Liqian Ma
Wentao Yao
Xingbang Liu

Checkpoint 5: NLP

OVERVIEW & PURPOSE

In this analysis, we would like to further explore the relationship between over-policing and economic status. Specifically, summarization and criminal report texts can be analyzed with NLP techniques. For example, is there a topic difference between different areas? Also, Bert can be applied to generalize embeddings for the texts and calculate similarities between them.

Question 1

Question 1: Compare the Bigram and Trigram in the top 5 highest and lowest socio-economy status communities.

Analytics

To understand the misconduct that happened in low socio-economy and high socio-economy areas, we need to understand the differences between the incidents that happened in each area. Topic modeling is a great way to understand the gist of each incident. To first get a general impression of the gist of each incident, we can use the statistical approach to calculate the frequency of each term.

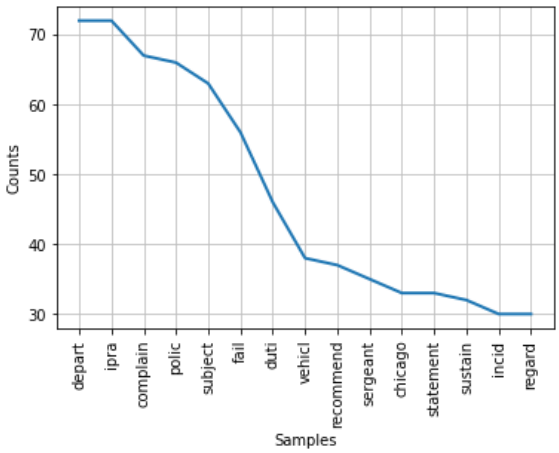
Discussion

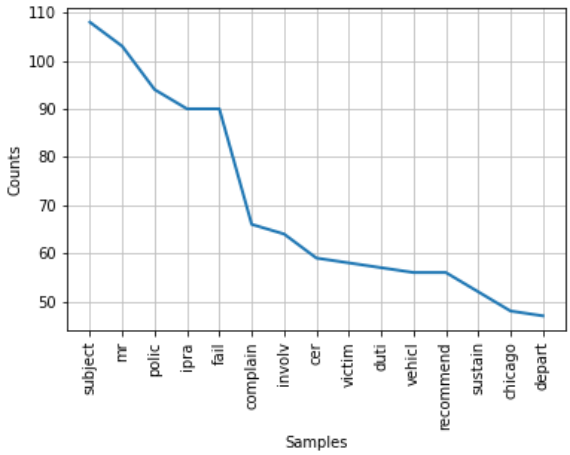
The summary of the incident can be found in the data_aligation table. We need to first split the data by the socio-economy status. The following SEQUEL was used to group the summary by connecting the beat area with the communities since only communities have income data.

```
SELECT summary from data_allegation,(SELECT name, id, median_income
FROM  data_area
WHERE median_income IS NOT NULL
ORDER BY CAST( replace(replace(median_income, '$',''),',','') AS INT )DESC
LIMIT 5)b, data_allegation_areas where data_allegation.crid = data_allegation_areas.allegation_id and
                        data_allegation_areas.area_id = b.id and not data_allegation.summary  = ''
```

