HW 5:Text Mining

Deniz Celik & Ryan Louie

March 2014

Abstract

A MediaWiki API based web crawler was used to grab the revision history of Wikipedia article which was placed through the sentiment analysis package of pattern.en in order to find polarity and subjectivity of each article as a function of time. The results for the top 20 most visited articles between February 22^{nd} and March 1^{st} were graphed using the built in python plotting library.

1 Project Overview

Our goal was to use sentiment analysis and web crawling to analyze Wikipedia pages. We used a MediaWiki API based web crawler written by Brian Keegan and Nick Bennett found on the ipython site, instead of the pattern tools. This was done since pattern does not support crawling of Wikipedia revision history. However in order to do sentiment analysis, the pattern.en module was employed. The results for each article are then plotted for polarity and subjectivity and article length vs time, polarity vs length of article and subjectivity vs length of article.

2 Implementation

Our code base relies on three scripts: 1) revision.py contains the functions that allows web scraping of Wikipedia revision history; 2) explore.py imports revision, uses it to scrape revisions of a partiucular Wikipedia article, and applies sentiment analysis for each of the articles; 3) visualize.py imports explore and visualizes how polarity, subjectivity, and article length vary with time.

Running the Code From the hw5 folder at the terminal, type ¿¿¿ python visualize.py. If you want to visualize for a different wikipedia article, open visualize.py, go to the "ifmain" statement at the bottom, and change the argument of SentimentTimePlot to another Wikipedia Article Title (i.e "Franklin W. Olin College of Engineering").

Data Structures Revisions.py uses a dictionary to store the revision history. It is keyed by timestamps, so that we could sort the keys in chronological order for our time series visualization.

Explore.py has its primary function sentimentreturn which returns a list of nested tuples, which story the values for sentiment, time, and length of articles. This is easy to iterate through, and we wanted to maintain the ordering of our chrnological sort.

3 Results

One result that is interesting is about the Pornography article, which appeared as one of the top 12 most recently viewed articles on Wikipedia. We found that polarity began at a negative value, but over revision history, increased and slowly leveled off at a mildly positive value. This description matches several intutions, namely that the first revisions of the Pornography started with a negative tone, but as a result of the "Wikipedia Authorship Neutrality" and maybe general comfortability with the topic over the decade, a more neutral sentiment is reached. Another result with interesting movement is the Deaths in 2014, which has large drops in article when each month is moved to a new page while the polarity goes to near zero values.

The views for every article are the number of non-unique page calls from 11:59PM on February 22^{nd}, 2014 to 11:59Pm on March 1^{st}, 2014. The list was created by user West.andrew.g and can be found here.

3.1 Olin College

Our testing page for our code, it is a low revision, 527, short article with differing sentiment over time.

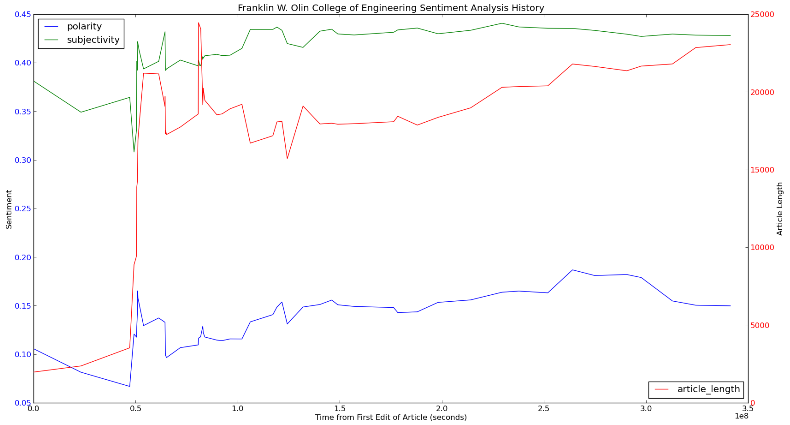

Figure 1: Franklin W. Olin College of Engineering Sentiment History

3.2 Main Page

The most visited Wikipedia page with 96,229,028 views in a single week.

