CHECKPOINT 5: NATURAL LANGUAGE PROCESSING

DATA PIRATES

> Kavya Jaganathan, Clarissa Cheam and Ayushi Mishra

Overview

> In this checkpoint we explore the text content available in the tables data_attachmentnarrative and data_allegation. We study both these tables since data_attachmentnarrative table's text content gives an account of the allegation, along with the victim perspective in terms of words uttered by the accused to them, tone/demeanor of accused that the observed, etc. whereas the data_allegation table's text content gives a report written by an officer without much of the victim's direct words.

The question we hope to answer:

What was the sentiment in the CPDB complaint data pre major reforms (before 2016) and post major reforms?

Note: Major reforms here indicate reforms like COPA and introduction of TRR

Installing packages

Show code

Connecting to CPDB database

Show code

FOR DATA ATTACHMENT NARRATIVE TABLE

Getting data and converting to dataframe

Saving the desired table output

Show code

```
shape is: 31555
(31555, 5)
```

Show code

Show code

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 31555 entries, 0 to 31554
Data columns (total 5 columns):
 #   Column         Non-Null Count  Dtype         
---  ------         --------------  -----         
 0   attachment_id  31555 non-null  int64         
 1   column_name    31555 non-null  object        
 2   text_content   31555 non-null  object        
 3   allegation_id  31555 non-null  object        
 4   incident_date  31275 non-null  datetime64[ns]
dtypes: datetime64[ns](1), int64(1), object(3)
memory usage: 1.2+ MB
```

Checking our dataframe output

Show code

```
   attachment_id                  column_name  \
0          51827  Initial / Intake Allegation
1          51827                   Allegation
2          51827  Initial / Intake Allegation
3          31479  Initial / Intake Allegation
4          31462  Initial / Intake Allegation

                                        text_content allegation_id  \
0  The complainant alleges that the accused\nDepa...       1056174
1        Victim/Offender\nSituation Aty Weapon Types       1056174
2   The complainant alleges that the accused\nDep...       1056174
3  The complainant alleged that the accused\noffi...       1051104
4   The complainant alleged that the accused\noff...       1051104

  incident_date
0    1971-02-28
1    1971-02-28
2    1971-02-28
3    1982-12-31
4    1982-12-31
```

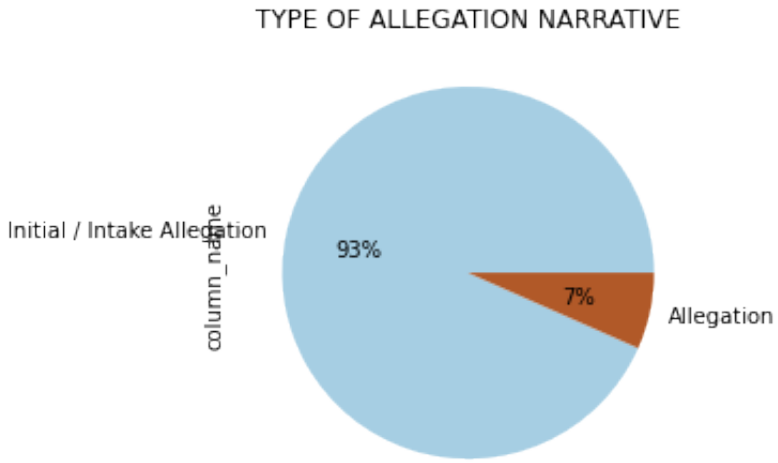
Applying pre processing to clean text with Text Hero

Show code

NOTE: We remove all rows where column_name is FINDINGS as the text content is sparse and doesn't contribute to analysis

Show code

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f9956c31f10>
```



Apply sentiment analysis to cleaned data with Vader

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

Show code

```
   attachment_id                  column_name  \
0          51827  Initial / Intake Allegation
1          51827                   Allegation

                                        text_content allegation_id  \
0  complain alleg accus depart member fail conduc...       1056174
1              victim offend situat ati weapon type        1056174

  incident_date  compound    neg    neu  pos
0    1971-02-28   -0.8442  0.503  0.497  0.0
1    1971-02-28   -0.6705  0.684  0.316  0.0
```

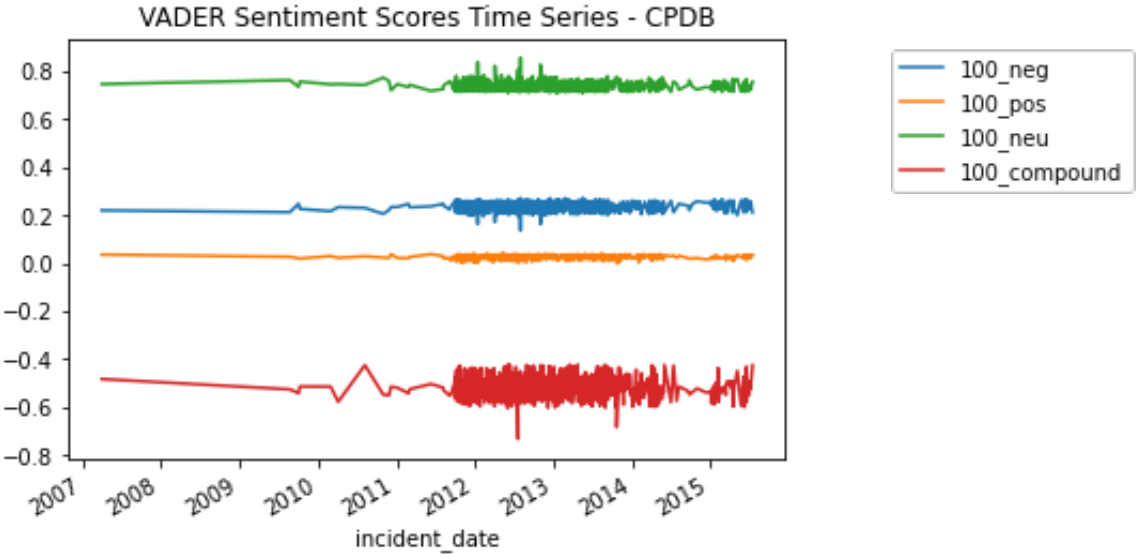
Time Series complaint analysis using incident date

Show code

Show code

incident_date	100_neg	100_pos	100_neu	100_compound
2012-01-13	0.1630	0.0000	0.8370	-0.542300
2011-11-07	0.2265	0.0000	0.7735	-0.542300
2013-10-20	0.2240	0.0000	0.7760	-0.681767
2012-07-19	0.1950	0.0000	0.8050	-0.731500
2012-10-31	0.1624	0.0116	0.8260	-0.539940

<Figure size 432x288 with 0 Axes>



Analysis

> Before analyzing the time series graph some of the terminology to be aware of:

1. **Sentiment:** Attitude to a certain situation or event
2. **Compound:** It is a 'normalized, weighted composite score computed by summing the valence scores of each word in the lexicon, adjusted according to the rules, and then normalized to be between -1 (most extreme negative) and +1 (most extreme positive).

In the data_attachmentnarrative table, the text content gives an account of the allegation along with the victim perspective in terms of words uttered by the accused to them, tone/demeanor of accused that they observed, etc. We perform sentiment analysis on this data to understand the prevalent emotion/sentiment.

