COMP30027 Report

Anonymous

1. Introduction

Geotagging using short texts is a powerful tool that can be applied in many fields. Whether it be for target advertising or defence (Pappas 2018), the fundamental problem of how to locate users using text remains the same. The rationale of this research is that there are differences in geography, culture, toponyms and linguistic variations of words between locations, which implies that geolocation using text should be possible (Roller 2012). In this research, tweets from twitter users are used as the short text from which we attempt to geolocate the location they were written in. Within this study, these locations were limited to the Australian states of Brisbane, Perth, Melbourne and Sydney, all of which have equal frequencies. The goal of this paper is to use machine learning methods to classify 108 148 tweets using the 103 364 training data.

2. Related literature

While generally most geolocation studies have mostly been conducted on blogs and tweets (Li 2012), there has been an many different of ways in which they were processed. Roller's research used kNN to calculate the similarity of each unlabelled tweet to a labelled tweet using cosine distance (Roller 2012). While Pappas introduced the Word Locality Heuristic (WLH) and Location Lexicon which assigned a weighted probability to each word used within a location (Pappas 2018). In contrast, Popescu and Grefenstette approached the geotagging problem using logistic regression (Popescu 2010). In this research, Logistic regression and a variation of the WLH and Location Lexicon are used.

3. Content based geotagging

3.1 Bassline classifier

To begin, the zero-R classifier was used as the bassline classifier in order to set a bare minimum standard of performance. Since all states in the training data are equally distributed, the bassline is 25% accuracy

3.2 Multinomial Naïve Bayes and Linear Regression

Next, the multinomial Naïve Bayes was applied on the University of Melbourne's frequency count of words that were in the top 10 and top 100 most 'present' word, determined by Mutual Information and Chi-Square, for each tweet. To test its performance, the data was split into 66% training set and 33% testing set and then building the model out of the training and applying it on the testing set. This process was repeated three times through repeated random subsampling and the mean accuracy was computed. The same process was repeated but now using logistic regression. The following are the accuracies from each model:

	Multinomial Naïve Bayes	Logistic Regression
Top 10	%28.5	%29.0
Top 100	%33.1	%33.6

As expected, the Multinomial Naïve Bayes model performed marginally better than the bassline classifier because the attributes were natural numbers corresponding to counts. However, its conditional independence assumption is not always met in this dataset, for example, the word 'status' may usually be used in conjunction with the word 'quo'. Similarly, Logistic regression also performed as expected with its all-round slightly better scores than Naïve Bayes. This is simply because logistic regression is a more powerful model than Multinomial Naïve Bayes and since there was sufficiently large amount of data to build the model (68 221 instances), this model was able perform well.

3.3 Location Lexicon

The variation of Pappas' Location Lexicon was also created to classify the tweets. This was done by firstly doing some pre-processing of removing non-alphanumeric characters in the tweets and then creating a variation of Pappas' WLH which is defined by "the probability of a word occurring in a state, divided by its probability to appear in any state" (Pappas 2018), but unlike Pappas' WLH, this version returns the WLH score and the state with the highest score and the frequency of

the word appearing. The formula for the WLH score is as follows:

$$\text{WLH score} = \frac{P|(w|s)}{P(w)}$$

Using the WLH, the Location Lexicon was created by filtering out words that were written at least x number of times and with a WLH score greater than h, with the respective intuition for each being that lexicons should be used by many users and be reasonably indicative of a state. The Location Lexicon can be used to predict by going through each pre-processed word in a message, classifying each word to a state and then classifying the whole text to the highest number of states classified, with tie breakers being settled by classifying to the state with the highest WLH score. On the other hand, if no words in the message are in the location lexicon, then the state is randomly guessed.

4. Hyperparameter tuning

The hyperparameters that were tuned were x and h from the location lexicon. This was done through calculating the mean accuracies using repeated random subsampling five times on a 80% training set and 20% testing set from the accuracies generated by using values of x between 5 to 100 in steps of 5 and values of 0.25 to 0.67 in steps of 0.03. Afterwards, the mean accuracies between the same x values and h values were computed respectively to get the following results:

