Michael Gryncewicz
April 21, 2018

Data 902 Sentiment Analysis Project

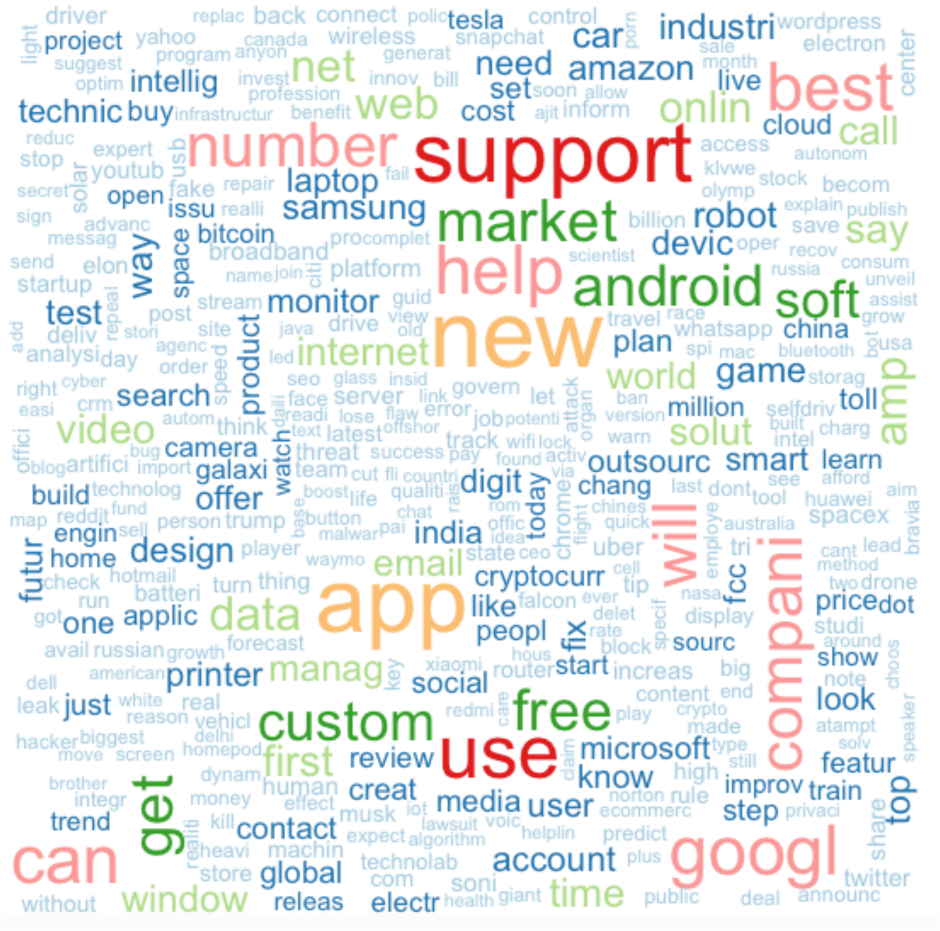
Project:

The goal of this project was to apply various natural language processing techniques to a corpus of documents. These techniques consisted of producing word clouds based on various n-grams, using a lexicon to apply sentiment to the documents, and running an emotional analysis on the corpus.

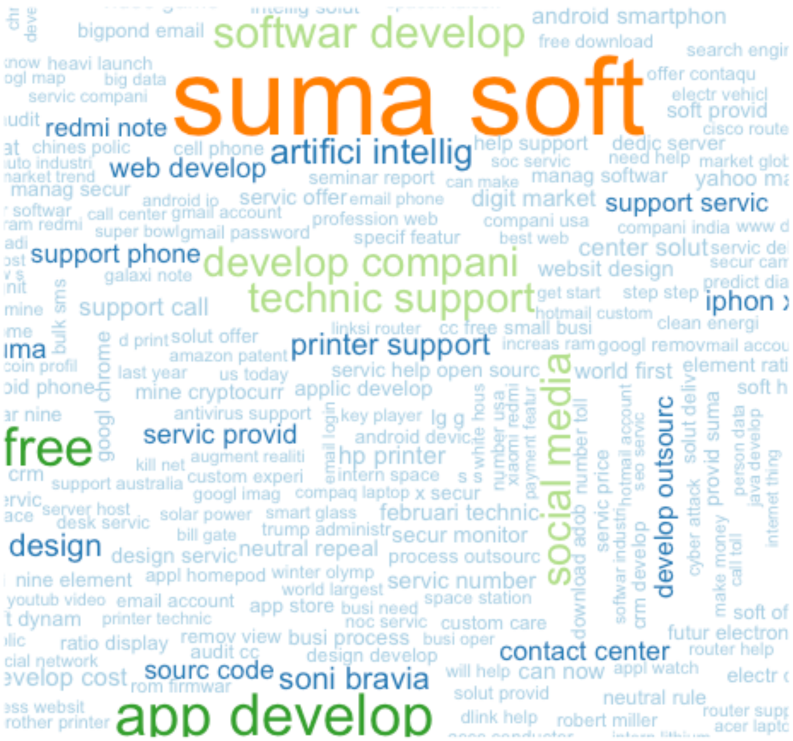
Data:

The data used in this project was taken from reddit.com. The data consisted of various features from reddit posts made in February of 2018. The full data contained about 10 million rows but a subset was used. The data used was a subset by one single section of reddit, the technology subreddit. The sentiment analysis was then performed on the titles of all of these posts. There were about 10,000 posts used in the analysis.

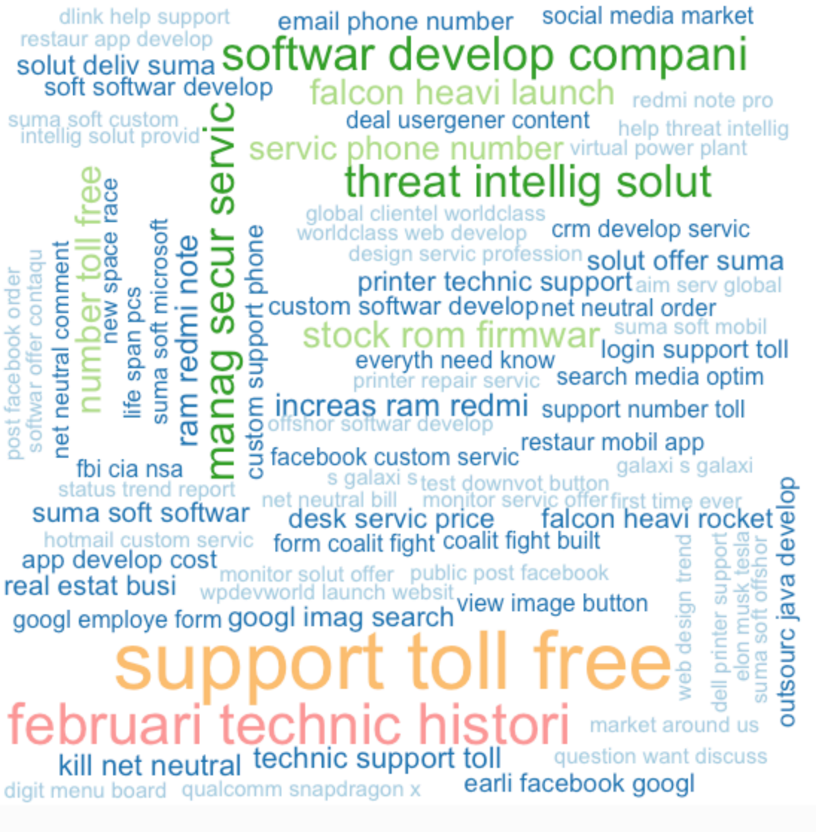
Term Frequency Unigram Word Cloud:



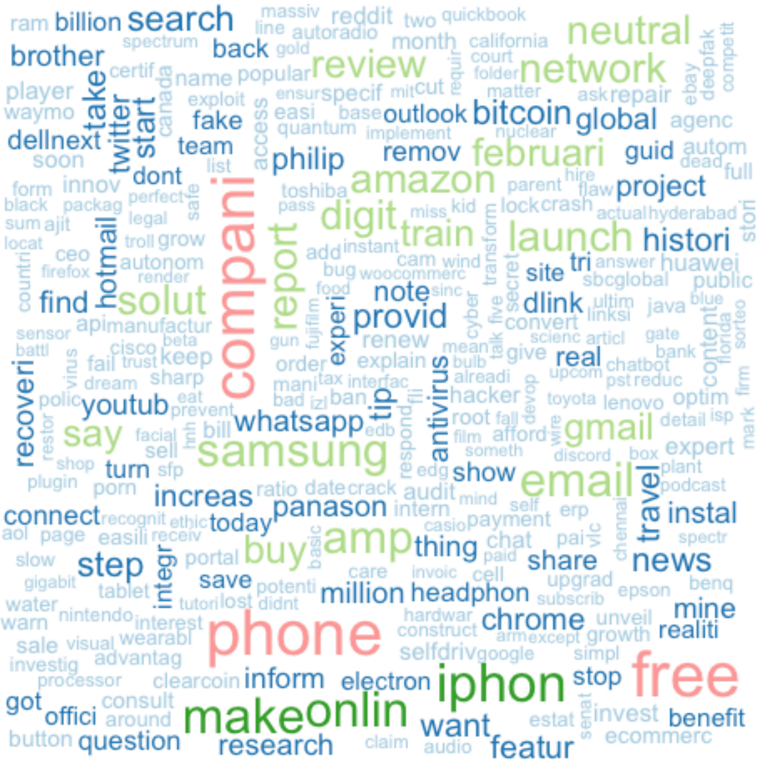
Term Frequency Bigram Word Cloud:



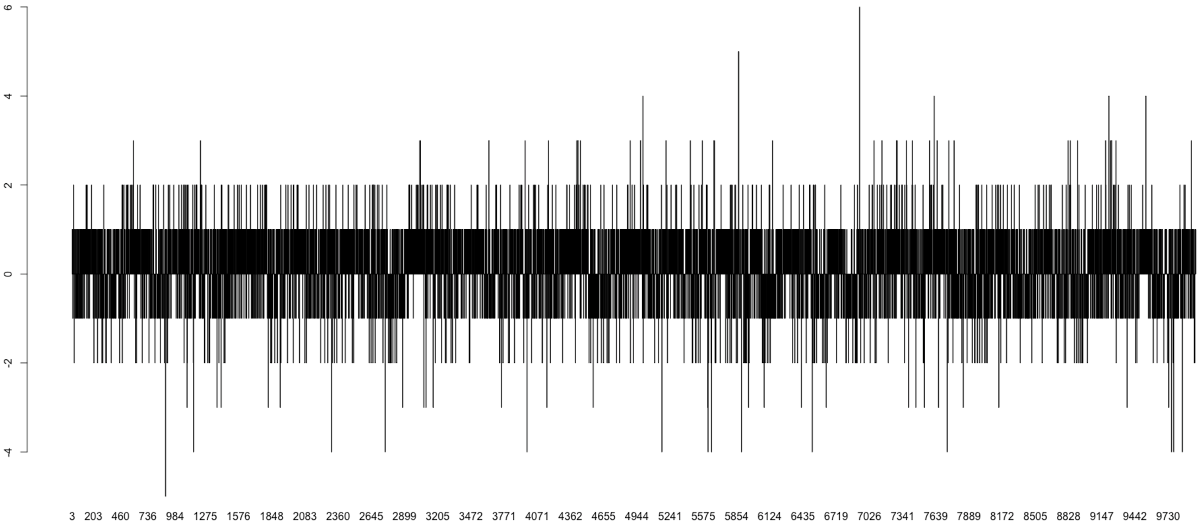
Term Frequency Trigram Word Cloud:

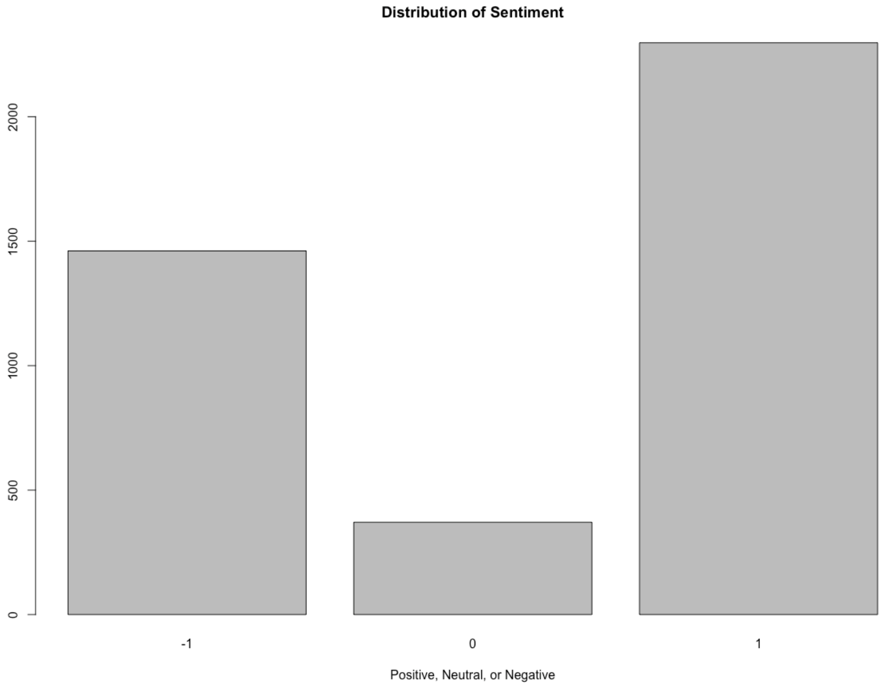


TF-IDF Word Cloud:



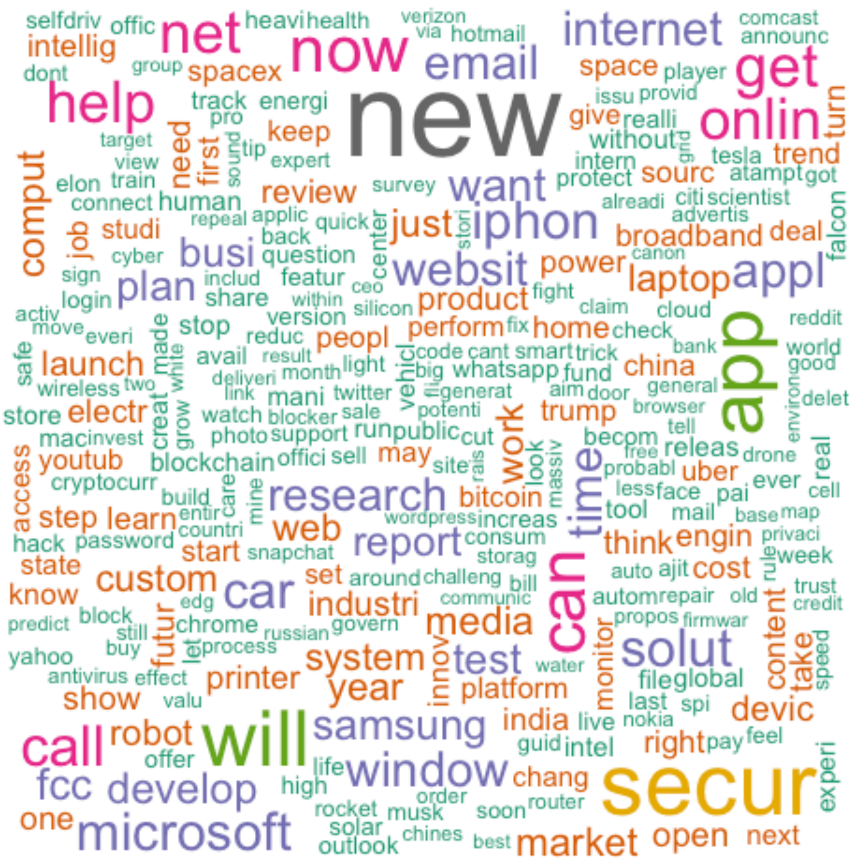
Sentiment Analysis using Bing Lexicon:



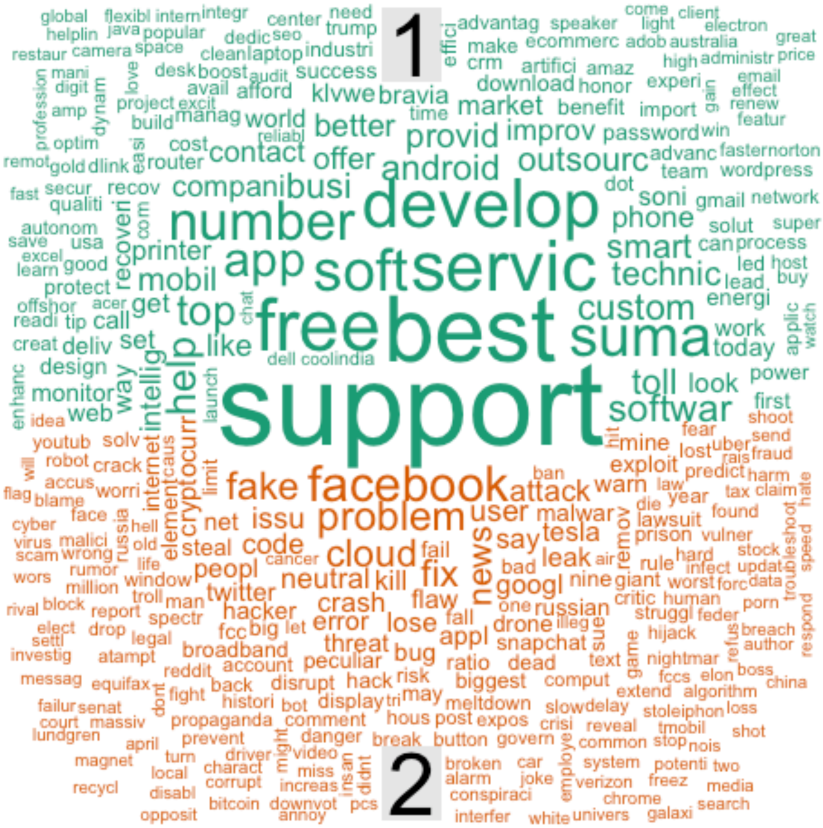


In general, the majority of the sentiment appears to be positive based on the Bing lexicon, but it not overwhelmingly positive.

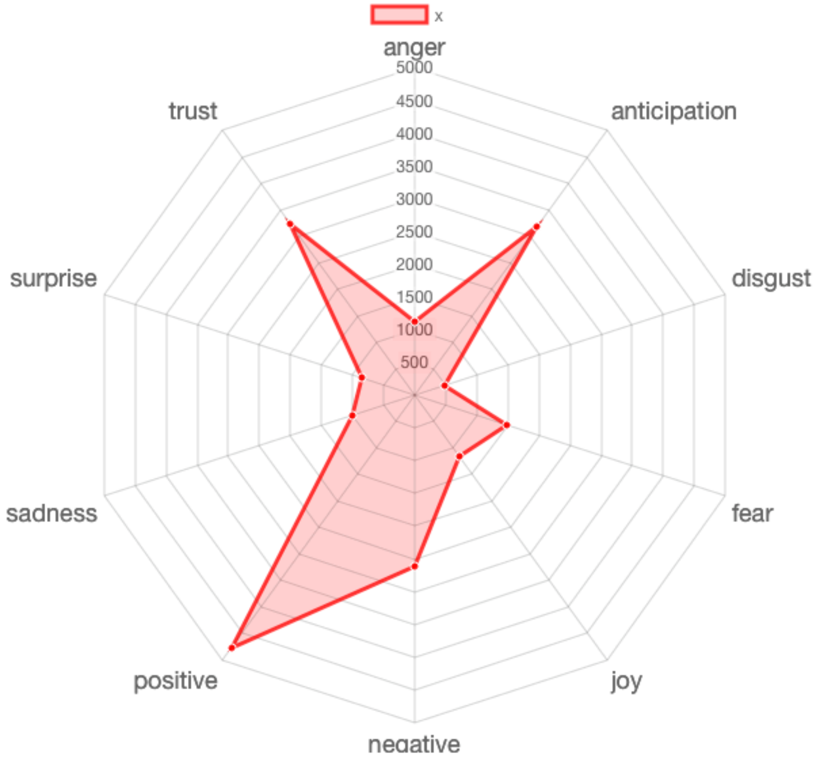
Comparison Word Cloud:



Contrast Word Cloud:



Emotion Analysis with NRC Lexicon:


x
anger
anticipation
disgust
fear
joy
negative
positive
sadness
surprise
trust