The Spectacular Sailors - Checkpoint 5 Natural Language Processing

SHOW CODE

Overview

- Evaluate allegation complaint narratives to improve predictions and explore data
- Can we improve crew prediction accuracy with sentiment scores?
- What else can we learn about CPDB complaint data?

Instructions

- To run this analysis from Google Colab: From the Runtime menu option, select Run All.
- All the dependicies will install and all anlaysis will run. The later stages of text analysis may take up to several minutes to complete.
- Most code blocks are hidden by default, click Show to expose code

SHOW CODE

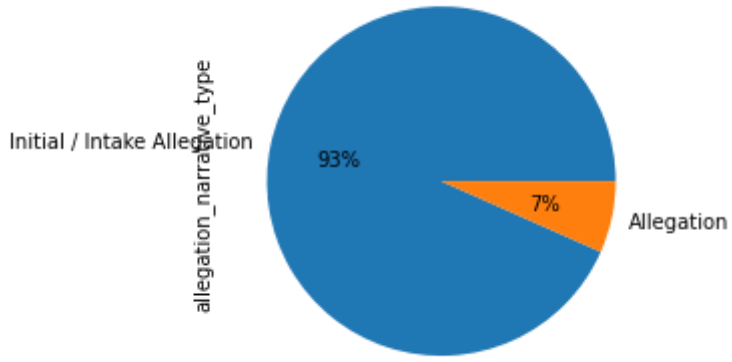
SHOW CODE

```
======================================== DataFrame Head for Allegation Narrative
   allegation_id    allegation_narrative_type  \
0          58067  Initial / Intake Allegation
1          83721  Initial / Intake Allegation

                                       text_content  officer_id cohort_num  \
0  It is alleged that three unknown white male\no...       21441          3
1  The reporting party alleged that the accused\n...       28429          3

  start_date  cohort_id  is_crew  is_community  is_unaffiliated
0 1995-07-03          0        0             0                1
1 1997-08-20          0        0             0                1
======================================== DataFrame Information
```

Apply Pre-Processing to Clean Text With Text Hero

SHOW CODE

```
======================================== Start Cleaning
======================================== Done Cleaning
======================================== DataFrame Head
<matplotlib.axes._subplots.AxesSubplot at 0x7f2850112630>
```



Apply Sentiment Analysis to Cleaned Text with Vader

Reference: https://medium.com/swlh/simple-sentiment-analysis-for-nlp-beginners-and-everyone-else-using-vader-and-textblob-728da3dbe33d

SHOW CODE

```
======================================== DataFrame Head
   allegation_id    allegation_narrative_type  \
0          58067  Initial / Intake Allegation
1          83721  Initial / Intake Allegation

                                        text_content  officer_id cohort_num  \
0  alleg three unknown white male offic state com...       21441          3
1  report parti alleg accus stop demand see ident...       28429          3

  start_date  cohort_id  is_crew  is_community  is_unaffiliated  compound  \
0 1995-07-03          0        0             0                1   -0.5719
1 1997-08-20          0        0             0                1   -0.9081

     neg    neu  pos
0  0.343  0.657  0.0
1  0.407  0.593  0.0
```

