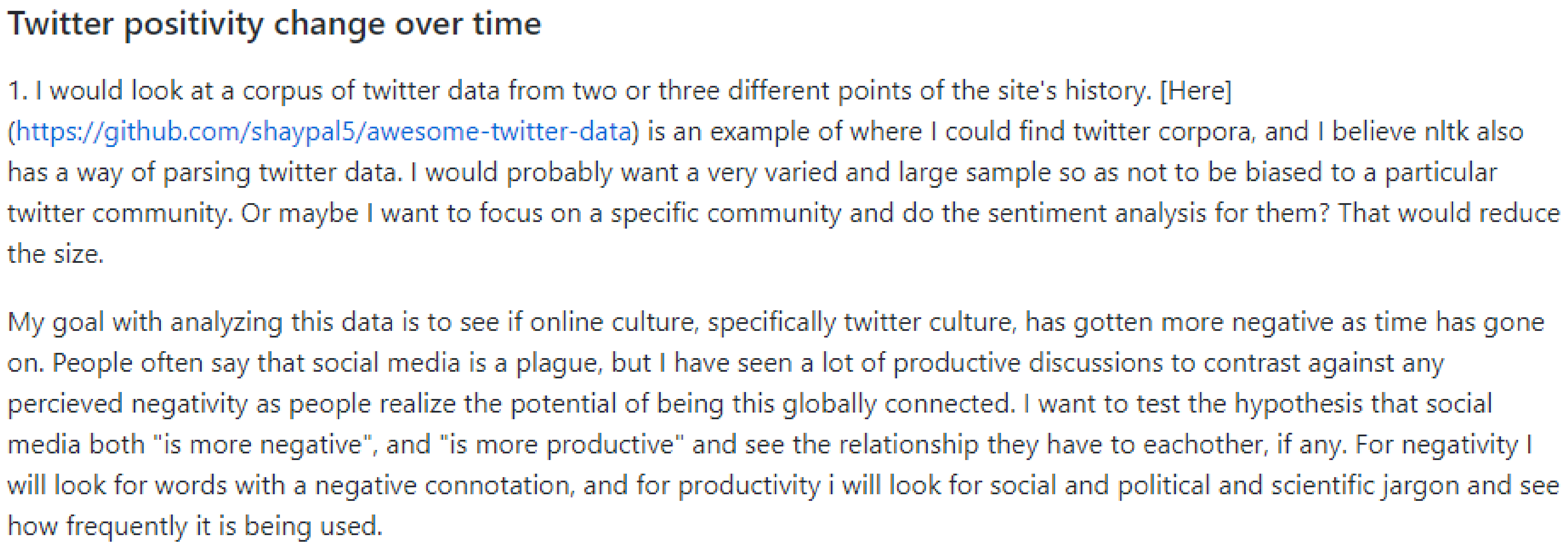
TWITTER SENTIMENT ANALYSIS OVER TIME

Natasha Kamtekar | LING 1340

MOTIVATION

- Saw how social media interactions had changed over the course of my time on the internet- wanted to examine in what ways it had changed.
- Wanted to see how different topics were talked about.

MOTIVATION - PROPOSAL



Twitter positivity change over time

1. I would look at a corpus of twitter data from two or three different points of the site's history. [Here](https://github.com/shaypal5/awesome-twitter-data) is an example of where I could find twitter corpora, and I believe nltk also has a way of parsing twitter data. I would probably want a very varied and large sample so as not to be biased to a particular twitter community. Or maybe I want to focus on a specific community and do the sentiment analysis for them? That would reduce the size.

My goal with analyzing this data is to see if online culture, specifically twitter culture, has gotten more negative as time has gone on. People often say that social media is a plague, but I have seen a lot of productive discussions to contrast against any percieved negativity as people realize the potential of being this globally connected. I want to test the hypothesis that social media both "is more negative", and "is more productive" and see the relationship they have to eachother, if any. For negativity I will look for words with a negative connotation, and for productivity i will look for social and political and scientific jargon and see how frequently it is being used.

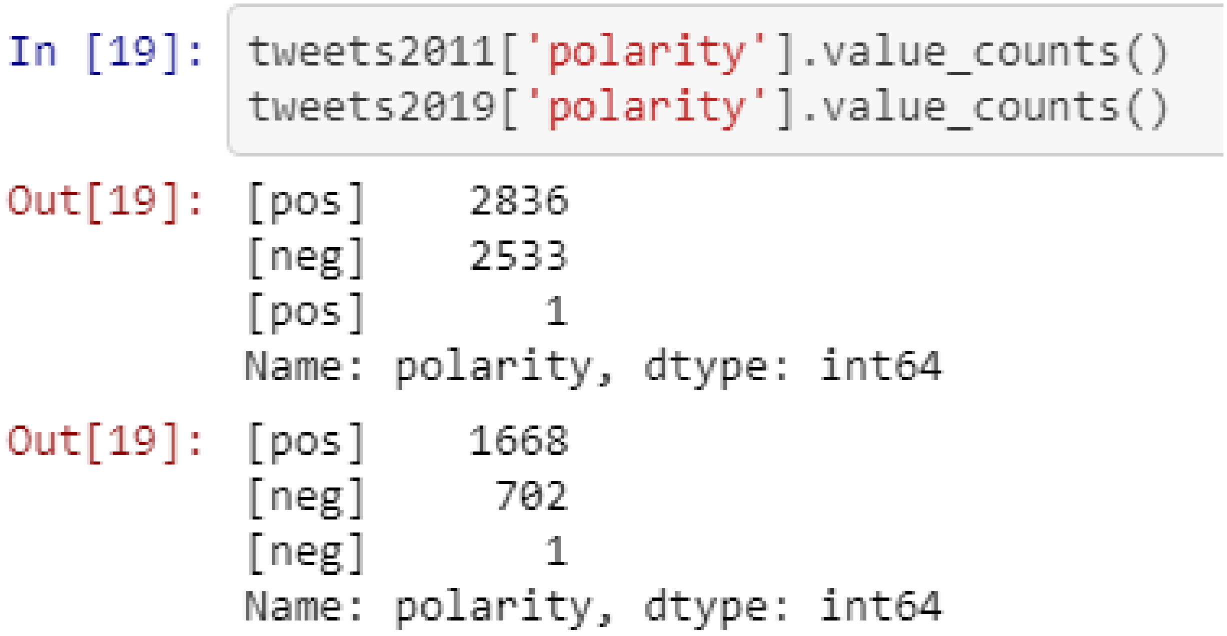
THE DATA

- Used a diachronic dataset from September 2011 and October 2019.
- JSON files
- Trained a classifier with data from Sentiment140, a classification tool with open source data. Data was organized into positive and negative sentiments.
- Initially data was very large and hard to work with, more recent techniques of big data came in handy.

INITIAL HYPOTHESIS

Twitter has gotten more negative with time.

VISUALIZATION



```
In [19]:  tweets2011['polarity'].value_counts()
           tweets2019['polarity'].value_counts()

Out[19]:  [pos]    2836
          [neg]    2533
          [pos]       1
          Name: polarity, dtype: int64

Out[19]:  [pos]    1668
          [neg]     702
          [neg]       1
          Name: polarity, dtype: int64
```