With a compound score in the range of (-0.8, -0.4) we can understand that the sentiment was majorly negative. We cannot perform further analysis on this data to understand the effect of reform since the data in this table dates up to 2015. Despite the lack of data, analyzing this data is essential as understanding the history of data is important in forming any new reforms.

Evaluating text for checkpoint

Show code

```
Prep Tables for NLP Analysis
   attachment_id                  column_name  \
0          51827  Initial / Intake Allegation
1          51827                   Allegation
2          51827  Initial / Intake Allegation
3          31479  Initial / Intake Allegation
4          31462  Initial / Intake Allegation

                                        text_content allegation_id  \
0  complain alleg accus depart member fail conduc...       1056174
1              victim offend situat ati weapon type       1056174
2  complain alleg accus depart member fail conduc...       1056174
3  complain alleg accus offic sexual abus time ac...       1051104
4  complain alleg accus offic sexual abus time ac...       1051104

  incident_date  compound    neg    neu  pos
0    1971-02-28   -0.8442  0.503  0.497  0.0
1    1971-02-28   -0.6705  0.684  0.316  0.0
2    1971-02-28   -0.8442  0.503  0.497  0.0
3    1982-12-31   -0.6124  0.357  0.643  0.0
4    1982-12-31   -0.6124  0.357  0.643  0.0
```

Show code

```
/usr/local/lib/python3.7/dist-packages/catalogue.py:138: DeprecationWarning

SelectableGroups dict interface is deprecated. Use select.

/usr/local/lib/python3.7/dist-packages/catalogue.py:138: DeprecationWarning

SelectableGroups dict interface is deprecated. Use select.

/usr/local/lib/python3.7/dist-packages/catalogue.py:126: DeprecationWarning

SelectableGroups dict interface is deprecated. Use select.

/usr/local/lib/python3.7/dist-packages/catalogue.py:138: DeprecationWarning

SelectableGroups dict interface is deprecated. Use select.

Show Filtered Tokens
['complain', 'alleg', 'accus', 'depart', 'member', 'fail', 'conduct', 'prop
```

Show code

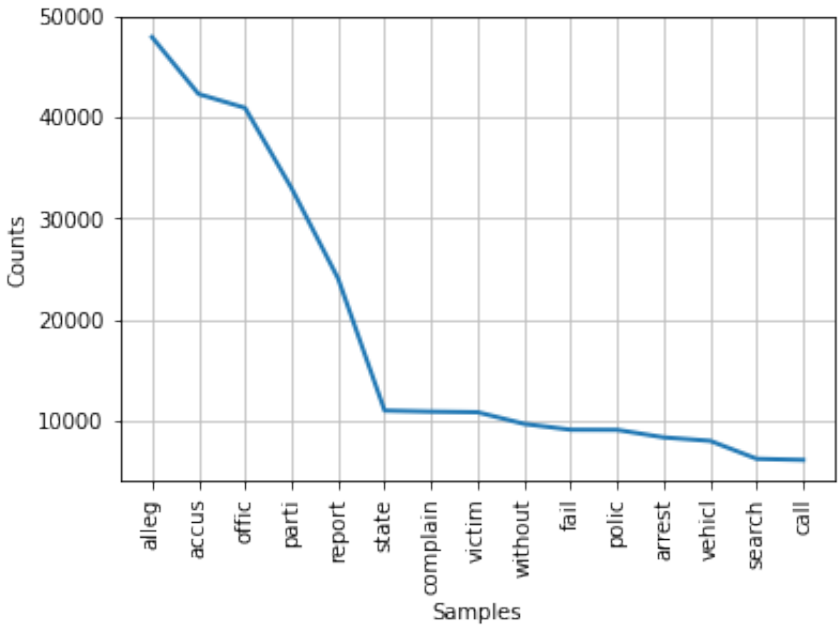
```
Number of Unique Words
21608
```

Show code

```
(4399, 1)
Shape of Corpus
[[(4399, 1)]]
```

Word Frequency

Show code




```
<matplotlib.axes._subplots.AxesSubplot at 0x7f996a3c0ad0>
```

Top 20 frequent words

Show code

```
Top 20 Most Frequent Words
[('alleg', 47932),
 ('accus', 42306),
 ('offic', 40937),
 ('parti', 32996),
 ('report', 24072),
 ('state', 11018),
 ('complain', 10912),
 ('victim', 10854),
 ('without', 9701),
 ('fail', 9123),
 ('polic', 9109),
 ('arrest', 8367),
 ('vehicl', 8023),
 ('search', 6247),
 ('call', 6138),
 ('``', 5836),
 ('stop', 4864),
 ('justif', 4839),
 ('male', 4526),
 ('refus', 4156)]
```

