Twitter Hashtag Prediction

Mukund Manikarnike
ASU ID: 1208597425

Social Media Mining
CSE 472/598 – Phase – 2
Fall 2015

school of computing, informatics, & decision systems engineering

ASU IRA A. FULTON SCHOOLS OF engineering

Contents

Tables and Figures

Figures

Abstract

As of today, there is a lot of data in the form of blogs, posts, and microblogs from users from varied communities, locations and coming from different backgrounds. Given the varied background of people that use all of these platforms and create such varied data sources, there is often a need to retrieve this data. One such mechanism that Twitter (a microblogging site) uses is a concept called Hashtags which are an ever growing set of indices for retrieval. The way it varies from the typical indices is that these indices are ever-growing, user created and similar such hashtags could be named differently although in reality, they represent the same thing. The problem this project addresses is of achieving prediction of hashtags for a tweet that doesn't have any hashtags by learning from a tweet dataset that contains hashtags. The basic principle that this project explores is learning based on the similarity of words in the tweets.

Keywords

#Twitter #HashtagPrediction

Introduction

Goal Description

The goals that this project sets out to achieve are as follows

1. Creation of an algorithm that is able to learn about similarities between tweets in the dataset that is provided.
2. Creation of an algorithm that can predict new hashtags for the other tweets in the dataset.

Assumptions

The project makes a few assumptions which are listed below.

1. The project assumes that the dataset is a static dataset that doesn't grow with time as stated in the abstract.
2. The project doesn't use information about the user, location from where the tweet was sent out to perform the classification. It is purely based on the words in the tweets and their similarities.

Dataset Description

The dataset used for this project was provided as part of the course. It consists of the following information

1. User Graph
    a. This is a table of all the user IDs and the user IDs that they follow.
2. Tweet Data
    a. This is a table of all the user IDs, the 140 character tweet that they've sent out along with latitude and longitude from which the tweet was sent out.

Dataset Statistics

The dataset contains

1. About 60,000 unique users
2. 677,732 Tweets

This project used 1000 tweets as the training set and the remaining 1000 tweets as the test set. The similarity between users was also calculated only for the users who were involved in these 1000 tweets. The reason for this is stated in further sections.

Overview

The solution to the problem is described in the following few sections as described below

1. The Requirements Specification section describes detailed requirements that the project aims to meet.

2. The Design Choices section describes details of the algorithm chosen and the reasons for the same.
3. The Implementation section describes certain choices made during implementation and how each of the algorithms were implemented, issues that came up during the same.
4. The Results section describes the tests that were carried out using the implementation and the outputs that were obtained.
3. The Conclusion section describes how the approaches compare against existing approaches and future enhancements that can be made.

Requirements Specification

This section introduces each requirement that the project aims to achieve.

[REQ_1] It is required to use the dataset provided as a basis for any decisions that are made.

[REQ_2] All decisions that are made using the data will have to be carried out using social media mining techniques like similarity calculation, classification and so on that were discussed as part of the course.

[REQ_3] In order to carry out any of the algorithms designed for this project, none of the machine learning libraries shall be used. Libraries to carry out mathematical or linear algebraic tasks are allowed.

[REQ_4] The algorithms chosen shall use the dataset provided to make any decisions required and shall not make assumptions outside of the dataset about the general trend in twitter datasets.

[REQ_5] All ideas behind the design choices made shall be clearly stated along with results obtained for those design choices.

Design Choices

[DSN_CH_1] This design choices made for this project were mainly inspired by the paper referenced in [1]

[DSN_CH_2] The ideas that were referenced by the paper were

a. To treat tweets as a set of function which varies according to the words that are present in the tweet.
b. Usage of the natural language processing library NLTK to compute the similarity of words in order to find how the tweets rank against each other.
c. Prediction of tags using the dominant tag approach.

[DSN_CH_3] The design provides two kinds of similarity matching techniques

a. Similarity matching using word similarities in tweets
b. Similarity matching using user based similarities.

[DSN_CH_4] In order to compute user based similarities, cosine similarity was used.

[DSN_CH_5] The procedure chosen to perform matching using similarities in words is detailed in the next sub-section

Similarity Matching using NLP

[DSN_CH_6] The approach used here is to find similarities between nouns in each of the tweets.

