# Checkpoint 5 Analysis

*The Freedom Donkeys*

*Initial Proposition: We're going to use a topic model on complaint report summaries and see if certain civilian report topics lead to officer discipline more than others. We may also use a keyword based model and see if there is a difference when compared to the topic model. Additionally, we would like to determine whether a given complaint report summary topic matches with that the reported allegation category is.*

**Civilian Complaints**

Below are the number of reports with summaries for each allegation category and allegation name. These reports are civilian reported allegations that result in police discipline of at least 1 of the reported policemen.

```
df_civ_complaints["category"].value_counts()
```

```
Operation/Personnel Violations    25
Lockup Procedures                  5
Conduct Unbecoming (Off-Duty)      4
Illegal Search                     3
Use Of Force                       3
Traffic                            2
Criminal Misconduct                2
Domestic                           1
Drug / Alcohol Abuse               1
Supervisory Responsibilities       1
Name: category, dtype: int64
```

```
df_civ_complaints["allegation_name"].value_counts()
```

```
Inadequate / Failure To Provide Service                17
Prisoners Property                                      4
Miscellaneous                                           4
Improper Search Of Vehicle                              3
Neglect Of Duty                                         2
Slow / No Response                                      2
Association With Felon                                  2
Improper Processing / Reporting / Procedures            1
Reports                                                 1
Conspiracy To Commit A Crime                            1
Damage / Trespassing To Property                        1
Indebtedness To City                                    1
Insubordination                                         1
Intoxicated Off Duty                                    1
Fail To Obtain A Complaint Register Number              1
Excessive Force / On Duty - No Injury                   1
Altercation / Disturbance - Neighbor                    1
Leaving Assignment (District, Beat, Sector, Court)      1
Domestic Altercation - Physical Abuse                   1
Arrest, Improper Procedures                             1
Name: allegation_name, dtype: int64
```

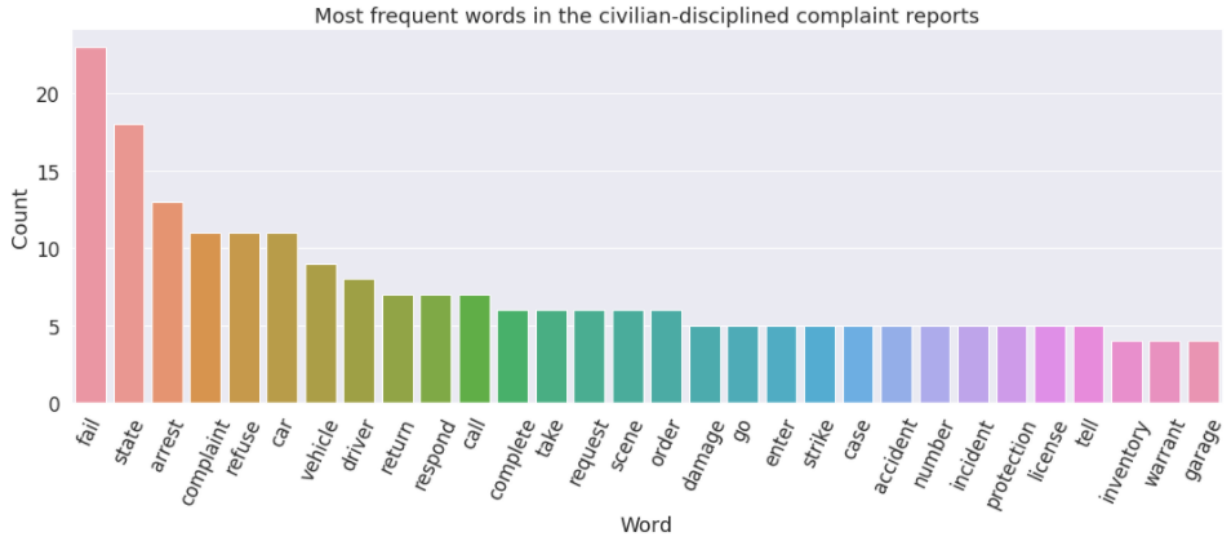
For topic modeling we used LDA. The best LDA model was chosen based on the maximum coherence value, which we found to be 5 topics.

```
Number of Topics 3
Coherence Value 0.3474739616186633
Number of Topics 5
Coherence Value 0.4423622050481605
Number of Topics 8
Coherence Value 0.38698400669130695
Number of Topics 10
Coherence Value 0.3140707553355784
Number of Topics 15
Coherence Value 0.4400048883739584
```