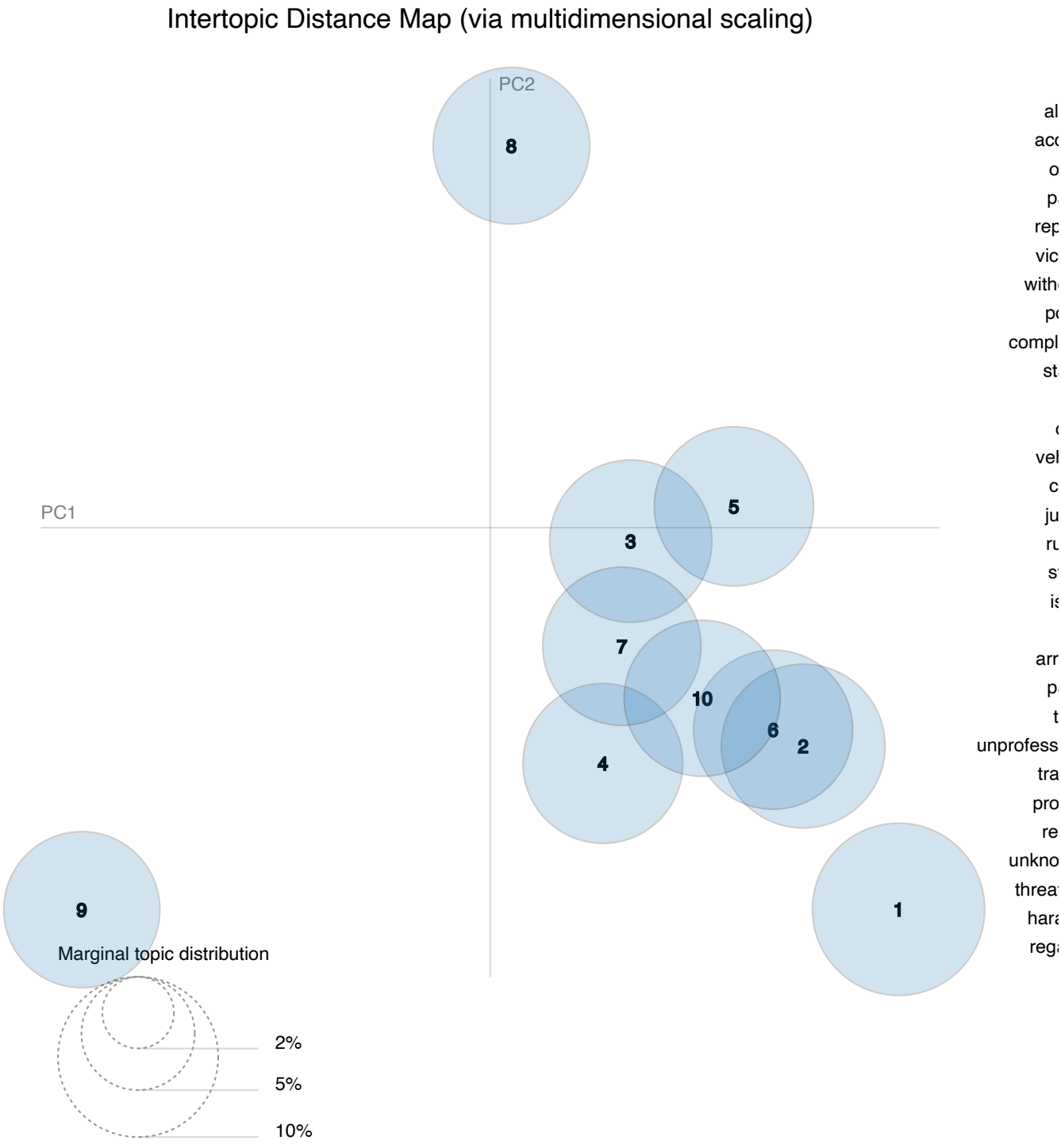
LDA

Show code

Interactive PCA graph

Show code

Selected Topic: 0 Previous Topic Next Topic Clear Topic



PCA ANALYSIS

The PCA tool was utilised since the high level view of frequent words do not give much insight.

1. PCA 1-3 displays aggressive, unprofessional behavior of the accused.
2. PCA 4-6 displays misconduct in terms of procedure highlighted by the frequent appearance of words like 'false', 'ignore', 'harrass', 'inappropriate','supervisor'
3. PCA 6-10 displays a lot of rude verbal abuse and unprofessional conduct including profanities.

Some topic wise analysis:

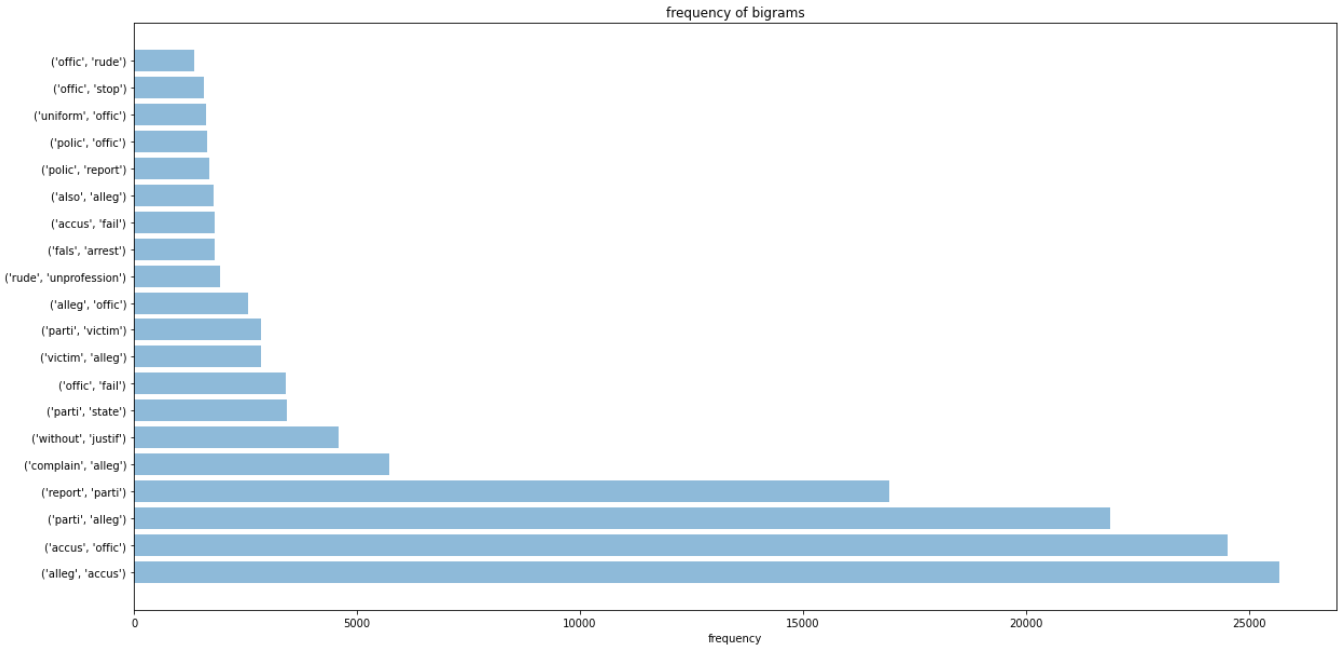
1. PCA 1,2 seems to be centered around domestic violence cases and cases against black civilians with a failure of appropriate response from officers
2. PCA 3 seems to be centered around stopping vehicles without probable cause and conversations seem to have been terse bordering on rude.

Show code

```
Total corpa:  813843
Total bigrams:  150450
Total trigrams:  280934
```

Bigram frequencies

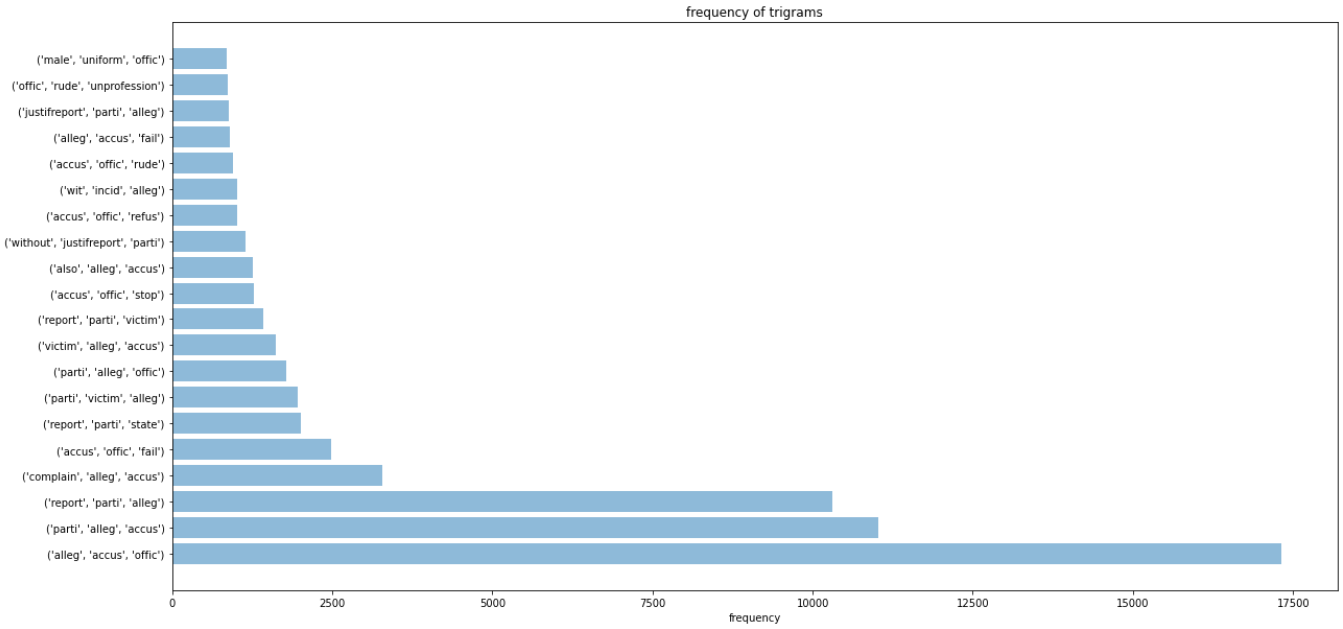
Show code



Ignoring common words that appear on all reports like alleged, accused, complaint, report etc. we can see the bigrams like "offic,rude","fals,arrest","without,justif" are all indicative of unproffessional and improper conduct with civilians and callousness towards procedure.

