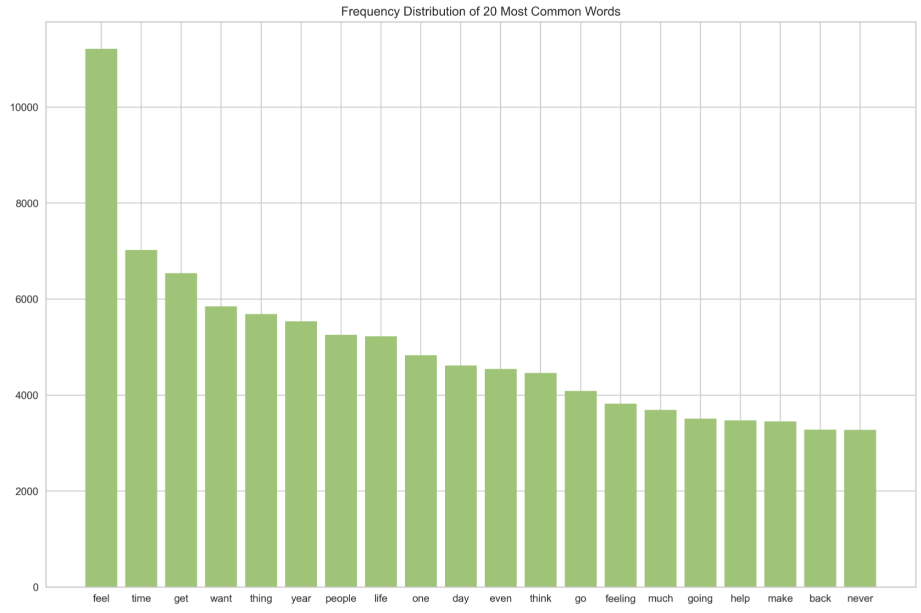
Introduction

One of the challenges of analyzing text documents is that it isn't naturally computer-friendly; that is, text documents require a thorough pre-processing before it can be analyzed. As shown in the Jupyter Notebook, the pre-processing steps are clearly documented. The pipeline is as follows:

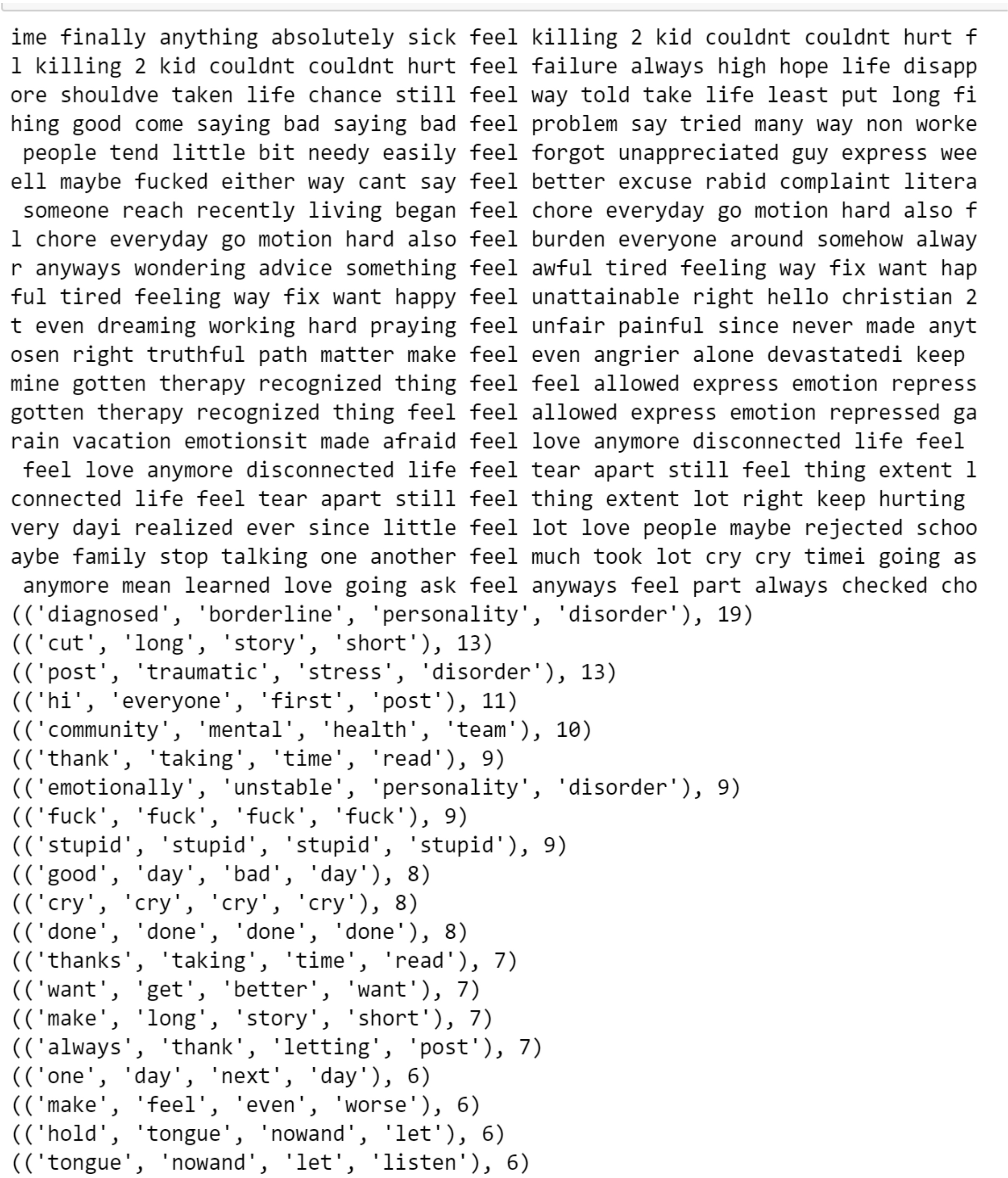
- Load the dataset and split off a holdout section for final testing.
- Process the documents by removing special characters, stopwords, and by lemmatizing.
- Tokenize the documents and construct a corpus.
- Explore with frequency distribution of top 20 words, POS tag distributions, concordances, wordcloud, and quadgram collocation finder.