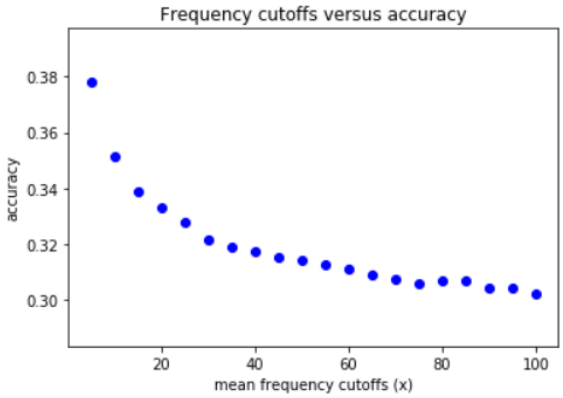

Figure 1: affects of minimum number of times a word was written on accuracy

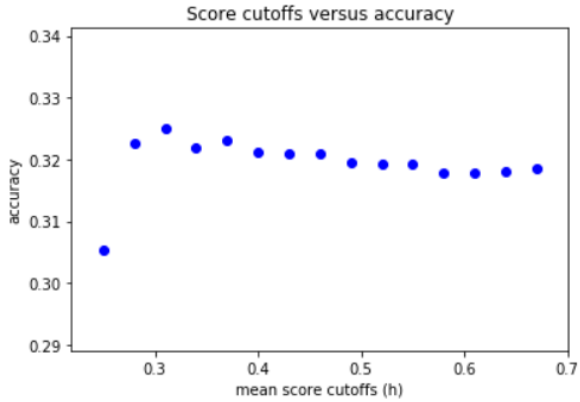

Figure 2: affects of minimum WLH score on accuracy

Figure 1 shows a seemingly logarithmic decrease in accuracy towards 0.25 as the x values decrease, where the model begins to randomly classify instances more often. Due to the sufficiently sized data and the even distribution of the classes, just randomly assigning classes will probabilistically give accuracies of around 0.25. Therefore, the effects of overfitting is not shown here as any values predicted by overfitting is similar to randomly assigning classes or zero-r in the case that there is a bias made from overfitting. Meanwhile, the small advances in knowledge from learning information about low frequency words appears to have had non-trivial effects on accuracy. In Figure 2, the scatter plot shows a significant drop in accuracy when lexicons have a h value of 0.25 as at this point, there is no information to be gained from the words. While at approximately 0.28-0.37 h values, the classifier remains balanced with having trimmed enough words to give meaningful information while having sufficient words such that many words are classified by Location Lexicon rather than by random classification. Using the accuracy computed by iterating over these x values and h values, the highest accuracy attained was 0.388 with x = 5 and h = 0.37, which will be used as hyperparameters for Location Lexicon

5. Error Analysis

Location Lexicon

Actual \ Predicted	Brisbane	Perth	Melbourne	Sydney
Brisbane	772	338	304	395
Perth	430	761	338	359
Melbourne	484	375	703	371
Sydney	479	376	335	644

Multinomial Naïve Bayes top 100

Actual \ predicted	Brisbane	Perth	Melbourne	Sydney
Brisbane	1217	6894	180	273
Perth	309	7657	324	199
Melbourne	255	6820	1162	300
Sydney	313	6872	291	1045

Logistic regression top 100

Actual \ Predicted	Brisbane	Perth	Melbourne	Sydney
Brisbane	1209	6765	190	273
Perth	352	7803	307	209
Melbourne	247	6806	1154	248
Sydney	336	6838	282	1092

In Location Lexicon, the errors have a relatively even distribution because of the random classification used when there are no words in the lexicon. In contrast, the Multinomial naïve Bayes and logistic regression appears to have higher precision in every class except for Perth predictions, which is assumed to occur whenever all values are 0. Other than this, there appear to be no irregularities or biases in predictions.

6. Conclusion

In this research, the geotagging of tweets was conducted to determine the state in which they were made. Through utilising the top 10 and top 100 data, it was seen that the Multinomial Naïve Bayes does reasonably well to classify the data, but logistic regression is still in overall better given that it had a sufficiently sized data. The Location Lexicon built out of the WLH appears to be an even better classifier given the choice of x is small and h is between the 0.28-0.37 zone. Despite the Location Lexicon's better overall accuracy, its precision accuracy is lower than both Multinomial Naïve Bayes and Logistic Regression across all classes excluding the class that was given to all zero vectors.

References

Pappas, K., Azab, M., & Mihalcea, R. (2018). A Comparative Analysis of Content-based Geolocation in Blogs and Tweets. *arXiv preprint arXiv:1811.07497*.

Roller, S., Speriosu, M., Rallapalli, S., Wing, B., & Baldridge, J. (2012, July). 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 (pp. 1500-1510). Association for Computational Linguistics.

R. Li, K. H. Lei, R. Khadiwala, K. C.-C. Chang, Tedas: A twitter-based
event detection and analysis system, in: Data engineering (icde), 2012 ieee
28th international conference on, IEEE, 2012, pp. 1273{1276.

A. Popescu, G. Grefenstette, et al., Mining user home location and gender
from ickr tags., in: ICWSM, 2010.