Document Classification of Tweets Using Naive Bayes

Christian Reotutar (creotut1, creotut1@jhu.edu), Michael Peven (mpeven???, mpeven@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

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
```

As you can see, the body of the tweet may not make much sense; however, it is inherently classified by construction. Our goal is to use the hashtag (# pizza) as the classifier representing the topic of the sentence.

3 Resources

We will use the Twitter Streaming API in order to access large amount of tweets. The API allows us to access a sample of a live tweets from the popular social media website 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.

4 Milestones

4.1 Must achieve

Implement a nave bayes binary classifier for certain hand picked topics (such as pizza), as well as a multiclass nave bayes classifier for any classes (and possibly multiple classes per line of text).

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 (Nave Bayes). We expect a tweet message summarizer that extends well into the chat message domain (well meaning the summaries are at least 75

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