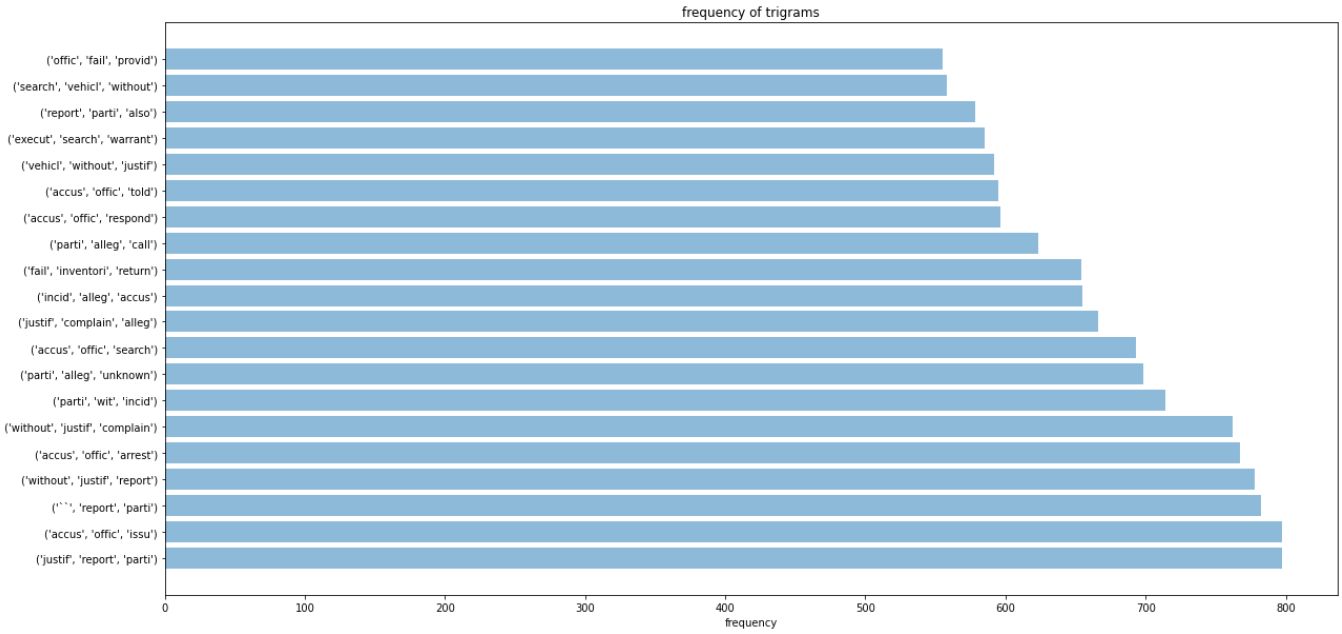
Trigram frequencies

Show code



Trigram frequency continued...

Show code



The trigrams "search,vehicl,without","vehic,without,justif","fail,inventori,return","without,justif,complain" all support the findings from bigrams indicating lack of attentiveness bordering on complete disregard of procedure.

FOR DATA ALLEGATION TABLE

Getting data and converting to dataframe

Show code

```
shape is: 1147
(1147, 5)
```

Show code

Checking our dataframe output

Show code

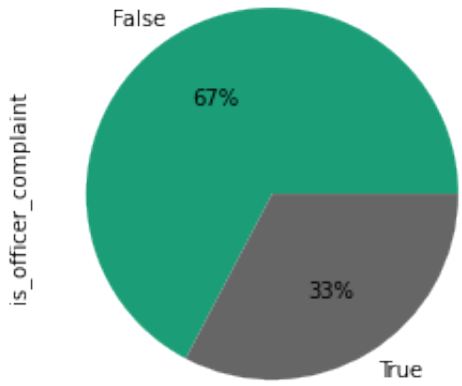
	crid	summary	incident_date	is_officer_complaint	location
0	1049123	In an incident involving three on-duty CPD mem...	2011-10-05	False	Private Residence
1	1049179	On October 10th, 2011, a complaint was registe...	2011-10-08	True	XX
2	1049208	In an incident involving an on-duty Lieutenant...	2011-10-10	False	Restaurant
3	1049273	On October 13, 2011, a complaint was registere...	2011-10-12	True	Public Way - Other
4	1049571	On October 25, 2011, a complaint was registere...	2011-10-24	False	Private Residence

Applying pre processing to clean text with Text Hero

Show code

Show code

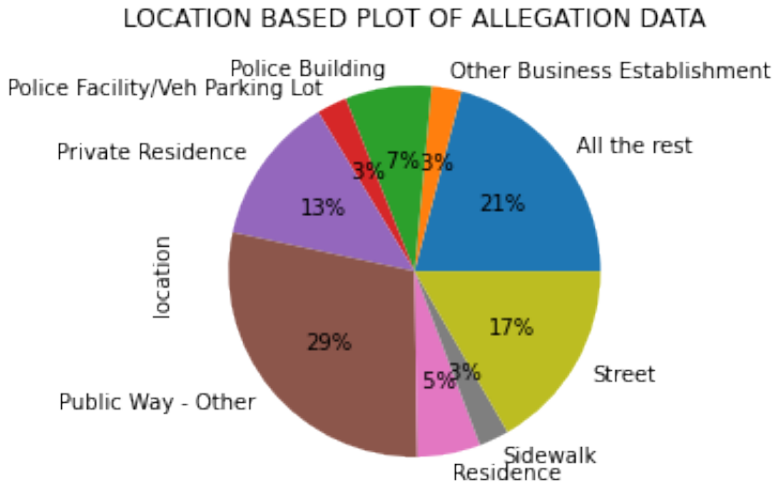
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f9950588a90>
```



NOTE: About 67% of the data used are complaints made by civilians and 33% made by fellow officers

Show code

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f995dd61c10>
```