[DSN_CH_7] Natural Language processing enables us to find similarities between any 2 words. By using the word similarity measure, the similarities between tweets is computed as a function of frequencies and semantic similarity as shown below.

$$I_{p,q} = S_{i,j} * (f_p - f_q)$$

$I_{p,q} \rightarrow importance\ of\ word\ p\ over\ q$

$S_{i,j} \rightarrow Semantic\ similarity\ obtained\ using\ NLP$

$f_p, f_q \rightarrow frequency\ of\ words\ p\ and\ q$

[DSN_CH_8] By using this metric, the importance of the noun is obtained. A summation of each such importance would give the rank of the noun with respect to the other nouns.

[DSN_CH_9] By using this metric, one would get the importance of each noun in the tweet. A summation of all the importance of each noun would give a measure of the importance of each tweet. The logic is that frequency and similarity put together would give an idea of what the tweets are talking about.

[DSN_CH_10] If the same procedure is carried out on a tweet for which hashtags need to be predicted, we would get an importance value for the tweet of interest as well.

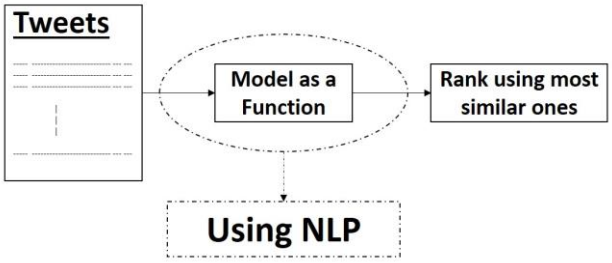


Figure 1 Similarity Matching - NLP

Prediction Using Dominant Tags

[DSN_CH_11] After the ranking of tweets is performed, the prediction is done using the 10 most similar tweets.

[DSN_CH_12] Out of the 10 most similar tweets, the average number of tags for all tweets is obtained. The final number of suggested tags would be as many as this number.

[DSN_CH_13] In order to actually predict the tags, the average number of dominant tags in the entire set of tags from the top 10 tweets in their vicinity. A summary of the design is as shown in **Figure 1**

Implementation

Overview

[IMP_CH_1] The implementation was carried out using Python 2.7. The libraries that were used were

a. NLTK
   i. To extract nouns from tweets and similarities between them.

Data Pre-processing

[IMP_CH_2] The given data contained tweet texts which contained “RT”, “@mentions”, special characters, numbers, URLs, #Tags and all of them were in different cases.

[IMP_CH_3] In order for the data to be easily processed by the NLP library, all of these special words had to be removed as part of pre-processing.

Finding Similarities

[IMP_CH_4] Similarities were found using the NLTK library by extracting nouns and computing semantic similarities and frequencies as highlighted in the design section.

[IMP_CH_5] Cosine Similarity between the users who were involved in the tweet dataset chosen was calculated using the graph of users provided as an adjacency list.

Prediction using Dominant tags

[IMP_CH_6] This was done by finding the most dominant tags according to tweet similarity as well as user-cosine similarity as highlighted in the design section.

Results

This section will briefly talk about the output file formats generated by this project and the results that were seen for each step.

There were several intermediate files that were generated and have been submitted as part of the package which will be talked about in the readme and not here.

Data Pre-processing

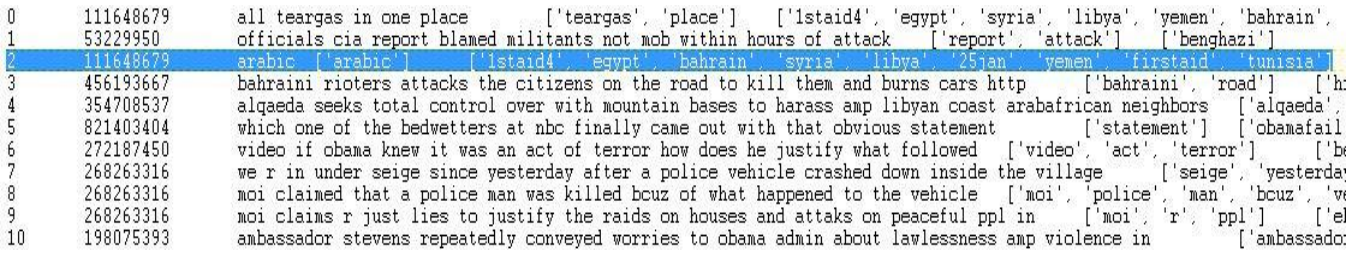
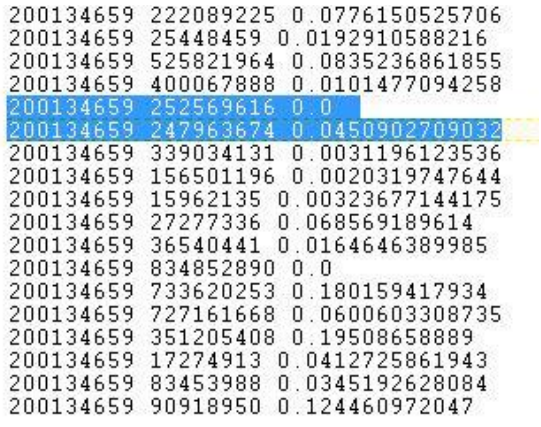