53% Positive in 2011

70% Positive in 2019

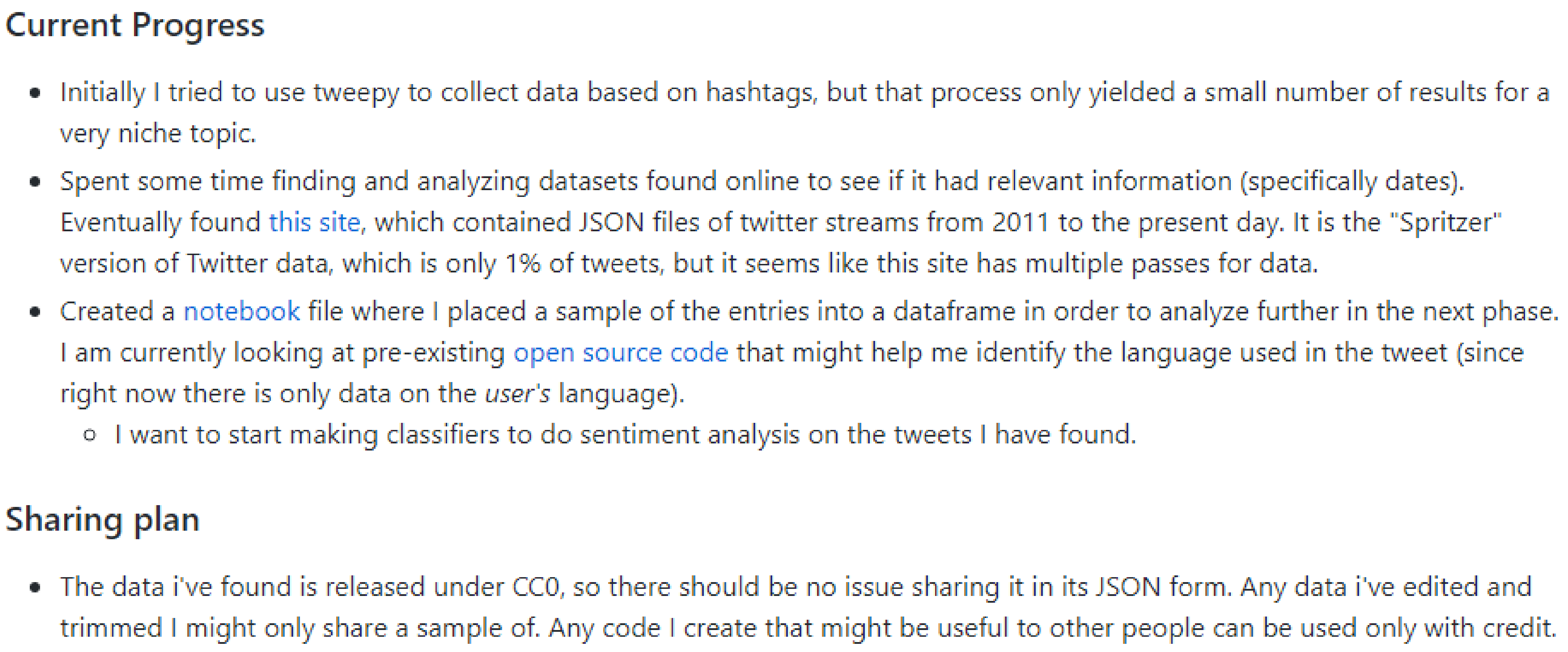
INITIAL THOUGHTS

- There was no clear way to contextualize this.
- The data ended up somewhat lending itself to the opposite. There were more tweets tagged with "negative" in 2011. There was also more profanity used in the 2011 negative tweets.
- This made me consider changing gears and look at different aspects of the data.

INITIAL DATA

- Tried scraping my own tweets but I would only get about 100 or 200 at a time and they were super messy. All of them were also based on a keyword or hashtag (which is not what I wanted at the time)
- When I found the diachronic files I could only use a small fraction of the total data because of my computers processing limits. This was made even smaller after getting rid of non-English tweets.

PROGRESS REPORT 1



Current Progress

- Initially I tried to use tweepy to collect data based on hashtags, but that process only yielded a small number of results for a very niche topic.
- Spent some time finding and analyzing datasets found online to see if it had relevant information (specifically dates). Eventually found this site, which contained JSON files of twitter streams from 2011 to the present day. It is the "Spritzer" version of Twitter data, which is only 1% of tweets, but it seems like this site has multiple passes for data.
- Created a notebook file where I placed a sample of the entries into a dataframe in order to analyze further in the next phase. I am currently looking at pre-existing open source code that might help me identify the language used in the tweet (since right now there is only data on the *user's* language).
    - I want to start making classifiers to do sentiment analysis on the tweets I have found.

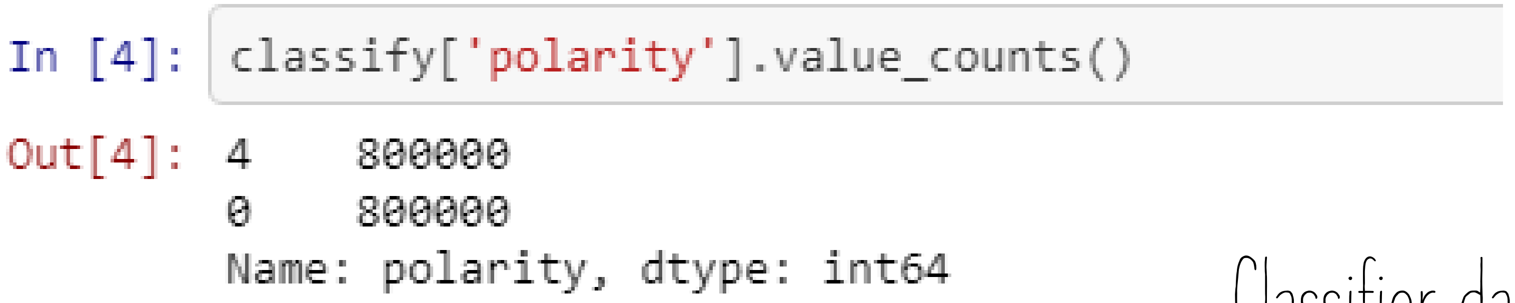
Sharing plan

- The data i've found is released under CC0, so there should be no issue sharing it in its JSON form. Any data i've edited and trimmed I might only share a sample of. Any code I create that might be useful to other people can be used only with credit.

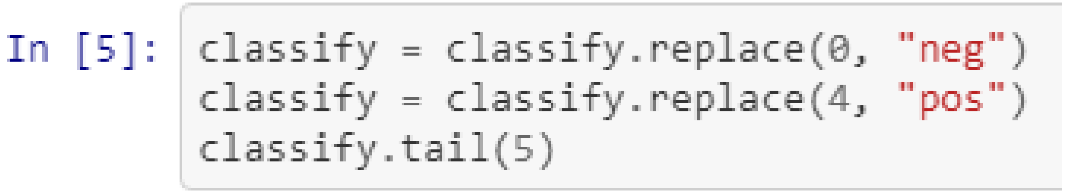
CLEANING THE DATA

- Removed non-English words using langdetect.
- Removed stopwords using nltk's English library.
- After first round of analysis, cleaned out any additional text lingo stop words that were very frequent such as "u"
- Removed punctuation and links.
- Kept RT because it indicated engagement and conversation.

BUILDING CLASSIFIER



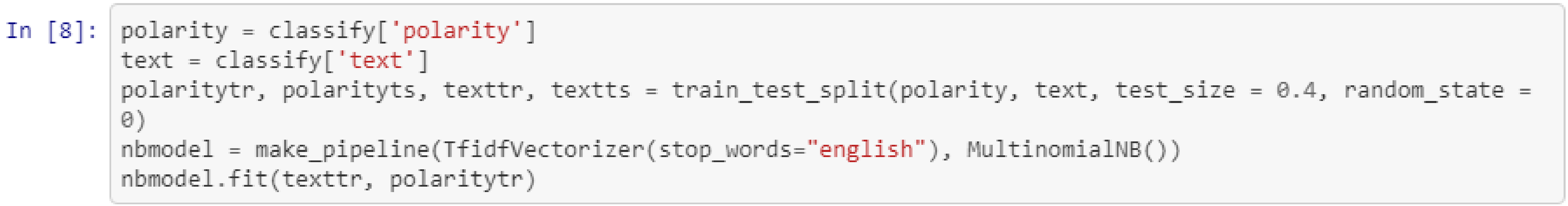
```
In [4]: classify['polarity'].value_counts()

Out[4]: 4    800000
        0    800000
        Name: polarity, dtype: int64
```



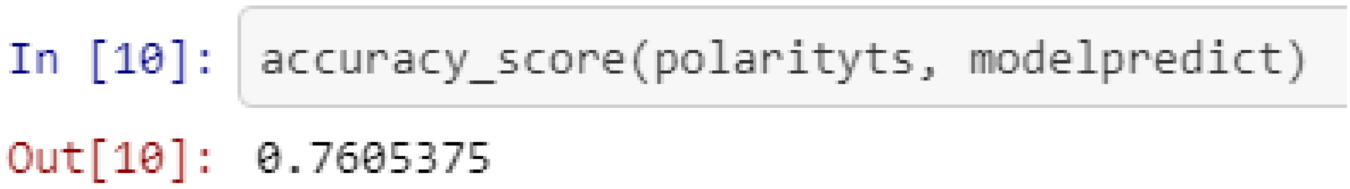
```
In [5]: classify = classify.replace(0, "neg")
        classify = classify.replace(4, "pos")
        classify.tail(5)
```

Classifier data was already spilt into "0" which was negative and "4" which was positive. I renamed them.

