Comparing NLP Algorithms

Multinomial Naïve Bayes, SVM, LDA, NER

Goals

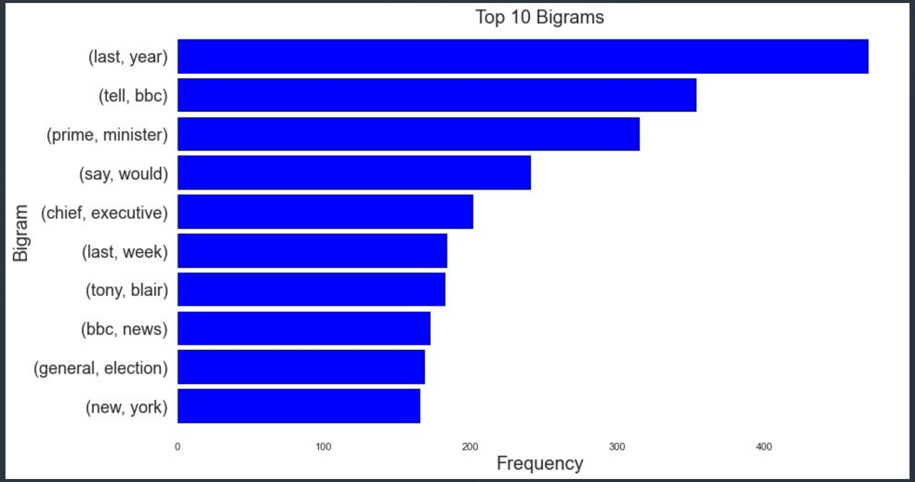
- Train NLP models for use in production data pipeline for unstructured scientific and technical data
  - Automate pre-processing
  - Text Classification
  - Topic Modeling
  - N.E.R. for metadata tagging
- Determine feasibility of LDA for Topic Modeling in production pipeline
- Use Named Entity Recognition to extract metadata

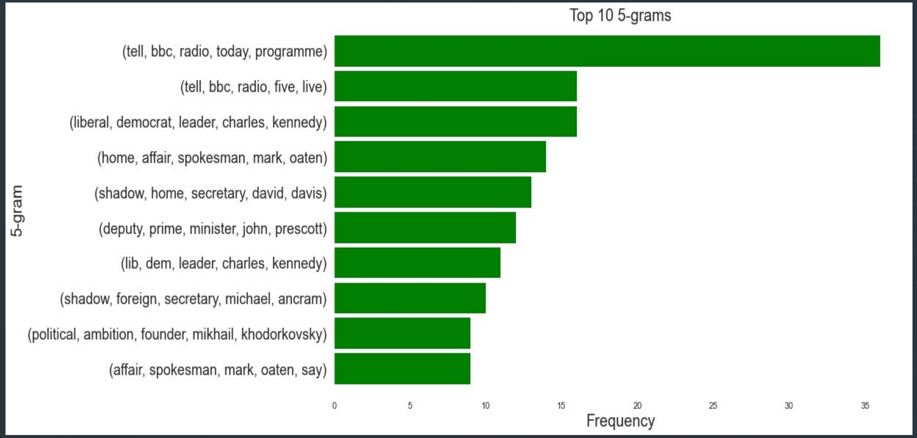
The Data Set

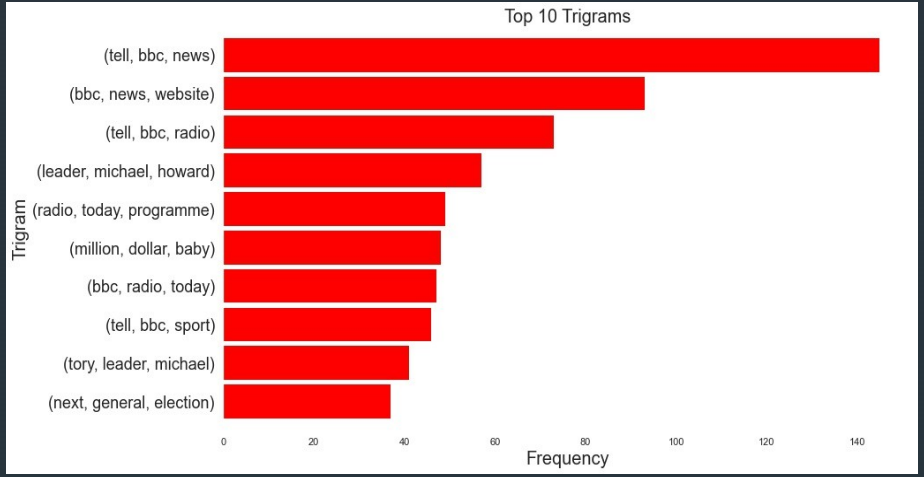
- Consists of 2225 documents from the BBC news website corresponding to stories in five topical areas from 2004-2005.Natural Classes: 5 (business, entertainment, politics, sport, tech)
- Data set came from an academic research project that I can’t even begin to comprehend
  - D. Greene and P. Cunningham. "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering", Proc. ICML 2006.
- Series of text documents in folders grouped by topic
- Required very little wrangling (read in text, strip 1st line as title, split filepath to get topic, store in dataframe)
- Data Set stats (after trimming some outliers)