Corpus Exploration

First, I take the frequency distribution of the 20 most common words in the corpus. I then plot a bar graph with those words and their frequencies. I noticed in the result that the frequencies of these words decline relatively gradually.

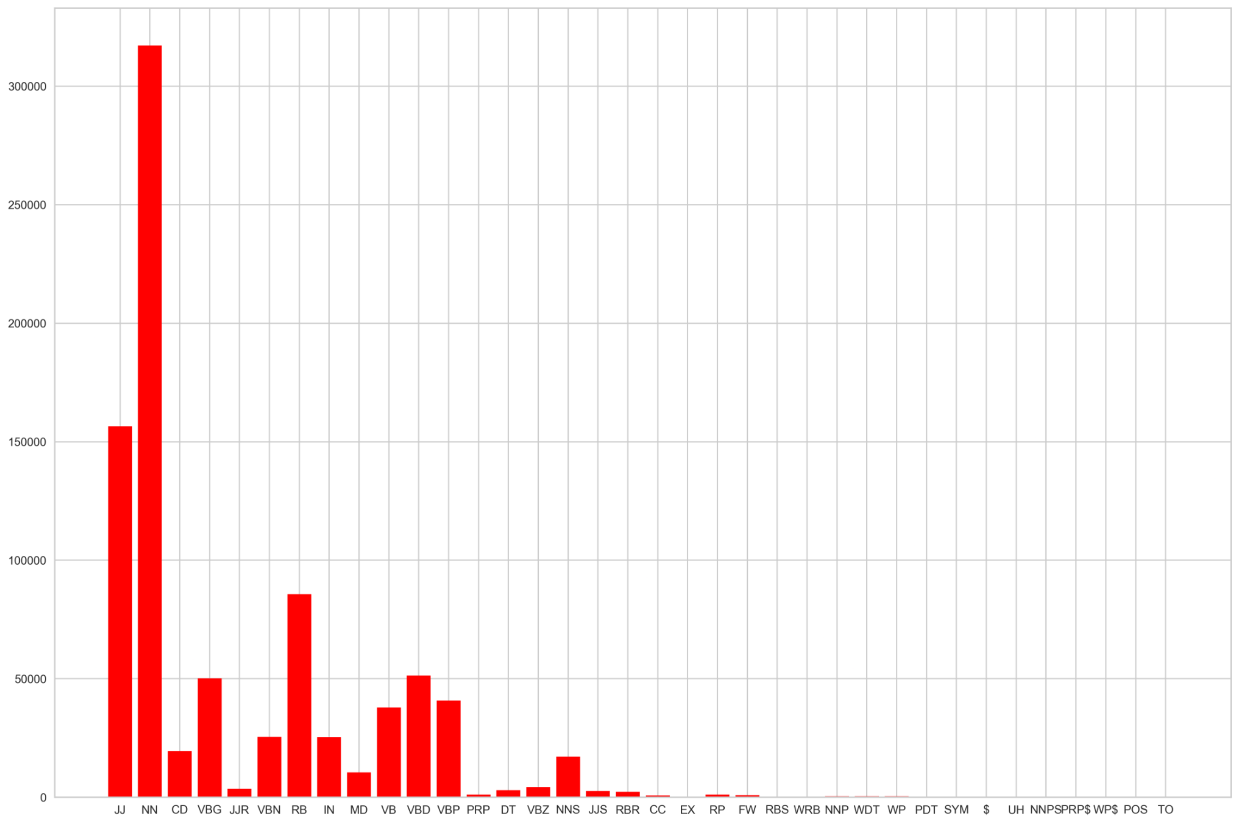


This could be due to concordances of the most frequent words, so I printed out a list of concordances based on the word, "feel", then used a quadgram collocation finder to examine which words frequently appear together. Result is presented on the page below.

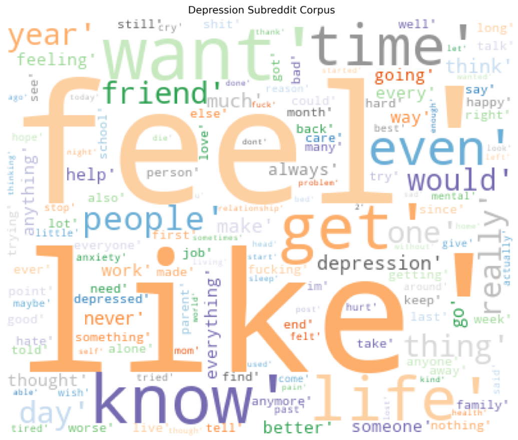


```
ime finally anything absolutely sick feel killing 2 kid couldnt couldnt hurt f
l killing 2 kid couldnt couldnt hurt feel failure always high hope life disapp
ore shouldve taken life chance still feel way told take life least put long fi
hing good come saying bad saying bad feel problem say tried many way non worke
 people tend little bit needy easily feel forgot unappreciated guy express wee
ell maybe fucked either way cant say feel better excuse rabid complaint litera
 someone reach recently living began feel chore everyday go motion hard also f
l chore everyday go motion hard also feel burden everyone around somehow alway
r anyways wondering advice something feel awful tired feeling way fix want hap
ful tired feeling way fix want happy feel unattainable right hello christian 2
t even dreaming working hard praying feel unfair painful since never made anyt
osen right truthful path matter make feel even angrier alone devastatedi keep