After that, we can see below the topics generated by the LDA model. A subset of the words found in each topic can be seen below. Additionally, to see a full list of each word in each topic, see the visualizations at "checkpoint-5/data/interactive_visualization_lda/ldaModel_civ_vis.html" and "checkpoint-5/data/interactive_visualization_lda/ldaModel_pol_vis.html".

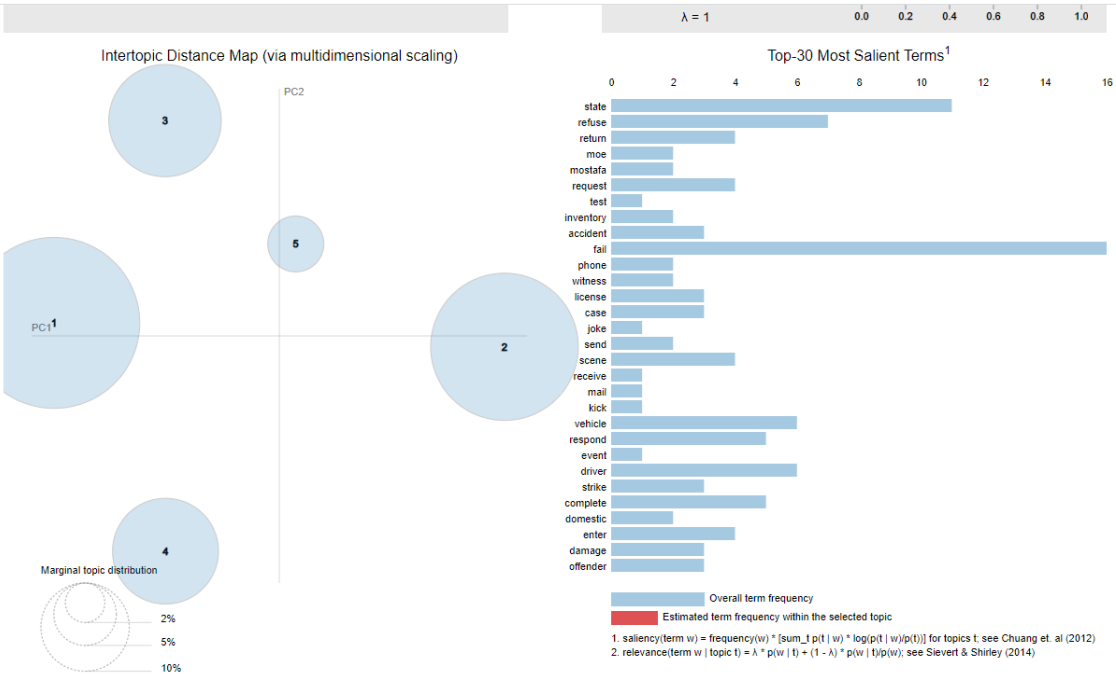
```
best_ldaModel_civ.show_topics()
```

```
[(0,
  '0.023*"fail" + 0.016*"case" + 0.014*"return" + 0.013*"arrest" + 0.012*"driver" + 0.012*"strike" + 0.011*"domestic" + 0.011*"event" + 0.011*"incident" + 0.011*"enter"'),
 (1,
  '0.032*"state" + 0.021*"complaint" + 0.017*"fail" + 0.011*"car" + 0.011*"arrest" + 0.011*"license" + 0.011*"driver" + 0.010*"enter" + 0.009*"refuse" + 0.008*"unprofessional"'),
 (2,
  '0.034*"refuse" + 0.020*"moe" + 0.020*"mostafa" + 0.020*"request" + 0.017*"accident" + 0.016*"car" + 0.016*"state" + 0.015*"phone" + 0.015*"scene" + 0.015*"witness"'),
 (3,
  '0.026*"state" + 0.017*"test" + 0.017*"joke" + 0.014*"return" + 0.014*"inventory" + 0.013*"fail" + 0.010*"receive" + 0.010*"send" + 0.010*"mail" + 0.009*"kick"'),
 (4,
  '0.033*"fail" + 0.016*"complete" + 0.014*"arrest" + 0.013*"order" + 0.013*"protection" + 0.013*"vehicle" + 0.012*"complaint" + 0.011*"howard" + 0.011*"franklin" + 0.010*"tell"')]
```

