Sungwoo Park
Subreddit Sentiment Analysis

Project Overview

The purpose of this project is to get a rough ballpark estimation of how positive or negative different subreddits are relative to each other. To accomplish this, I fetched a list of the post titles from a number of subreddits and ran sentiment analysis in Pattern library on those titles to get an estimation of the positivity of subreddits. I was hoping to see if there is a quantitative way of determining whether one subreddit is more positive/negative compared to others.

Implementations

The program uses PRAW (Python Reddit API wrapper) package to fetch data from Reddit that I would be using to run sentiment analysis on. The program would take in a list of subreddit names in strings and PRAW package was used to get the titles of top [n] current “hot” posts from requested subreddits. N is hardcoded to be 100 in my current iteration of the program, because I thought 100 posts would provide good sample size to run a sentiment analysis on. However, the program can be easily modified to take in variable n number of “hot posts”

After using PRAW to fetch the data of first 100 current “hot posts” of a subreddit, the program loops through the list to generate a list of only the title of the posts. With the list of all fetched titles, Pattern API’s natural language processing function (more specifically the sentiment analysis function) is used to calculate the sentiment value of each title. Lastly, the program calculates the average of all sentiment values and returns the average as the overall sentiment value of a subreddit.

One critical design decision made during the implementation of this program is to not use Pickle to store data on a disk. I wanted to get a live data from Reddit as “hot posts” are constantly changing. Also, I wanted the program to be able to calculate the sentiment value for any subreddit that an user wants to get the result of, so I didn’t hard code the subreddits that I ran the sentiment analysis on. By using Pickle, I would be storing a dataset from certain time and only for certain subreddits that I ran the code on. I wanted my program to work on any subreddits so I did not use pickle to store data.

Results

The program was ran on 5 different subreddits that I thought would have significantly differing positivity: ‘aww’, ‘science’, ‘worldnews’, ‘nottheonion’, and ‘nba’. Following table summarizes the overall sentiment value of each subreddit according to my program:

Subreddit	Sentiment Value
Aww	0.26125586219336216
Science	0.16943228872576699
Worldnews	0.069422162804515769
nottheonion	0.0017583239805461987
nba	-0.02636593204775025

These values seem reasonable on a number of fronts. Subreddit ‘aww’ is a compilation of pictures that people will find cute or adorable so the titles from that subreddit would be positive. On the other hand, subreddit ‘nottheonion’ is a collection of real news articles that are as ridiculous as Onion articles so it would be more negative, as confirmed by the program. Surprisingly, the subreddit ‘science’ was relatively positive while subreddit ‘world news’ and ‘nba’ were leaning toward negative.

Reflection

There is a number of areas of improvements for my project. First of all, I will get more accurate estimation of the sentiment value of a subreddit if I also analyze the comments in the subreddit. Secondly, I can separate my project into two parts: one for fetching data from reddit and another for processing the data. By doing so, I can store my data using Pickle, eliminating the need to make API call every time when I run the program and still getting live data from Reddit by using data fetch script that updates pickle data.