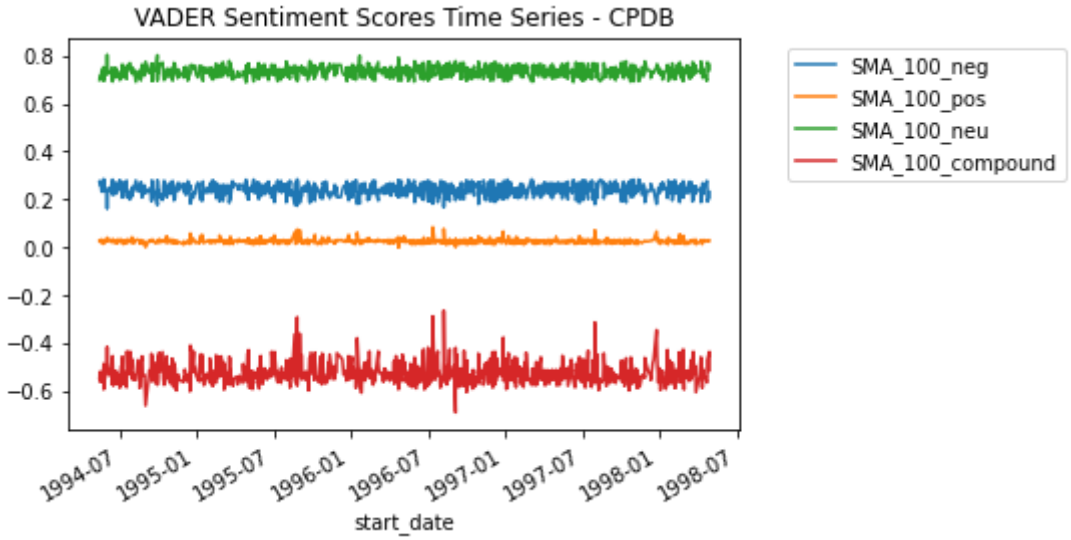
Time Series Sentiment Analysis by Complaint Start Date

Exploratory analysis of sentiment analysis over time.

Question: Why is data limited to complaints between 1994 and 1998? Missing data or bad query?

SHOW CODE

```
<Figure size 432x288 with 0 Axes>
```



SHOW CODE

Data with sentiment scores

	officer_id	allegation_id	cohort_id	is_crew	is_community	is_unaffiliated	
0	4	53006.5	2283.0	0.0	1.0	0.0	
1	11	70839.0	0.0	0.0	0.0	1.0	

Join Sentiment Scores with officers_crews_ml Table

Merge sentiment scores to data used in Machine Learning Checkpoint and re-assess prediction scores.

Create merged tables with and without missing values.

SHOW CODE

```
Created Joined Tables
```

Export csv data to Google Drive

SHOW CODE

Assess ML Checkpoint 4 with Sentiment Scores

SHOW CODE

```
Table 1: 2    5146
3    3943
1     663
Name: cohort_id, dtype: int64

Table 2: 3    7329
2    4900
1     493
Name: cohort_id, dtype: int64
Prepare Tables for ML Predictions
```

SHOW CODE

```
Scale input values
```

SHOW CODE

```
========== View Scaled Feature Values
[[0.005 1.    0.49  0.    0.213 0.095]
 [0.011 0.    0.469 0.143 0.18  0.771]
 [0.044 1.    0.523 0.143 0.317 0.057]
 ...
 [0.011 0.    0.409 0.    0.184 0.358]
 [0.013 0.    0.452 0.1   0.256 0.26 ]
 [0.    0.    0.348 0.857 0.142 0.158]]
```

SHOW CODE

```
Split Train Test Data
```

SHOW CODE

```
Fit Linear Regression
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)
```

SHOW CODE

```
Make Predictions
array([2., 2., 3., ..., 2., 3., 3.])
```

SHOW CODE

```
Accuracy of Linear Regression Model on Predictions
```

```
Model Accuracy Score
0.6985645933014354
```

SHOW CODE

```
Coefficients of the Linear Model
array([-1.04396633, -0.03218181, -1.35695313,  0.67179191, -1.34054077,
        0.05687459])
```

SHOW CODE

```
The logistic regression
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=100,
                   multi_class='ovr', n_jobs=None, penalty='l2',
                   random_state=None, solver='lbfgs', tol=0.0001, verbose=0,
                   warm_start=False)
```

SHOW CODE

```
Accuracy of Logistic Regression Model on Predictions
0.7112098427887902
```

Evaluate Text for NLP Checkpoint 5

SHOW CODE

```
Prep Tables for NLP Analysis
```

SHOW CODE

```
Show Filtered Tokens
['alleg', 'three', 'unknown', 'white', 'male', 'offic', 'state', 'complain', 'tra
```