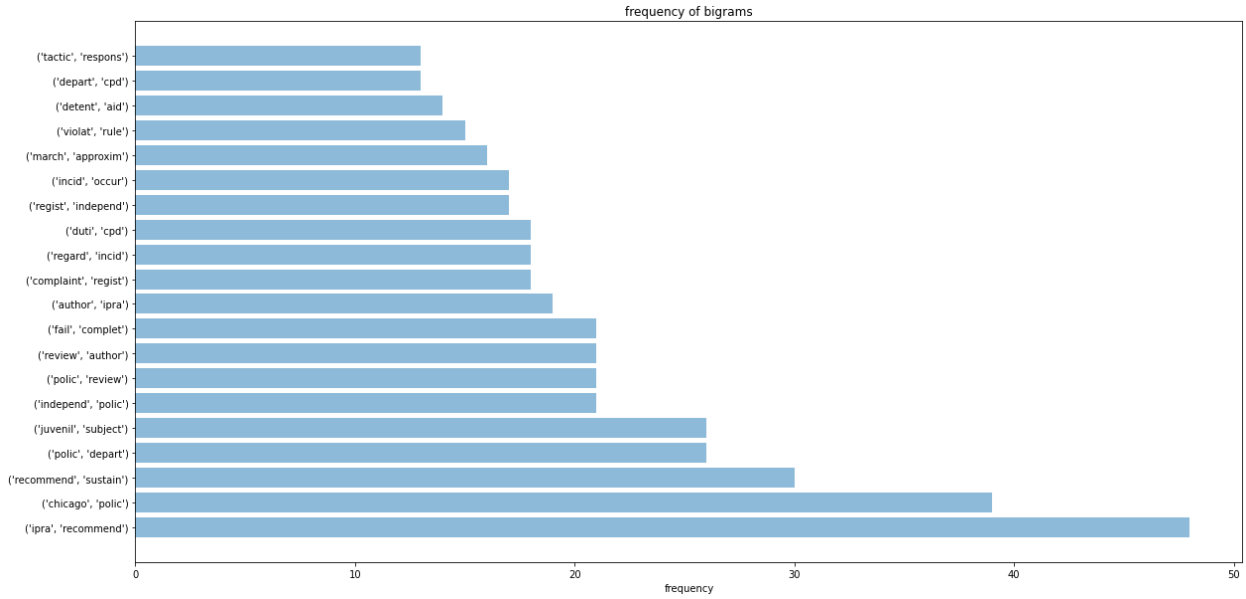
```
SELECT summary from data_allegation,(SELECT name, id, median_income
FROM  data_area
WHERE median_income IS NOT NULL
ORDER BY CAST( replace(replace(median_income, '$',''),',','') AS INT )ASC
LIMIT 5)b, data_allegation_areas where data_allegation.crid = data_allegation_areas.allegation_id and
                                    data_allegation_areas.area_id = b.id and not data_allegation.summary  = ''
```

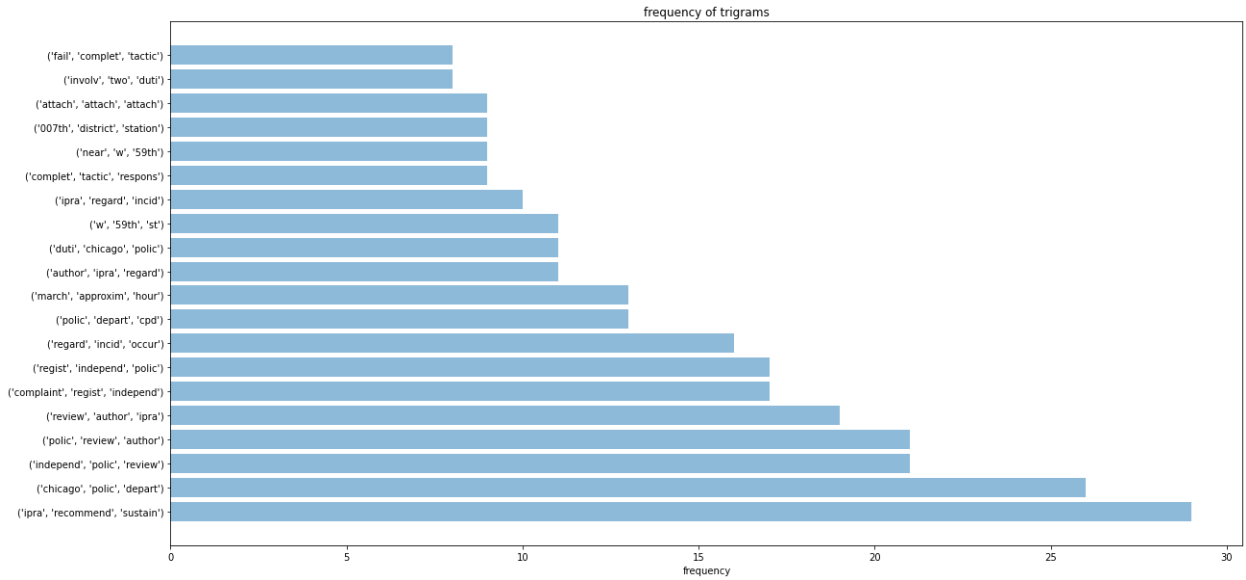
To capture the topic, the most frequent term can provide a general idea. After the pre-process for the text, the following graphs shows the term frequency in high-income areas and the term frequency in low-income areas.