BUILDING CLASSIFIER



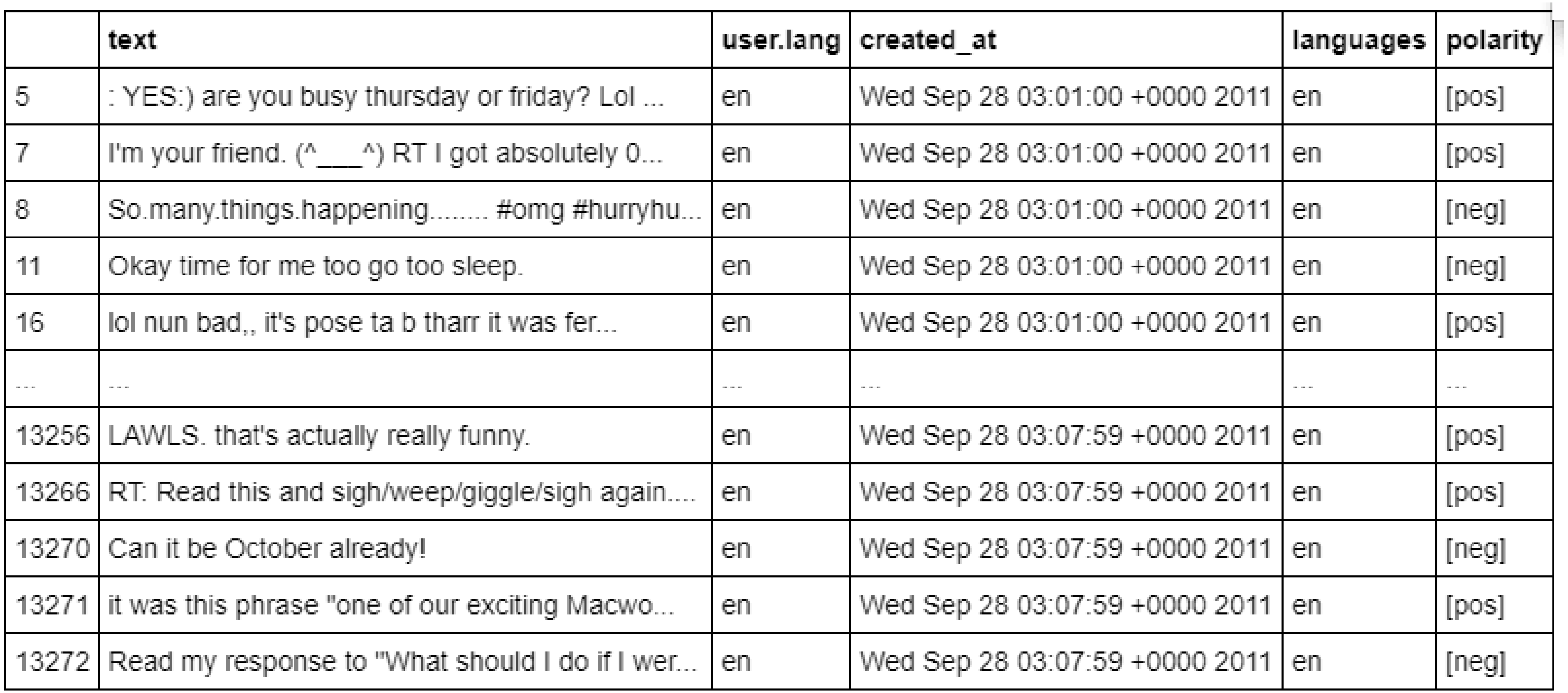
```
In [8]: polarity = classify['polarity']
        text = classify['text']
        polaritytr, polarityts, texttr, textts = train_test_split(polarity, text, test_size = 0.4, random_state =
        0)
        nbmodel = make_pipeline(TfidfVectorizer(stop_words="english"), MultinomialNB())
        nbmodel.fit(texttr, polaritytr)
```



```
In [10]: accuracy_score(polarityts, modelpredict)

Out[10]: 0.7605375
```

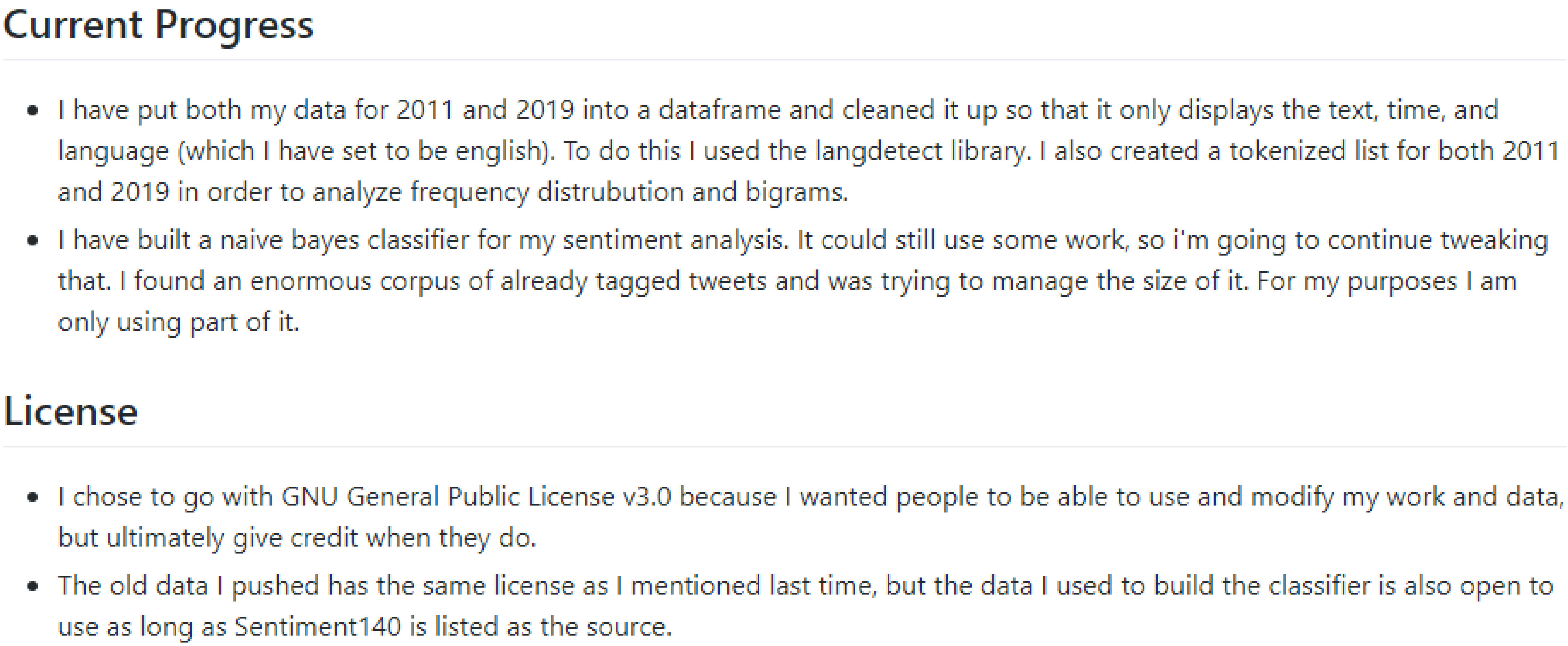
Using pipeline to build a multinomial naïve bayes model with a 76% accuracy.