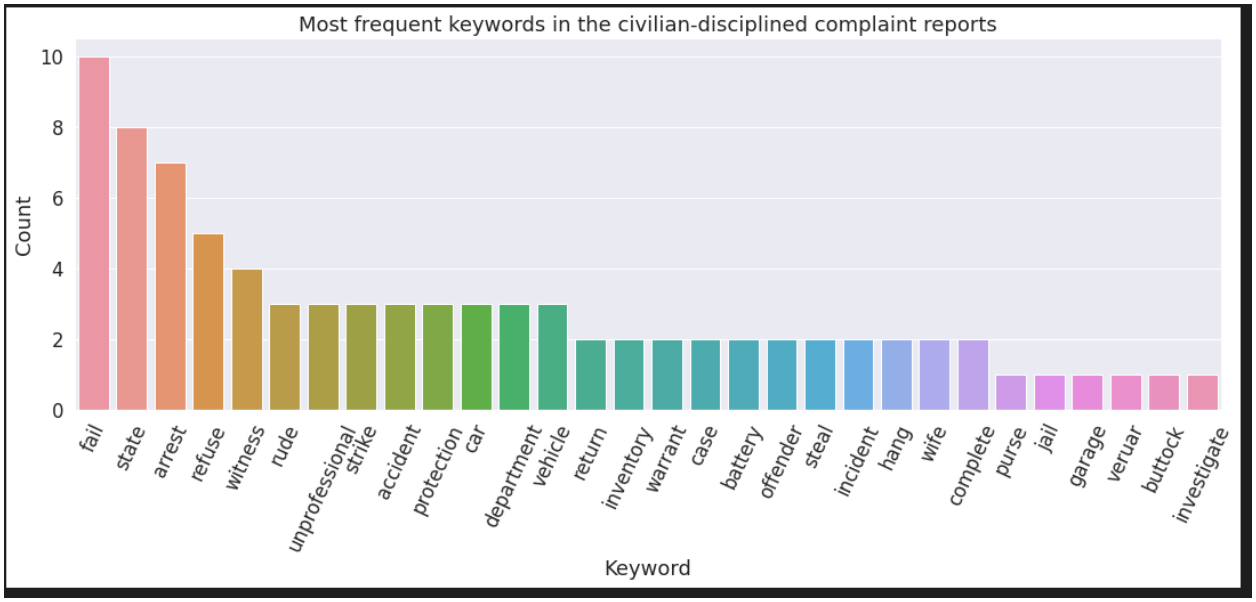


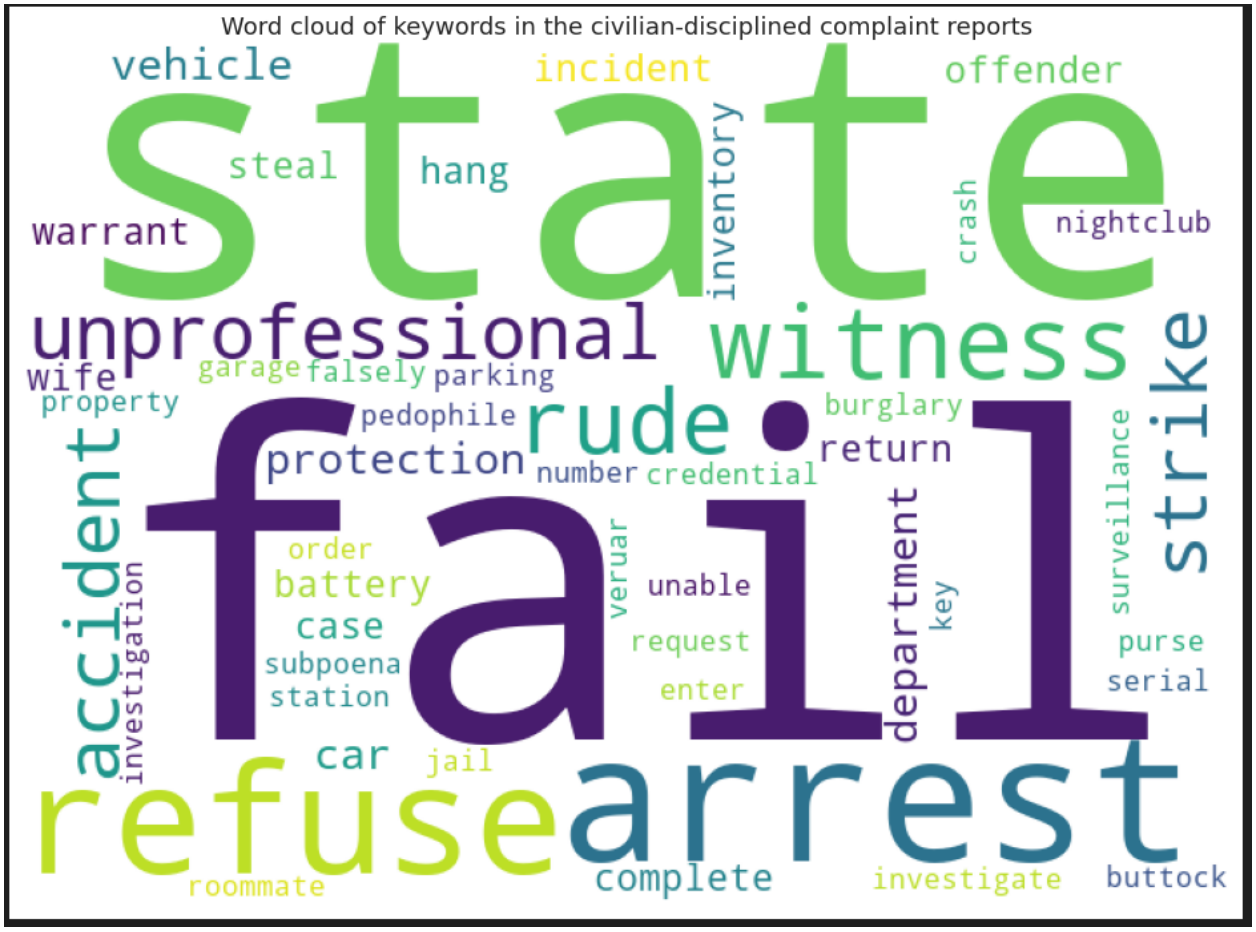

From the intertropic distance map, we can see that our topics are distinct and unique; there is little to no overlap between the topics.
The topic 1 has keywords such as “vehicle, howard, garage, driver, door, apartment, desk, nephew, complaint” and more, so we determined that this topic is related to traffic violations and domestic disturbances.
Additionally, topic 2 has keywords such as “fail, car, arrest, driver license, enter, refuse, unprofessional” and others, so we quickly realized that there was some overlap between the topics despite the tool reporting that there is not.

Overall, we found that several of the topics the LDA model found had similar contextual words in them, which made it impossible to determine whether they matched up with what the reported allegation category was for a given civilian-complained allegation report summary.




Our keyword analysis of the summaries can be seen below. We used the keybert package to generate our keyword models. The words such as “state, fail, arrest” were found as prominent key words in most of the report summaries. However, these words do not fit into any specific allegation category again, so our keyword topics are equally as bad as our topic model.


