A STUDY OF VECTORIZATION OPTIONS

HOW WE FEEL ABOUT ARTIFICIAL INTELLIGENCE (and cats)

Photo via Brett Jordan | Unsplash

Introduction

What is 'Artificial Intelligence?' At the nexus of machines and humans is this strange hard-to-grasp, even-harder-to-quantify blanket term called Artificial Intelligence. Once a Hollywood blockbuster depicting one of the many strange futures and concepts that is Artificial Intelligence, it is now a silicon valley buzzword, like bitcoin or blockchain, used to excite stakeholders and artificially increase valuations.

In reality, Artificial Intelligence is considerably less glamorous. Artificial Intelligence is simply taking advantage of computers (no, not in that way iRobot enthusiasts) by utilizing their computing power across many different things that would be far too tedius (and error prone) for a human to do.

For example, let’s say we want to know how the world feels about the President of the United States. In the olden days, before things like mass communication, computers and the internet, we might have to walk door to door, ring the doorbell, interview the inhabitants, take notes, and return to our university where we would manually sift through the notes pulling out words that might seem more “positive” or “negative” in nature. This could be manageable for one 2nd grader on his/her cul de sac, (I’d venture she’d disagree, though) but on a large scale, this is nearly impossible.

Let's pretend for a minute that we can magically snap our fingers and get a sentence from each person. If each person in the United States simply wrote one sentence about the President, we’d have over 300 million sentences to review. Even if it magically (call Hogwarts) took us one second to review and categorize each sentence, and we worked around the clock, it would take us over 9 years to do this -- and by then, we’d have a different president! Not only is this nearly impossible, it is quite ineffective. Computers, on the other hand, are quite effective at tasks like this.

Computers are absolutely amazing at menial tasks -- especially counting things. Computers are also very good at doing math quickly and efficiently with numbers too large even for our very expensive T.I. calculators. Computers have a lot of other skills but that is slightly (ahem, well) beyond the scope of this research paper. In short, Artificial Intelligence is using computers and machines to do things humans can’t do as well, and often using things like counting and math to train computers to do even more amazing things.

Analysis & Models

ABOUT THE DATA:

Four different datasets were used in this study -- five if the ‘Example Analysis’ sentence is included (however, this sentence was not used for analysis, only for education). The initial non-educational-only data set contained a single folder with two different text documents -- Cats.txt and Spinach.txt. As one might correctly assume, the Cats.txt document had a small blurb about cats, while the Spinach.txt document had a small blurb about spinach.

Cats.txt

```
"Kendra loves cats. In fact, she has TEN cats. If she didn't have a house, a
husband and a graduate degree in data science, she'd be a cat lady!"
```

Spinach.txt

```
'Wow. Spinach is great. Not just for cartoon sailors. Interestingly, one of my
cats loves spinach, too! So does my husband.'
```

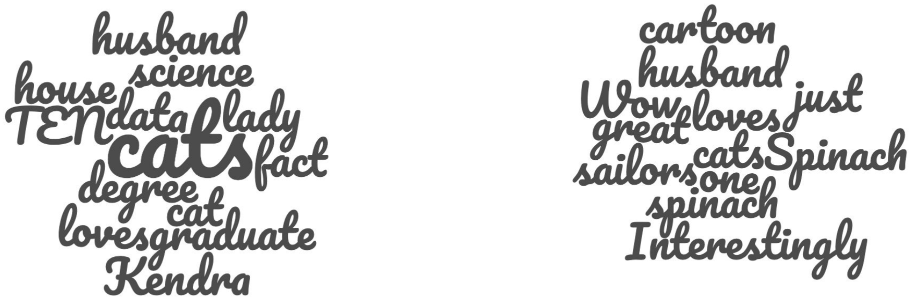


FIG 1: WordClouds -- Very verbiage. Much cumulus. Wow.

The second dataset was created to mimic social perceptions of Artificial Intelligence in the form of tweets or comments. Two distinct corpuses were created -- one for negative sentiment, one for positive sentiment.

Here they are in word cloud form:

NEGATIVE

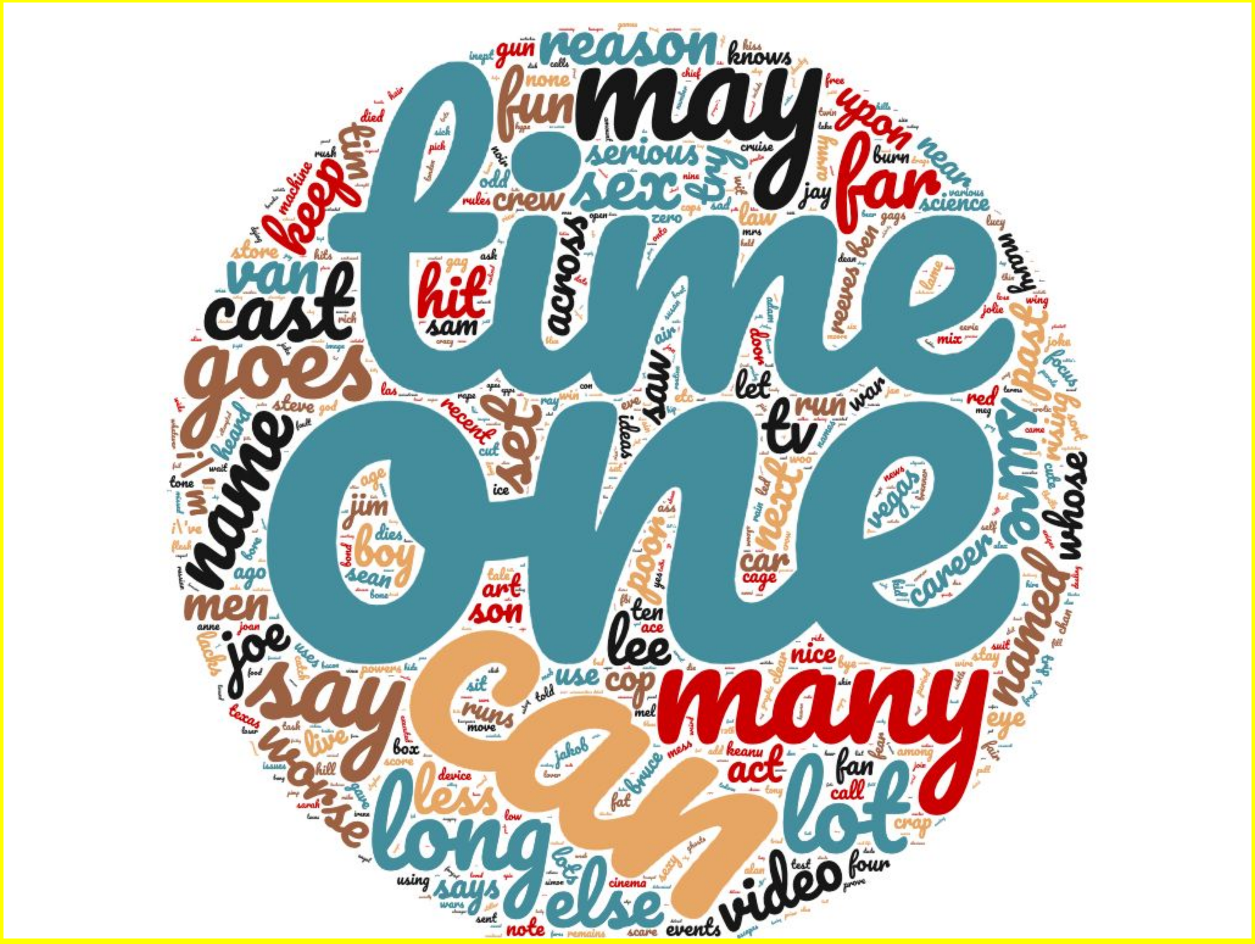
TOOK
terrified
melt
hack
JOBS
can
RIGHT
trust
even
hate
intelligence
brains
cars
artificial
drive
children
dog
going
captcha