EXAMPLE OF DATA AFTER CLASSIFICATION



	text	user.lang	created_at	languages	polarity
5	: YES:) are you busy thursday or friday? Lol ...	en	Wed Sep 28 03:01:00 +0000 2011	en	[pos]
7	I'm your friend. (^___^) RT I got absolutely 0...	en	Wed Sep 28 03:01:00 +0000 2011	en	[pos]
8	So.many.things.happening........ #omg #hurryhu...	en	Wed Sep 28 03:01:00 +0000 2011	en	[neg]
11	Okay time for me too go too sleep.	en	Wed Sep 28 03:01:00 +0000 2011	en	[neg]
16	lol nun bad,, it's pose ta b tharr it was fer...	en	Wed Sep 28 03:01:00 +0000 2011	en	[pos]
...	...	...	...	...	...
13256	LAWLS. that's actually really funny.	en	Wed Sep 28 03:07:59 +0000 2011	en	[pos]
13266	RT: Read this and sigh/weep/giggle/sigh again....	en	Wed Sep 28 03:07:59 +0000 2011	en	[pos]
13270	Can it be October already!	en	Wed Sep 28 03:07:59 +0000 2011	en	[neg]
13271	it was this phrase "one of our exciting Macwo...	en	Wed Sep 28 03:07:59 +0000 2011	en	[pos]
13272	Read my response to "What should I do if I wer...	en	Wed Sep 28 03:07:59 +0000 2011	en	[neg]

PROGRESS REPORT 2



Current Progress

- I have put both my data for 2011 and 2019 into a dataframe and cleaned it up so that it only displays the text, time, and language (which I have set to be english). To do this I used the langdetect library. I also created a tokenized list for both 2011 and 2019 in order to analyze frequency distrubution and bigrams.
- I have built a naive bayes classifier for my sentiment analysis. It could still use some work, so i'm going to continue tweaking that. I found an enormous corpus of already tagged tweets and was trying to manage the size of it. For my purposes I am only using part of it.

License

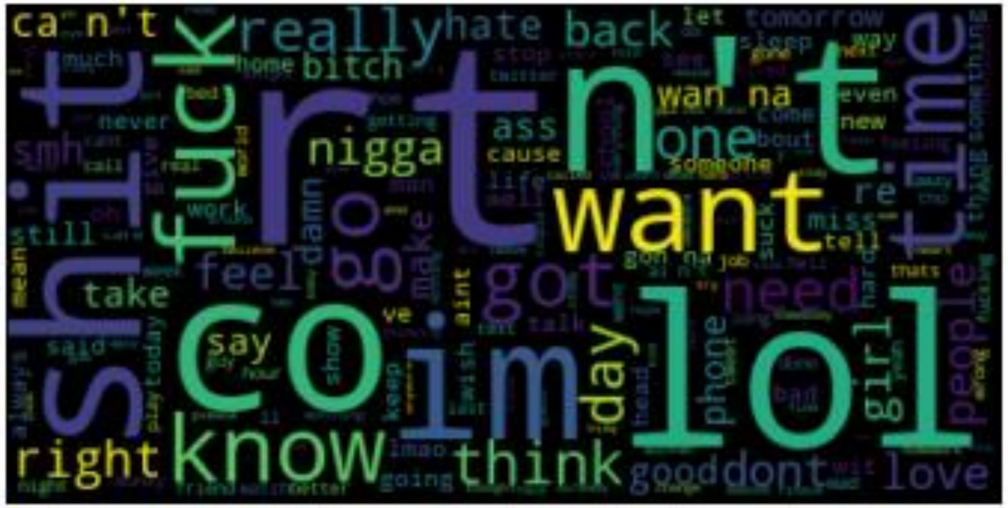
- I chose to go with GNU General Public License v3.0 because I wanted people to be able to use and modify my work and data, but ultimately give credit when they do.
- The old data I pushed has the same license as I mentioned last time, but the data I used to build the classifier is also open to use as long as Sentiment140 is listed as the source.

ANALYSIS

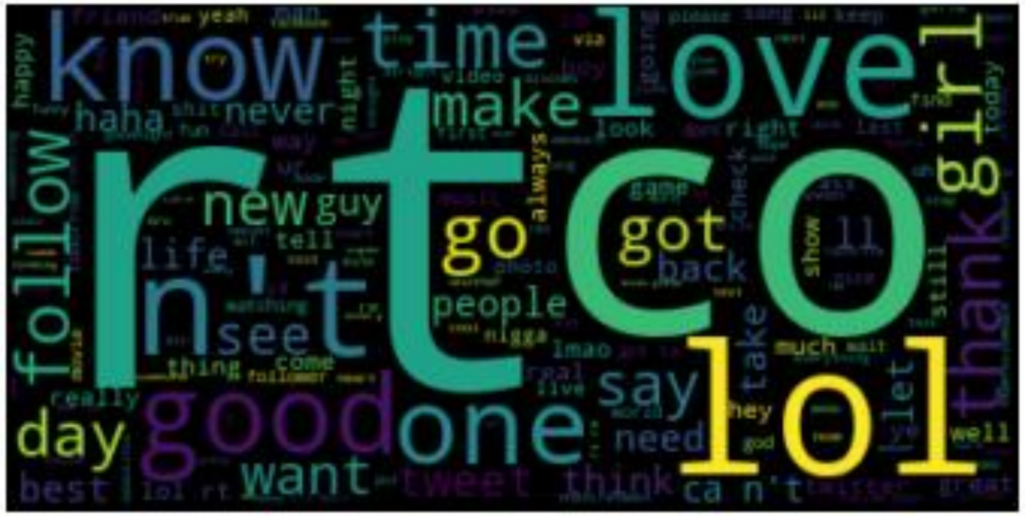
- Frequency Distribution and most common words
- Looked at n-grams of the tokens in order to contextualize certain frequently occurring words within their positive/negative contexts
- Looking at time of day- when do negatively connotated tweets occur
- Looking at content- negative tweets in 2019 are more political, negative tweets in 2011 are about not being able to get a burrito.

VISUALIZATION

- Still working on that part...
- Created word clouds and showed frequency distribution both overall and then with a positive or negative tilt.



