Project overview:

My primary source for this project was GitHub. I began by querying GitHub’s API but in order to get a more extensive set of data I used GHTorrent’s database. I tried querying their web databases for parts of the data, but I ended up downloading their complete set of GitHub transactions. The most accessible part was their pull request comments, which I’ve analyzed for this project. I primarily looked at word frequencies for analysis because of the size of this project. I hoped to learn something about what people

Implementation:

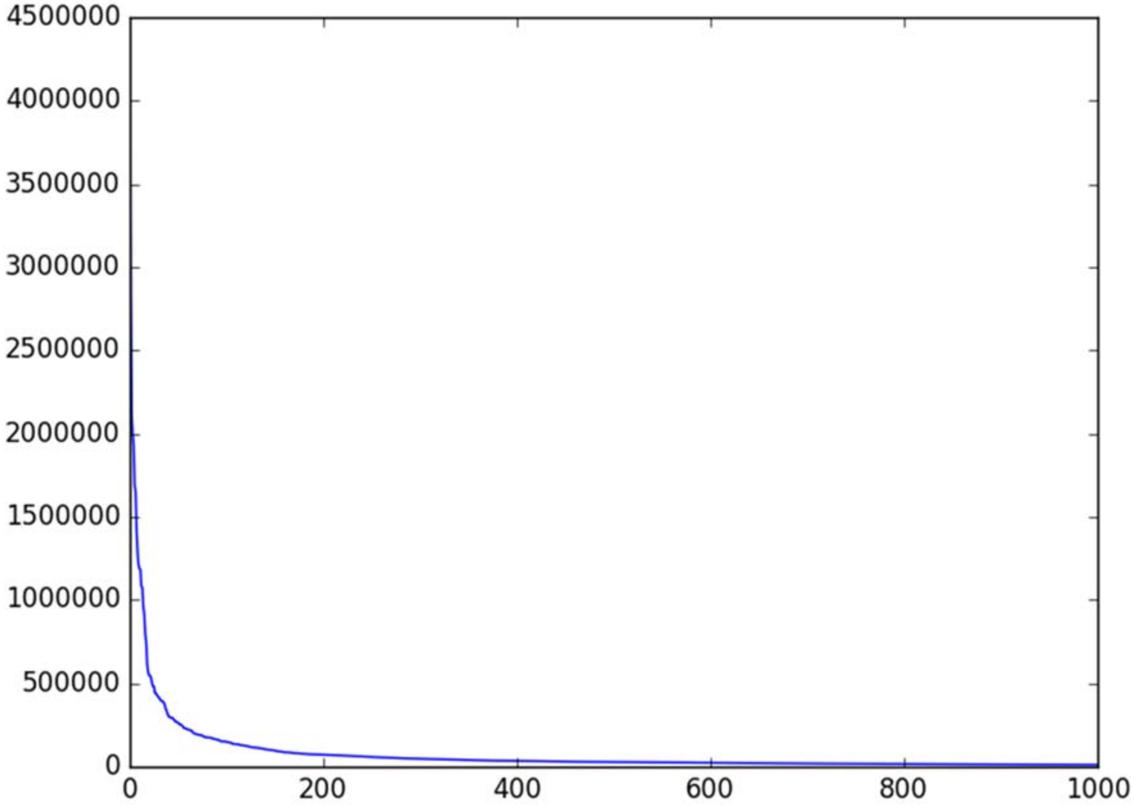
There were some interesting challenges in dealing with my dataset. The initial bank I started with was more than a gigabyte in size. Very quickly I chopped off all the fluff and dropped that down to 600 megs, by removing other columns and things like user meta-data. Then by condensing the data into a dictionary of word frequencies I reduced the size to less than a hundered megabytes. This is the set that I did most of my analysis on.

A big decision I made was to do my analysis on a subset of this data. I only looked at the most common words. My graph is the most 1000 common. I used a lot of list comprehensions when I could figure them out because they are faster. I used pyplot to plot because it’s a pretty fast plotter. I made the decision to grab my frab strings from the file unedited and edit them in small batches. I made all text lower and stripped characters before playing with it though, it meant that less errors would happen.

Results:

At first glance the data from every pull request comment on github until February first of this year seems to be very plain and not at all interesting. Upon closer inspection it’s only mildly interesting. The majority of the common words from these comments are just common words.

The word ‘the’ appears about twice as often as the second most common word. The word frequency decays exponentially. Here is a graph showing that. This is close to words in the average set of English, but it decays much faster.



The 50 most common words beyond the common ones are :use, think, need, would, like, line, don't, it's, code, 1, one, com, could, also, add, test, make, github, https, change, method, name, file, instead, get, see, want, using, i'm, function, remove, new, used, sure, case, 0, good, comment, 2, since, way, please, class, something, better, right, check, return, maybe, doesn't,

These words start to show the subject matter of my data (coding). As you go into more obscure words the trend remains the same, but often there is more noise on top.