Bill Wong
Thu Feb 26 2015
Software Design Spring 2015
Mini Project 3: Text Mining Reflection

Project Overview

I used reddit as a data source, analyzing the word frequencies of comments in any reddit thread. Reddit has an API, but I couldn't find a way to use it that didn't feel like cheating, so I decided to scrape the html instead. The output of the reddit-comments.py script could be copy-pasted into worditout.com directly to generate word clouds. I hoped to learn the basics of web scraping and to create word clouds that represented different reddit threads.

Implementation

I first used reddit's built-in URL functions to only display the first 500 top-level comments. I then used pattern's Document Object Model (DOM) function to search for instances where the div class was 'usertext-body' and grab the text of the comments. The count_words function then took the giant block of text and removed punctuation and unimportant words before counting the word frequencies and displaying the top 30 most used words in a worditout.com-friendly format.

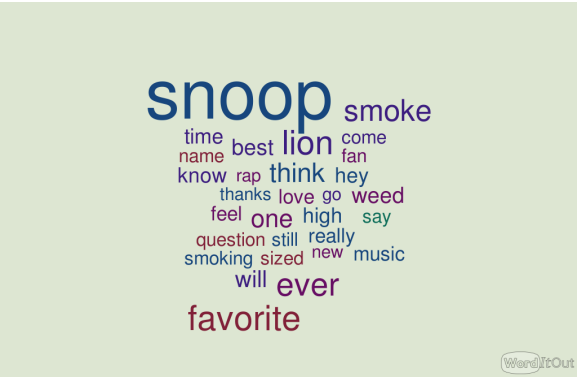
One design decision I had to make was how to block “uninteresting” words from making the final list. When I initially ran the word frequencies, only small words such as articles and pronouns made the top 30 list. This was not very interesting, so I had to figure out a way to not count those unimportant words. First, I tried only counting words that are at least five letters long, but then I missed some interesting four-letter words. My final code contains a list of uninteresting words called STOPWORDS that I found on the internet, and it skips over these uninteresting words when analyzing the word frequency.

Results

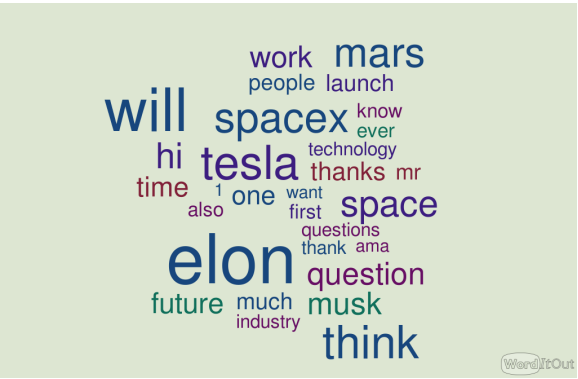
My code returns a big string that is formatted in such a way that it can be directly copy-pasted into worditout.com. Using this method, I made word clouds out of seven different reddit threads.

In the AMA threads, you can see that the comments address each celebrity by name most often, and the word clouds are relevant to each celebrity's occupation. Snoop Dogg's AMA features a lot of mentions of 'smoking', while Elon Musk's mentions 'technology' and 'space' more often. People usually say stupid things to 'girls' or at 'school' quite often, and their creepy stories usually involve 'house's or 'dream's. Also, Tom Cruise really need to come out already.

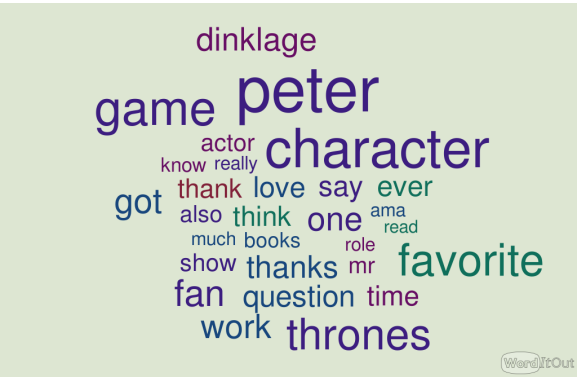
Snoop Dogg's AMA:



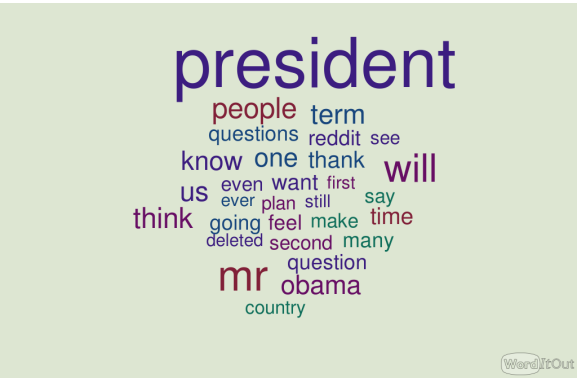
Elon Musk's AMA:



Peter Dinklage's AMA:

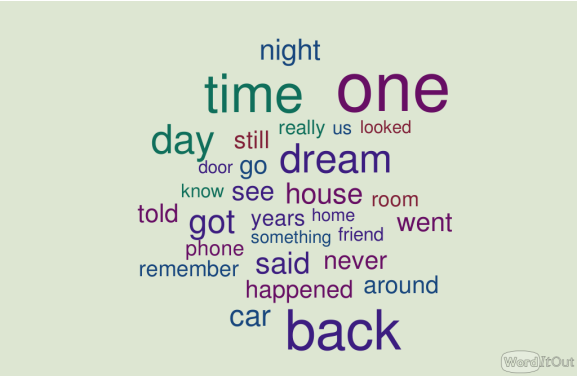


Obama's AMA:

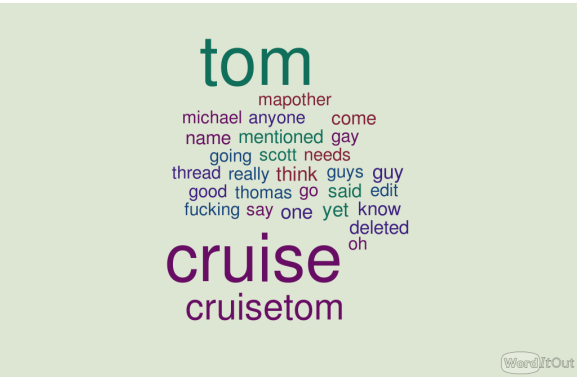


Stupidest thing you've said AskReddit:

Glitch in the matrix AskReddit:



Who needs to come out already AskReddit:



Reflection

Overall, I think this project was pretty well scoped. Originally, I had hoped to be able to generate my own word clouds from within python, but I recognized the possibility that it would be too difficult for this assignment. If I had to do this project again or had more time, I would probably make more functions, splitting up the code more logically. I never thought about the total organization of the code when I wrote it, so I work more on that aspect of it.