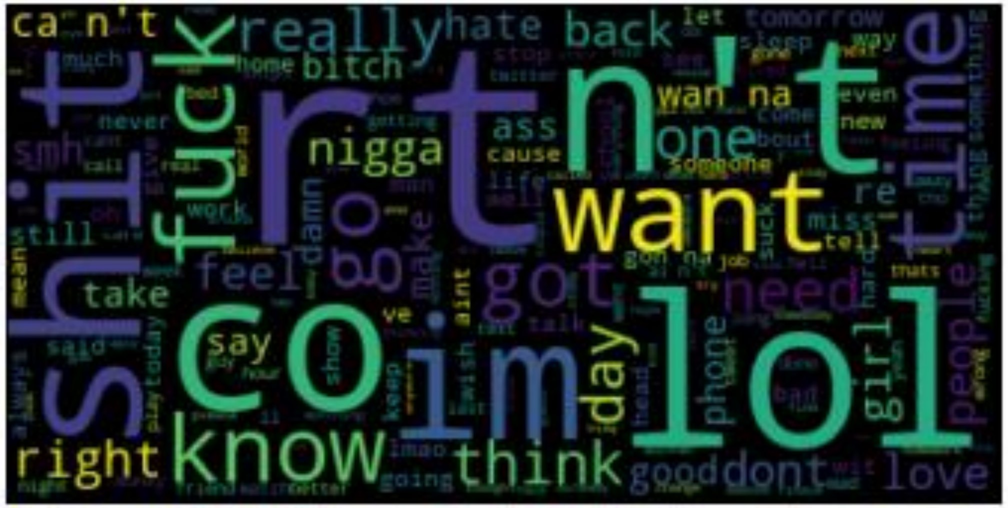
2011 Negative Words



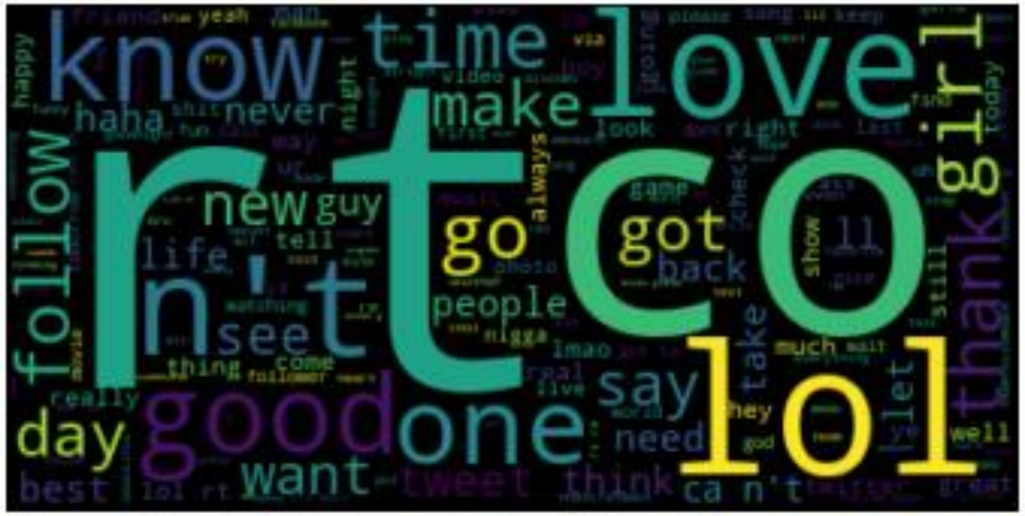
2011 Positive Words

VISUALIZATION

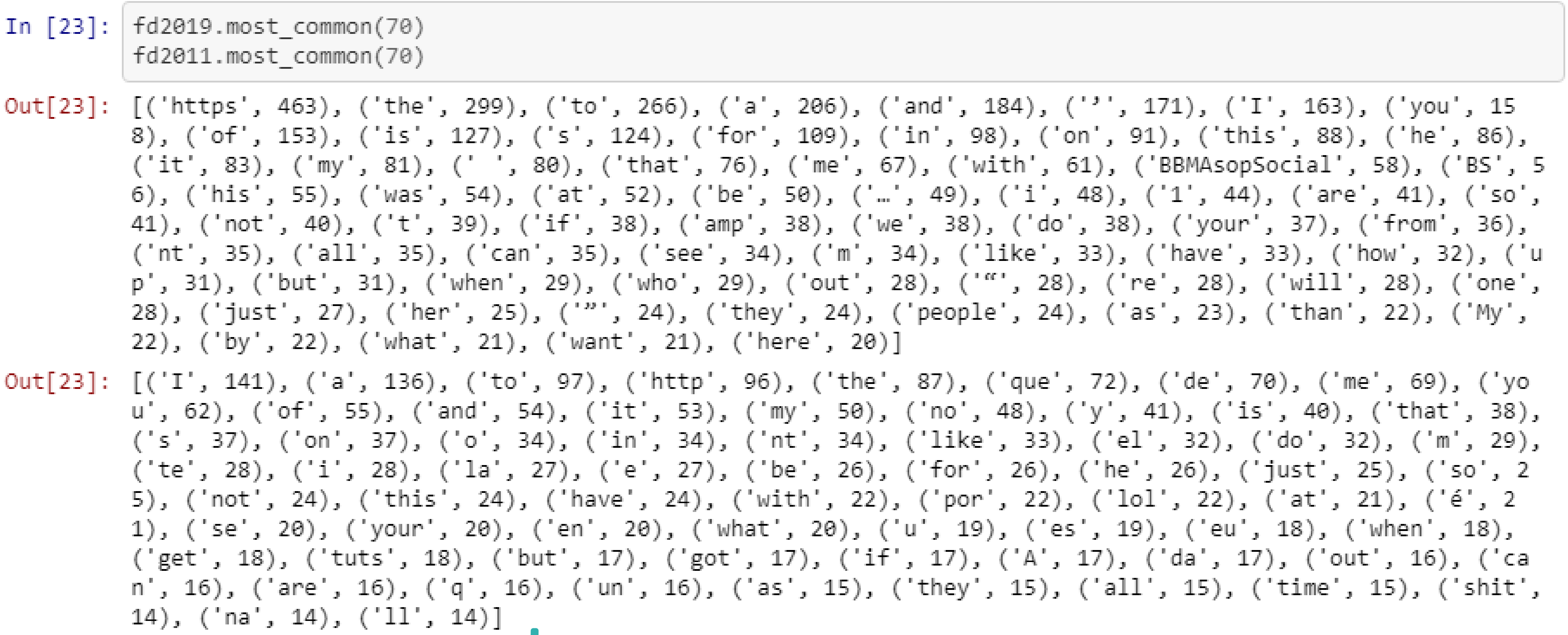
- Positive engagements had more "RT" interactions.
- Negative engagements had more profanity.
- They both "LOL" a lot



2011 Negative Words



2011 Positive Words

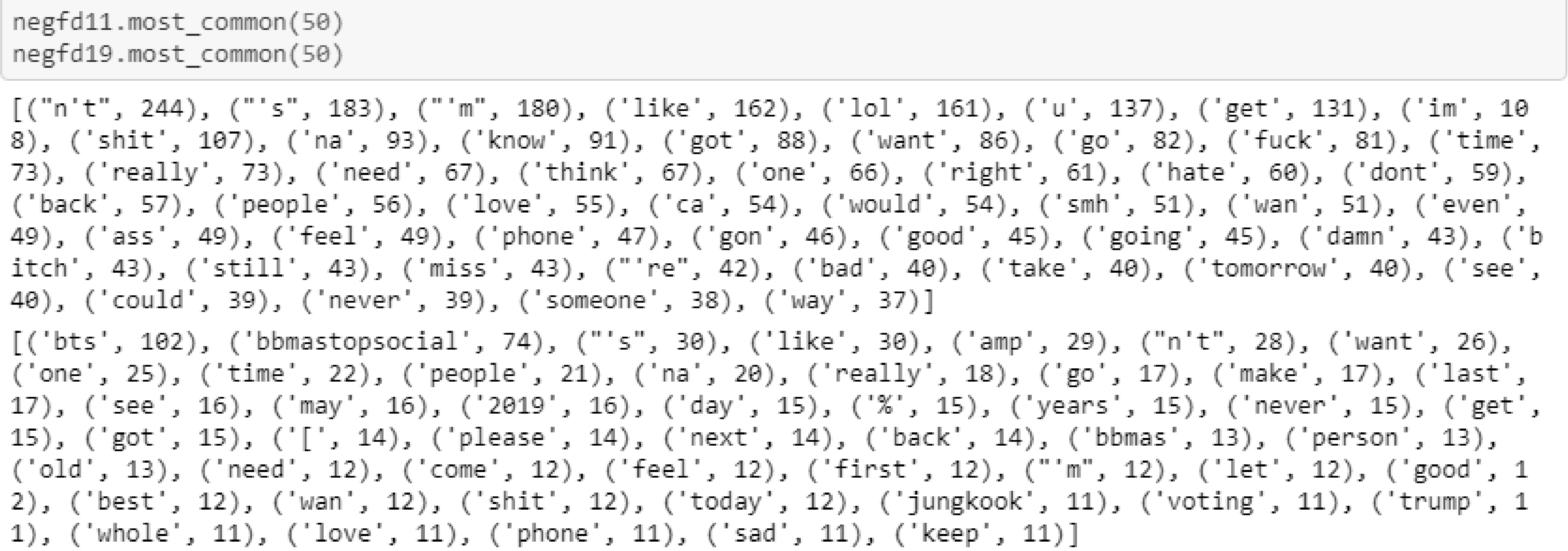


```
In [23]: fd2019.most_common(70)
         fd2011.most_common(70)

Out[23]: [('https', 463), ('the', 299), ('to', 266), ('a', 206), ('and', 184), ('’', 171), ('I', 163), ('you', 15
         8), ('of', 153), ('is', 127), ('s', 124), ('for', 109), ('in', 98), ('on', 91), ('this', 88), ('he', 86),
         ('it', 83), ('my', 81), (' ', 80), ('that', 76), ('me', 67), ('with', 61), ('BBMAsopSocial', 58), ('BS', 5
         6), ('his', 55), ('was', 54), ('at', 52), ('be', 50), ('…', 49), ('i', 48), ('1', 44), ('are', 41), ('so',
         41), ('not', 40), ('t', 39), ('if', 38), ('amp', 38), ('we', 38), ('do', 38), ('your', 37), ('from', 36),
         ('nt', 35), ('all', 35), ('can', 35), ('see', 34), ('m', 34), ('like', 33), ('have', 33), ('how', 32), ('u
         p', 31), ('but', 31), ('when', 29), ('who', 29), ('out', 28), ('“', 28), ('re', 28), ('will', 28), ('one',
         28), ('just', 27), ('her', 25), ('”', 24), ('they', 24), ('people', 24), ('as', 23), ('than', 22), ('My',
         22), ('by', 22), ('what', 21), ('want', 21), ('here', 20)]

Out[23]: [('I', 141), ('a', 136), ('to', 97), ('http', 96), ('the', 87), ('que', 72), ('de', 70), ('me', 69), ('yo
         u', 62), ('of', 55), ('and', 54), ('it', 53), ('my', 50), ('no', 48), ('y', 41), ('is', 40), ('that', 38),
         ('s', 37), ('on', 37), ('o', 34), ('in', 34), ('nt', 34), ('like', 33), ('el', 32), ('do', 32), ('m', 29),
         ('te', 28), ('i', 28), ('la', 27), ('e', 27), ('be', 26), ('for', 26), ('he', 26), ('just', 25), ('so', 2
         5), ('not', 24), ('this', 24), ('have', 24), ('with', 22), ('por', 22), ('lol', 22), ('at', 21), ('é', 2
         1), ('se', 20), ('your', 20), ('en', 20), ('what', 20), ('u', 19), ('es', 19), ('eu', 18), ('when', 18),
         ('get', 18), ('tuts', 18), ('but', 17), ('got', 17), ('if', 17), ('A', 17), ('da', 17), ('out', 16), ('ca
         n', 16), ('are', 16), ('q', 16), ('un', 16), ('as', 15), ('they', 15), ('all', 15), ('time', 15), ('shit',
         14), ('na', 14), ('ll', 14)]
```