Figure 2 Pre-processed Data

As indicated in ***Figure 2***, the pre-processed data contains the Tweet ID, User ID, Tweets stripped of the data that wad mentioned in the design section, nouns in the tweet, hashtags in the tweet. The tweet ID is an ID that was assigned by the pre-processing step for easy identification.

As shown in ***Figure 3***, the similarities between each user ID involved in the Tweets that are being processed is provided in cosine Similarity. As part of the Similarity using NLP, the importance of each of the words as rated according to the other words in the whole dataset is shown.

Cosine Similarity



Similarity Using NLP

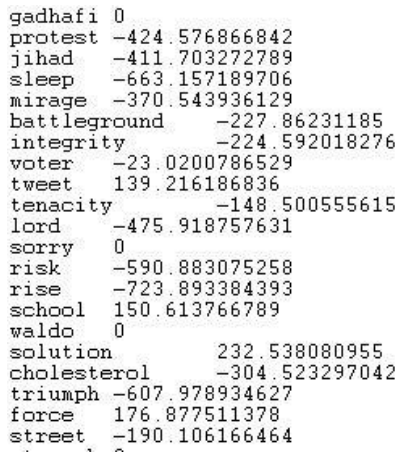


Figure 3 Similarity Matching

Prediction using NLP

Figure 4 Predicted Outputs

As shown in ***Figure 4***, the predicted tags and their actual counterparts are indicated for 4 different input tweet IDs. A detailed analysis of the accuracy of the results aren't available. But, a theoretical analysis is provided as part of the conclusion.

Conclusion

This section talks about the analysis of the model that was used, the things that were learned as part of this project and further enhancements that could be made.

Theoretical Analysis

Prediction using Cosine Similarity

This method considers the neighborhood of the users and how similar they are. Using such a measure for prediction of tweets wouldn't be a very good measure because of the following reason reasons

Similar neighborhoods don't indicate similar users because different kinds of people could have similar neighborhoods around them.

However if they were similar, given that the dataset is centered around a particular topic, i.e. the tweets are all talking about something very specific, it might be very wise to suggest that similar users following similar people might want to tweet about a particular thing.

Prediction using NLP

Given that we've just established that the data in the tweets is centered on a topic, this is a very good mechanism because it factors in the frequencies of words and the semantic similarities between them to establish similarity between tweets.

However, if the dataset were to have material covering different topics, it would be better to combine the User similarity and NLP technique to predict suitable hashtags.

The Learning Curve

The learning curve for this project was very steep since it involved

1. Understanding usage of Natural Language Processing libraries
2. Reading up on existing techniques to perform Hashtag prediction
3. Carrying out complex algorithms on large datasets and ensuring that they run fast.

Future Enhancements

As on 5th of December, 2015, the following things weren't implemented which were originally part of the plan.

1. Prediction Using Cosine Similarity
    a. Cosine similarity was calculated. The last part where matching using the similarities has to be done wasn't carried out.
2. Testing the existing implementation on the entire dataset.

As on 5th of December, 2015, the following things could be done to make the project better

1. Changing the Cosine Similarity to a better measure giving a better user similarity metric
2. Correlating the 2 similarities to factor both in.
3. Enhancing the speed of the algorithm by using a distributed setup like Hadoop.

Notes to the Grader

1. Since no participation from both of the other group members (Harshil Shah and Diego Henrique Ferreira), their names have been omitted from this submission.
2. Due to this, several things that were planned weren't completed. However, 90% of the project has been implemented and tested on a small dataset.

Bibliography

[1]	Tianxi Li, Yu Wu, Department of Computer Science Stanford University and Yu Zhang, Department of Computer Science Trinity University , *Twitter Hash Tag Prediction Algorithm*
[2]	Python NLTK Documentation, *http://www.nltk.org/index.html*