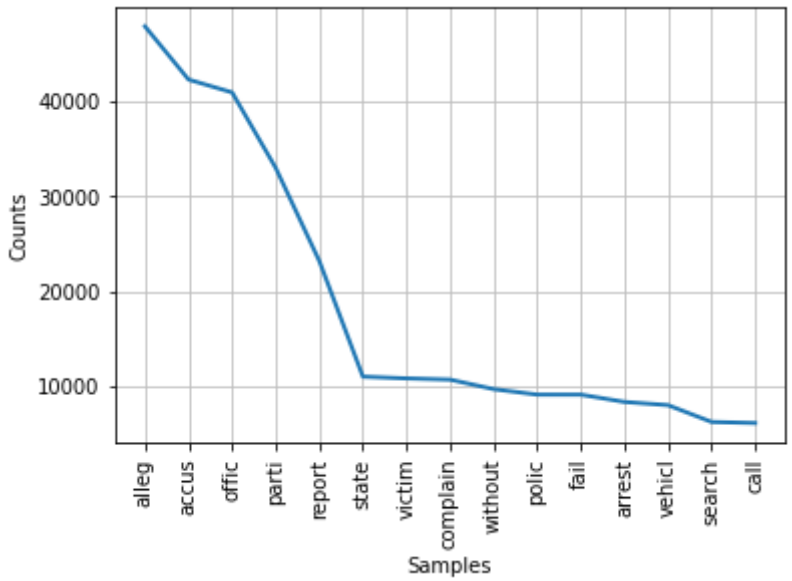
SHOW CODE

```
Number of Unique Words
21640
```