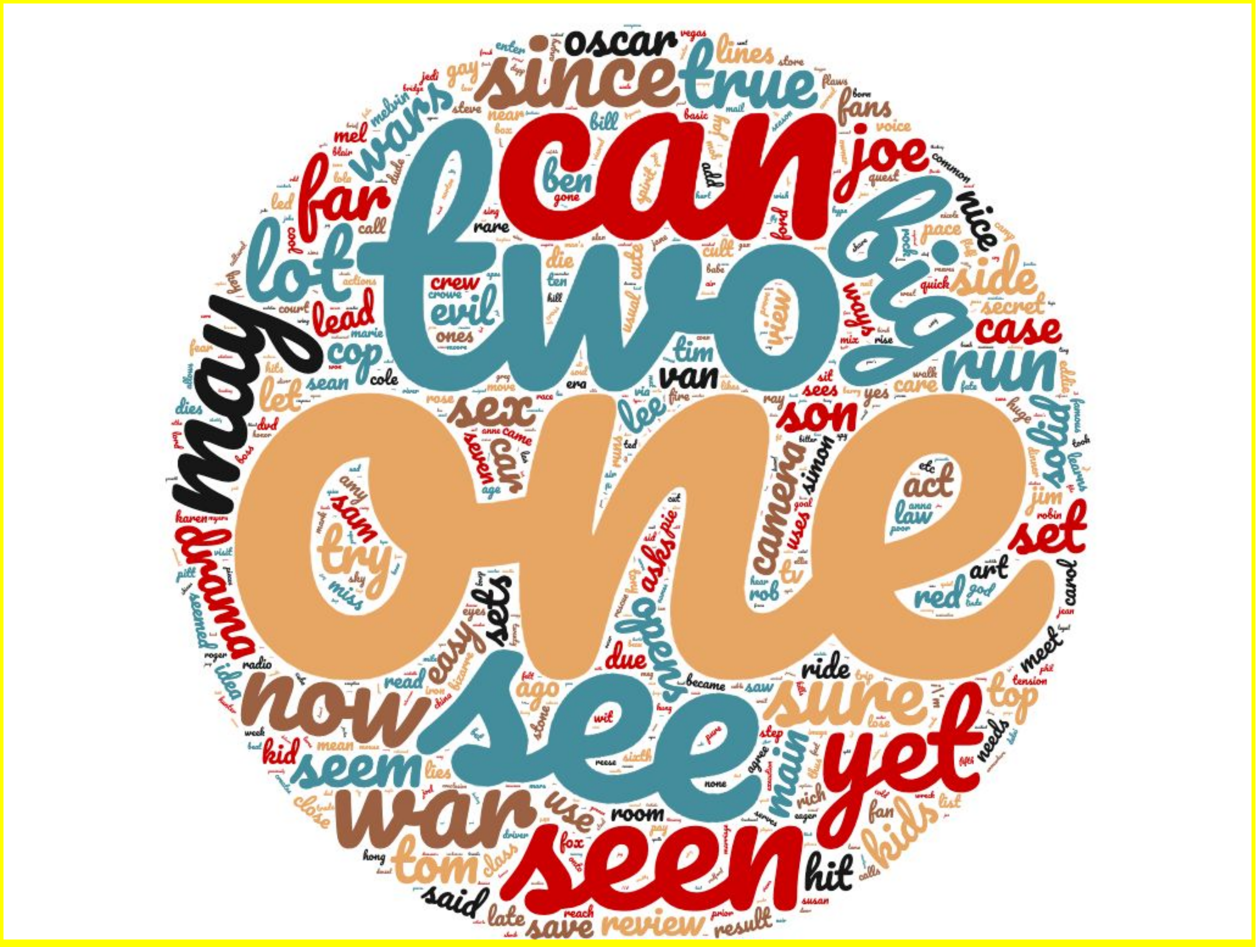
advancements
Order
groceries
taxes
kids
day
child
school
world
grifty
grow
grateful
nifty
pay
excited
life
love
Yes
thanks
please
Artificial
likely
take
Intelligence
will
back
cancer-free
taken
dog
every
daydream
artificial
given
care
time
intelligence

