CSE 140 Assignment 9 - Reddit Username Analysis

By Wesley Wolanski and Jeff Gevaert

Research Questions:

1. Is there any correlation between reddit usernames and voting patterns?

Given reddit usernames, we are trying to compute statistics that represent voting patterns. Specifically, we are looking for voting patterns that are based on words that are present in different usernames. This information can help us Identify a correlation between user types, or personalities, and the type of language they use.

- Result: According to our data, there is no clear correlation between inappropriate usernames and voting trends. With a larger dataset, it could be more evident.

2. If there is any correlation, can different categories or types of language affect user data differently?

The point of checking different categories of language (inappropriate language, for example), is that we can then more easily check for correlations. For example, if all usernames that use inappropriate language vote, on average, lower, we might be able to say that users who would use inappropriate language are more critical of content.

-Result: The data we processed shows no clear relationship between the trends, and thus the null hypothesis is the most likely based off of our small dataset.

Motivation and Background:

Society has always had inappropriate language. Using reddit, which allows this language we can see if there is a difference between obscene or "socially acceptable" language. This is also allowed in a person's alias, which makes it easier to see if there is an effect on the quality of information provided. Many websites filter out names and words, however this could show that using these words could have little or no effect on the quality of links provided.

By showing that there is a correlation between the type of words used, we can see how society as a whole is right or wrong to condemn this language as a sign of ignorance or otherwise.

Dataset:

Reddit is a website where users can post links in different categorized boards, and other users can vote on the links depending on what they think of them. Users can also vote on other users comments that are posted to links.

The dataset that we are using is a large CSV (Comma-separated values) file containing user voting data from the website 'www.reddit.com'.

- The format it is presented in is '[username], [link ID], [Their vote (+1/-1)]'.
- This file contains data for a total of 31,927 usernames.
- The total number of votes that are represented by all of the users is 7,405,561.

The other dataset that is used is a small amount of text files containing inappropriate words and information about inappropriate words. Most of these inappropriate words are sourced from banned word lists of assorted websites, articles on inappropriate words, and studies involving them. They come in a variety of formats, but the most common is a text file with an individual word on each line.

- To access the Reddit voting data:

1. Follow the link http://redditketralnis.s3.amazonaws.com/publicvotes.csv.gz?torrent in order to download the torrent file.
2. Use your torrent client of choice to open the torrent file. Choose to save to your directory of choice.
3. Use an unpacking program (7zip, available at: http://www.7-zip.org/download.html) to open the downloaded file, and extract the file 'publicvotes.csv' to the program directory

- To access the inappropriate word data:

1. Download the .7z file available at the following link (the word_list_dirty_words.7z uploaded alongside this document on the catalyst dropbox. https://catalyst.uw.edu/collectit/file/mernst/24992/103969/455546/0837b2e3df4d78ae722c3baf7e9f25d80d92a9dffbc064c890f37ce55675ec00
2. Use an unpacking program (7zip, available at: http://www.7-zip.org/download.html) to open the downloaded file, and extract the folder 'word_list_dirty_words' to the program directory.

Methodology:

Here are the steps we plan to use to get the desired results.

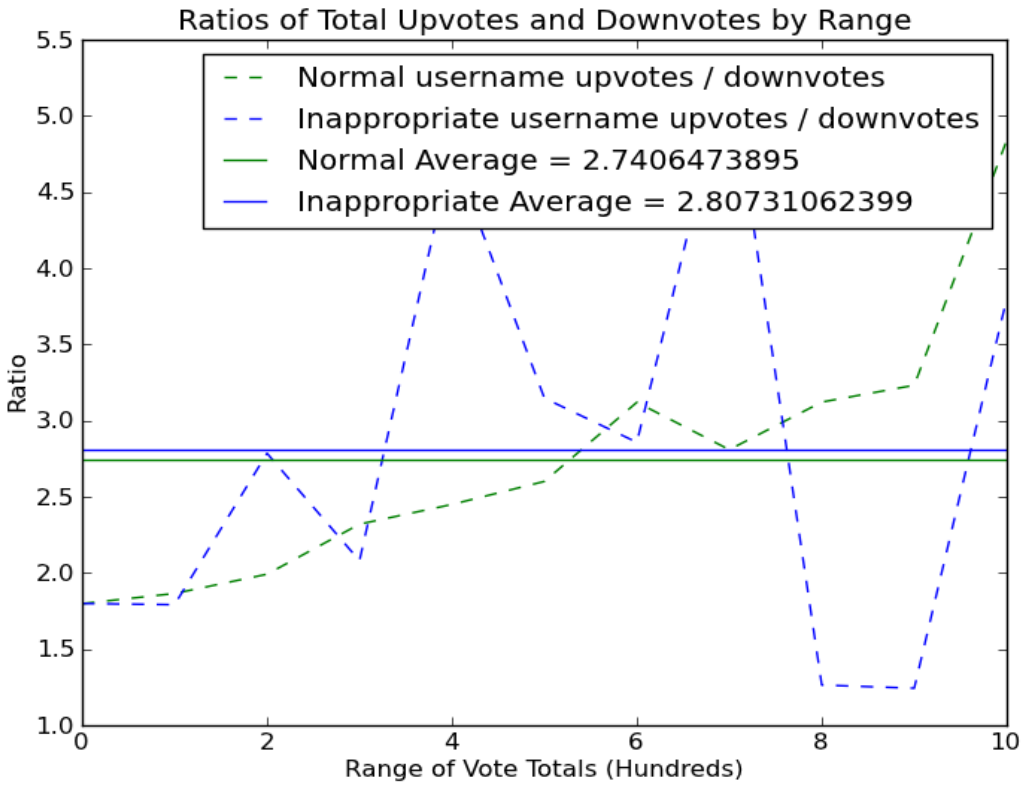
1. The file containing reddit voting data first needs to be turned to a form that is usable for analysis, particularly mapping each username to its total upvotes and downvotes.
2. The inappropriate words files then need to be read and the group of words we are looking for is then put together (in a set) to be more easily analyzed.
3. Find the overall average of all of the +1 and -1 votes and ratios between each of all of the usernames.
4. Use the "inappropriate words" set to analyze each name in the reddit data, and sort them into two lists.
5. Using each list, create a set of the values from these votes.
6. Compare the average of the upvotes and the downvotes between the inappropriate usernames and the appropriate usernames.
    a. This will allow us to perform statistical analysis to examine the correlation.
1. The data for each is then made into a graph for easy analysis.