Apply sentiment analysis to cleaned data with Vader

Show code

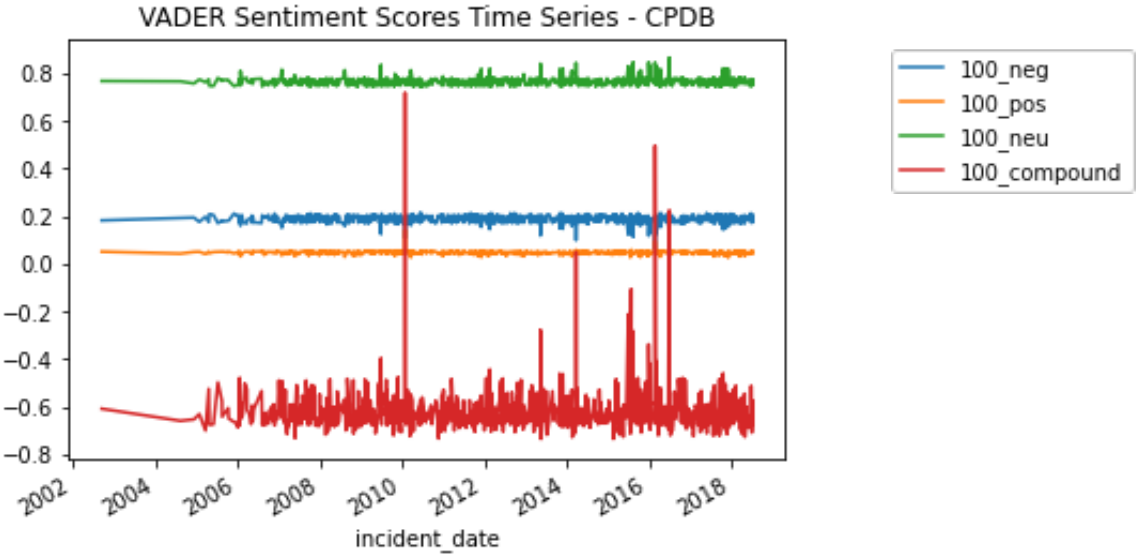
```
      crid                                                summary incident_date
0  1049123  incid involv three duti cpd member includ two ...    2011-10-05
1  1049179  octob 10th complaint regist independ polic rev...    2011-10-08

   is_officer_complaint          location  compound    neg    neu    pos
0                 False  Private Residence   -0.9744  0.419  0.581  0.000
1                  True                XX   -0.9313  0.203  0.746  0.051
```

Time Series complaint analysis from incident date

Show code

Show code

```
                100_neg   100_pos   100_neu  100_compound
incident_date
2010-01-26       0.1080  0.141000  0.752000      0.718400
2016-02-25       0.0615  0.094000  0.845000      0.495800
2016-07-01       0.0580  0.077333  0.865333      0.224467
2014-03-25       0.0985  0.058000  0.844000      0.049175
2015-07-29       0.1236  0.058400  0.818400     -0.106040
<Figure size 432x288 with 0 Axes>
```



NOTE: This table has data pre and post major reforms The random spikes in compound scores towards positive value encourage further analysis. To understand the effects of reform on these spikes, we split the data to pre and post major reforms using incident_date.

Pre Major reform and Post major reform analysis (pre and post 2016)

> The text content in the table data_allegation is from the CPD perspective which gives us a better understanding of sentiment within the police force pre-reform and post-reform. Understanding this will allow us to obtain inferences on how receptive was the CPD of reforms and the effectiveness of reform.

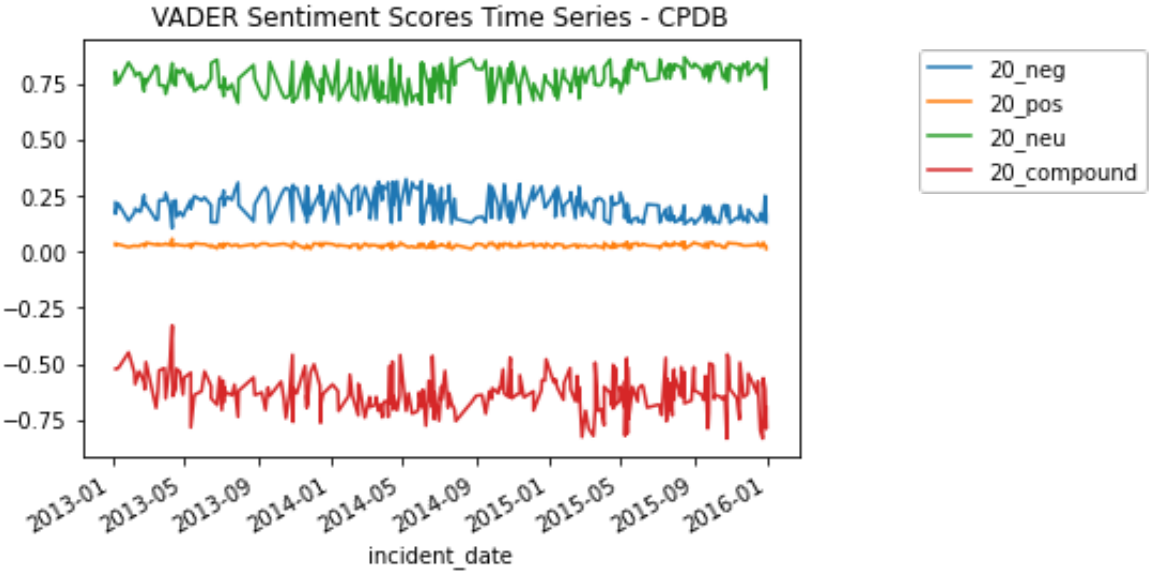
Show code

```
Pre COPA data:
       crid                                               summary incident_da
112  1078866  januari chicago polic depart cpd member assist...    2016-01-
113  1078888  januari chicago polic offic respond report thr...    2016-01-
114  1079748  incid involv offic b c alleg offic b c enter s...    2016-03-
115  1079542  complain itter walk street view chicago polic ...    2016-03-
116  1079908  complain alleg offic sit vehicl remov taser pl...    2016-03-

     is_officer_complaint            location  compound    neg    neu    po
112                 False  Public Way - Other   -0.9260  0.104  0.808  0.08
113                  True       Bar Or Tavern   -0.5267  0.159  0.841  0.00
114                 False           Residence    0.0000  0.000  1.000  0.00
115                 False  Public Way - Other   -0.3612  0.185  0.815  0.00
116                  True              Street   -0.6369  0.157  0.843  0.00

 Post COPA data:
       crid                                               summary incident_dat
42   1061399  incid involv one duti cpd offic offic alleg of...    2013-04-1
43   1061722  subject juvenil alleg offic b c detain without...    2013-04-2
44   1061779  incid involv offic alleg offic yell obscen hel...    2013-04-1
45   1061941  incid involv four duti cpd offic b c alleg off...    2013-05-0
46   1062191  incid involv offic alleg offic damag subject v...    2013-01-2

    is_officer_complaint                location  compound    neg    neu
42                  True                Sidewalk   -0.3291  0.104  0.839
43                 False                  Street   -0.7430  0.249  0.751
44                 False                  Street   -0.4939  0.114  0.844
45                  True  Jail / Lock-Up Facility   -0.6808  0.219  0.781
46                 False                  Street    0.0000  0.000  1.000

      pos
42  0.057
43  0.000
44  0.042
45  0.000
46  0.000
```

