A-News-Ment Project Proposal (formerly spot-a-bot)
CS122-Win-17-resrar-ayaspan-jarroyom
January 24, 2017

Abstract:
Public engagement can take many forms. For every event that takes place, positive or negative sentiment is expressed in a myriad of ways and from a number of different sources, some purporting to be unbiased and others intentionally opinionated. News outlets produce articles that serve one or the other purpose. Yet, even in unbiased reporting, one can find words that express the sentiment of a writer or someone that is being interviewed for a story. We intend to create a web application that scrapes particular news agencies' websites for articles based on a user-given keyword or phrase and analyzes the type of sentiment expressed in regards to the keyword or phrase. Our application will operate in both English and Spanish (using sites from the United States and Mexico, respectively).

Goal:
The aim of our project is to develop a sentiment measure that allows us to classify writer sentiment accurately at least 60% of the time. In order to achieve this goal, our team will implement various methods to gather, clean, and analyze news article data. In particular, our team will perform this sentiment analysis on article web sites both from Mexico and the United States.

Data Sources:
Mexican newspaper sites that allow web scraping:

- http://www.jornada.unam.mx/ultimas
- http://www.milenio.com/
- http://www.reforma.com/

US newspaper sites that allow web scraping:

- www.usatoday.com/
- propublica.org/

News API

Deliverable Description:
A web application that will take user input, parse the input for keywords, and then return a sentiment analysis for articles about applicable keywords. Using beautiful soup and the request library, we plan to scrape text from given article. The sentiment analysis will be conducted by

referencing news article text to standard lists of positive and negative words.[1] We have found such a list for English but not yet for Spanish. If we are able to scrape and process the data in a timely manner, we would like to extend our data analysis by correlating news sentiment to historical financial indicators, specifically the USD/Peso exchange rate.

Timeline:

1. Data Collection (week 5)
    a. Scrape news websites
    b. News API to gather additional news
    c. Yahoo API to gather information on financial indicators
2. Data Cleaning (week 7)
3. Sentiment index development (week 8)
    a. Scrape text of newspaper sites and count presence of words in the positive and negative word lists
    b. Construct positive and negative word lists for Spanish if necessary and repeat step one for Mexican newspapers
4. User Interface (week 9)
5. Tentative correlations with financial indicators

[1] Minqing Hu and Bing Liu. "Mining and Summarizing Customer Reviews."Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), Aug 22-25, 2004, Seattle, Washington, USA