Project Proposal: r/AmItheAsshole Sentiment Analysis

Team Information:

Captain (solo project): Samantha Ernst (sdernst2)

Free Topic Project Summary:

course topics used: web scraping, data cleaning, linguistic processing, sentiment analysis

For my project, I will create a program that takes the link to a r/AmItheAsshole post and renders a color-coded word cloud based on sentiment analysis (negative/neutral/positive each corresponding to a certain color), a chart with the top emotions in replies, and a simple pie chart of the percentage of each of the 5 possible judgments.

This project is interesting because of the popularity, emotional intensity, and divisive nature of the AITA subreddit. Scrolling through the replies to a post is fun because it challenges your initial instinct and forces you to think beyond the boundaries of your own life and experiences. But sometimes after reading and forming your own opinion, you want a good summary to get the gist of what other people are thinking about the post, especially if a consensus hasn't yet been reached. This program will give an overarching idea of what other people in the replies are thinking and feeling about the post.

My planned approach is to take in the url for an r/AmItheAsshole post from the user, web scrape the page, clean up the data, and perform sentiment analysis on the replies (both negative/neutral/positive sentiment analysis for the top words and emotion detection on the overall set of replies). I'll use a word cloud and charts to display the two forms of sentiment analysis and a simple analysis of percentages for each possible judgment (NTA, YTA, ESH, NAH, INFO).

Language: Python
Platform/Service: Jupyter Notebooks / Google Colab
Libraries: Selenium (for web scraping), WordCloud, NLTK (VADER) (for negative/neutral/positive sentiment analysis), NRCLex (for emotion detection sentiment analysis), numpy, pandas, matplotlib

Evaluation:

I will evaluate my work by comparing the results with my own judgment of what the most common responses and emotions in replies were. Specifically, I will also compare the results of a

majority consensus NTA post with a majority consensus YTA post. Theoretically, the NTA should have much more positive sentiment and emotion, particularly if the audience was sympathetic with the OP.

Workload:

N = 1

Here are the tasks I aim to complete, along with estimated time cost for each task. These time costs include time spent learning about the topics, tools, and libraries.

Task	**Estimated Time Cost**
Web scraping	8 hours
Pre-processing/data cleaning	4 hours
Lexicon sentiment analysis	6 hours
Rendering visualizations (word cloud, charts)	3 hours
	Total: ~ 21 hours