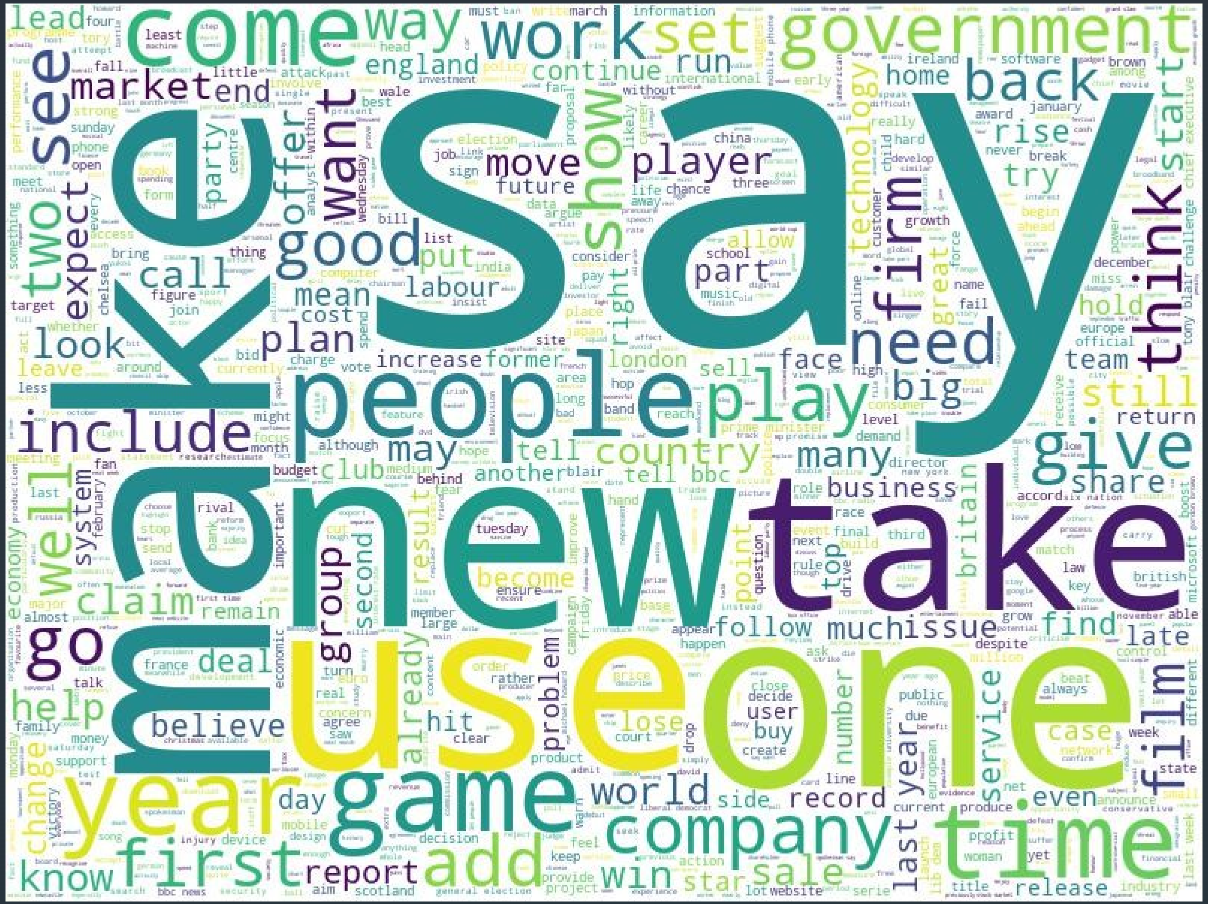
```
count    2200.000000
mean      198.202273
std        86.017350
min        46.000000
25%       132.000000
50%       178.000000
75%       251.000000
max       499.000000
```

```
business         510
sport            505
politics         413
tech             392
entertainment    380
```