mine gotten therapy recognized thing feel feel allowed express emotion repress
gotten therapy recognized thing feel feel allowed express emotion repressed ga
rain vacation emotionsit made afraid feel love anymore disconnected life feel
 feel love anymore disconnected life feel tear apart still feel thing extent l
connected life feel tear apart still feel thing extent lot right keep hurting
very dayi realized ever since little feel lot love people maybe rejected schoo
aybe family stop talking one another feel much took lot cry cry timei going as
 anymore mean learned love going ask feel anyways feel part always checked cho
(('diagnosed', 'borderline', 'personality', 'disorder'), 19)
(('cut', 'long', 'story', 'short'), 13)
(('post', 'traumatic', 'stress', 'disorder'), 13)
(('hi', 'everyone', 'first', 'post'), 11)
(('community', 'mental', 'health', 'team'), 10)
(('thank', 'taking', 'time', 'read'), 9)
(('emotionally', 'unstable', 'personality', 'disorder'), 9)
(('fuck', 'fuck', 'fuck', 'fuck'), 9)
(('stupid', 'stupid', 'stupid', 'stupid'), 9)
(('good', 'day', 'bad', 'day'), 8)
(('cry', 'cry', 'cry', 'cry'), 8)
(('done', 'done', 'done', 'done'), 8)
(('thanks', 'taking', 'time', 'read'), 7)
(('want', 'get', 'better', 'want'), 7)
(('make', 'long', 'story', 'short'), 7)
(('always', 'thank', 'letting', 'post'), 7)
(('one', 'day', 'next', 'day'), 6)
(('make', 'feel', 'even', 'worse'), 6)
(('hold', 'tongue', 'nowand', 'let'), 6)
(('tongue', 'nowand', 'let', 'listen'), 6)
```