Pre major reforms sentiment analysis

Show code

```
                  20_neg    20_pos    20_neu  20_compound
incident_date
2013-04-11      0.104000  0.05700  0.839000     -0.32910
2013-04-24      0.176500  0.02850  0.795000     -0.53605
2013-04-18      0.155667  0.03300  0.811333     -0.52200
2013-05-01      0.171500  0.02475  0.803750     -0.56170
2013-01-28      0.137200  0.01980  0.843000     -0.44936
<Figure size 432x288 with 0 Axes>
```



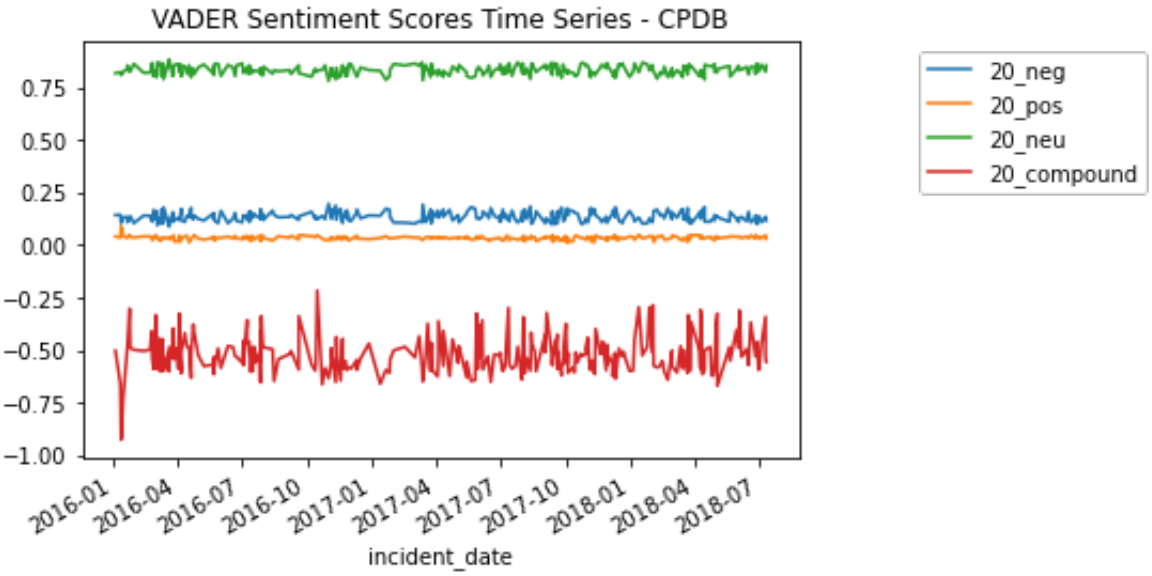
Pre major reform Analysis

> The range of compound score lies between (-0.25, -0.80). Standalone, this data displays a pretty negative atmosphere. There are also high neutral scores generic of reports. This data combined with future data provides a better perspective. COPA and enforcement of TRR were initiated in 2016 and we want to see whether it was received well. If it was received well the compound scores should increase tending closer to +1.

Post major reform analysis

Show code

```
                20_neg    20_pos  20_neu  20_compound
incident_date
2016-01-12    0.104000  0.088000  0.8080    -0.926000
2016-01-15    0.131500  0.044000  0.8245    -0.726350
2016-03-20    0.087667  0.029333  0.8830    -0.484233
2016-03-06    0.112000  0.022000  0.8660    -0.453475
2016-03-31    0.121000  0.017600  0.8614    -0.490160
<Figure size 432x288 with 0 Axes>
```



Post major reform Analysis

> Contrary to expectations the compound scores are observed to be further negative, going as low as 0.92. The range of the compound score has shifted to (-0.25, -1.0) indicative of excessive negativity.

This highly negative score could be indicative of harsher allegations and tonality of reports. We further look into the words to pick out contributors to this score.

Evaluating text for checkpoint

Show code

```
Prep Tables for NLP Analysis
      crid                                                summary incident_date
0  1049123  incid involv three duti cpd member includ two ...    2011-10-05
1  1049179  octob 10th complaint regist independ polic rev...    2011-10-08
2  1049208  incid involv duti lieuten sergeant unknown off...    2011-10-10
3  1049273  octob complaint regist independ polic review a...    2011-10-12
4  1049571  octob complaint regist independ polic review a...    2011-10-24

   is_officer_complaint            location  compound    neg    neu    pos
0                 False   Private Residence   -0.9744  0.419  0.581  0.000
1                  True                  XX   -0.9313  0.203  0.746  0.051
2                 False          Restaurant   -0.9690  0.237  0.763  0.000
3                  True  Public Way - Other   -0.4019  0.115  0.828  0.057
4                 False   Private Residence   -0.8555  0.247  0.703  0.049
```

Show code

```
/usr/local/lib/python3.7/dist-packages/catalogue.py:138: DeprecationWarning

SelectableGroups dict interface is deprecated. Use select.

/usr/local/lib/python3.7/dist-packages/catalogue.py:138: DeprecationWarning

SelectableGroups dict interface is deprecated. Use select.

/usr/local/lib/python3.7/dist-packages/catalogue.py:126: DeprecationWarning

SelectableGroups dict interface is deprecated. Use select.

/usr/local/lib/python3.7/dist-packages/catalogue.py:138: DeprecationWarning

SelectableGroups dict interface is deprecated. Use select.

Show Filtered Tokens
['incid', 'involv', 'three', 'duti', 'cpd', 'member', 'includ', 'two', 'off
```