Results and Answers to questions:

Due to the relatively small size of the data, it is difficult to obtain a true result. Even with

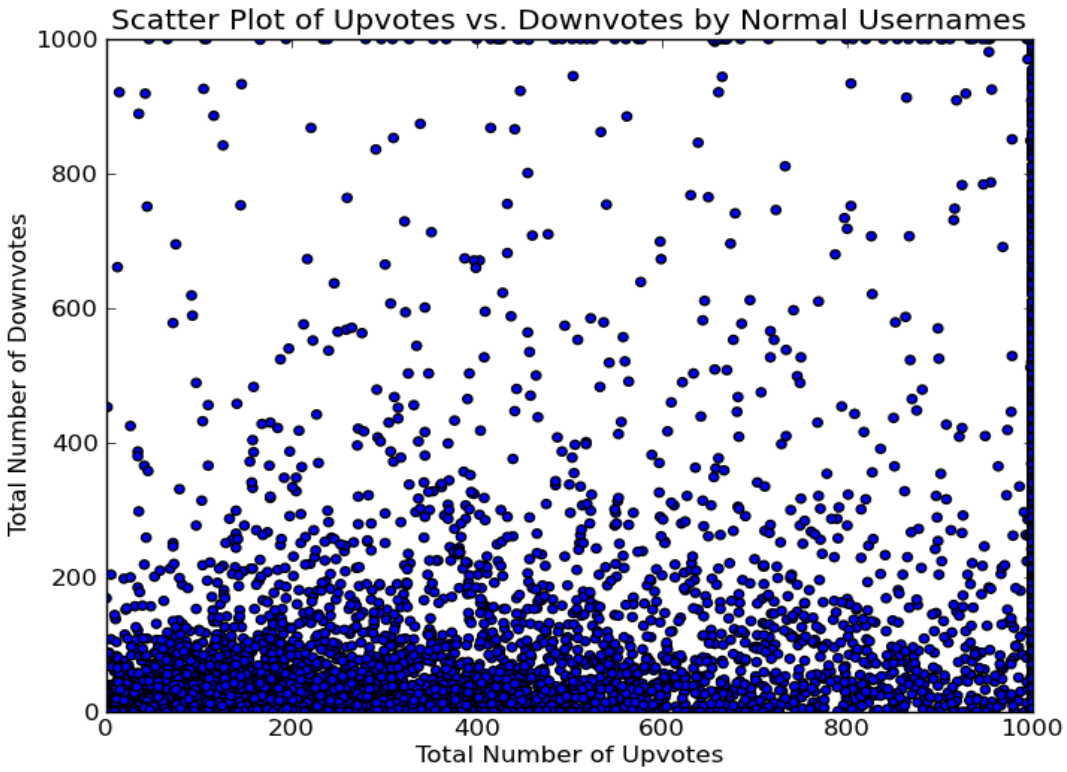
over 30,000 names, we could get much more accurate results with a larger amount of data. Our results show that as users vote more often, they tend to vote positively in a higher proportion. Both datasets follow this trend, but the inappropriate data is more erratic because of limited data points.