SHOW CODE

```
(1180, 1)
Shape of Corpus
[[(1180, 1)]]
```

SHOW CODE

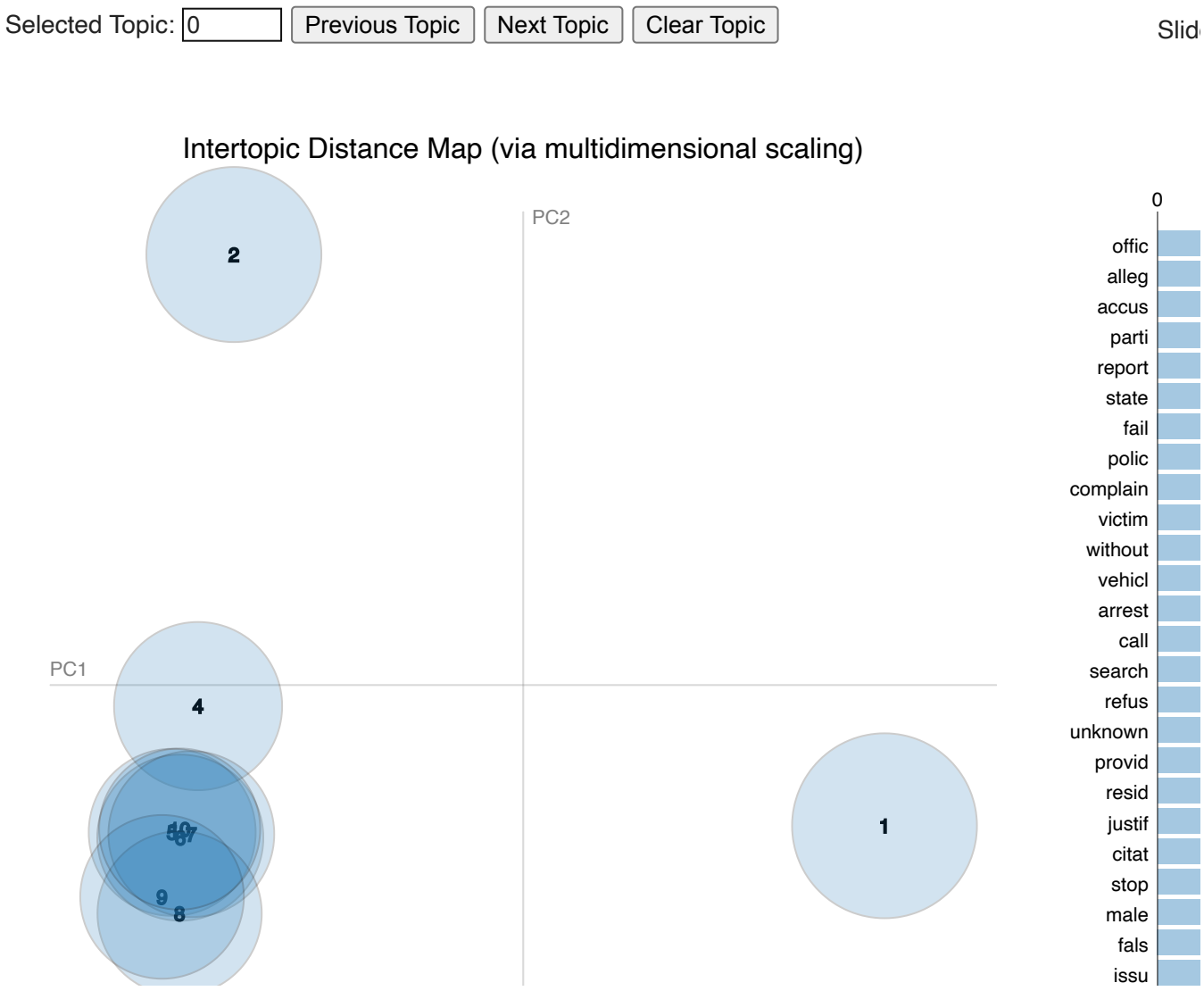


SHOW CODE

```
Top 10 Most Frequent Words
[('alleg', 47899),
 ('accus', 42297),
 ('offic', 40942),
 ('parti', 32994),
 ('report', 23197),
 ('state', 11017),
 ('victim', 10832),
 ('complain', 10705),
 ('without', 9699),
 ('polic', 9124)]
```

SHOW CODE

SHOW CODE



Analysis

Once investigating from a sentiment analysis, n-gram perspective we found inconclusive results, as a result we continued our exploration utilizing an additional NLP methond for exploration.

Principal Component Analysis (PCA) tool for analyzing text data data to facilitate interpretation and the summaries or topics we found were inline with past behavior of crews.

Clear narrative for topics 1 and 2, the negative sentiment are reflective of what we would assume from the worst police officers.

And we are extremely curious, as to why and additional 80% of the corpus closely aligns together in PCA3.

We thought analyzing sentiment would be really important

The big idea - we assessed our ML algo and added our sentiment score as a feature to officers and it turned out that sentiment had little effect on our prediction accuracy. We thought analyzing sentiment would be really important to our predictions as we initially thought.

PCA 1

- The 1st topic / theme is centered around continued overall unprofessional and bad conduct.

PCA 2

- The 2nd topic seems to be getting stopped without any reason where complaintaints felt they are getting continually harrassed by officers. More importanly, the words are descrbing adverseral conduct. Given the history of the CPD, we can confirm that what the issues might be...unprofessional search and agressive, violent behavior is occuring so frequent that it has its own bubble of thousands of words supporting this type of agressive conduct.

