ECS 289 Scalable Machine Learning

Vivek Dubey
Yagnik Suchak

10 November 2015

1 Overview

The main aim of this project is to mine the data from Twitter feed and classify the tweets into one of the following categories:

1. Business
2. Education
3. Entertainment
4. Technology
5. Environment

We would also like to use the geo-location information contained in each tweet to visualize where each category of tweet is produced across the world.

2 Related work

1. Luciano Barbosa and Junlan Feng. 2010. Robust Sentiment Detection on Twitter from Biased and Noisy Data. Coling 2010.
2. Christopher Burges. 1998. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, 2(2):121-167.
3. S. Kumar, G. Barbier, M. A. Abbasi, and H. Liu. Tweettracker: An analysis tool for humanitarian and disaster relief. In Proceedings of the 2011 International Conference on Weblogs and Social Media, 2011.
4. R. Li, K. H. Lei, R. Khadiwala, and K.-C. Chang. TEDAS: a twitter-based event detection and analysis system. In Proceedings of the 2012 IEEE International Conference on Data Engineering (ICDE), pages 12731276. IEEE, 2012.

3 Idea and Approach

1. Gathering Data : We will use the twitter API to interact with their service. There is also a bunch of Python based clients (like Tweepy) that we can use without reinventing the wheel. The twitter streaming API outputs the tweets in JSON format, which can be arbitrarily complex. At this point, depending upon the size and complexity of data that we gather, we may/may not use the Hadoop Ecosystem.
2. Pre-process and tokenize the tweet text.
3. Process further to generate the tf-idf weights for the bag-of-words type approach for text classification.
4. We would be using a Naive Bayes Multinomial Classifier.
5. To visualize the data, we aim to work with and modify a platform to analyze tweets like the TweetTracker.

4 Experiments

We would use the following for evaluation purposes :

1. Split our total data into training and test set, the ratio of which will be decided when we gather the data.
2. Use a 10-fold cross validation to evaluate the performance.
3. Manually annotate a small section of the dataset in order to provide a rough estimate to the test set performance.