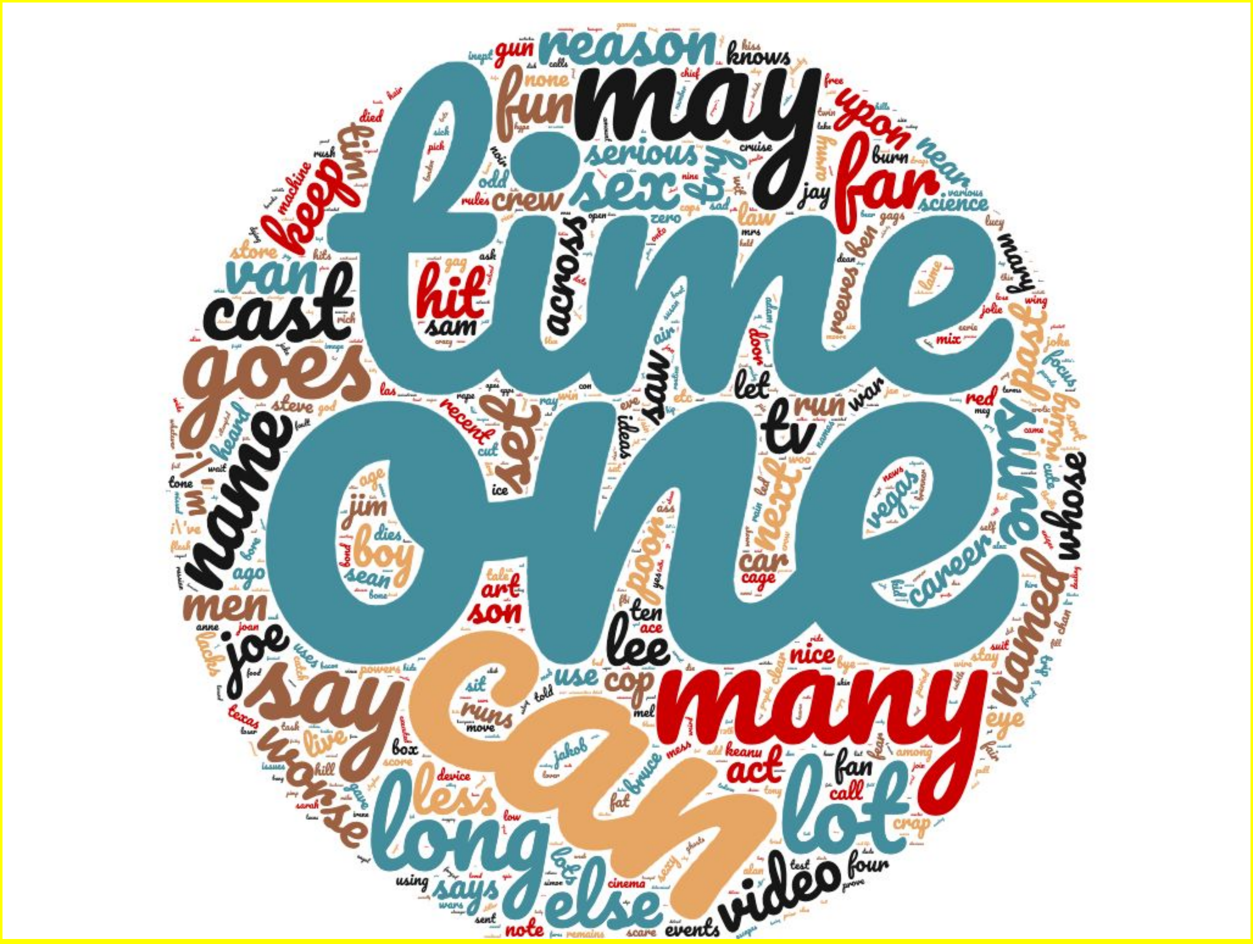
POSITIVE

The fourth data set was “Ami’s Data,” as referenced throughout. This dataset was used as an additional benchmark when comparing the accuracy of different Sentiment Analysis tools. Similarly to the second (and main) dataset, Ami’s Data consists of two labeled mini-corpuses of 5 txt files each, one for positive movie reviews, one for negative movie reviews.

The final data set was “Cornell Data,” as referenced throughout. This dataset was found via a Google search [“free sentiment analysis data sets”]. The data set contains two corpuses, (one positive, one negative) each with 1000 txt files containing the text of either a negative or positive review. The data was eventually labeled and combined to form a data frame of 2000 labeled reviews.


time
one
can
many
may
reason
lot
long
else
video
say
goes
name
cast
van
keep
far
sex
fun
upon
hit
across
set
tv
run
next
named
past
army
whose
lee
men
joe
worse
less
act
fan
boy
poor
career
vegas
jim
son
art
cop
use
runs
live
says
bruce
serious
crew
tiny
near
science
store
steve
heard
ago
red
war
let
saw
recent
ideas
jakob
device
note
events
cinema
scare
remains
four
crap
call
among
eye
stay
nice
lacks
texas
powers
kill
score
issues
sarah
using
wars
sent
box
move
told
sit
ten
mel
ace
keanu
mess
fat
lover
low
tony
test
alan
yes
cage
car
dies
sean
bone
age
cut
ice
ass
woo
main
news
rising
focus
joke
mix
eerie
jolie
wing
reeves
ben
law
lame
door
air
mrs
zero
open
jay
sad
nine
caps
rules
odd
sick
pick
died
machine
rush
tim
gag
ask
sam
rich
image
crazy
las
rape
god
fault
wait
tone
flash
anne
joan
uses
kids
bacon
lady
cruise
burn
twin
lake
chief
free
kiss
knows
calls
none
gun
inept
noir
hype
hits
gags
beer
dean
lucy
various
suit
fill
chan
bye
wire
fair
joe
flick
prove
sexy
plenty
bob


one
two
can
see
seen
yet
war
now
may
big
since
true
drama
run
lot
far
sure
wars
oscar
joe
side
nice
case
evil
crew
lead
cop
let
sex
lee
son
van
solid
camera
act
law
set
sam
try
opens
easy
sets
red
art
rob
asks
seem
kids
tom
hit
said
review
save
late
result
room
fox
use
main
ride
top
meet
needs
fan
rich
kid
mean
close
lies
read
ago
due
pie
cute
usual
cult
ben
bill
melvin
mel
fans
lines
gay
jedi
steve
voice
common
quest
basic
view
secret
pace
fear
court
marie
ones
tim
sean
cole
dies
dvd
karen
amy
miss
sky
eyes
seemed
idea
radio
roger
pitt
visit
cool
led
lots
blair
call
rare
die
ten
gone
era
via
fire
ray
sees
yes
care
jim
robin
poor
carol
jean
god
eddie
huge
famous
learns
took
simon
uses
tv
agree
step
tension
trip
became
none
reese
sixth
saw
stone
wit
driver
class
hong
susan
reach
prior
eager
list
wreck
lose
rise
walk
quick
mix
store
vegas
enter
near
box
mail
dude
actions
crowe
seven
car
came
anne
age
ted
killer
goal
runs
boss
fate
etc



PREPROCESSING THE DATA:

What are words? What is data? How can we turn data into words? The computer doesn't know what a cat is. How can we tell the computer what a cat is?! There is so much the computer doesn't know and so much we need to tell the computer. How can we turn words into something the computer can understand? Well, first, we count them. How can we count words you ask? Excellent question! We first take the words in regular sentence form (just like this here paragraph) and we turn the words into **tokens**. That's correct -- we **Tokenize** the words. This is simply turning each word into a data point. And it doesn't stop just at words. We can capture punctuation, too! Let's look at an example.

VECTORIZATION via TOKENIZATION:

Tokenization is simply breaking text down into "tokens" which, in this case, is words! *NOTE: Tokens can include things like punctuation, but we'll get to that a little later.*

EXAMPLE ANALYSIS:

Let's start with a simple (but highly relatable) sentence.