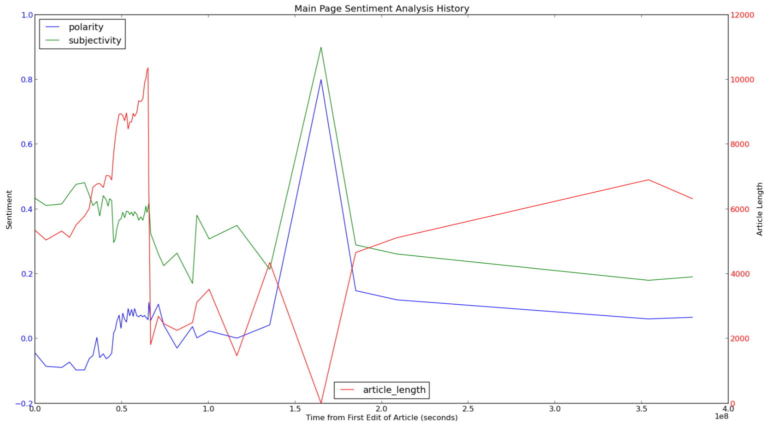

Figure 2: Main Page (Wikipedia Landing Page) Sentiment History

3.3 Harold Ramis

The second most visited Wikipedia page with 1,158,070 views in a single week.

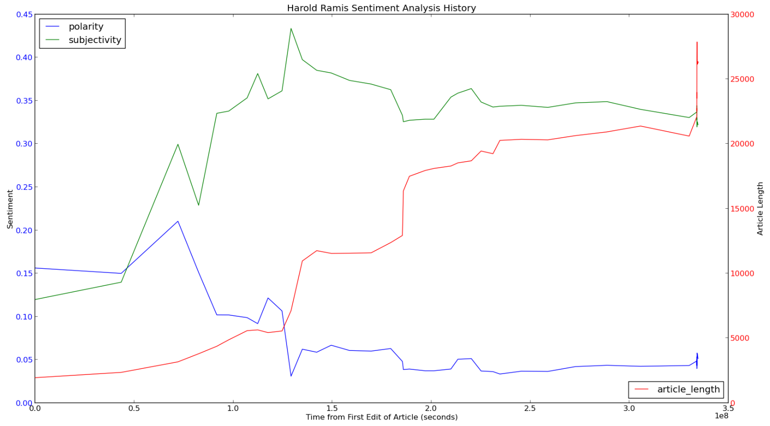

Figure 3: Harold Ramis Sentiment History

3.4 Java

The third most visited Wikipedia page with 1,084,764 views in a single week.

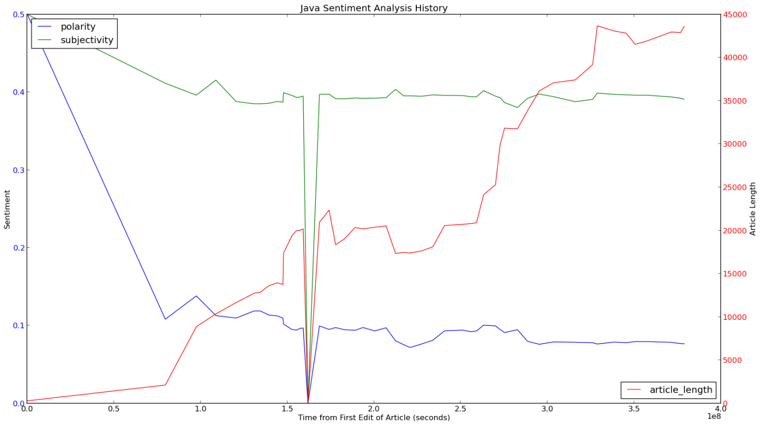

Figure 4: Java (indonesian island) Sentiment History

3.5 Climatic Research Unit email controversy

The fourth most visited Wikipedia page with 763,640 views in a single week.