Most frequent keywords in the civilian-disciplined complaint reports
Count
Keyword
0
2
4
6
8
10
fail
state
arrest
refuse
witness
rude
unprofessional
strike
accident
protection
car
department
vehicle
return
inventory
warrant
case
battery
offender
steal
incident
hang
wife
complete
purse
jail
garage
veruar
buttock
investigate


Word cloud of keywords in the civilian-disciplined complaint reports
vehicle
incident
offender
state
steal
hang
inventory
warrant
crash
nightclub
unprofessional
witness
wife
garage
falsely
parking
property
rude
burglary
pedophile
protection
return
number
credential
strike
accident
fail
surveillance
order
battery
veruar
unable
department
key
investigation
case
request
purse
subpoena
enter
serial
station
car
jail
refuse
arrest
roommate
complete
investigate
buttock

**Police Complaints**

Below are the number of reports with summaries for each allegation category and allegation name. These reports are police reported allegations that result in police discipline of at least 1 of the reported policemen.

```
df_pol_complaints['category'].value_counts()
```

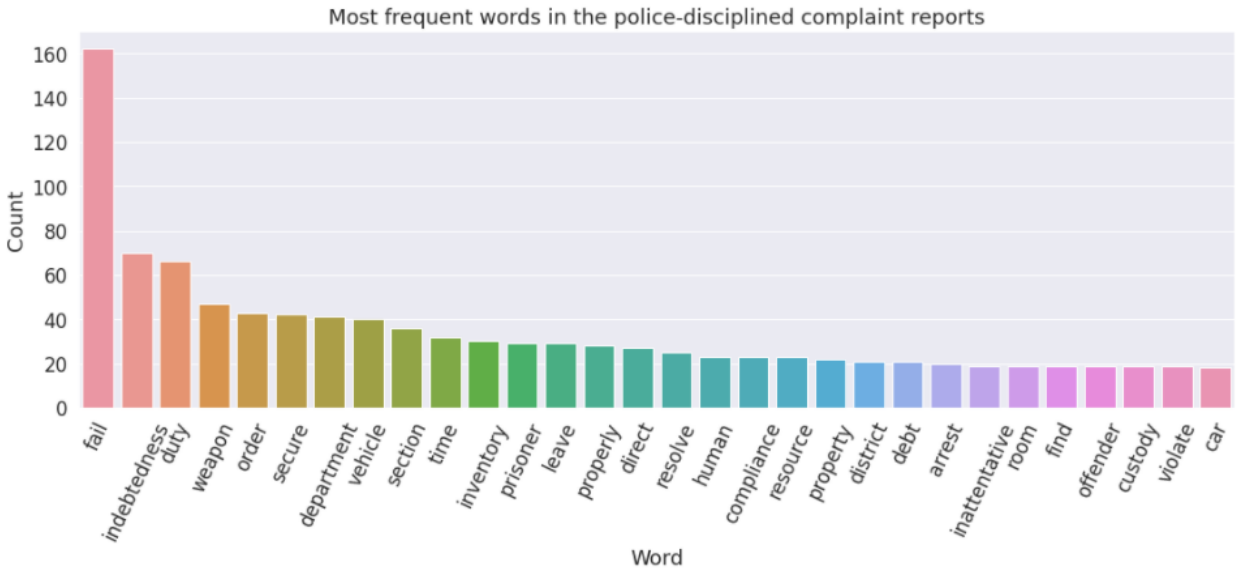
```
Operation/Personnel Violations    209
Conduct Unbecoming (Off-Duty)      87
Lockup Procedures                  24
Drug / Alcohol Abuse               11
Criminal Misconduct                 5
Bribery / Official Corruption       3
Use Of Force                        2
Traffic                             1
Illegal Search                      1
Supervisory Responsibilities        1
False Arrest                        1
Name: category, dtype: int64
```

