Checkpoint 5 Findings

The Wise Lobsters

Question 1

Based on an allegation type/name, can the probability of the same type of allegation to occur be predicted?

Our goal was to predict a model that can determine the probability of the same type of allegation for officers having first or early allegation. It was observed from cpdp data that a particular officer has been offended with same type of allegation and there were high chances of him to be a repeater.
To answer this question we have included 4 features as input passed in machine learning model. These features are category, years on force of officers, count of same type of allegations since they were appointed and the count of the total allegations down the line. Our label was count of same type of allegations after their early allegation. In order to predict the probability of the allegation type we have:

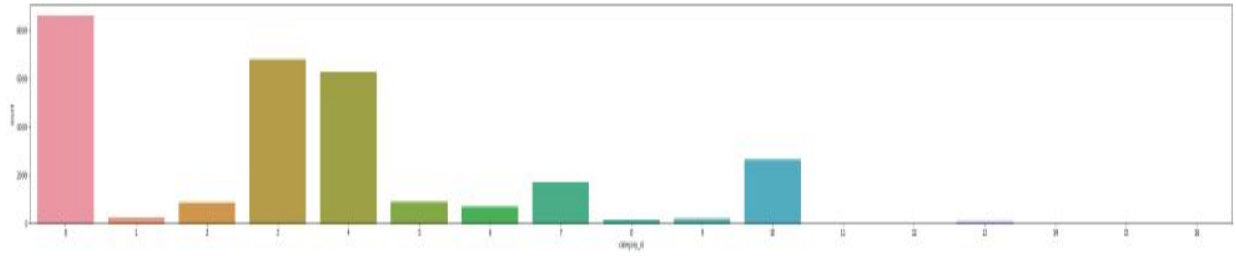
1) Determining data by Sql query (uploaded in src folder). We created view of subset of officers having first/early allegations having appointed_date BETWEEN '2000-01-01' AND '2007-12-31'. And extracted our features from it (as stated above).
2) We have used Sklearn to train our model and performed logistic regression algorithm for training the model.

Analysis:

With our data(csv), we had all the features(stated above) to determine the output. However, we formed column named ‘category_id’ which gave index to each different categories. For example 0 means Operations/Personal violations, 1 stands for Criminal Misconduct etc.

	category	years_on_force	number_since_same_type	number_since_total	category_id
0	Operation/Personnel Violations	0	0	0	0
1	Operation/Personnel Violations	0	0	0	0
2	Operation/Personnel Violations	0	0	0	0
3	Criminal Misconduct	0	0	2	1
4	Operation/Personnel Violations	0	0	0	0
5	Operation/Personnel Violations	0	0	0	0
6	Operation/Personnel Violations	0	0	2	0
7	Operation/Personnel Violations	0	0	0	0
8	Operation/Personnel Violations	0	0	0	0
9	Domestic	0	0	0	2
10	Operation/Personnel Violations	0	0	0	0
11	Use Of Force	0	0	0	3
12	Operation/Personnel Violations	0	0	4	0
13	Operation/Personnel Violations	0	0	0	0

Next, in order to analyze the data, we have plotted graph for count of category_id using matplot lib and seaborn. It was clearly observed that category_id 1 was most used by the officers having max count which is Operations/Personal violations.



Our label was count of same type of allegations after their first allegation. So we now have our training data for both input and output. We have tested on 30-70 ratio for training and test data. And printed our predictions and f1 scores for all categories.

```
predictions= logmodel.predict(X_test)
print(predictions)
```

```
[0 1 0 ... 0 0 0]
```

```
print(classification_report(y_test , predictions))
```

```
              precision    recall  f1-score   support

           0       0.66      0.99      0.79      5767
           1       0.17      0.02      0.03      1650
           2       0.00      0.00      0.00       690
           3       0.00      0.00      0.00       307
           4       0.00      0.00      0.00       156
           5       0.00      0.00      0.00        86
           6       0.00      0.00      0.00        50
           7       0.00      0.00      0.00        28
           8       0.00      0.00      0.00        24
           9       0.00      0.00      0.00        15
          10       0.00      0.00      0.00         5
          11       0.00      0.00      0.00         2
          12       0.00      0.00      0.00         1
          13       0.00      0.00      0.00         1
          15       0.00      0.00      0.00         1

    accuracy                           0.65      8783
   macro avg       0.06      0.07      0.06      8783
weighted avg       0.47      0.65      0.53      8783
```

It gave 65.28 accuracy which was best among other classifiers which we have tried. Since from the above graph and data we can see that allegation of category id 0 was mostly done by officers followed by 1. Thus, this shows a small amount of actual information predicted by the model.