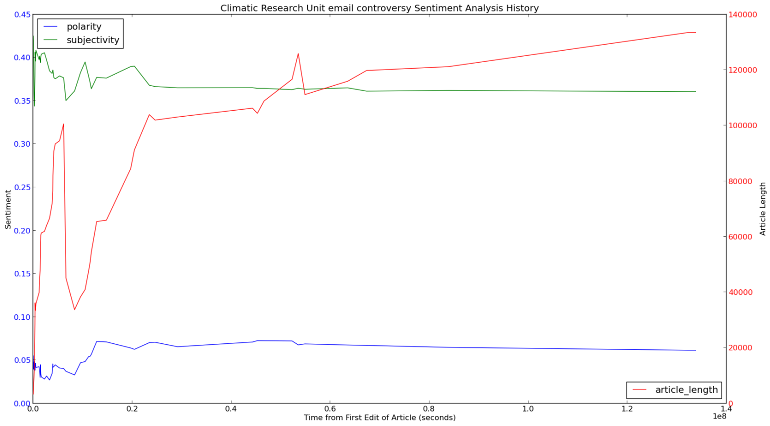

Figure 5: Climatic Research Unit email controversy Sentiment History

3.6 Ukraine

The fifth most visited Wikipedia page with 700,513 views in a single week.

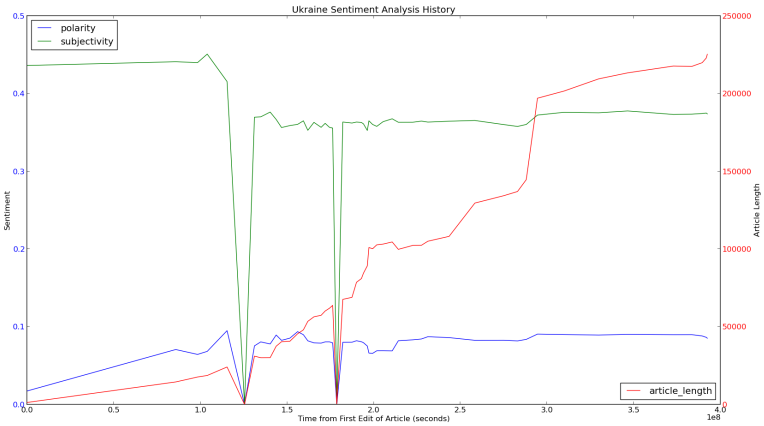

Figure 6: Ukraine Sentiment History

3.7 True Detective (TV Series)

The sixth most visited Wikipedia page with 633,432 views in a single week.

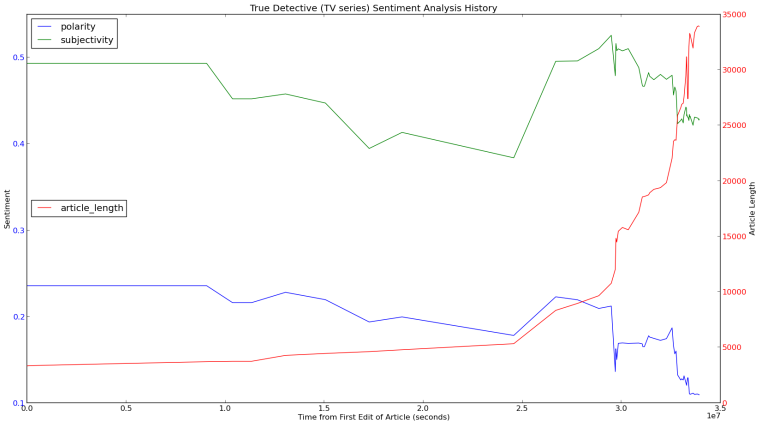

Figure 7: True Detective (TV Series) Sentiment History

3.8 Crimea

The seventh most visited Wikipedia page with 606,887 views in a single week.