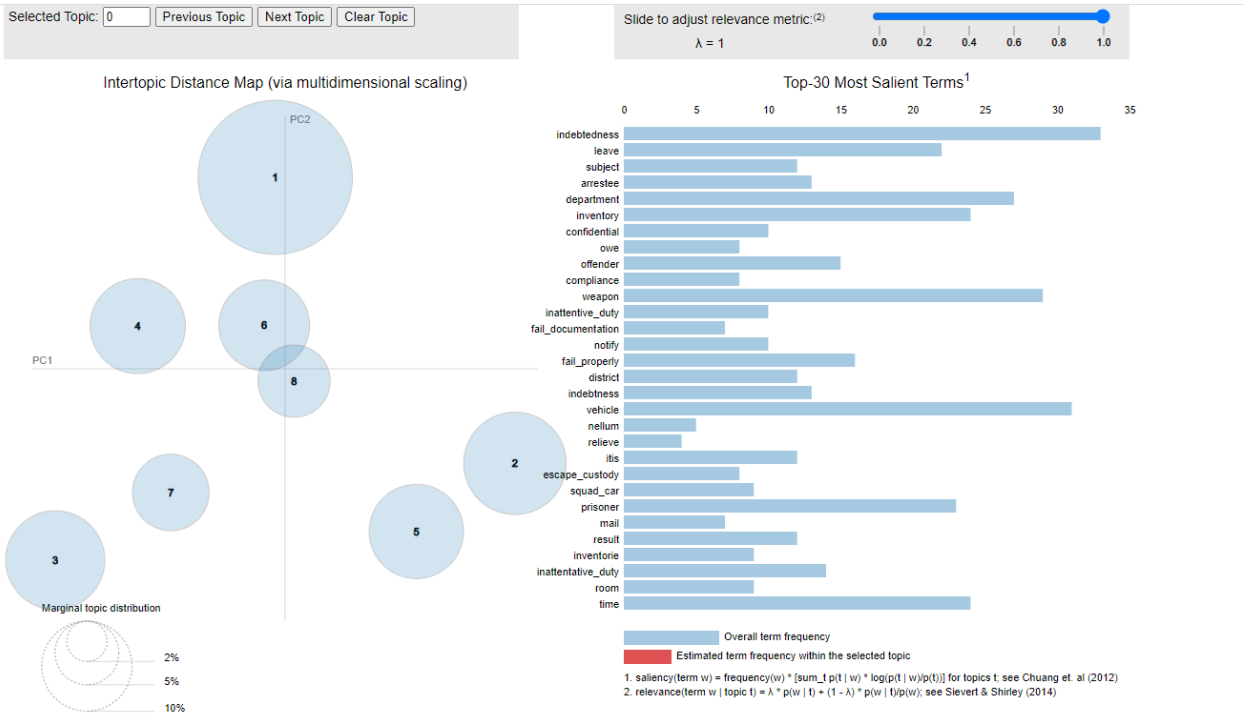
```
df_pol_complaints['allegation_name'].value_counts()[0:25]
```

```
Neglect Of Duty                                          77
Indebtedness To City                                     70
Miscellaneous                                            29
Weapon / Ammunition                                      23
Insubordination                                          22
Misuse Of Department Equipment / Supplies                15
Association With Felon                                   13
Escape                                                   13
Inventory Procedures                                     10
Leaving Assignment (District, Beat, Sector, Court)        9
Seat Belts                                                9
Search - Person / Property                                4
Intoxicated On Duty                                       4
Absent Without Permission                                 4
Prisoners Property                                        4
Court Attendance Irregularities                           3
D.U.I. - Off Duty                                         3
Gang Affiliation                                          2
Inadequate / Failure To Provide Service                   2
Conspiracy To Commit A Crime                              2
Secondary/Special Employment                              2
Excessive Force - Use Of Firearm / Off Duty - No Injury   2
Reports                                                   2
Bonding/Booking/Processing                                2
Intoxicated Off Duty                                      2
Name: allegation_name, dtype: int64
```

We used LDA to perform topic modeling. The best LDA model was chosen based on the maximum coherence value, which we found to be 8 topics.