INPUT:

```
example_text = "Kendra loves cats. In fact, she has TEN cats. If she didn't
have a house, a husband and a graduate degree in data science, she'd be a
cat lady!"
```

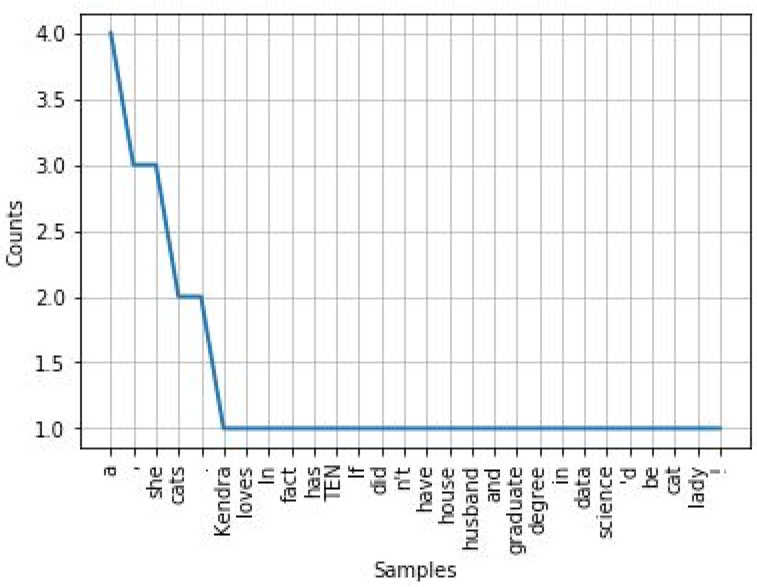
OUTPUT TOKENS:

```
['Kendra', 'loves', 'cats', '.', 'In', 'fact', ',', 'she', 'has', 'TEN',
'cats', '.', 'If', 'she', 'did', "n't", 'have', 'a', 'house', ',', 'a',
'husband', 'and', 'a', 'graduate', 'degree', 'in', 'data', 'science', ',',
'she', "'d", 'be', 'a', 'cat', 'lady', '!']
```

These just look like the words in the **INPUT** sentence, yes? Well, currently, they are. But now, we can treat the tokenized words as data points and do fun (and illuminating!) things to -- like counting! -- and with them!

FREQUENCY DISTRIBUTIONS:

What if I asked you to find the most frequent word in our super fun example sentence? You'd have to manually count, likely using your finger or even a pencil if you printed it out like a luddite, each word and somehow keep track of both the word and the number of times it occurred in the sentence. Now, this is a fairly simple, albeit tedious, task for this small sentence. However, what if you had many sentences? What if you had hundreds of millions of sentences? Supposing it takes you 3 minutes to catalog our superfun sample sentence, you'd be stuck cataloging our hundreds of millions of sentences until you were gray in the hair and face! Thankfully, computers are wizards when it comes to counting words and we can have the computer quickly do this for us using a handy function called **FREQUENCY DISTRIBUTIONS** (which is just a fancy way of saying "counting" (the frequency)).

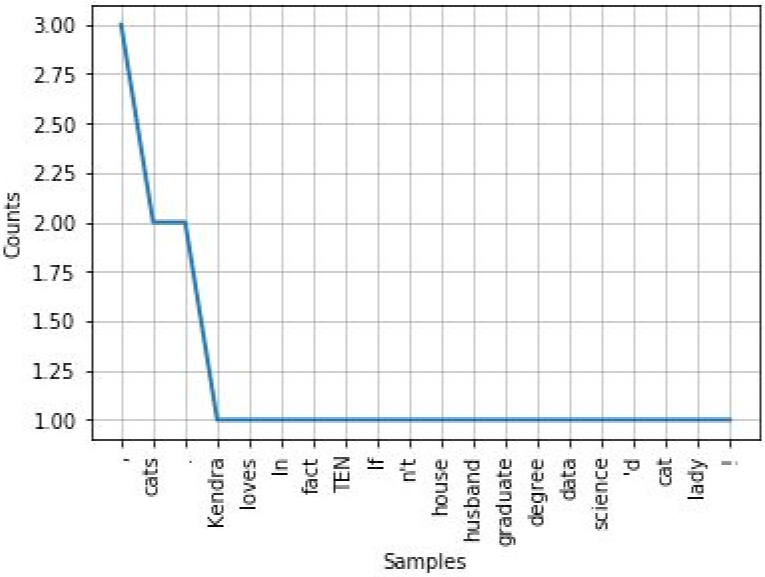


To answer our question above **What is the most frequently occurring word in this collection of sentences** we can see from our graph that it's the word "a."

But wait, "a" isn't a super helpful word for us when analyzing sentences. Words like "a" and "the" aren't very helpful when it comes to determining things like content or sentiment, so these words are called **STOPWORDS**.

STOPWORDS:

After we remove the stopwords, we can get a better understanding of what this sentence is about.



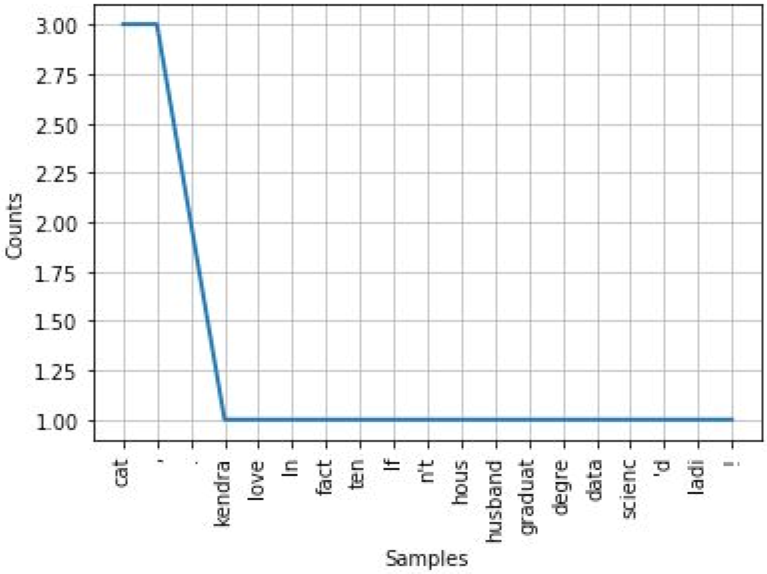
Interestingly, it looks like a punctuation mark is what’s most frequently used in these sentences. That doesn’t really help us so we can add that to our list of stop words OR we can use an even handier method **`isalpha()`** which we discovered slightly late in the game.

Also, wait, doesn't the word "cat" appear more than 2 times?