VISUALIZATION

Looking at most common words in general

(includes stopwords)



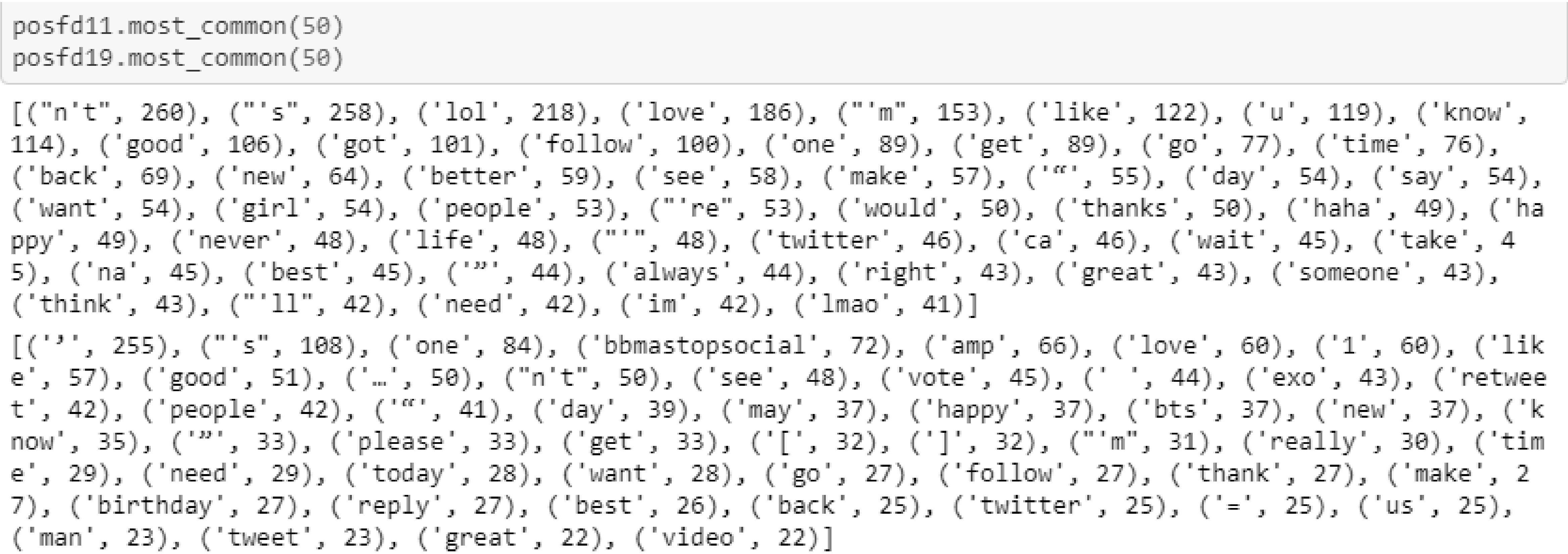
```
negfd11.most_common(50)
negfd19.most_common(50)
```

```
[("n't", 244), ("'s", 183), ("'m", 180), ('like', 162), ('lol', 161), ('u', 137), ('get', 131), ('im', 10
8), ('shit', 107), ('na', 93), ('know', 91), ('got', 88), ('want', 86), ('go', 82), ('fuck', 81), ('time',
73), ('really', 73), ('need', 67), ('think', 67), ('one', 66), ('right', 61), ('hate', 60), ('dont', 59),
('back', 57), ('people', 56), ('love', 55), ('ca', 54), ('would', 54), ('smh', 51), ('wan', 51), ('even',
49), ('ass', 49), ('feel', 49), ('phone', 47), ('gon', 46), ('good', 45), ('going', 45), ('damn', 43), ('b
itch', 43), ('still', 43), ('miss', 43), ("'re", 42), ('bad', 40), ('take', 40), ('tomorrow', 40), ('see',
40), ('could', 39), ('never', 39), ('someone', 38), ('way', 37)]
```

```
[('bts', 102), ('bbmastopsocial', 74), ("'s", 30), ('like', 30), ('amp', 29), ("n't", 28), ('want', 26),
('one', 25), ('time', 22), ('people', 21), ('na', 20), ('really', 18), ('go', 17), ('make', 17), ('last',
17), ('see', 16), ('may', 16), ('2019', 16), ('day', 15), ('%', 15), ('years', 15), ('never', 15), ('get',
15), ('got', 15), ('[', 14), ('please', 14), ('next', 14), ('back', 14), ('bbmas', 13), ('person', 13),
('old', 13), ('need', 12), ('come', 12), ('feel', 12), ('first', 12), ("'m", 12), ('let', 12), ('good', 1
2), ('best', 12), ('wan', 12), ('shit', 12), ('today', 12), ('jungkook', 11), ('voting', 11), ('trump', 1
1), ('whole', 11), ('love', 11), ('phone', 11), ('sad', 11), ('keep', 11)]
```

VISUALIZATION

Looking at only most common negative words



```
posfd11.most_common(50)
posfd19.most_common(50)
```

```
[("n't", 260), ("'s", 258), ('lol', 218), ('love', 186), ("'m", 153), ('like', 122), ('u', 119), ('know',
114), ('good', 106), ('got', 101), ('follow', 100), ('one', 89), ('get', 89), ('go', 77), ('time', 76),
('back', 69), ('new', 64), ('better', 59), ('see', 58), ('make', 57), ('“', 55), ('day', 54), ('say', 54),
('want', 54), ('girl', 54), ('people', 53), ("'re", 53), ('would', 50), ('thanks', 50), ('haha', 49), ('ha
ppy', 49), ('never', 48), ('life', 48), ("'", 48), ('twitter', 46), ('ca', 46), ('wait', 45), ('take', 4
5), ('na', 45), ('best', 45), ('”', 44), ('always', 44), ('right', 43), ('great', 43), ('someone', 43),
('think', 43), ("'ll", 42), ('need', 42), ('im', 42), ('lmao', 41)]
```

```
[('’', 255), ("'s", 108), ('one', 84), ('bbmastopsocial', 72), ('amp', 66), ('love', 60), ('1', 60), ('lik
e', 57), ('good', 51), ('…', 50), ("n't", 50), ('see', 48), ('vote', 45), ('  ', 44), ('exo', 43), ('retwee
t', 42), ('people', 42), ('“', 41), ('day', 39), ('may', 37), ('happy', 37), ('bts', 37), ('new', 37), ('k
now', 35), ('”', 33), ('please', 33), ('get', 33), ('[', 32), (']', 32), ("'m", 31), ('really', 30), ('tim
e', 29), ('need', 29), ('today', 28), ('want', 28), ('go', 27), ('follow', 27), ('thank', 27), ('make', 2
7), ('birthday', 27), ('reply', 27), ('best', 26), ('back', 25), ('twitter', 25), ('=', 25), ('us', 25),
('man', 23), ('tweet', 23), ('great', 22), ('video', 22)]
```

