Data Selection Proposal

Twitter Sentiment Analyzer

Cindy Wang
https://github.com/cindywang3299/sentiment_analyzer

1 Objective

Twitter Sentiment Analyzer will "computationally" determine whether a piece of writing is positive, negative or neutral. It analyzes topics by parsing the tweets fetched from Twitter using Python.

2 Methodology

This application will use a Python-based client called *Tweepy* [1] for data access in general and *TextBlob* [2] for processing textual data.

Datasets used for this application will be collected from Twitter's API.

1. Data Preprocessing
   The objective of this step is to clean noise those are less relevant to find the sentiment of tweets such as punctuation, special characters, numbers, and terms which don't carry much weight in context to the text.

   Tweets will be preprocessed in the following ways:

   - Removing Twitter handles (@user)
   - Removing punctuations, numbers, and special characters
   - Removing short words
   - Tokenization
   - Stemming

2. Machine Learning Model
   logistic regression will be used to build the model for this application. It predicts the probability of occurrence of an event by fitting data to a logit function. Predictive models on the dataset will use two feature set: Bag-of-Words and TF-IDF.

3. Final conceptualization
   Ideally, the model will be working well enough for a live demo. The program will take audiences' prompts and make real-time topic analysis. If not, the project will be only demonstrated for conceptualization purposes.

4. Application
   Ideally, the project will be available as a webapp. The user will be able to see the sentiment analysis on Tweets divided by topics they wish to investigate.