Conclusion:
By performing the above experiment, we were able to see the probabilities of particular category of allegation. However, according to data it was fruitful to find outstanding results. In future we could consider other features with new severity data sets released recently which we think can blow up our analysis.Along with this, we could go for bernoulli's algorithm to train our model.

Question 2

Given the first allegation type/name, can we predict the likelihood of an officer becoming a repeater?

Overview
We are framing this as a categorization problem where our output is a binary classification of some repeater threshold, but not necessarily the simplest definition of a repeater. While this is a simplistic experiment, the overall goal of an analysis like this would be to flag officers as a risk early in their career given a prediction of their behavior. If an officer could be put on some type of probation because of early behavior, perhaps the bad behavior could be stopped.

Many parameters were tried to see what type of prediction accuracy we could get. First, we need to define what is the time frame of an early career. Initially we were looking at 1 year, but found too few records. We expanded the search to 2 years and decided to aggregate information from the first 2 years. SKlearn was used to train our models

Analysis
There was a lot of tweaking to find an experiment that offered any type of meaningful results. Our final experiment training a model used the following features:

- Category of first allegation (any category)
- Tactical Response Report count for the first 2 years of employment
- Total allegation count for the first 2 years of employment
- Number of disciplined allegations in the first 2 years of employment

We labeled a repeater as an officer that has 5 severe allegations in 7 years of employment. Along with tweaking features we trained and compared a number of models for the binary classification. Logistic regression, decision tree classification, multilayer perceptron classifier and support vector classifier were explored. We measured our models using primarily the f1 score. In general, we had a hard time getting great results, but in our exploration we found that

multilayer perceptron performed the best and with some slight tweaking to the probability threshold for classification, we were able to get the following results:

	precision	recall	f1-score	support
0	0.82	0.86	0.84	203
1	0.73	0.67	0.70	117
accuracy			0.79	320
macro avg	0.77	0.76	0.77	320
weighted avg	0.78	0.79	0.79	320

This model is predicting non-repeater with 84% f1-score and repeater with 70% f1-score with an overall accuracy of 79%. We had a few test runs perform better on accuracy, but this is the best overall f1-score performance. These are results are not astounding, but it does show what a small amount of aggregate information can predict. While a binary result is used to measure performance of the model, a probability would be more desired when trying to use such a model for planning. For example officer *x* has a 40% probability of repeat behavior in the next *y* years. These classification models output these probability then predict based on a threshold, for example above 50%.

Conclusion and Future

After such an experiment and after spending time working in the dataset, we do not believe the data is present in an officers early career for predictive analysis of an officer's behavior. Even if a model performed well, the prediction would be extremely risky if used to enforce action on the officer. Instead, we believe that generally statistical analysis is more suited. It doesn't take much to prove that any type of allegation early in an officers career causes in the uptick in the probability of a repeater. Having any type of allegation in the first year gives that officer a 20% chance of being in the 10% percentile of all allegations in the 10-15 year mark.

We specifically wanted to keep our features simple in this experiment to see what results we could get. Obviously there are many attributes we could have brought in. We think it would be interesting to train a number of different decision tree models and make a custom random forest ensemble for repeat behavior. Decision trees also performed about as well as perceptrons in our training. If more data is pulled in, perhaps from arrest and location and a number of models are trained outputting probability perhaps a better outcome could be achieved.

Question 3

Mapping document tags to a scale of physical violence (mentions of injury, verbal abuse, firearm use, etc) are there an increasing trend of violence over time for an officer with multiple reports starting with the first 'use of force' report?

