Text Mining Project
Tatiana Anthony
23 February 2017

Project Overview:

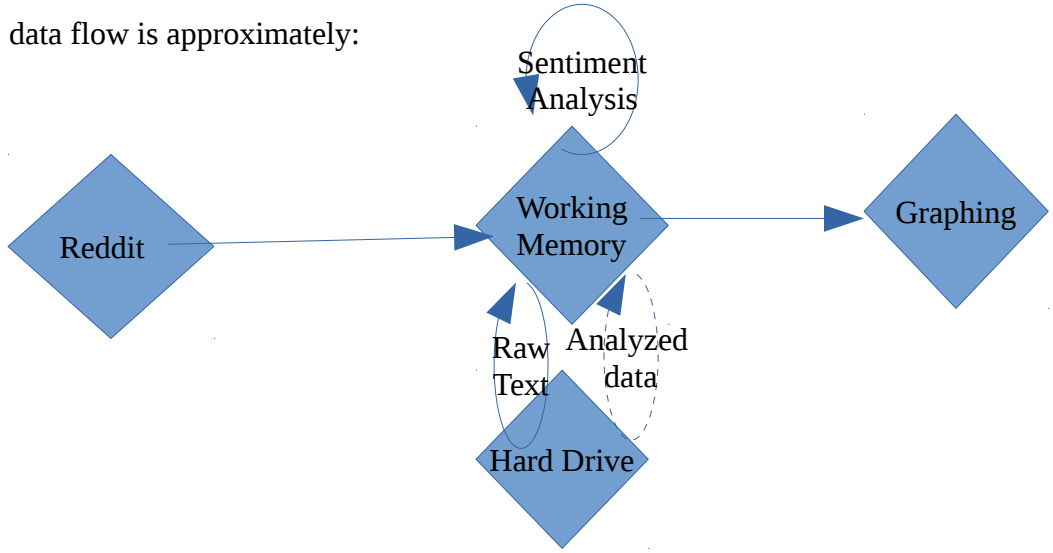
I used reddit's default subreddits, 15 largest communities, and about 7 communities I was interested in getting data about for my data. I analyzed the sentiment of each comment, scaled by the score of the comment, to get the overall positivity or negativity of the community in each subreddit. I hoped to compare subreddits with the defaults, to see which subreddits were more or less positive than reddit as a whole.

Implementation:

I split the code into several processes:

- Pulling the data from each subreddit, and saving into a pickle file.
- Loading the data from pickle file
- Analyzing the sentiment of the subreddits and saving into a pickle file
- loading the sentiments of the subreddits from a pickle file
- graphing the sentiments of each subreddit.

The data flow is approximately:



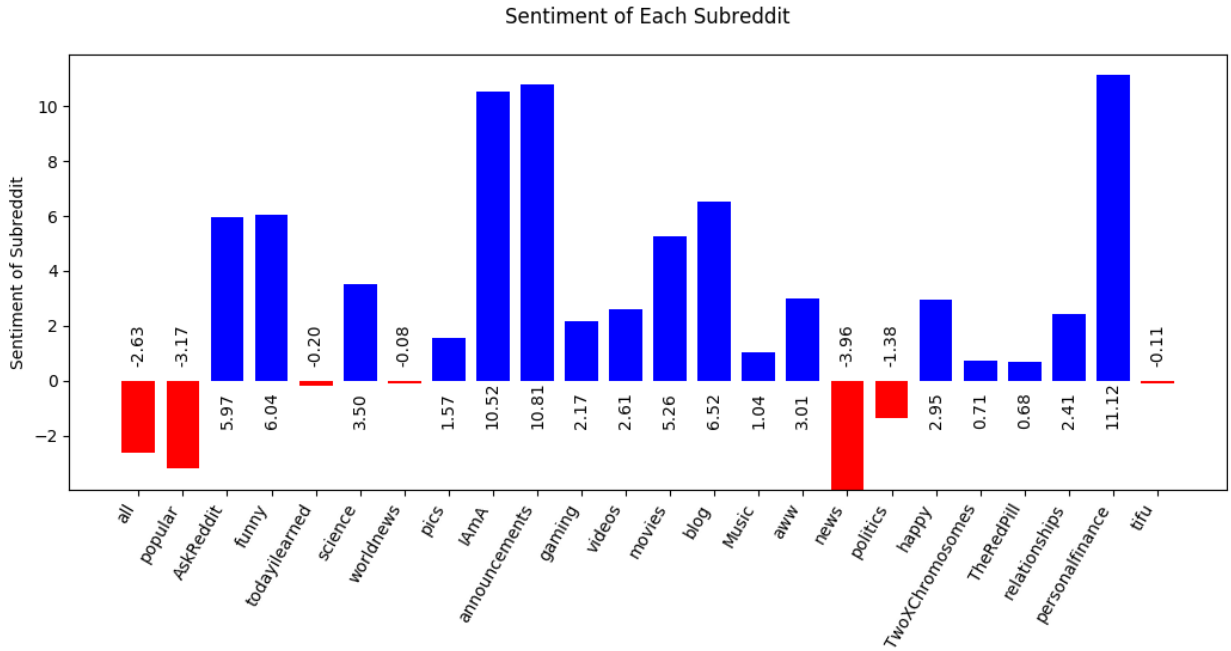
One particular design decision was saving the sentiments into a pickle file and being able to load them back up again. Initially, I just had the sentiment analysis script outputting the sentiments, and the graphing script using them directly. They were already separate scripts, but I realized I needed a way to save the data and use it later, so I pickled it and added an extra script to unpickle it, to save time while working on formatting the graphs.

I used a lot of looping through subreddits in more than one function. I think maybe I should have made the analysis and pickling functions take one subreddit rather than an array, and loop through the subreddits in the main function. That would have meant fewer inputs to the analysis function, and simpler code overall.

Results:

I found that the large subreddits are more positive than I expected. In the figure below, AskReddit through News are the 15 largest subreddits, as of February 23, 2017 at 9:30 AM. Note that most of them are significantly positive. The negative ones – todayilearned, Worldnews, and and news – aren't particularly surprising. TodayILearned is mostly neutral – it tends to have a lot of people complaining about reposts, but also has cool stuff that people enjoy. Worldnews is also mostly neutral, as a subreddit that tends to have fairly good discussion. News is the most negative subreddit in our data set. This subreddit is fairly US-centric, and most of the stories are what becomes popular in the news – bad things happening, financial crises, etc. The positive stories tend to be relegated to the Uplifting News subreddit, which was not investigated here.

The ones that were surprising to me were /r/announcements, TheRedPill, and relationships. /r/announcements is the subreddit where the admins announce changes and new things happening. Most people don't like change – I expected it to be neutral or slightly negative. Being so positive, I guess more people like the changes being made than I expected. TheRedPill is a subreddit of anti-feminists. I expected the vitriol that comes out of there to cause a negative result. I suppose that since the community tends to be in agreement, they support each other more positively than I expected. The same happened with /r/relationships – usually the advice is "break up," but I forgot to consider that people tend to be fairly gentle with their advice, given that the people who posters are usually in a bad situation. This result may have been different if I'd included the body of the original post instead of just the comments.



Reflection:

Overall, my process worked. The project was appropriately scoped. I wish I'd started a bit earlier so I could've done more with the analysis (attempting multiple heuristics), but overall, it was fine. I found

that I made too many changes without testing, and didn't initially consider how to prevent needing to analyze data with every test, so I'll probably consider those better. I think next time I'll do a better job making a full process diagram before starting my code, to prevent some of the layered changes that make my code feel a bit hack-y.