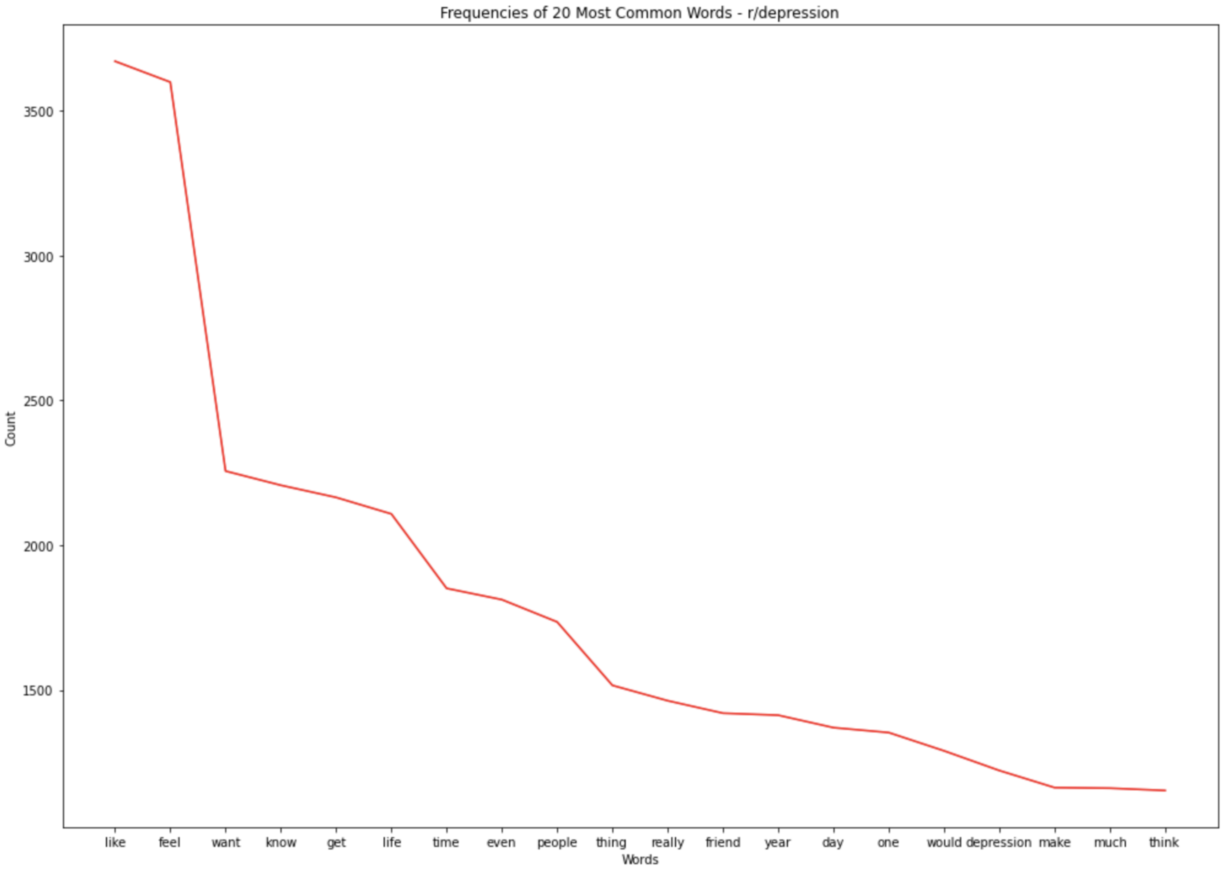
Moving on, I also wanted to explore the distribution of POS (Part of Speech) tags. I used NLTK's POS tagger to perform this task and visualized it into another bar graph. Two disproportionately frequent POS tags were, "NN" - noun - and, "JJ" - adjective. This makes sense when thinking about the origination of the text documents, all of which came from mental health support forum sites. The assumption I make from this result is that posts contain a lot of nouns to track, and they get rather descriptive about them.



WordCloud


Depression Subreddit Corpus

Top 20 Words


Frequencies of 20 Most Common Words - r/depression
Count
3500
3000
2500
2000
1500
like
feel
want
know
get
life
time
even
people
thing
really
friend
year
day
one
would
depression
make
much
think
Words