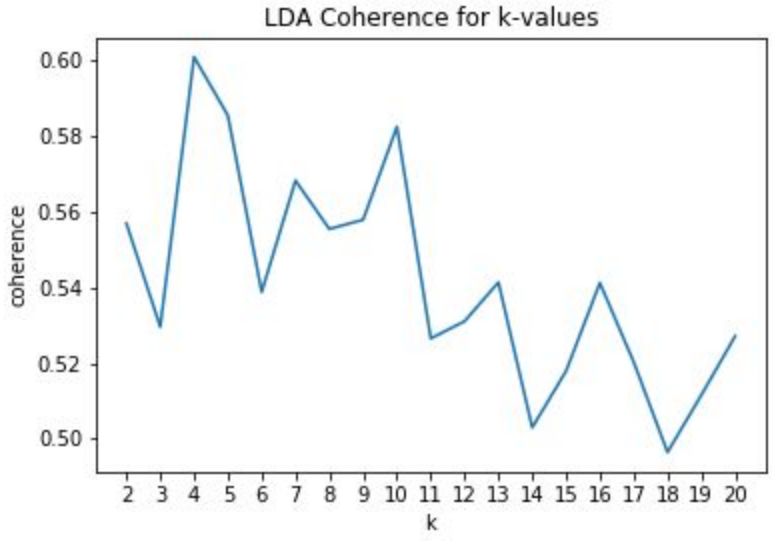
Overview

This question assumes that documents are labeled consistently with underlying allegations or details about the allegations. After exploring the document tags dataset we found that this was not the case, the data is incredibly sparse. After attempting some active learning in challenge 3 we also noticed that many documents are drastically different in format and content. For example some are allegation reports, as one would expect, explaining the details of the investigation and outcome. Others, for example, are arrest reports written by the officer that we assume are used an evidence for the officers account of the arrest. So, general data was not what we expected from the start and our question became harder to answer.

Keeping with the theme of natural language processing, we decided to attempt a different method of tagging. A tagging to explore how we could categorize these documents to answer our question. We used a Latent Dirichlet allocation (LDA) model to create new unsupervised categorization of the use of force documents. LDA is a topic model that groups documents based on words that are related to their similarity. LDA outputs topics and the words associated with each topic and their distributed probability. LDA only has corpus of the documents to train with. To do this we preprocessed the document text by tokenizing, removing non-english words, removing stop words finding bigrams, and lemmatization. Standard english stop words were removed, but also words that are dominant in all the documents and clearly do not relate to the content (e.g. *summary*, *name*, *police report*, etc.) Once cleaned tokens were used to create a bigram of ~160,000 vocab words. The bigram is used to create a corpus from all document texts. Next the LDA models is trained with the corpus and topics are used. We then can visualize the topics on a Intertopic Distance Map to see how they are separated. The main parameters of the model are the number of categories which we experimented with to view degrees of separation. Once trained we can explore the topics for severity and tag each document with these new categories. To learn this model and implement in python we used gensim, spacy and pyLDAvis (visualization). This tutorial was adapted to our problem.