VISUALIZATION

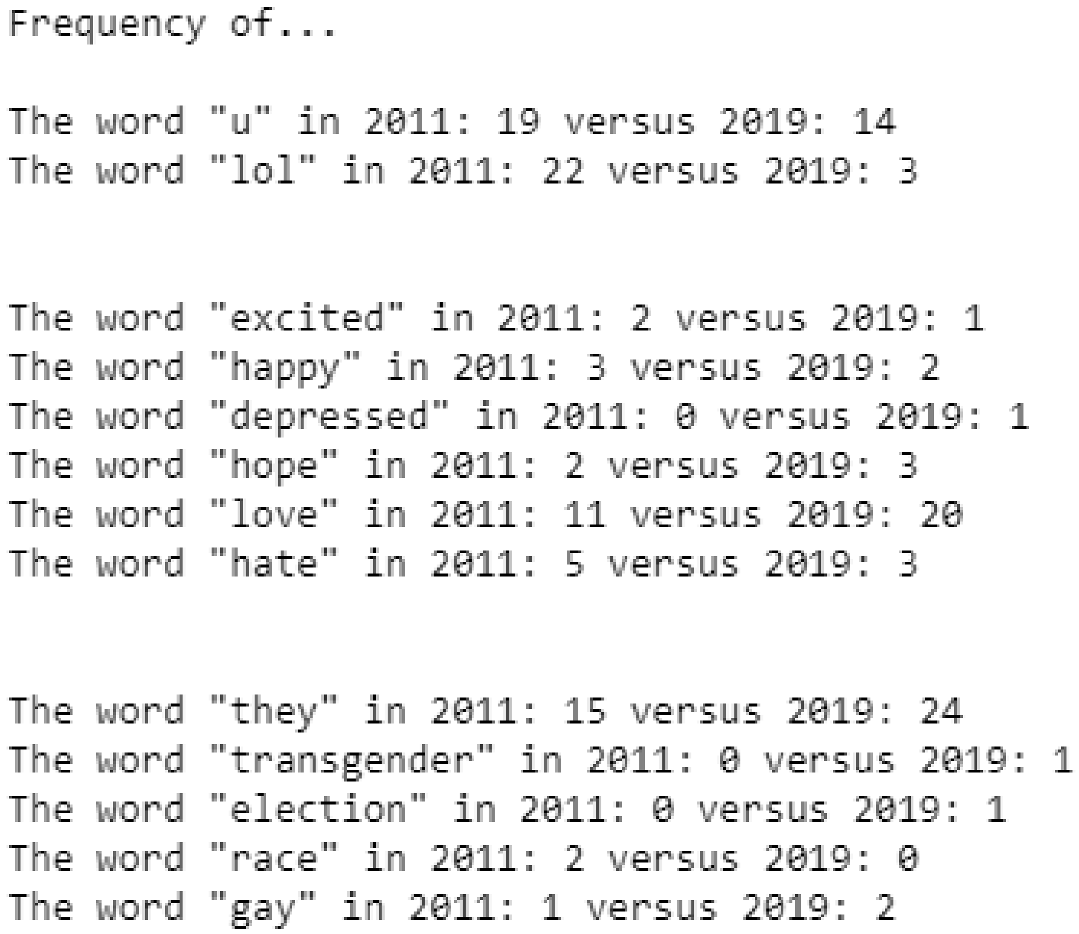
Looking at only most common positive words

VISUALIZATION

This is from my smaller dataset:

I divided them up into a few groups:

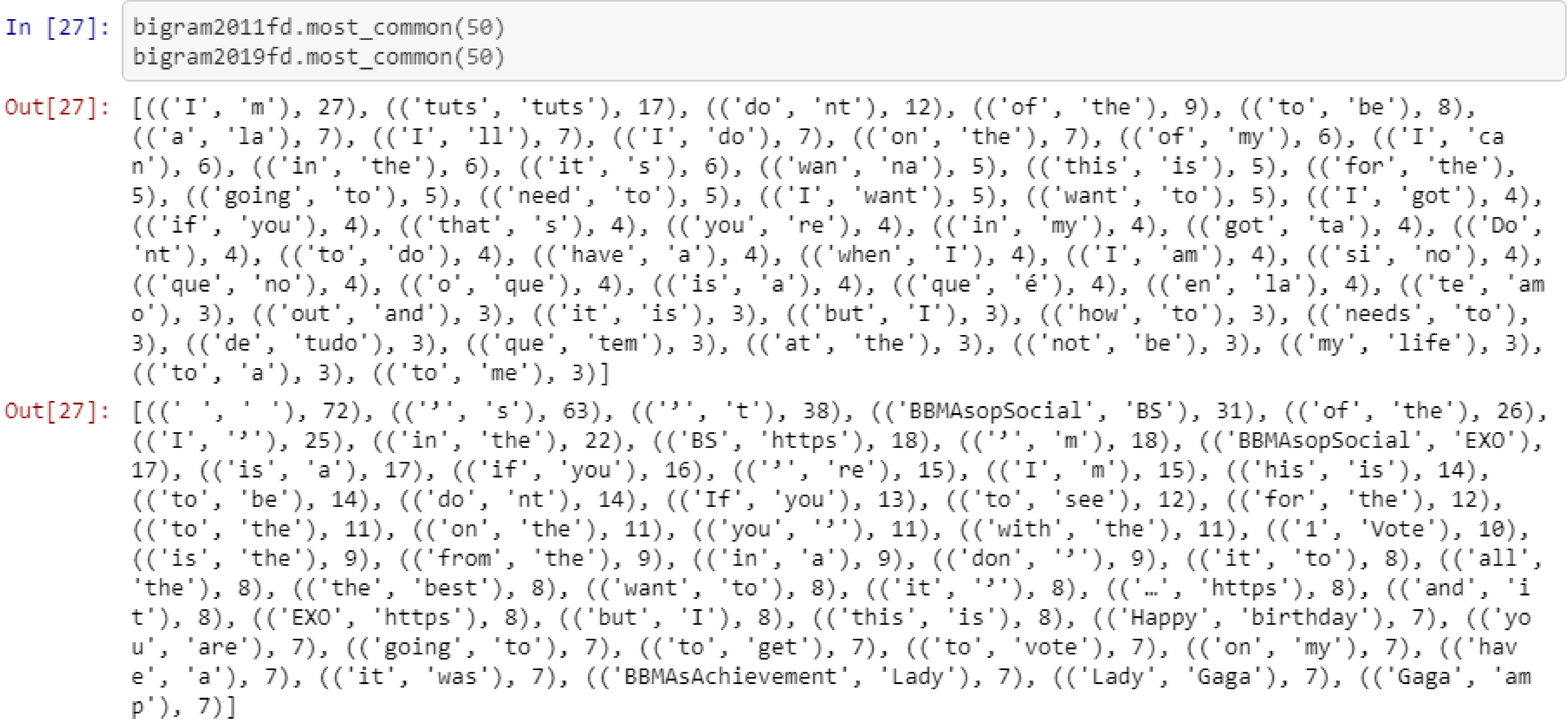
- Texting lingo
- "Feeling" words
- Current topics of conversation in pop culture



```
Frequency of...

The word "u" in 2011: 19 versus 2019: 14
The word "lol" in 2011: 22 versus 2019: 3


The word "excited" in 2011: 2 versus 2019: 1
The word "happy" in 2011: 3 versus 2019: 2
The word "depressed" in 2011: 0 versus 2019: 1
The word "hope" in 2011: 2 versus 2019: 3
The word "love" in 2011: 11 versus 2019: 20
The word "hate" in 2011: 5 versus 2019: 3


The word "they" in 2011: 15 versus 2019: 24
The word "transgender" in 2011: 0 versus 2019: 1
The word "election" in 2011: 0 versus 2019: 1
The word "race" in 2011: 2 versus 2019: 0
The word "gay" in 2011: 1 versus 2019: 2
```

