Document Classification of Tweets Using Naïve Bayes

Reotutar, Christian
creotut1, creotut1@jhu.edu

Peven, Michael
mpeven1, mpeven1@jhu.edu

1 Abstract

Summarizing chat is a notoriously hard problem to solve because of the inherent noise from slang, lack of sentence structure, and incorrect grammar when humans chat with one another. However, we believe that there are still ways to extract meaningful information if there is a massive dataset, no matter how unstructured, because the construction of twitter data is pre-classified and we can run a supervised learning algorithm over the data.

In this project, our aim is to classify lines of human chat by topic. We will be implementing a machine learning algorithm to run over twitter data to learn a model for multi-class classification.

2 Methods

Twitter is a social networking application where users can send 140 character tweets comprised of text and hashtags. Here is a (handcrafted) example of a tweet: ***Jimmy and I luv munchin on Dominos #pizza***

The goal of this project is to learn hashtags as classifications for tweet message body in a document classification model. Our input to learn the model is a word label and then the words of the tweet body as its features. The input for testing is just the tweet body words. The output will be a classification of the words as one word.

While a body of a tweet may not make sense, we are learning the collection of words and their probabilities that lead to the hashtags. We make the assumption that the order of the words does not matter but rather only the content. We believe this will correlate well to a chat message topic classifier because the message content and styles are similar.

We can measure success directly by looking at the accuracy of hashtag prediction. If we can guess a hashtag for a given tweet even a small percentage of the time (5-10

3 Resources

We will use the Twitter Streaming API (`https://dev.twitter.com/streaming/firehose`). The API allows us to access a sample of a live tweets with information on tweet messages, the

users who posted them, and most importantly, the hashtags the users attached to the tweet message. The data is a random sample of all tweets on the website. We will also be using Google's Word2Vec library to calculate distancs between words.

We have done all the necessary work to access this API and were able to collect about 2,000 tweets in a relatively short amount of time. We will be collecting more to have as large of a dataset as needed.

4 Milestones

4.1 Must achieve

Implement a naïve bayes binary classifier for certain hand picked topics (such as pizza), as well as a multiclass naïve bayes classifier for any classes.

4.2 Expected to achieve

We are expected to achieve another type of classifier (like a Support Vector Machine or Linear Regression). Then, we would like to compare the accuracies, running times, and output of these other models against our base model (Naïve Bayes). We expect a tweet message summarizer that extends well into the chat message domain (well meaning the summaries are at least 75% accurate).

4.3 Would like to achieve

We would like to achieve a very accurate text summarizing model that can be used with everyday chat messages for input. We would also like to achieve a comparison between 3 different models to predict this chat message summarization.

5 Final Writeup

In the final report, we will be delivering an analysis on the machine learning algorithms that were implemented and their relative successes and accuracy.

We will try to compare our own algorithm against other popular chat and document summarization algorithms that are out there. We will also be spending time trying to test and visualize the results and see if they can be applied in a real world context.

6 Bibliography

Vishwakarma et al., "Text Stream Classification Techniques And Research Issues: A Review"

Ouyang et al., "Applying regression models to query-focused multi-document summarization"

Li et al, "Twitter Hash Tag Prediction Algorithm"