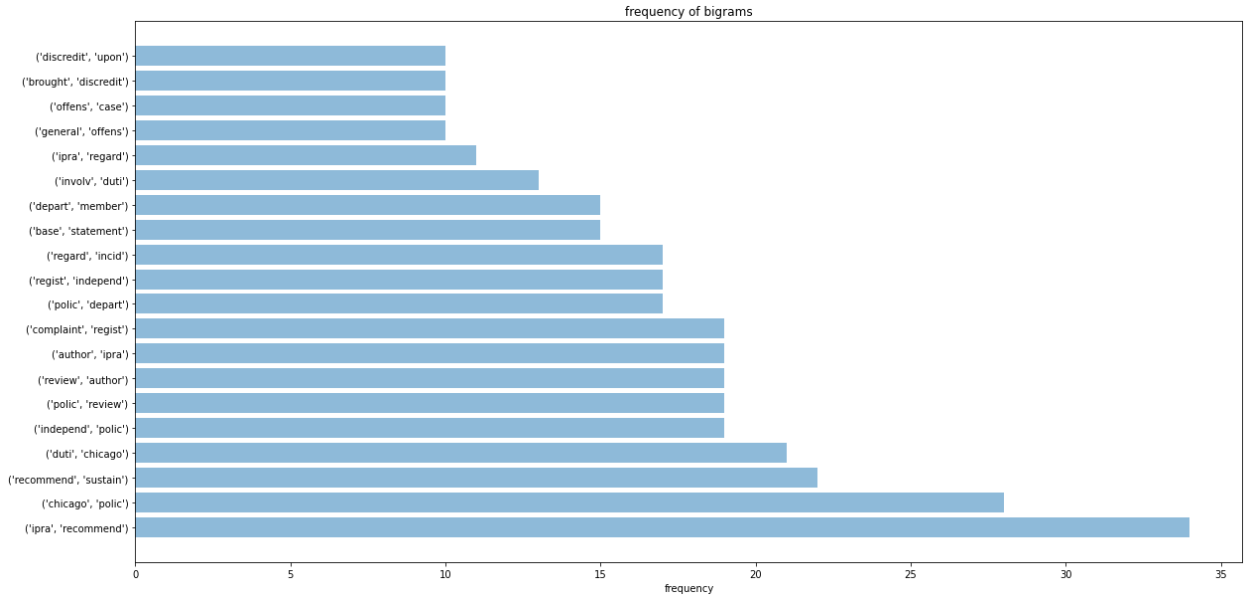


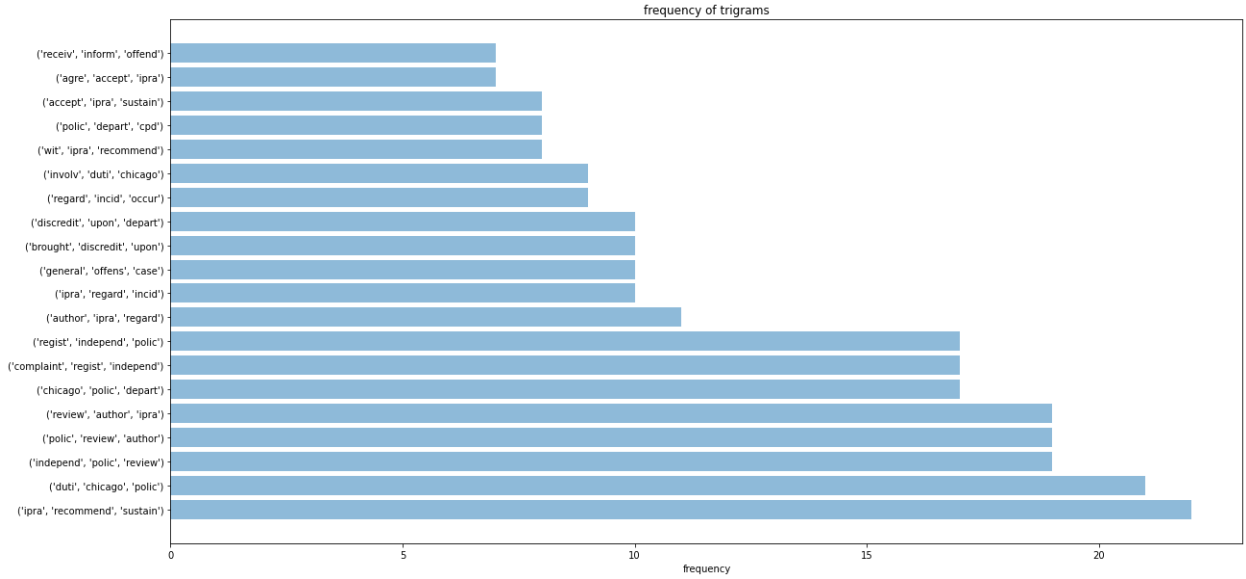
The problem with term frequency is that a single word cannot provide enough information for the context. For example, depart and subject in each plot represents the most frequent term. However, they cannot provide enough information for people to understand the context. To solve this problem, we can calculate the frequency for a sequence of words, which is the continuous words in a sentence. The high-frequency sequences are hard to view since there can be too many terms in similar frequencies, we inverted the graph so that the shortest bar represents the highest frequency. The followings are the Bigram and Trigram in low socio-economy status communities.