Show code

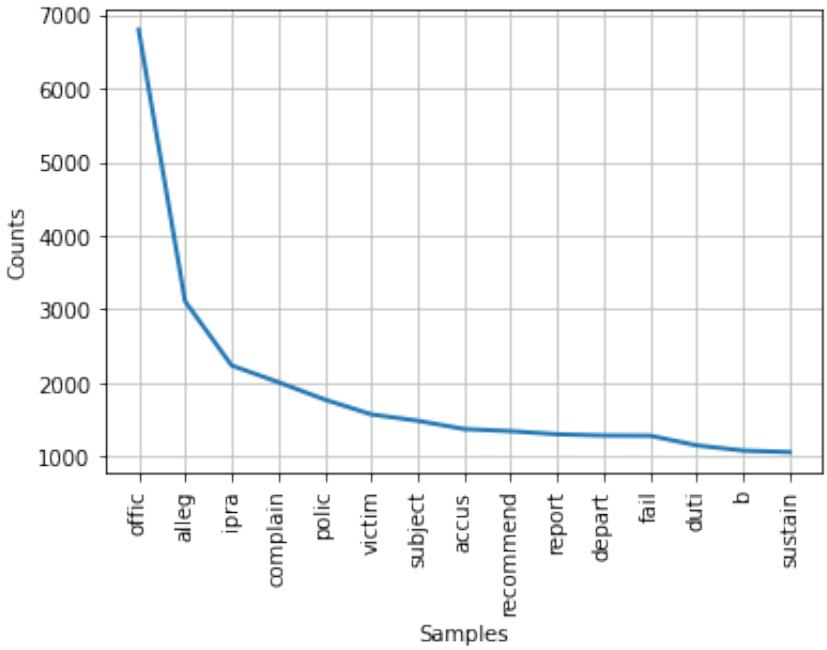
```
Number of Unique Words
4543
```

Show code

```
(2102, 1)
Shape of Corpus
[[(2102, 1)]]
```

Word Frequency

Show code




```
<matplotlib.axes._subplots.AxesSubplot at 0x7f995e3c2ad0>
```

20 most frequent words

Show code

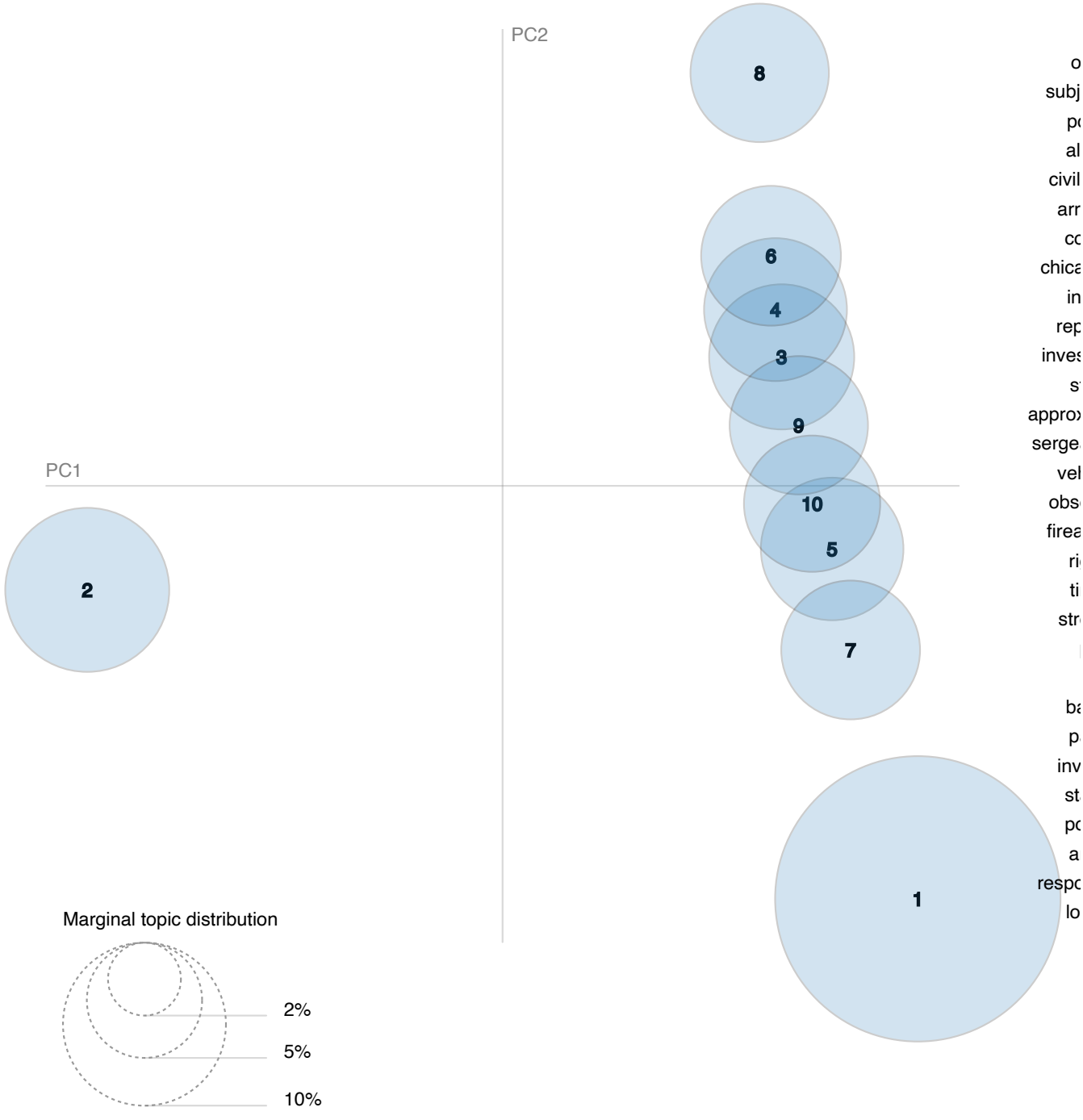
```
Top 20 Most Frequent Words
[('offic', 6797),
 ('alleg', 3108),
 ('ipra', 2238),
 ('complain', 2012),
 ('polic', 1775),
 ('victim', 1574),
 ('subject', 1487),
 ('accus', 1375),
 ('recommend', 1346),
 ('report', 1304),
 ('depart', 1287),
 ('fail', 1284),
 ('duti', 1153),
 ('b', 1082),
 ('sustain', 1061),
 ('incid', 840),
 ('chicago', 831),
 ('involv', 803),
 ('sergeant', 802),
 ('vehicl', 754)]
```

LDA

Show code

Show code

Selected Topic: 0 Previous Topic Next Topic Clear Topic



PCA Analysis

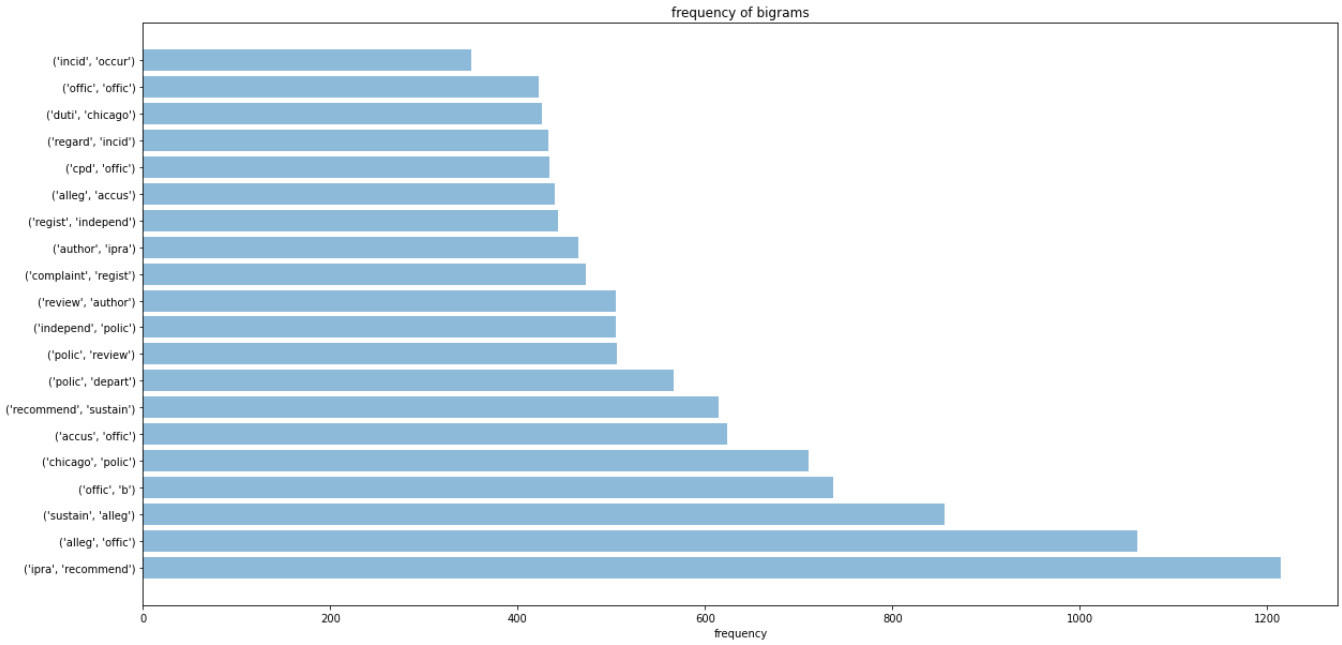
All PCA topics in general seem to indicate excessive violence with frequent words being strike, discharge, head, foot, offend, alterc, control, punch, knee, gun, knife etc.