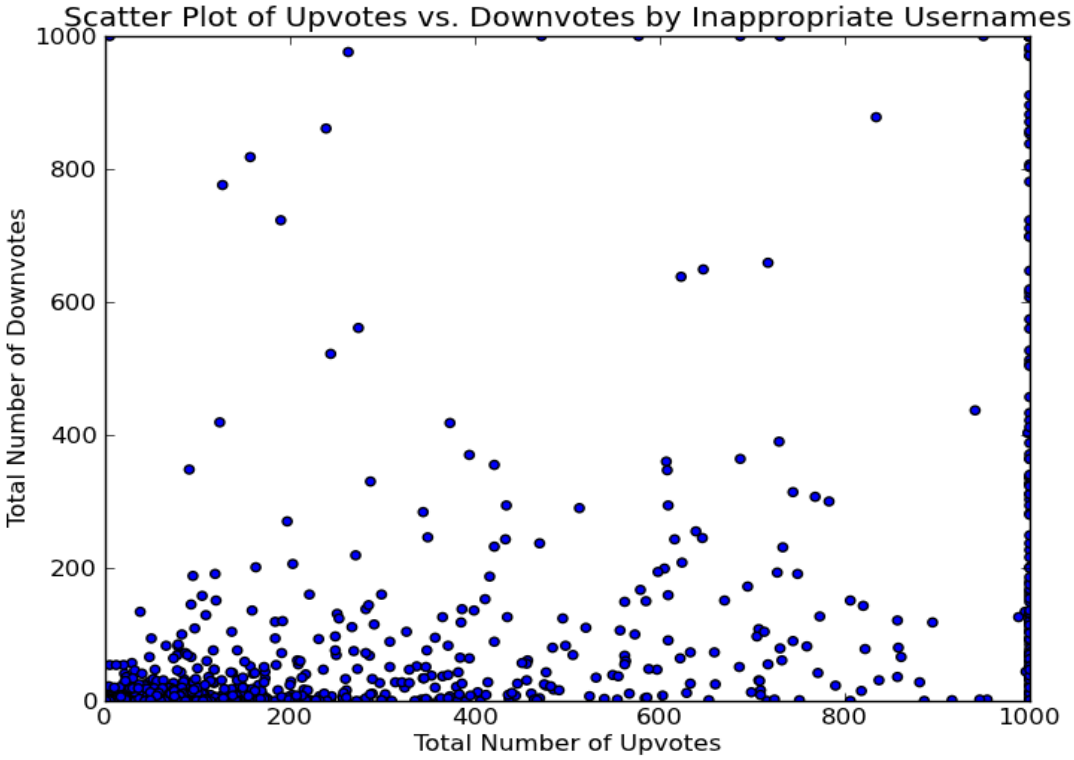
The following is a ratio plot, which displays the ratio between total upvotes and downvotes in ranges of 100. (e.g. 800 upvotes would be in the 8 range). This shows voting trends that are present with both datasets.



As you can see, the ratio between the averages over the entire vote range is higher with inappropriate language than without. However, the difference is relatively insignificant as the difference is less than 0.1.
A ratio of 1.8 means there are 1.8 upvotes for every downvotes, and so on.

The next plots are scatter plots that show two sets of data, one of which is the average +- votes for users with improper words, the other without.


Scatter Plot of Upvotes vs. Downvotes by Normal Usernames
Total Number of Downvotes
Total Number of Upvotes
0
200
400
600
800
1000


Scatter Plot of Upvotes vs. Downvotes by Inappropriate Usernames
Total Number of Downvotes
Total Number of Upvotes
0
200
400
600
800
1000

Summary

Overall, in the scatter plots, the normal username and inappropriate username data follows the same trend, with a higher proportion of positive votes. The amount of data points decrease as the number of votes increas on both, and the amount of downvotes decrease at a much higher rate. Because the difference is so small, we can be relatively confident there is no correlation between types of usernames and voting data. With a larger dataset, we could be more confident.

Reproducing results:

In order to reproduce our results, the file 'reddit_username_analysis.py' needs to be saved to a directory. Following the instructions in the 'Dataset" section, the reddit voting data needs to be downloaded and 'publicvotes.csv' needs to be extracted to the same directory. Finally the inappropriate word data needs to be downloaded from catalyst dropbox with this file. The folder 'word_list_dirty_words' needs to be extracted to the same directory.

In order to run the program, open the command prompt or terminal, then navigate the the directory that 'reddit_username_analysis.py' is in, then type:
python reddit_username_analysis.py

The program will then produce and save images of the plots (.png format) in the same directory.

Collaboration:

None

Reflection:

We learned that there are many things in a project that might not be easily considered, or acknowledged. Upon analysis of the different aspects of the project, we gained a much deeper understanding of how we will accomplish our goals.

In part 2, there were many algorithms we had to use to filter out the extraneous data from the inappropriate words. As there are .txt and .yaml files which came in different styles, this made it a bit more difficult to process. Ideas such as making all letters lowercase in both the names and in the words increased the amount of words able to be used, and thus able to provide a much better data analysis.

Our obscene word set could have also had better values, as some of the words had double meanings that are unknown to most people. For example, "ass" is a bad word, whereas assassin is not. There are many more examples of this, however deciding what and what isn't a proper obscene word is difficult and depends on the reviewer. Naturally, because of this, there is going to be a larger amount of error. In the future, we would find ways to reduce error and check words with more consistency.