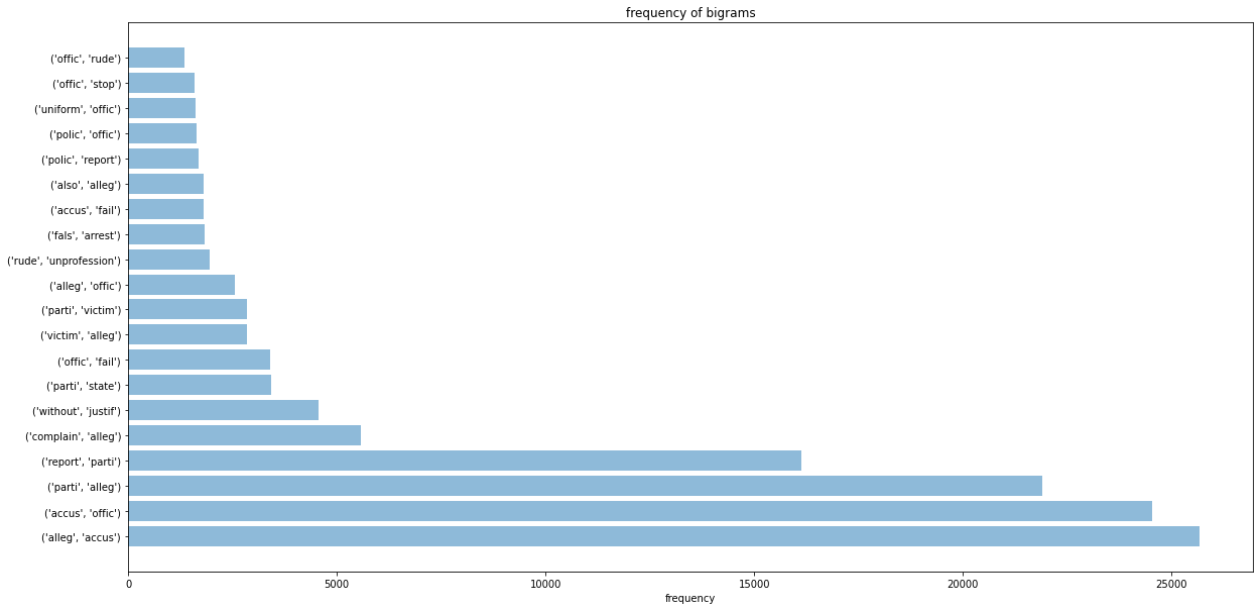
PCA 3 The majority of topics 3-9 are unbelievable: bitch, threw, rude, handcuff, apart, without, permissions, etc...

- The 3rd topic seems to be getting stopped without justification and they feel theh officers treatment was rude and used degoratory lauguage, etc.
- The 4th topic seems to be about alleged white officers are potentially accused them of a crime
- The 5th topic seems to be about a traffic stop coming from a party and stoped because they were black
- The 6th topic seems to be about a parent making complaints that officers threatened their son, stoped one or two times.
- The 7th topic seems to be about a complaint that they were stopped for no reason at strange hours
- The 8th topic seems to be about more agressive behavior yell, push, ground, detain, lock.
- The 9th topic seems to be about while sitting in their vehicle however I was..
- The 10th topic seems to be about unprofessional and clearly apparent that they were coming to my place to arrest me

SHOW CODE

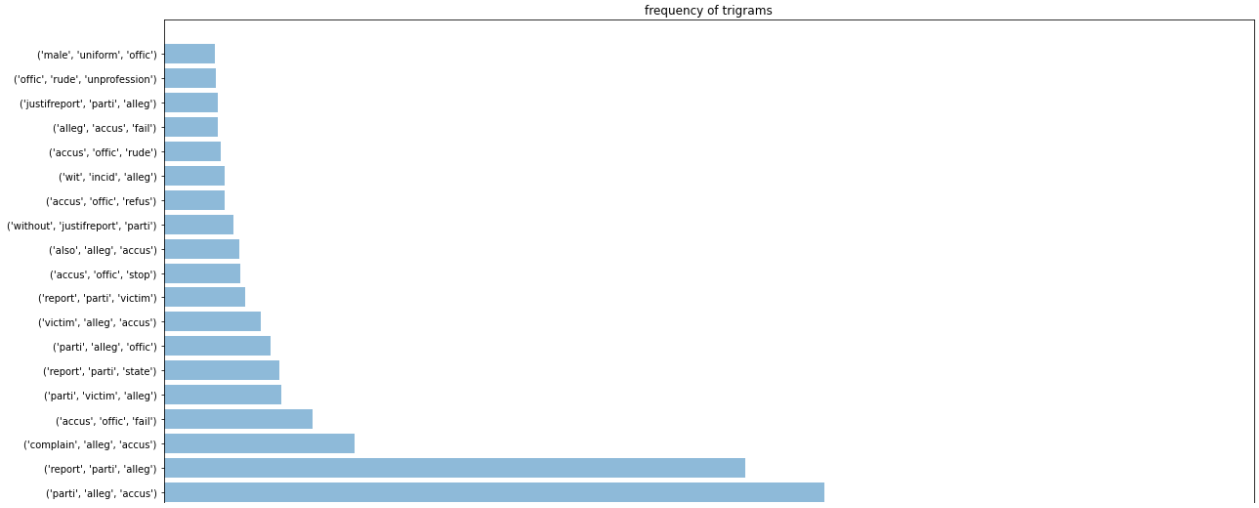
```
Total corpa:  804830
Total bigrams:  149269
Total trigrams:  276608
```