Top 10 Bigrams
Bigram
(last, year)
(tell, bbc)
(prime, minister)
(say, would)
(chief, executive)
(last, week)
(tony, blair)
(bbc, news)
(general, election)
(new, york)
0
100
200
300
400
Frequency

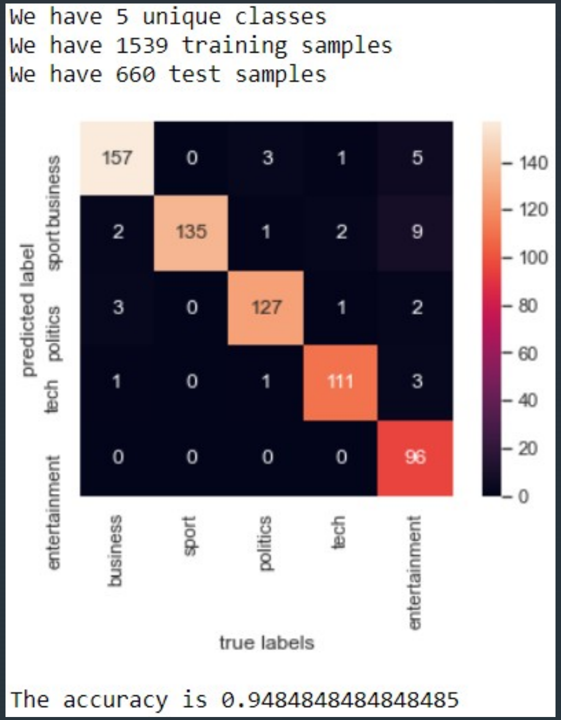

Top 10 5-grams
5-gram
(tell, bbc, radio, today, programme)
(tell, bbc, radio, five, live)
(liberal, democrat, leader, charles, kennedy)
(home, affair, spokesman, mark, oaten)
(shadow, home, secretary, david, davis)
(deputy, prime, minister, john, prescott)
(lib, dem, leader, charles, kennedy)
(shadow, foreign, secretary, michael, ancram)
(political, ambition, founder, mikhail, khodorkovsky)
(affair, spokesman, mark, oaten, say)
0
5
10
15
20
25
30
35
Frequency


Top 10 Trigrams
Trigram
(tell, bbc, news)
(bbc, news, website)
(tell, bbc, radio)
(leader, michael, howard)
(radio, today, programme)
(million, dollar, baby)
(bbc, radio, today)
(tell, bbc, sport)
(tory, leader, michael)
(next, general, election)
0
20
40
60
80
100
120
140
Frequency


say
make
new
take
use
one
year
game
time
company
people
government
come
way
work
set
back
see
show
think
film
need
give
include
play
first
add
know
good
go
want
offer
two
well
market
player
firm
still
share
world
win
report
sale
star
country
business
many
labour
look
plan
call
mean
claim
deal
believe
help
move
technology
start
service
number
late
case
find
issue
much
follow

Text Classification – MNB & SVM

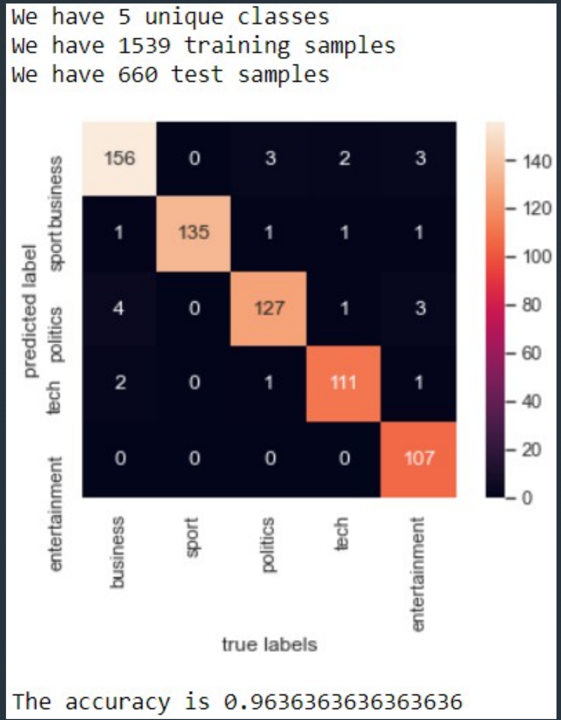
- Text classification works…too well?
- Pre-processing options:
  - Manual, TfidfVectorizer, PorterStemmer, Lemmatizer, others.
- Multinomial Naïve Bayes: 94.8% - 96.8%
- SVM: 97.5%



```
We have 5 unique classes
We have 1539 training samples
We have 660 test samples
```

	business	sport	politics	tech	entertainment
business	157	0	3	1	5
sport	2	135	1	2	9
politics	3	0	127	1	2
tech	1	0	1	111	3
entertainment	0	0	0	0	96