Oh! That's because we have "cat" AND "cats" which the computer is counting as two different words! Introducing…

STEMMING:

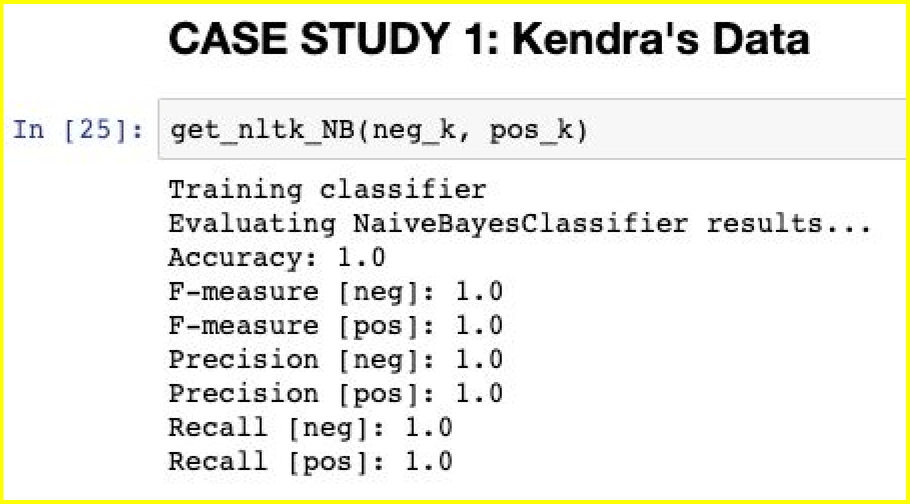
Stemming (and Lemmatization, which we will visit a little later) are both ways we reduce words down to the base word. So in this instance, changing “cats” to “cat” (because in this example, the plural of cats doesn’t impact the meaning or our overall goal of counting the occurrences of references to “cat”)



Ah, finally. We can clearly see that the subject of this example sentence is cats. Hooray! Now to incorporate our actual data.

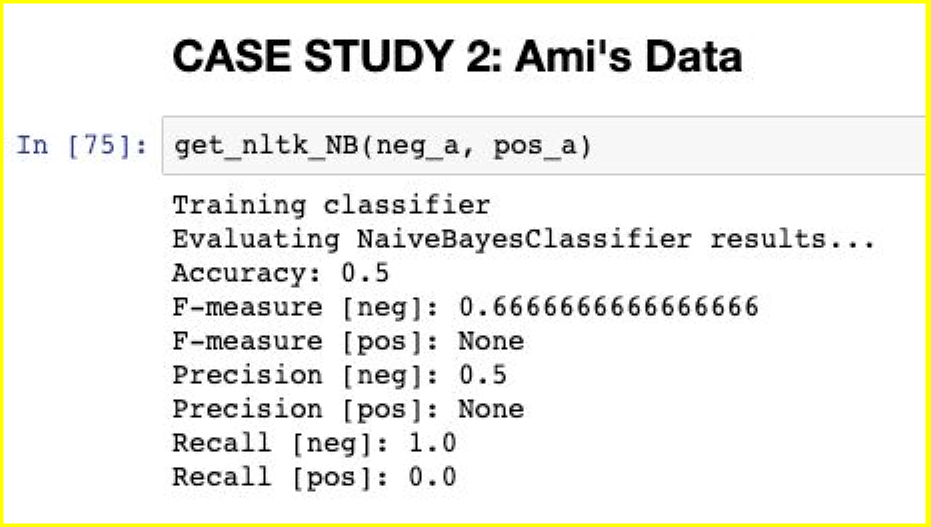
MACHINE LEARNING:

NLTK SENTIMENT ANALYSIS:



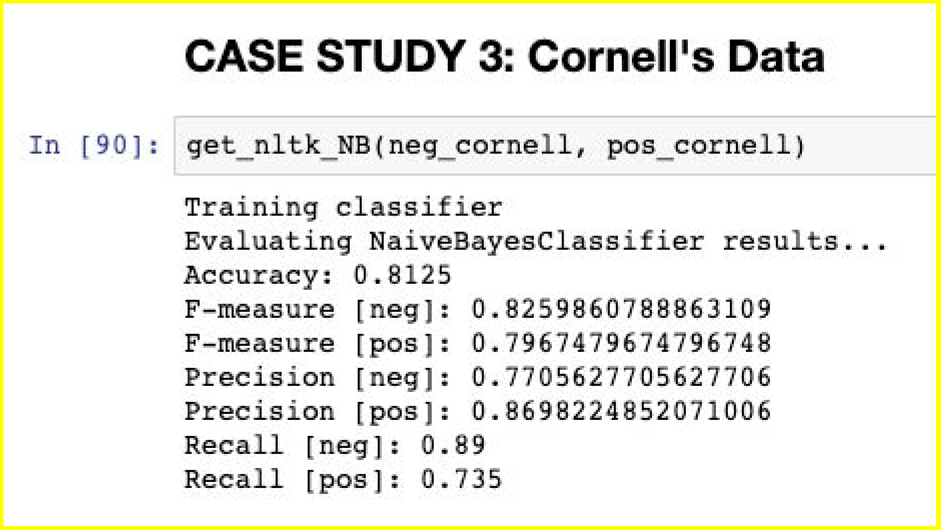
CASE STUDY 1: Kendra's Data

```
In [25]: get_nltk_NB(neg_k, pos_k)

         Training classifier
         Evaluating NaiveBayesClassifier results...
         Accuracy: 1.0
         F-measure [neg]: 1.0
         F-measure [pos]: 1.0
         Precision [neg]: 1.0
         Precision [pos]: 1.0
         Recall [neg]: 1.0
         Recall [pos]: 1.0
```



CASE STUDY 2: Ami's Data

```
In [75]: get_nltk_NB(neg_a, pos_a)

         Training classifier
         Evaluating NaiveBayesClassifier results...
         Accuracy: 0.5
         F-measure [neg]: 0.6666666666666666
         F-measure [pos]: None
         Precision [neg]: 0.5
         Precision [pos]: None
         Recall [neg]: 1.0
         Recall [pos]: 0.0
```



CASE STUDY 3: Cornell's Data

```
In [90]: get_nltk_NB(neg_cornell, pos_cornell)

         Training classifier
         Evaluating NaiveBayesClassifier results...
         Accuracy: 0.8125
         F-measure [neg]: 0.8259860788863109
         F-measure [pos]: 0.7967479674796748
         Precision [neg]: 0.7705627705627706
         Precision [pos]: 0.8698224852071006
         Recall [neg]: 0.89
         Recall [pos]: 0.735
```