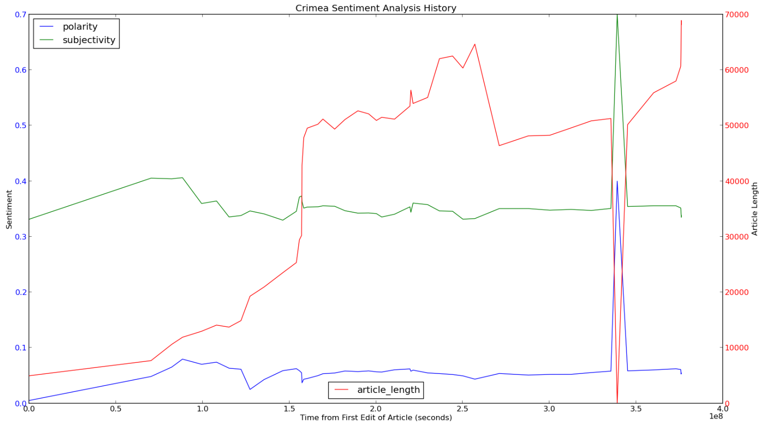

Figure 8: Crimea Sentiment History

3.9 IPv6

The eighth most visited Wikipedia page with 541,223 views in a single week.

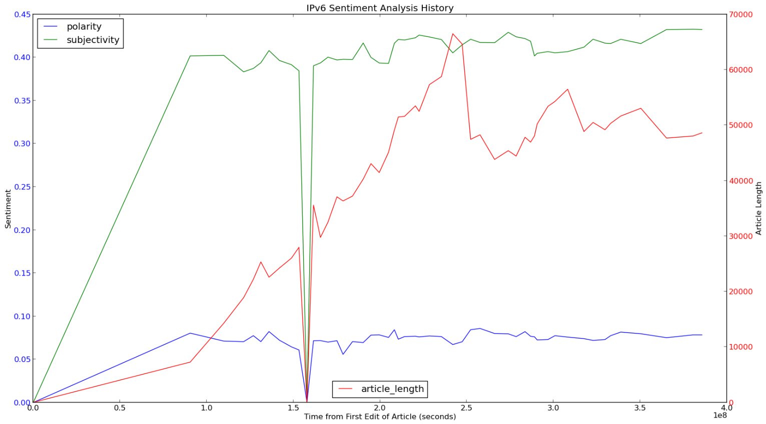

Figure 9: IPv6 Sentiment History

3.10 Pornography

The ninth most visited Wikipedia page with 528,194 views in a single week.

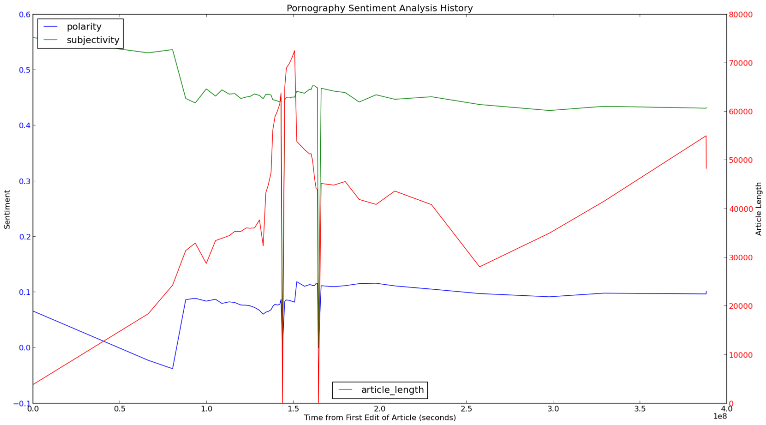

Figure 10: Pornography Sentiment History

3.11 Internet

The tenth most visited Wikipedia page with 506,413 views in a single week.