Show code

```
Total corpa:  113459
Total bigrams:  39456
Total trigrams:  65917
```

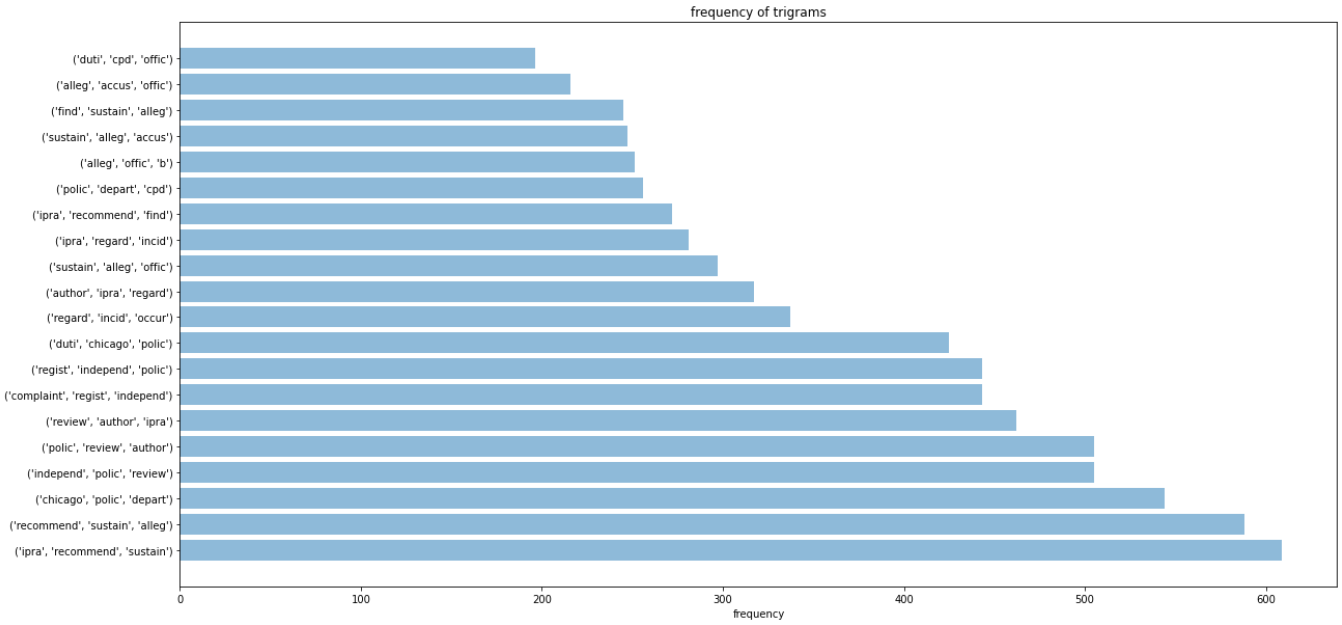
Bigram frequencies

Show code



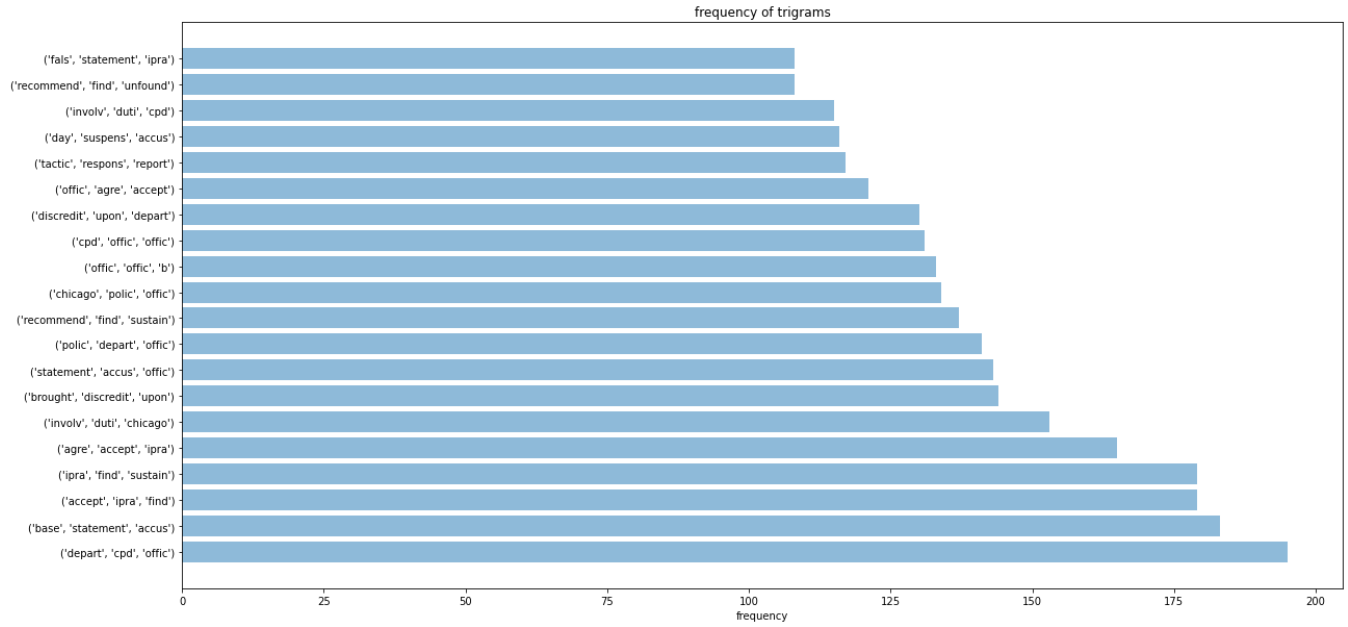
Trigram Frequencies

Show code



Trigram Frequencies continued...

Show code



The bigrams and trigrams unlike the previous dataset do not give a lot of insight on the sentiment but it shows clearly the parties involved in all allegations.

Parties involved

- CPD
- IPRA
- COPA

Conclusion

From this Checkpoint, we can infer that major reforms fostered further negativity due to the prevalence of more reports displaying escalated aggression. We cannot claim the reform completely failed or succeeded due to minimal data post-reform but, the spike in negativity due to reports with escalated aggression could indicate that the reforms fostered more reporting of misconduct. This can be a partial success of the reform. Although, there was barely any downward trend in the negative sentiment. The only success of the reform could be encouraging reporting but not correcting the officers/preventing misconduct.