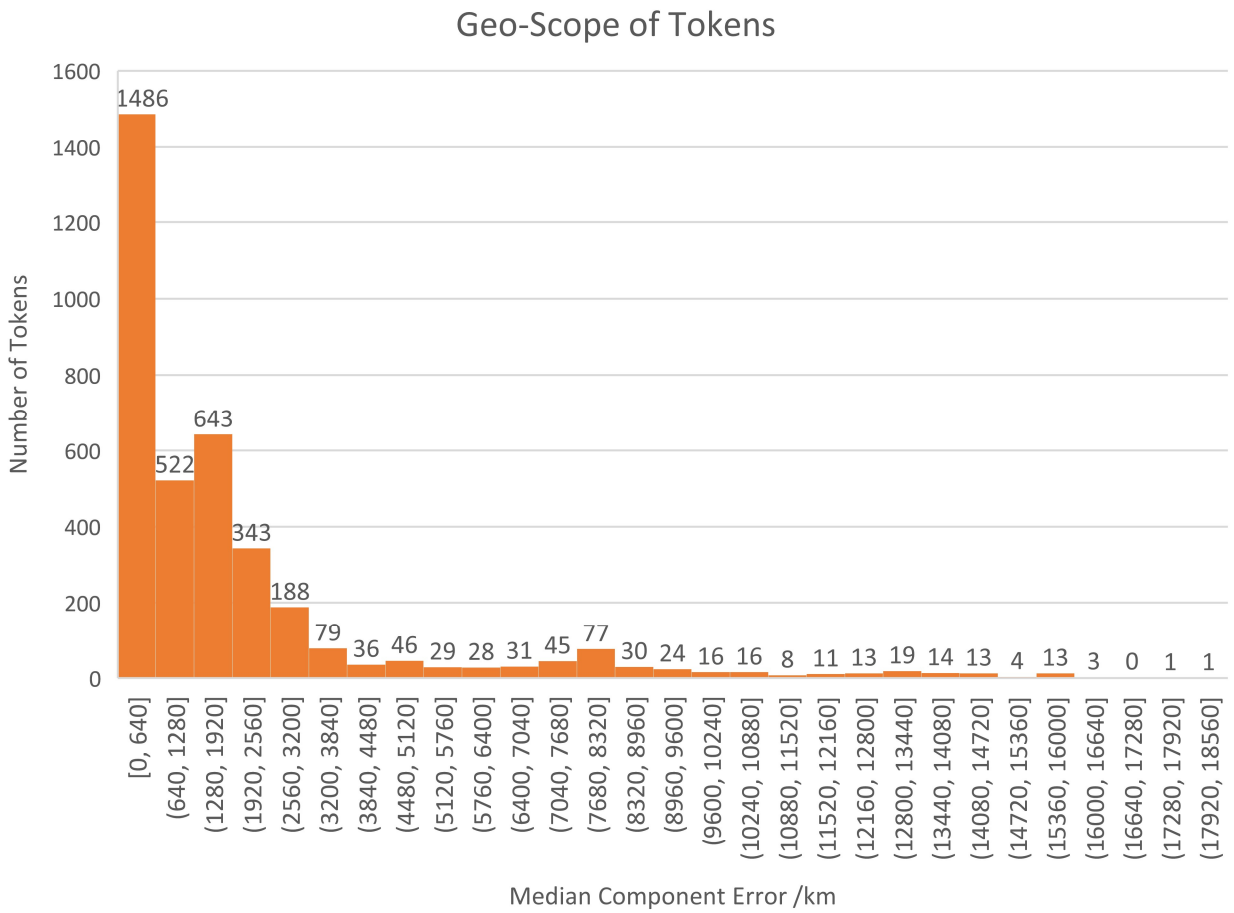
High mean errors

In most of the results above, there was a significant difference of at least one order of magnitude between the median and mean errors. This could be explained by the observation that the tweet corpus contains tokens like ‘would’, ‘tweet’, ‘sorry’, ‘every’ which do not store any

location information in them. GMMs of these tokens only add noise to the tweet GMM resulting in high average error.

To analyze the location-specificity of each word in the tweet corpus, the GMM library had a function to calculate the median component error of each GMM. It is the median of distances between the means of each component in GMM and the mean of the component with the highest weight. The smaller the median component error, the more geographically localized the token is. These errors were calculated for a small corpus of 46785 tweets, and the histogram of the resulting errors was plotted as shown in Graph 7. As we can see, many words have a broad distribution over the globe and cannot be good indicators of location, explaining high mean errors. Further work can be done in filtering these words out. However, this would reduce the size of vocabulary, and the GMM would not be able to estimate any location for a proportionate number of tweets.

On the other hand, low median errors can be attributed to the tall peak on the left with 1486 words which have median component errors less than 640 km. This means that there is a group of tokens with sufficient geo-scope in it. This group comprises of almost $2/5^{th}$ of all the unique words in the corpus, and helps keep the median and mean errors low.



Graph 7. Geo-scope of Tokens.

CHAPTER IV

CONCLUSION

Social media contains massive amounts of data that can be used to power many beneficial machine learning applications. However, these applications depend heavily on data that has location associated with it, and most of the data on social media is not geo-tagged. This paper has analyzed the use of Gaussian Mixture Model to predict the geographical location of tweets based purely on content. The results of the experiments look very promising in making real-time personalized social media information service applications a reality.

We learned that by removing punctuations and adding phrases through GMM's tokenization function, we can help expose the location semantics hidden in the training sample and improve results. We further observed that increasing the training sample size can substantially reduce errors, while the size of test sample does not make any difference in the results. A trained GMM can last for a long time. We saw that a GMM trained on three years old tweet sample can estimate locations of the latest tweets with surprisingly low errors, and that training GMM on different languages can increase or decrease errors depending on the language.

Further work can be done on GMM to get even better results. Words with little location scope in them could be filtered which can reduce mean errors significantly. Also, each morphologically different language can be tokenized with its specialized tokenization principle such that the location scope of words in that language are most exposed.

REFERENCES

Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: a content based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (CIKM '10). ACM, New York, NY, USA, 759-768. DOI=http://dx.doi.org/10.1145/1871437.1871535

Jeffrey McGee, James Caverlee, and Zhiyuan Cheng. 2013. Location prediction in social media based on tie strength. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management* (CIKM '13). ACM, New York, NY, USA, 459-468. DOI=http://dx.doi.org/10.1145/2505515.2505544

David Jurgens, Tyler Finethy, James McCorriston, Yi Xu, and Derek Ruths. 2015.Geolocation Prediction in Twitter Using Social Networks: A Critical Analysis and Review of Current Practice. *International AAAI Conference on Web and Social Media 2015 (ICWSM '15)*. Available at: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10584.

Lars Backstrom, Eric Sun, and Cameron Marlow. 2010. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceeding of the 19th international conference on World Wide Web*. ACM, New York, NY, USA, 61-70.