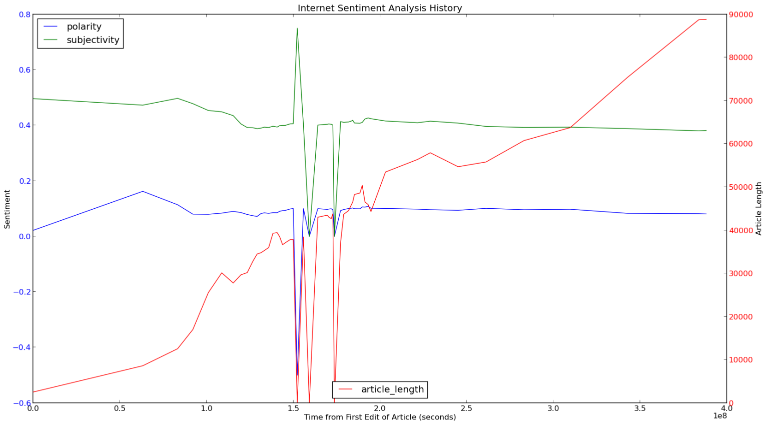

Figure 11: Internet Sentiment History

3.12 Facebook

The eleventh most visited Wikipedia page with 482,478 views in a single week.

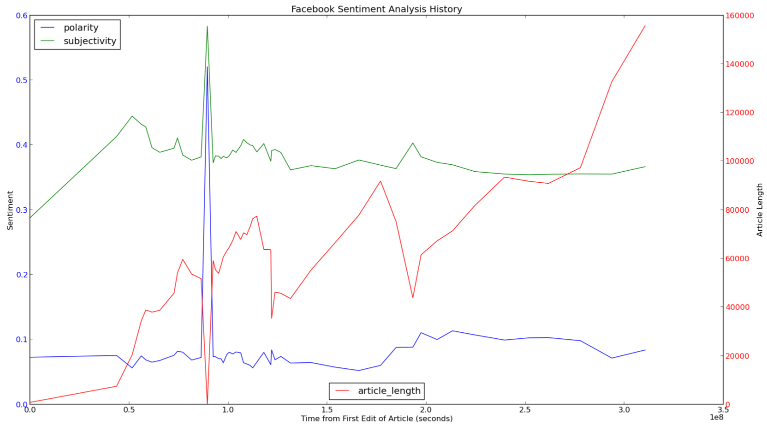

Figure 12: Facebook Sentiment History

3.13 Deaths in 2014

The twelfth most visited Wikipedia page with 460,121 views in a single week.

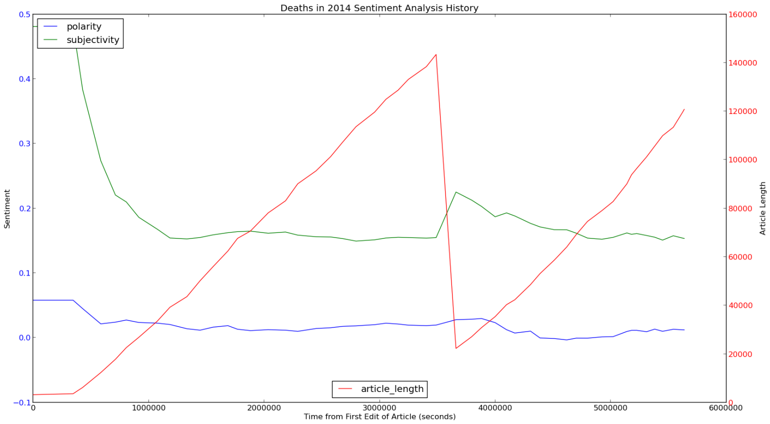

Figure 13: Deaths in 2014 Sentiment History

4 Reflection

4.1 Things that Went Well

- Maintained a facination and enthusiasm for what we were doing
- Used a project organizer (Trello) to propopse task and carry them to completion
- Stayed strong to our early dreams for the project. Did not back off just because tools like Pattern did not have the neccessariy functionality. We were resourceful and altered open source functions to aid our pursuits.

4.2 Things that could have been improved

- Prioritizing time – project felt like a feverish hackathon at the very end of deadline.
- No formal unit testing functions were used.
- More time spent hacking away at making code work. Ignored edge cases (for example no intuition why the deep downward fractures in graph occurr).