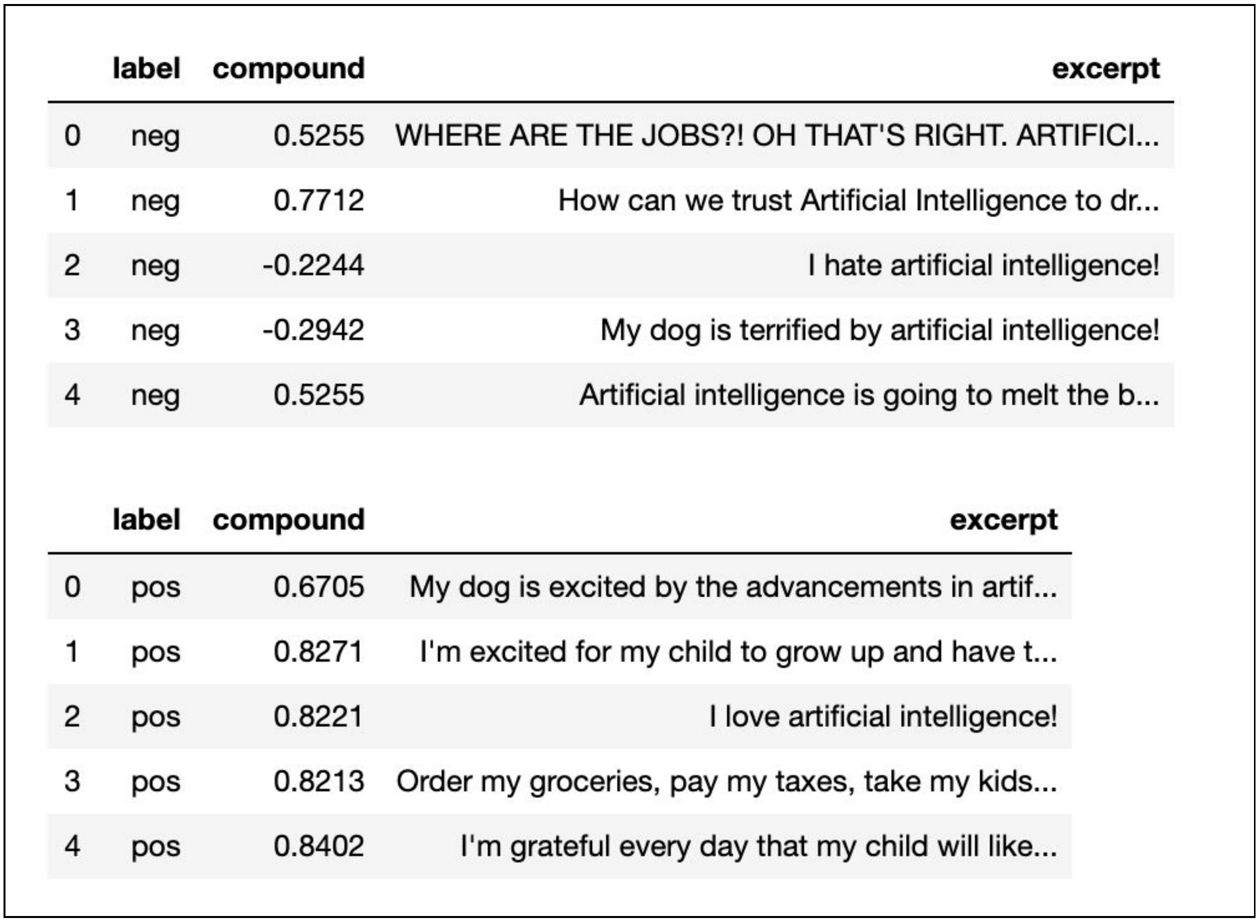
As seen above, this classifier is not nearly as useful on smaller datasets as some of the other classifiers covered below (see: Vader). Because both the "Kendra Data" and "Ami Data" contain only 5 examples of positive or negative, the separation of test and train is extremely challenging. If we split it into 3 and 2 for train and test, we don't have a lot of data to give our model. If we split it into 4 and 1 for train and test, we are literally only testing one thing meaning our results will either be an aggressive over-exaggeration (100% correct!) or the deflating 0% accuracy. However, with a larger data set (the Cornell data), it's clear to see this classifier, right OOTB, is doing something right because we're starting with an 81% accuracy. In summation: Use this classifier on larger data sets.

VADER:

See full overview (including code) here

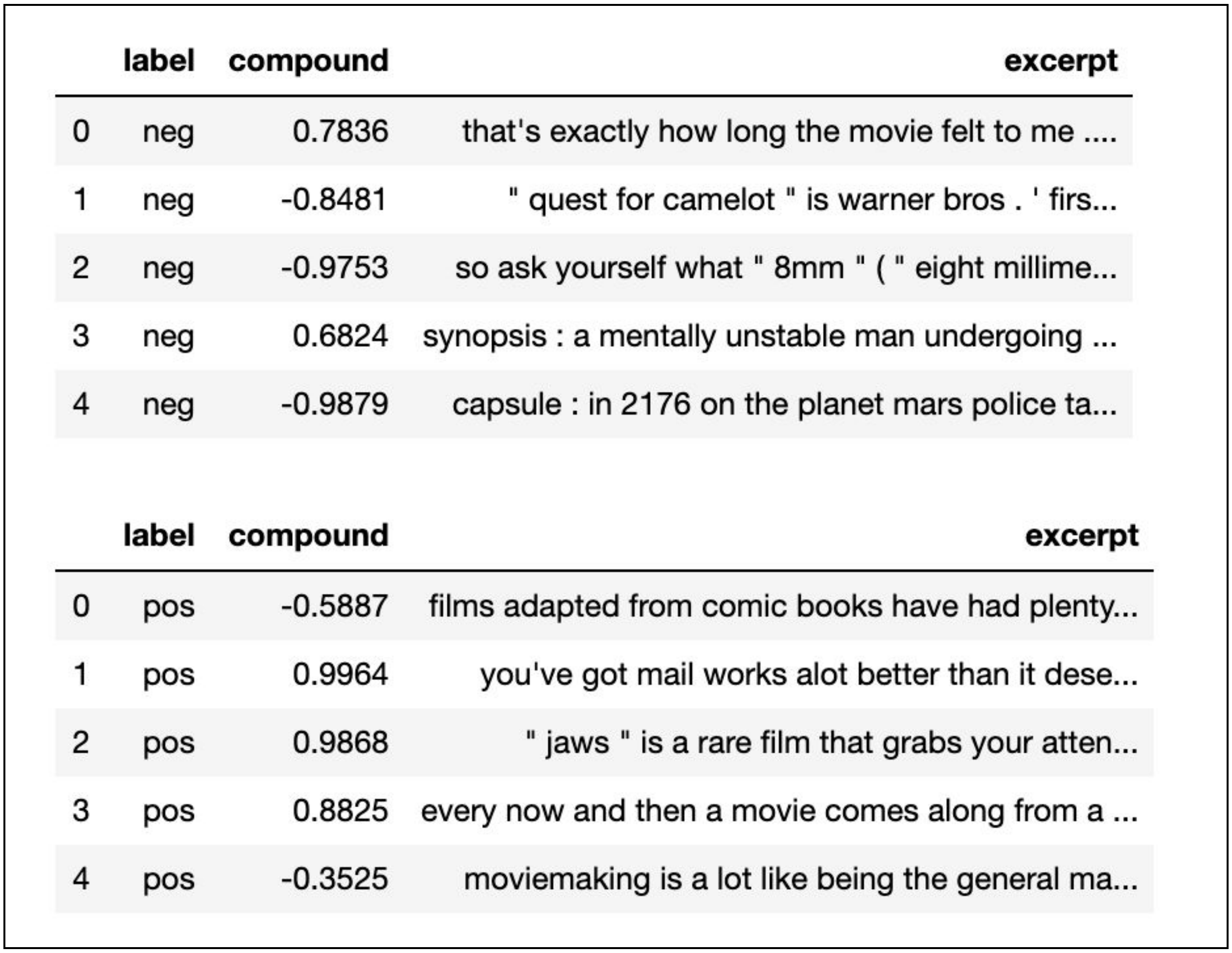
https://danielcaraway.github.io/html/HW1_viathedocs_vader_kdata.html

