

Alexander Swanson

January 31, 2018

Professor Lisa Dion

CS 124 – Data Structures and Algorithms

Project 1

Project 1 Report

For over seven months now, I have been working on a project fundamentally aimed at evaluating and influencing online public opinion. There are many platforms to work on this problem – *Facebook*, *Twitter*, *Instagram*, etc. However, attempting to access all of these platforms immediately is not a practical approach. Thus, working with the social, information aggregation forum *Reddit*, a large amount of data has been collected to begin evaluating public opinion. A portion of this data has been selected for use with this project in order to complete fundamental analysis.

The data represents basic metadata from Reddit submission, their comments, and the comments' authors. Using a Python wrapper for the Reddit API, the data has been in constant collection since December 2017. It is important to note that the original datasets contained the comment text body for each entry, but this has been omitted due to the vulgarity expressed in some of these comments. This particular portion of the data was collected in the beginning of January 2018. The pieces of data that were included for this set are as follows:

Index

The index of the row entry.

ID

The unique ID of the Reddit *comment* object.

Parent ID

The ID of the comment the entry was a response to.

Submission ID

The unique ID of the Reddit *submission* the entry was a response to.

Subreddit Name: Prefixed

The common identifier of a categorical forum on Reddit.

Ups

The comment's total resulting *upvotes*.

Downs

The comment's total resulting *downvotes*.

Score

The overall score, a function of the comment's upvotes and downvotes.

Controversiality

A measure of how controversial a comment is, a metric which is a function of the number of up- and downvotes the comment receives, and how evenly distributed they are.

Category

The general topic category of the comment's text body.

Sentiment Score

The score of the sentiment, rational to the sentiment magnitude. That is, the type of opinion expressed in the comment text body.

Sentiment Magnitude

The magnitude of the sentiment, rational to the sentiment score. That is, the weight of the opinion expressed in the comment text body.

Created

The UTC unit measure of the date the comment was created.

Date Created

The *Month/Day/Year* date the comment was created.

Time Created

The time the comment was created.

There are some important factors to note... The up- and downvotes produce the comment's score. This figure is simply the difference of the upvotes and the downvotes. Moreover, it is simple to understand the significance of sentiment score and magnitude: the magnitude represents the total weight of a sample of text's expressed opinion. Thus, if, for example, a comment has a sentiment magnitude of, say, 4.0 and a sentiment score of -1.0, then the comment was clearly a negative remark. Considering the same example sentiment magnitude, if the score is 4.0 as well, then such a comment is considered to be an overwhelmingly positive remark.