SHOW CODE



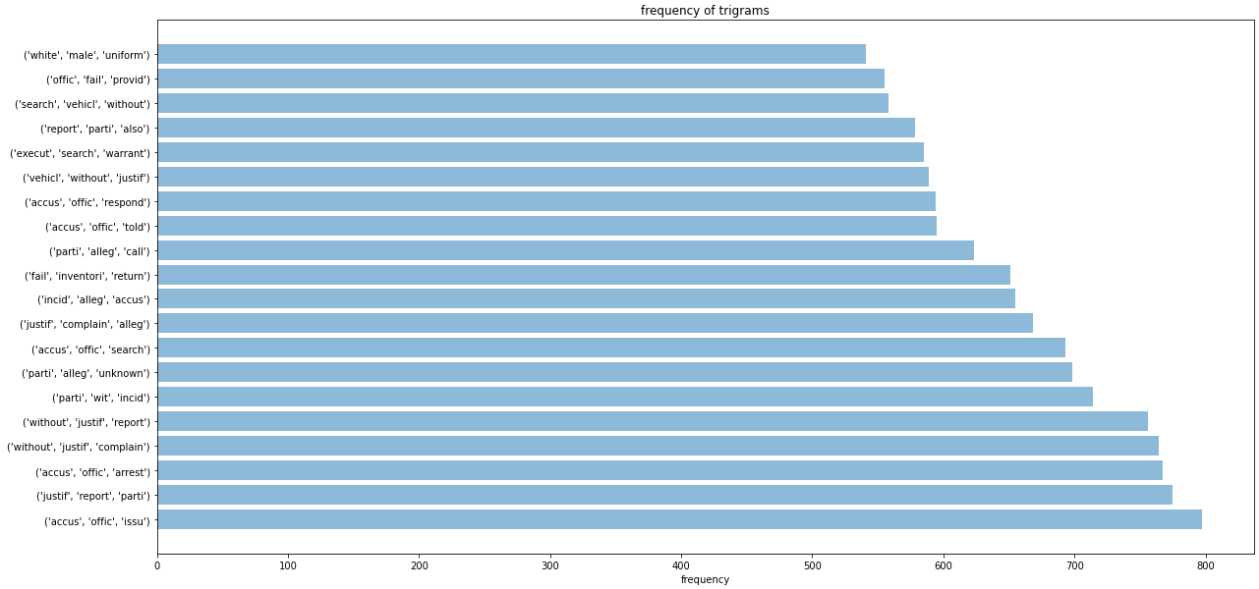
The bigrams are indicative of agressive behavior "rude, unprofession", "fals, arrest", "offic, rude", that we commonly see in crew behaviors.

SHOW CODE



Re-running trigrams again to filter out the the top standard string of words "alleg, accus, offic", "parti, alleg, accus", "report, parti, alleg", since they are mostly stardard in any investigation narrative.

SHOW CODE



We see above that the most common string of trigrams change (once commong narratives are removed) and are more inline of what we commonly see "without, justif, complain", "parti, alleg, unknown", "accus, offic, search", which is more reflective of what we would expect from officer associated with crew like behavior. For instance, "vechicl, without, justif" is clearly a refection of agressive police behavior.

> Indented block