VISUALIZATION

Looking at bigrams



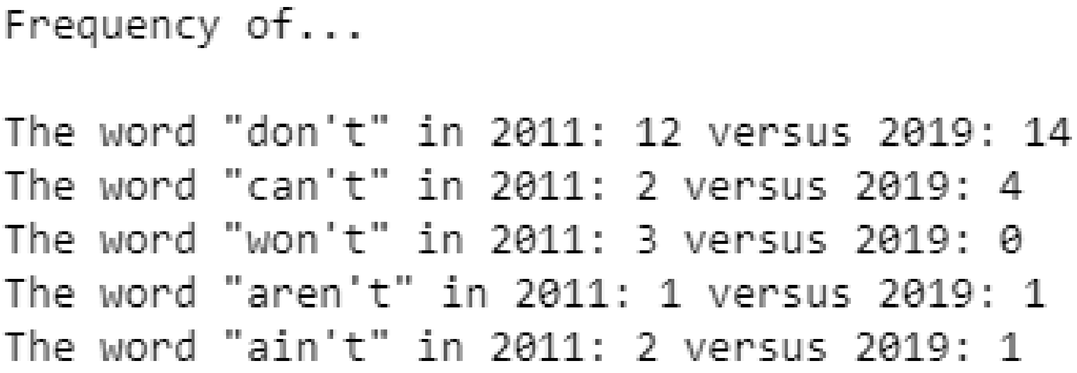
```
In [27]:  bigram2011fd.most_common(50)
          bigram2019fd.most_common(50)
```

```
Out[27]: [(('I', 'm'), 27), (('tuts', 'tuts'), 17), (('do', 'nt'), 12), (('of', 'the'), 9), (('to', 'be'), 8),
          (('a', 'la'), 7), (('I', 'll'), 7), (('I', 'do'), 7), (('on', 'the'), 7), (('of', 'my'), 6), (('I', 'ca
          n'), 6), (('in', 'the'), 6), (('it', 's'), 6), (('wan', 'na'), 5), (('this', 'is'), 5), (('for', 'the'),
          5), (('going', 'to'), 5), (('need', 'to'), 5), (('I', 'want'), 5), (('want', 'to'), 5), (('I', 'got'), 4),
          (('if', 'you'), 4), (('that', 's'), 4), (('you', 're'), 4), (('in', 'my'), 4), (('got', 'ta'), 4), (('Do',
          'nt'), 4), (('to', 'do'), 4), (('have', 'a'), 4), (('when', 'I'), 4), (('I', 'am'), 4), (('si', 'no'), 4),
          (('que', 'no'), 4), (('o', 'que'), 4), (('is', 'a'), 4), (('que', 'é'), 4), (('en', 'la'), 4), (('te', 'am
          o'), 3), (('out', 'and'), 3), (('it', 'is'), 3), (('but', 'I'), 3), (('how', 'to'), 3), (('needs', 'to'),
          3), (('de', 'tudo'), 3), (('que', 'tem'), 3), (('at', 'the'), 3), (('not', 'be'), 3), (('my', 'life'), 3),
          (('to', 'a'), 3), (('to', 'me'), 3)]
```

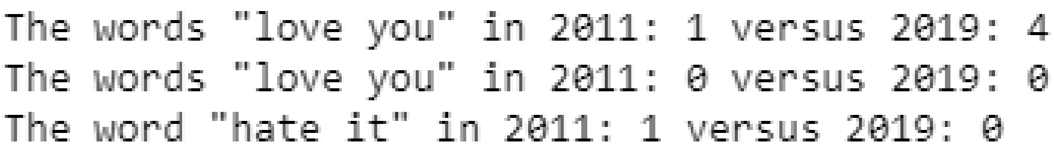
```
Out[27]: [((' ', ' '), 72), (('’', 's'), 63), (('’', 't'), 38), (('BBMAsopSocial', 'BS'), 31), (('of', 'the'), 26),
          (('I', '’'), 25), (('in', 'the'), 22), (('BS', 'https'), 18), (('’', 'm'), 18), (('BBMAsopSocial', 'EXO'),
          17), (('is', 'a'), 17), (('if', 'you'), 16), (('’', 're'), 15), (('I', 'm'), 15), (('his', 'is'), 14),
          (('to', 'be'), 14), (('do', 'nt'), 14), (('If', 'you'), 13), (('to', 'see'), 12), (('for', 'the'), 12),
          (('to', 'the'), 11), (('on', 'the'), 11), (('you', '’'), 11), (('with', 'the'), 11), (('1', 'Vote'), 10),
          (('is', 'the'), 9), (('from', 'the'), 9), (('in', 'a'), 9), (('don', '’'), 9), (('it', 'to'), 8), (('all',
          'the'), 8), (('the', 'best'), 8), (('want', 'to'), 8), (('it', '’'), 8), (('…', 'https'), 8), (('and', 'i
          t'), 8), (('EXO', 'https'), 8), (('but', 'I'), 8), (('this', 'is'), 8), (('Happy', 'birthday'), 7), (('yo
          u', 'are'), 7), (('going', 'to'), 7), (('to', 'get'), 7), (('to', 'vote'), 7), (('on', 'my'), 7), (('hav
          e', 'a'), 7), (('it', 'was'), 7), (('BBMAsAchievement', 'Lady'), 7), (('Lady', 'Gaga'), 7), (('Gaga', 'am
          p'), 7)]
```

VISUALIZATION

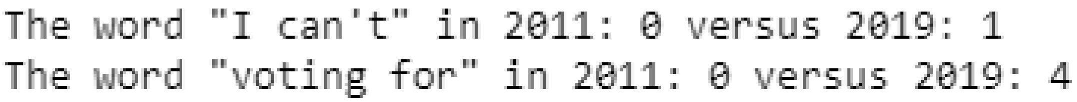
Looking at bigrams



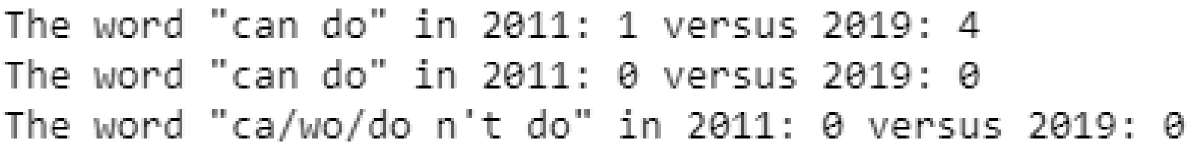
```
Frequency of...

The word "don't" in 2011: 12 versus 2019: 14
The word "can't" in 2011: 2 versus 2019: 4
The word "won't" in 2011: 3 versus 2019: 0
The word "aren't" in 2011: 1 versus 2019: 1
The word "ain't" in 2011: 2 versus 2019: 1
```



```
The words "love you" in 2011: 1 versus 2019: 4
The words "love you" in 2011: 0 versus 2019: 0
The word "hate it" in 2011: 1 versus 2019: 0
```



```
The word "I can't" in 2011: 0 versus 2019: 1
The word "voting for" in 2011: 0 versus 2019: 4
```



```
The word "can do" in 2011: 1 versus 2019: 4
The word "can do" in 2011: 0 versus 2019: 0
The word "ca/wo/do n't do" in 2011: 0 versus 2019: 0
```

PROJECT EVOLUTION

- Initially wanted to look at "positivity", but that was too vague which made it hard to pin down what kind of analysis I should do.
- Decided to shift focus on comparing the two different time periods in general. Zeroed in on topics of the tweets and contextualized them with the positivity analysis.

WHAT'S NEXT?

Now that I feel comfortable wrangling messy tweets it would be cool to look into…

- User demographics when it comes to some of the types of tweets that I analyzed.
- There's always more improve with how I cleaned data and how I classified it.
- Analyze texting lingo specifically instead of ignoring it.
- Lot's of things!