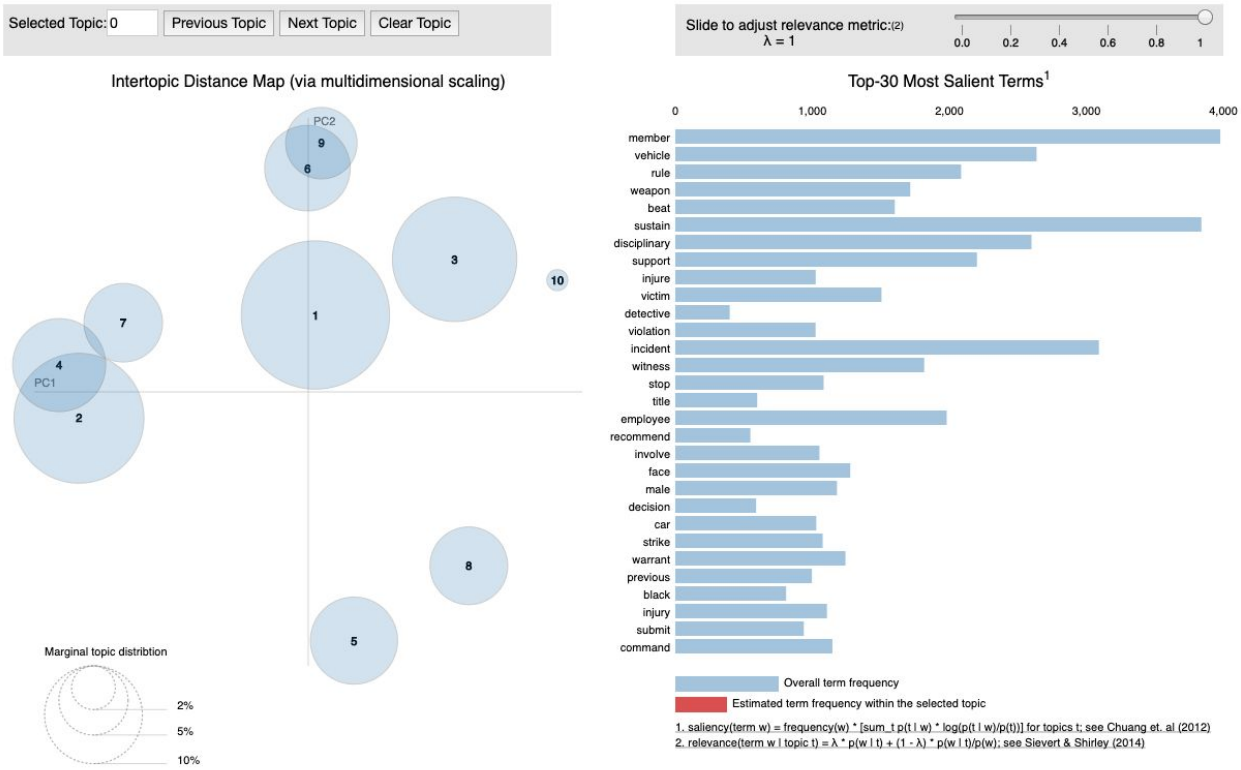
Analysis

From our experimentation the LDA model performed well in separating the documents. We experimented with 2-20 *k* values (number of categories) and measured the perplexity and coherence score. The coherence score is the main measure used for for how well the topics are extracted. Here is our results for training the model with different *k* values:



The training shows that seems that *k=4* results in the best coherence score of ~0.61. We decided to choose *k=10* which is the peak before falling into the mid to low .50s, *k=4*, while slightly better separation, was too low for our analysis.

Using [pyLDAvis](#) the topic separation can be visualized (interactive example can be found in `ida.html`):



Hovering over each bubble in the intertopic distance map shows the probability of words that exist in that topic. There is little overlap of the documents which is a good indication that the model separated properly. Below is a table format of the top 10 words in each topic:

	id	term1	term2	term3	term4	term5	term6	term7	term8	term9	term10
0	0										
1	1	injure	member	support	disciplinary	sustain	professional	medical	property	injury	visible
2	2	weapon	member	discharge	battery	fire	incident	force	firearm	injury	first
3	3	detective	homicide	supplementary	attorney	murder	inspector	taxi	government	emma	strip
4	4	rule	decision	follow	hear	hearing	conduct	write	guilty	suspension	agreement
5	5	go	incident	leave	male	sustain	back	strike	take	face	door
6	6	rule	violation	involve	copa	accountability	write	work	would	give	find
7	7	sustain	member	disciplinary	support	warrant	employee	incident	command	witness	file
8	8	vehicle	stop	car	drive	take	driver	individual	arrest	involve	seat
9	9	incident	witness	employee	face	audio	victim	command	preliminary	weapon	sustain
10	10	beat	victim	title	recommend	concur_recommend	black	unknown	age	event	male

One can see we are starting to get human readable separation with some of these topics. 2 is especially interesting to our question, seemly categorizing the report with some type of weapon. Also topic 3 seems to involve homicide.

The model can label each document with a probability of membership to one of the topics, we labeled each document with their top 3 documents and loaded into our database. LDA outputs a probability of topics membership, so documents can belong partly to multiple topics.

To answer the question about an officer's first allegation, unfortunately given the data, we only have a few examples of officers in our subset. There are about 1800 documents and many more officer allegation within our subset. So the data is very sparse. Although we believe this document tagging can still serve a purpose. While it doesn't necessarily tag the type of underlying allegations, it does provide a sense of severity and an insight into the type of documents. We have identified 1, 2, 4 and 10 as documents that are outlining some type of allegation report and likely with degrees of severity starting with homicide and ending with vehicle stops. Other categories 4, 5, and 6 for example outline a type of document most likely related to outcome or ruling within in the document. In the next question we will start to look at how these documents relate to allegation categories.

Conclusion and Future

While there is not enough data to find trends over time starting with an officer first use of force allegations, we were able to label the documents in a meaningful way. Given more data we believe this could serve as a baseline for mapping severity and to flag documents for further labeling. Relating to our theme, just having a documented allegation report is a good indication that something is wrong, but also having a clear concrete example of violent behavior is

something that could trigger further action. Retrieving more of these documents and having the ability to automatically tag them could be very important in identifying at risk officers.

The document tagging is a good start but there can much more tuning and enhancements to the process if put into use. It would be interesting to spend more time combing through commonly used words and removing more that do not relate to the overall theme of the document. We spend some time with this but it is a tedious process. Also it would be an interesting goal of this tagging to first attempt to categories documents by type, then identifying allegation reports and using a new corpus to separate them. Filtering out arrest reports and other reports that do not outline the allegation may allow for a better separation of violence.