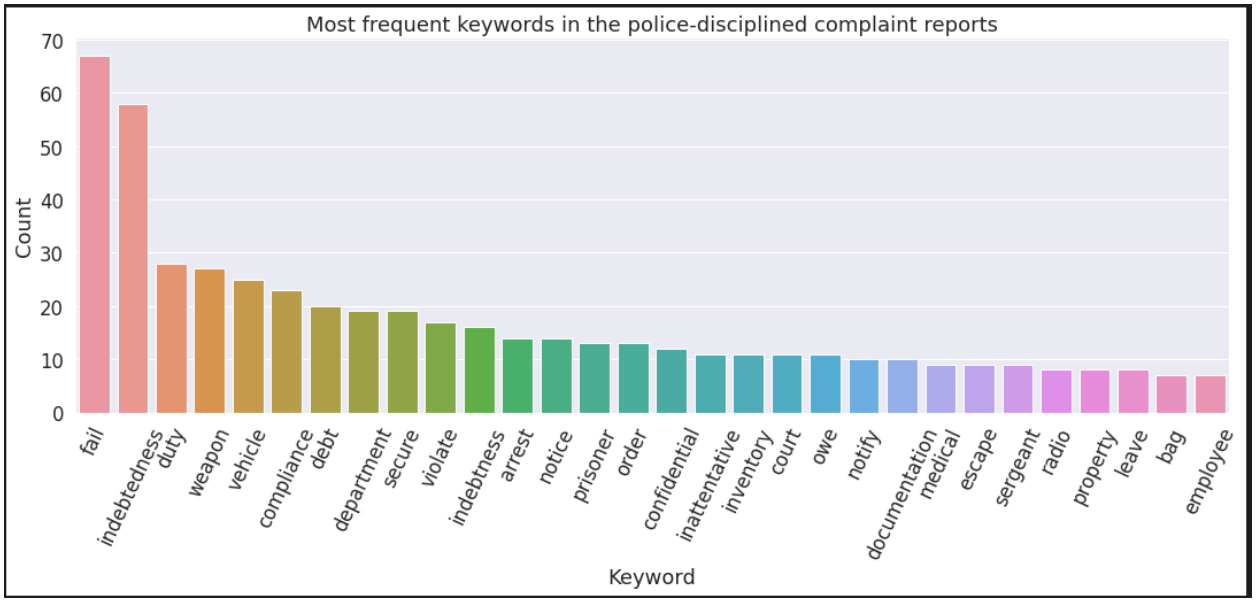
```
Number of Topics 3
Coherence Value 0.4642141918227911
Number of Topics 5
Coherence Value 0.41050223788609497
Number of Topics 8
Coherence Value 0.46513365822195474
Number of Topics 10
Coherence Value 0.44409424866042
Number of Topics 15
Coherence Value 0.4293151886679532
```






We again found that many of the topics shared keywords such as indebtedness, inventory, and compliance. However, we achieved better overall topics that were similar to the allegation categories, as topic 1 seemed to relate the most with indebtedness, although the keyword indebtedness appeared in several topics, other words like “resolve_indebtedness” and “compliance_indebtedness” showed up in this category. Topic 4 seemed to relate to insubordination, which is a part of “personnel violations”. However, none of the other categories seemed to divide evenly into the allegation categories. We saw that some allegation categories such as ‘drug/alcohol abuse’ were distributed in several categories, with keywords like “narcotics, substance, and alcohol” showing up in several topics(1, 3 and 5 respectively).

We also attempted to search for any keywords using keyBERT to determine whether the keywords found using that method were any different from the keywords found using the GENSIM library. However, these keywords were all similar to the ones found using GENSIM.





**Discussion**

We planned on using this checkpoint to verify that the data we are seeing in each allegation category is accurate. After analyzing the summaries, we determined that they contain a lot of unstructured data that provided little value when doing a topic analysis. Additionally, our

keyword analysis did not provide any better results/insights. We set out to answer the question of “does a summary match up with what the reported allegation category is?”, but the topics found by LDA seemed to have little semantic meaning. Further exploration of the topics of these summaries is necessary.