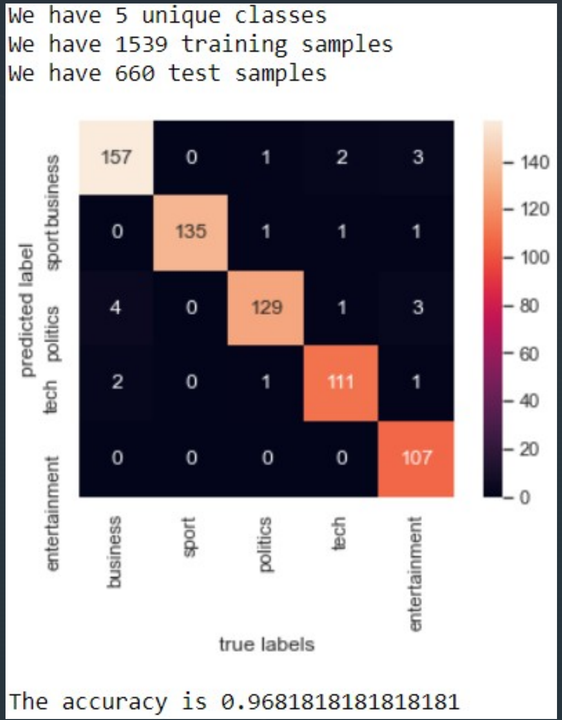
```
The accuracy is 0.9484848484848485
```



```
We have 5 unique classes
We have 1539 training samples
We have 660 test samples
```

	business	sport	politics	tech	entertainment
business	156	0	3	2	3
sport	1	135	1	1	1
politics	4	0	127	1	3
tech	2	0	1	111	1
entertainment	0	0	0	0	107

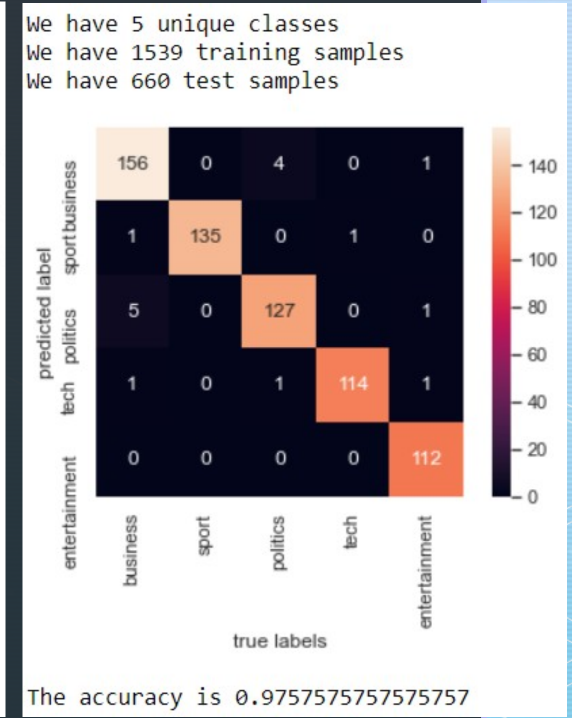
```
The accuracy is 0.9636363636363636
```



```
We have 5 unique classes
We have 1539 training samples
We have 660 test samples
```

	business	sport	politics	tech	entertainment
business	157	0	1	2	3
sport	0	135	1	1	1
politics	4	0	129	1	3
tech	2	0	1	111	1
entertainment	0	0	0	0	107

```
The accuracy is 0.9681818181818181
```



```
We have 5 unique classes
We have 1539 training samples
We have 660 test samples
```

	business	sport	politics	tech	entertainment
business	156	0	4	0	1
sport	1	135	0	1	0
politics	5	0	127	0	1
tech	1	0	1	114	1
entertainment	0	0	0	0	112

```
The accuracy is 0.9757575757575757
```

Topic Modeling – LDA

- Can you be a library purist?
  - Leveraging what works in the moment
  - nltk, gensim, spaCy, sklearn, etc…
- What is LDA good for anyway?
  - People with a ton of time and patience…
  - Or for determining topics from bulk unstructured data
  - Or for clustering classified documents to gain new insights
  - Or … Or … Or …
- Can LDA make predictions?
  - Why, yes. Yes, it can.
    - Just not as easy as Text Classification…

