GEOLOCATION THROUGH LANGUAGE RECOGNITION

Will Potter

Adviser: Professor David Kauchak

A Thesis

Presented to the Faculty of the Computer Science Department

of Middlebury College

in Partial Fulfillment of the Requirements for the Degree of

Bachelor of Arts

May 2014

ABSTRACT

With an increasing amount of text being shared on the web, through blogs, social media, websites pictures, it is becoming increasingly more difficult to translate the text in these mediums into geographic coordinates and physical locations. While geolocated-devices are becoming more popular, many people with mobile phones would prefer to not share their locations with applications and companies. Additionally, IP geolocation lacks the precision that GPS-enabled devices have. Yet, while internet users don't explicitly share their GPS-location, they often will share information about their location in the form of textual status updates. Using geotagged tweets and other geotagged information, it should be possible to identify similarities between non-geotagged text and classify someone's location by the words included in their tweet. With this information, more intelligence can be gathered about people tweeting, even if they haven't included their specific geographic coordinates with the tweet.

This thesis will focus specifically on classifying text to a variety of regions, including countries, states, counties and towns. It will use a variety of supervised learning classification techniques including SVM's and Naive Bayes. While the classifier is importance, the study will also focus on feature preprocessing as that will most likely have a great impact on the results of the final product.

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

CHAPTER 2

EXPLORING THE USES OF GEOTAGGED SOCIAL MEDIA

CHAPTER 3

DATA AND PREPROCESSING

3.1 Data

Data was collected from the Twitter API by querying their streaming API for all tweets with an attached pair of geographic coordinates or an attached geofenced region. Tweets with a region attached were assigned the midpoint of the region. A program, running on a personal computer would run for a period of time downloading new tweets during that period and storing them in a database.

The dataset is comprised of 1,656,146 tweets with 612,728 unique users tweeting within that period.

CHAPTER 4

METHODS OF CLASSIFICATION

CHAPTER 5

EXAMPLES/RESULTS

CHAPTER 6

CONCLUSION

BIBLIOGRAPHY

[1] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.

[2] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 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*, pages 759–768. ACM, 2010.

[3] Shinya Hiruta, Takuro Yonezawa, Marko Jurmu, and Hideyuki Tokuda. Detection, classification and visualization of place-triggered geotagged tweets. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 956–963. ACM, 2012.

[4] Sheila Kinsella, Vanessa Murdock, and Neil O'Hare. I'm eating a sandwich in glasgow: modeling locations with tweets. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 61–68. ACM, 2011.

[5] Yusuke Nakaji and Keiji Yanai. Visualization of real-world events with geotagged tweet photos. In *Multimedia and Expo Workshops (ICMEW), 2012 IEEE International Conference on*, pages 272–277. IEEE, 2012.

[6] Sharon Myrtle Paradesi. Geotagging tweets using their content. In *FLAIRS Conference*, 2011.

[7] Jay M Ponte and W Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281. ACM, 1998.

[8] Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. 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. Association for Computational Linguistics, 2012.

[9] Keiji Yanai. World seer: a realtime geo-tweet photo mapping system. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, page 65. ACM, 2012.