As a quick litmus test to the accuracy of Vader, two different datasets were used.



	label	compound	excerpt
0	neg	0.5255	WHERE ARE THE JOBS?! OH THAT'S RIGHT. ARTIFICI...
1	neg	0.7712	How can we trust Artificial Intelligence to dr...
2	neg	-0.2244	I hate artificial intelligence!
3	neg	-0.2942	My dog is terrified by artificial intelligence!
4	neg	0.5255	Artificial intelligence is going to melt the b...

	label	compound	excerpt
0	pos	0.6705	My dog is excited by the advancements in artif...
1	pos	0.8271	I'm excited for my child to grow up and have t...
2	pos	0.8221	I love artificial intelligence!
3	pos	0.8213	Order my groceries, pay my taxes, take my kids...
4	pos	0.8402	I'm grateful every day that my child will like...



	label	compound	excerpt
0	neg	0.7836	that's exactly how long the movie felt to me ....
1	neg	-0.8481	" quest for camelot " is warner bros . ' firs...
2	neg	-0.9753	so ask yourself what " 8mm " (" eight millime...
3	neg	0.6824	synopsis : a mentally unstable man undergoing ...
4	neg	-0.9879	capsule : in 2176 on the planet mars police ta...

	label	compound	excerpt
0	pos	-0.5887	films adapted from comic books have had plenty...
1	pos	0.9964	you've got mail works alot better than it dese...
2	pos	0.9868	" jaws " is a rare film that grabs your atten...
3	pos	0.8825	every now and then a movie comes along from a ...
4	pos	-0.3525	moviemaking is a lot like being the general ma...

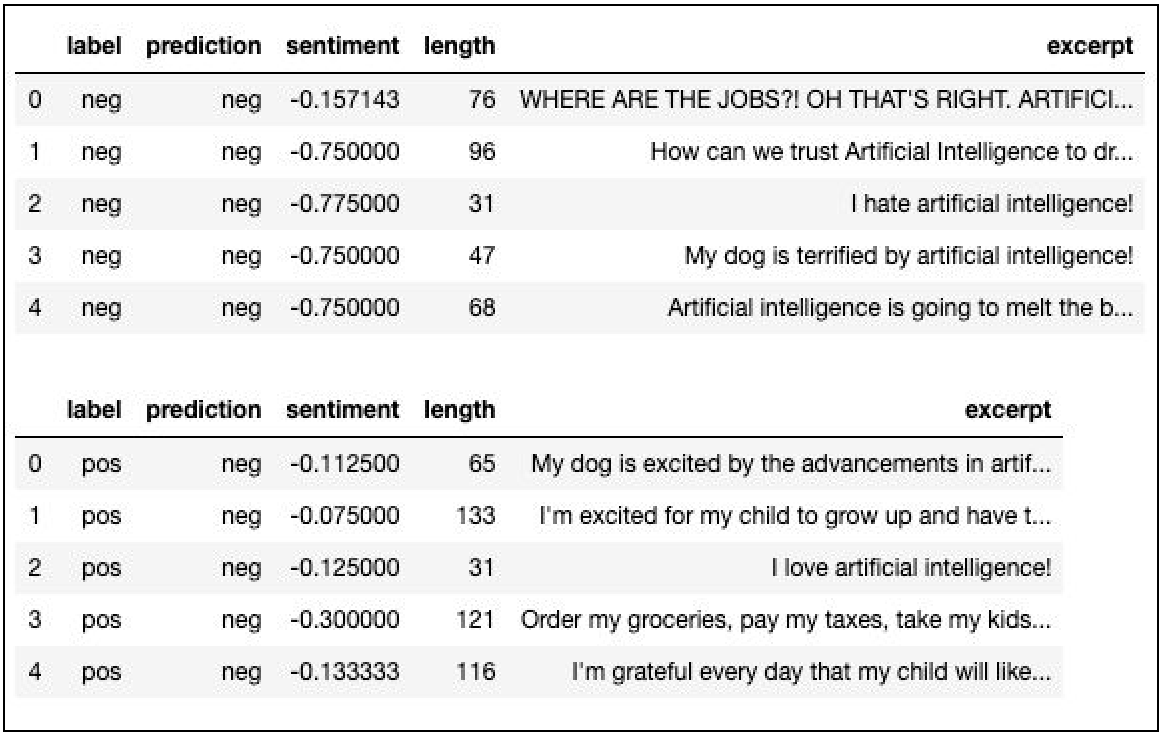
In summation: Vader isn't the most accurate at classification. However, it is a useful tool for creating labels out of unlabeled data as we will continue to see as we pull the "sentiment analysis" thread over the next few papers.

TextBlob:

See full overview (including code) here
https://danielcaraway.github.io/html/HW1_textblob_v3.html

As a quick litmus test to the accuracy of TextBlob, two different datasets were used.