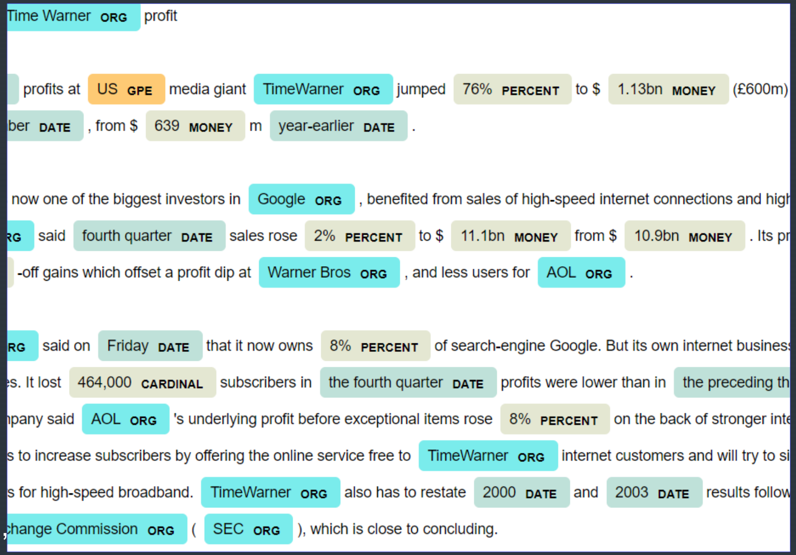
ldavis_5_topics_060321185419.html

ldavis_10_topics_060321192703.html

ldavis_5_topics_070321212952.html

Named Entity Recognition

- Not much to say here.
- I used spaCy, but I’m open to Stanford’s stuff too
- NER just works for metadata extraction
  - CAVEAT – manual process
- The pre-trained model is pretty fantastic
- I really need to learn how train a custom NER model



Time Warner ORG profit

profits at US GPE media giant TimeWarner ORG jumped 76% PERCENT to $ 1.13bn MONEY (£600m)
ber DATE , from $ 639 MONEY m year-earlier DATE .

now one of the biggest investors in Google ORG , benefited from sales of high-speed internet connections and high
RG said fourth quarter DATE sales rose 2% PERCENT to $ 11.1bn MONEY from $ 10.9bn MONEY . Its pr
-off gains which offset a profit dip at Warner Bros ORG , and less users for AOL ORG .

RG said on Friday DATE that it now owns 8% PERCENT of search-engine Google. But its own internet business
s. It lost 464,000 CARDINAL subscribers in the fourth quarter DATE profits were lower than in the preceding th
npany said AOL ORG 's underlying profit before exceptional items rose 8% PERCENT on the back of stronger inte
s to increase subscribers by offering the online service free to TimeWarner ORG internet customers and will try to si
s for high-speed broadband. TimeWarner ORG also has to restate 2000 DATE and 2003 DATE results follow
change Commission ORG (SEC ORG), which is close to concluding.

```
{'DATE': ['fourth quarter',
          '2005',
          'the preceding three quarters',
          '2003',
          'Quarterly',
          'the full-year',
          'Friday',
          'our full-year',
          '2000',
          'the fourth quarter',
          'year-earlier',
          'the three months to December'],
 'GPE': ['US'],
 'NORP': ['German'],
 'ORG': ['AOL Europe',
         'Rings',
         'Time Warner',
         'Bertelsmann',
         'the US Securities Exchange Commission',
         'TimeWarner',
         'AOL',
         'Google',
         'SEC',
         'Warner Bros'],
 'PERSON': ['Alexander', 'Richard Parsons',
'Catwoman']}
```

Outcomes

- Text classification models work very well
  - MNB vs SVM vs ?? … How good is too good
  - Need more (diverse) data
- LDA looks like it could be useful with a large enough data set and good analytic insight…and patience
- NER is a beautiful thing for metadata extraction
  - Need to customize it for specific mission sets
- Pipeline and make_pipeline are … beautiful
- Holy crap I’m glad I learned about pickle
  - Pickling trained models!?!? Yes please and thank you

Next Steps

- Read/write to MongoDB vs Pandas Dataframe
- Customized stop words list to remove things like ‘say’ and ‘bbc’
- Incorporate foreign language handling
  - Machine Translation (AWS Translate, Google Cloud Translation, etc)
  - NLP in foreign languages
- Customize N.E.R. for tailored results
- Incorporate sentiment analysis for social media data
- Web application for data interaction
  - Django / Flask / ??
  - Visualize existing data
    - word cloud, histograms, ngrams, sentiment spikes over time, heatmaps, etc
  - Upload new document(s)
    - pre-process, predict, update metadata, correct errors, add to training set, retrain model