Question 4

Are there new findings that can be derived from comparing our checkpoint 2 question: "How as the number of complaints by allegation type/name changed over time?" replacing allegation type with potential underlying misconduct from the document tagging?

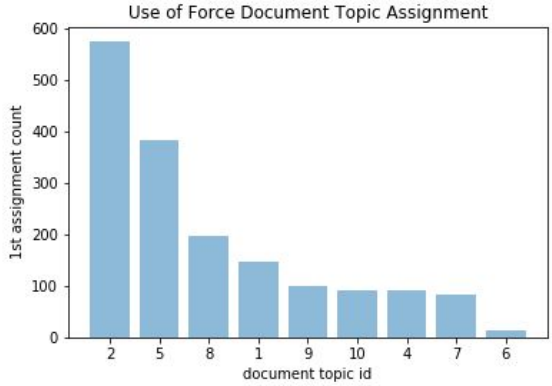
Overview

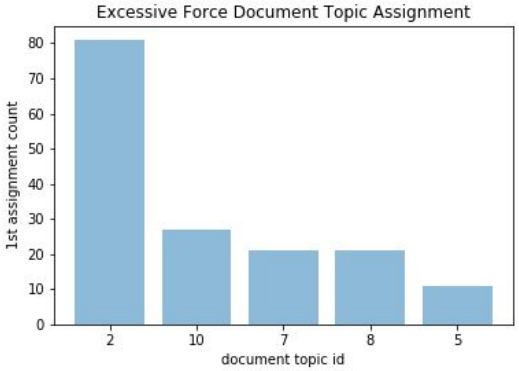
Again, the labels that we are given for allegation do not have the extent of labeling we expected when first formulating this question. Luckily, we can use our new labeling from above to see if there are changes in the type of reports over time. An example of where this could be insightful is if our LDA model is creating topics based on time of the incident. This could be in result to logistic changes in the way allegations or reported over time. If the topics seem to remain consistent we can use this as a confirmation of the separation of our model.

Along with a test of our LDA, we will look generally at any shifts in allegation category to document labeling in our dataset. Given our results from checkpoint 2, we generally saw a downtrend in allegations, but nothing sufficient to report in regards to change over time. It will also be interesting to see mapping of topics to allegation types over time with our LDA model.

Analysis

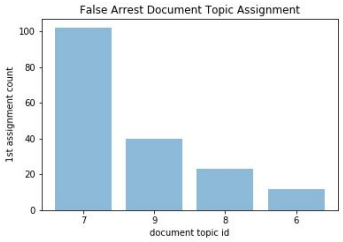
The first step in our analysis is to see how our LDA model tags specific categories (overall not constrained by time). We focused on 5 categories: Use of Force, Excessive Force, Illegal Search, and Operation Personnel Violations. Below are the first two:

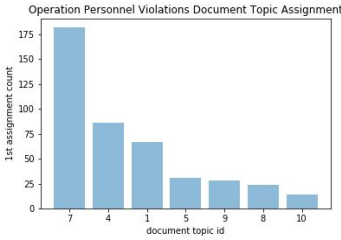


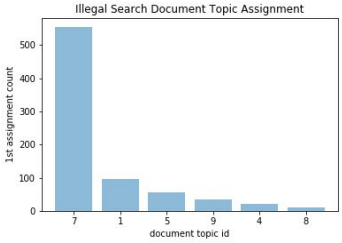


As you can see these have similar tagging, which is expected and a good sign for our tagging model. For reference topic 2 relates to words like weapon, member, discharge, battery, fire incident, force, firearm, and injury. This seems like a very relatable topic for these two categories.

Here are the later 3 categories:

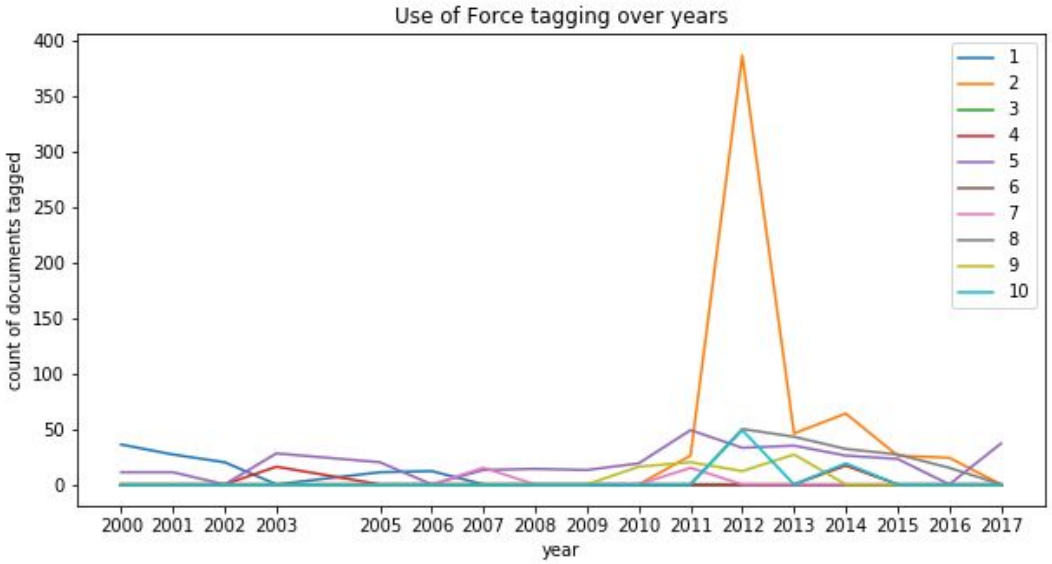


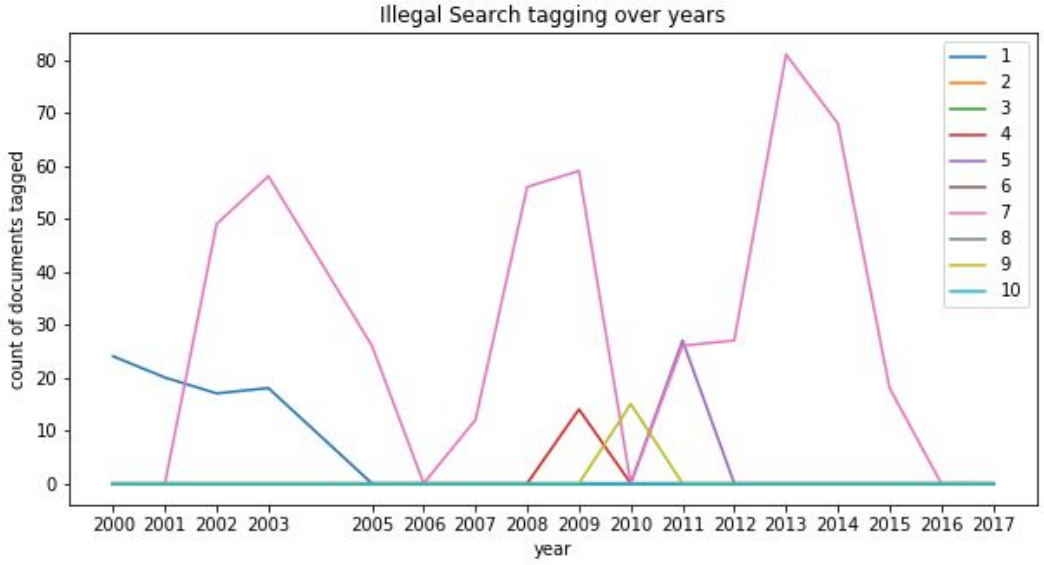




These also have similar primary tagging, but different after the first category of 7. For reference topic 7 relates to sustain, member, disciplinary, support, warrant, employee, incident command, witness. Again this topic seems reasonable given the above (no violence).

Next we looked at these topics over time. This we will explore the use of force and illegal search since they have the most documents in the database. We found some interesting results:





The time of the allegation incident date (year) does seem to have some effect on how the document is tagged. Use of force has an extreme spike in the number of documents labeled as category 2 in 2012 and it makes up a large portion of the dataset. Also 2004 is missing any documents. Illegal search has peaks and valleys for category 7, which also makes up the majority of the documents.

Conclusion and Future

This shows analysis continues to show the inconsistency in the documents dataset and this is proven by the type of document tagging at specific years. Unfortunately these documents don't provide enough data to extract insight for an officer early career, but we can further extend our

point from checkpoint 2 that the time in which these allegations are reported can have a big effect on the data. This causes problems when trying to look at patterns over time because results are not consistent throughout the 18 year span of the data. Documents and categories of allegations are more “popular” at different times. We assume that the actual allegations categories if labelled properly would follow some normal distribution. In the future if may be beneficial to work on a separate tagging system for allegation, perhaps more general than allegation categories, to try and normalize this data.