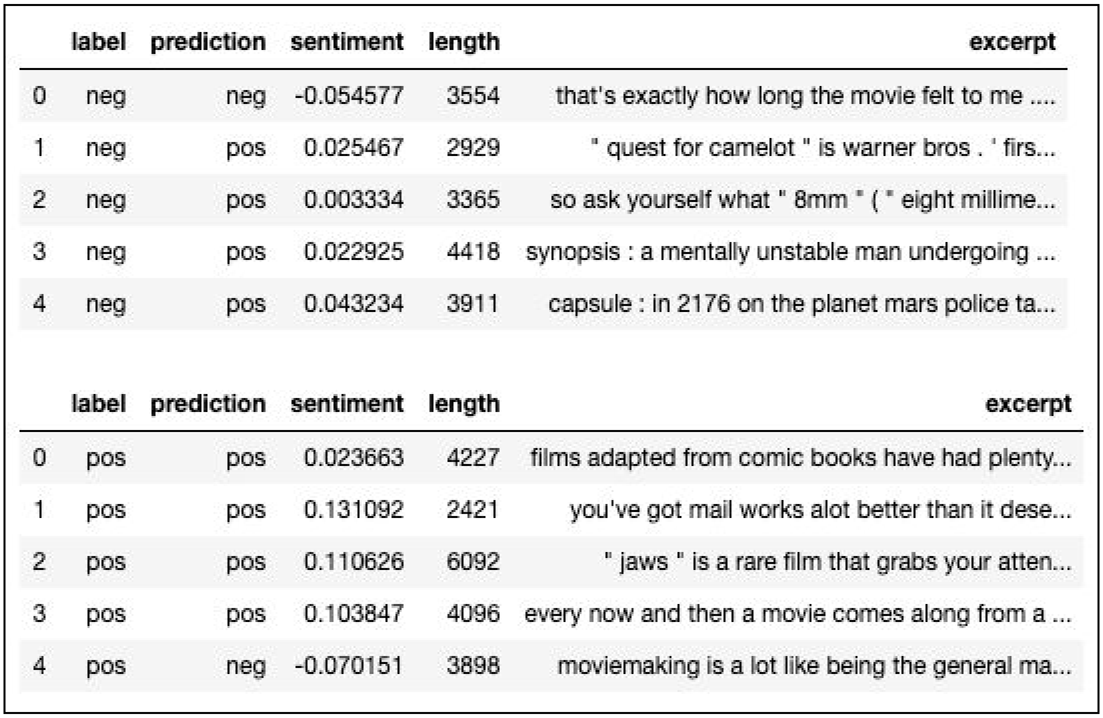
CASE STUDY 1: Kendra's Data:



	label	prediction	sentiment	length	excerpt
0	neg	neg	-0.157143	76	WHERE ARE THE JOBS?! OH THAT'S RIGHT. ARTIFICI...
1	neg	neg	-0.750000	96	How can we trust Artificial Intelligence to dr...
2	neg	neg	-0.775000	31	I hate artificial intelligence!
3	neg	neg	-0.750000	47	My dog is terrified by artificial intelligence!
4	neg	neg	-0.750000	68	Artificial intelligence is going to melt the b...

	label	prediction	sentiment	length	excerpt
0	pos	neg	-0.112500	65	My dog is excited by the advancements in artif...
1	pos	neg	-0.075000	133	I'm excited for my child to grow up and have t...
2	pos	neg	-0.125000	31	I love artificial intelligence!
3	pos	neg	-0.300000	121	Order my groceries, pay my taxes, take my kids...
4	pos	neg	-0.133333	116	I'm grateful every day that my child will like...

CASE STUDY 2: Ami's Data:



	label	prediction	sentiment	length	excerpt
0	neg	neg	-0.054577	3554	that's exactly how long the movie felt to me ....
1	neg	pos	0.025467	2929	" quest for camelot " is warner bros . ' firs...
2	neg	pos	0.003334	3365	so ask yourself what " 8mm " (" eight millime...
3	neg	pos	0.022925	4418	synopsis : a mentally unstable man undergoing ...
4	neg	pos	0.043234	3911	capsule : in 2176 on the planet mars police ta...

	label	prediction	sentiment	length	excerpt
0	pos	pos	0.023663	4227	films adapted from comic books have had plenty...
1	pos	pos	0.131092	2421	you've got mail works alot better than it dese...
2	pos	pos	0.110626	6092	" jaws " is a rare film that grabs your atten...
3	pos	pos	0.103847	4096	every now and then a movie comes along from a ...
4	pos	neg	-0.070151	3898	moviemaking is a lot like being the general ma...

As the data shows, TextBlob does **not** do an accurate job of determining sentiment. In Kendra's data, literally everything (from both sets!) was categorized as negative. In Ami's data, regardless of the set, 4 of the 5 were predicted to be positive and one loner was negative. In Kendra's data, it correctly predicted 50% of the time. In Ami's data, it correctly predicted 50% of the time as well. This is not a good accuracy.

COUNTVECTORIZER:

Countvectorizer a sparse matrix. We will use this at a later date (HW2) and evaluate it then.

Results

PREDICTIVE ANALYTICS:

This analysis found very little useful information. Due to the small sample size, the researchers were unable to come up with a reliable predictive model. Any actionable suggestions based on the results would be misguided as the sample size was simply too small to be of any use to the client.

PRESCRIPTIVE ANALYTICS:

If the goal is to be able to predict sentiment, more "training data" is needed for the models. However, if the goal is simply ascribing sentiment to existing data (or if the only data available is a small data set), something like vader can be used to judge/ascribe/prescribe sentiment (but again, not predict sentiment).

TL;DR: (Should this be part of the conclusion? I still talk about data... as in not having enough of it)

To truly get an accurate litmus test of the current zeitgest's feelings towards artificial intelligence, the client needs to increase her budget to allow the researchers to gather more data. With additional data, the researchers would be better equipped to run more thorough predictive models and equip the client with a more comprehensive list of "action items."

Speaking of “action items,” the researchers need to better understand the goals of the client --is the client trying to launch a new startup in the AI space? Is the client attempting to rebrand their currently-market-as-AI-product? Is the client simply looking to gather data about this field for a research project with a larger scope? -- in order to truly help the client, the researchers would need answers to these or similar questions.

What to do with this information: Regarding predictive analytics, more data is needed from all over the web. This data might come from twitter feeds, Facebook posts, comments on news articles and google search alerts. The researchers suggest the client hire an outside contractor to build benign bots to track certain keywords in each of these social pools and return to the researchers with the more robust dataset.

Conclusion

The current public sentiment regarding Artificial Intelligence is very polarized -- the sentiment is either extremely favorable and hopeful or highly wary and suspicious. Additionally there is a clear divide in how each party (each pole) choose to define both “Artificial Intelligence” and the perceived “benefits” or “risks” of the wide adoption and usage of this [insert their definition of “Artificial Intelligence”]. As this topic is fraught with intermingling definitions and is tightly connected to many more high-profile red button issues (e.g. job creation/loss, fake news, election tampering etc.), the researchers suggest avoiding this topic at family gatherings or Thanksgiving dinners.