The followings are the Bigram and Trigram in low socio-economy status communities.





Conclusion

From the above images, we can see that for the high-income community, we got “involv duti”, “polic review” and etc., or other normal behavior that we would expect from officers. However, in the low-income community, we find “tactic respond”, “violat rule”, and etc. The sequences in low-income areas indicate the high rate of policing in the area. However, it is hard to tell if there is any misconduct in the action.

Question 2

Question 2: For similar(in contents) complaint summary texts, are they having similar eco-socio status to each other?

Analytics

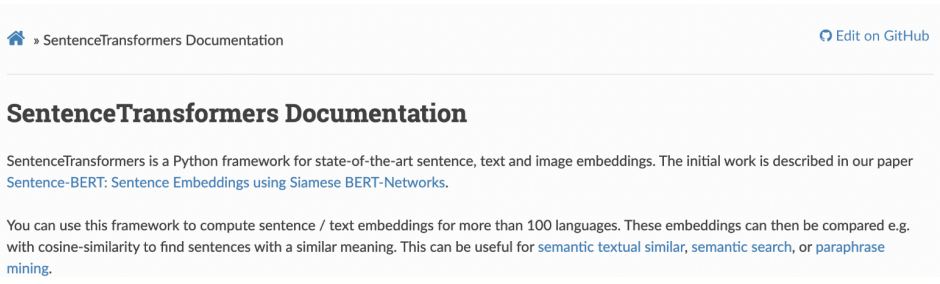
To understand the texts, one way is to make use of word frequency counts and topic modeling methods like we did in question1. The other feasible method is Bert. Bert is a language model which can be used to generate representational embeddings for input texts. We can further calculate the similarity between each of them. Moreover, we make use of the top 3 closest texts as edges between each other and create a graph. Finally, connected components are found to calculate the mean income for the cluster.

Discussion

In the original data allegation table, there is no clue for income in the area. Fortunately, we can make use of the data_area table to get it. But, it is hard to join two tables directly. Here we make use of the polygon relationship to create a mapping between community id and beat id and thus join texts with median income.

```
select a.*, c.median_income::money::numeric as income from data_allegation a join
(SELECT DISTINCT ON(1) table1.id as beat_id, table2.id as community_id FROM
                (SELECT * FROM  data_area WHERE data_area.area_type ='beat')table1,
                (SELECT * FROM  data_area WHERE data_area.area_type ='community')table2
   WHERE ST_Contains( geom1: table2.polygon,  geom2: table1.polygon)
      or st_intersects( geog1: table2.polygon,  geog2: table1.polygon) ) b on a.beat_id=b.beat_id
join data_area c on b.community_id = c.id;
```

SentenceTransformer is an implementation of the paper Sentence-Bert to generate embeddings for sentences, texts, and paragraphs. Here we made use of a lite version of the pre-trained model.



To get the embedding of the text, we just make use of one line of API from input text to 128 dimensional embedding like this.

```
from numpy import dot
from numpy.linalg import norm

embeddings = model.encode(np.array(train))
```

Next, to calculate the similarities between each other, we have two loops to apply cosine similarities calculations to make a 1105x1105 similarity score matrix.

```
res = []

for a in embeddings:
  for b in embeddings:
    res.append(dot(a, b)/(norm(a)*norm(b)))
res = np.array(res)
```

```
res = res.reshape(-1, len(train))
print(res.shape)
```

```
(1105, 1105)
```

Then, we create a graph based on every summary text as nodes and top 2 similar nodes as edges.

```
import networkx as nx
G = nx.Graph()
G.add_nodes_from(list(range(res.shape[0])))
```

```
for i in range(res.shape[0]):
  neighbors = df.loc[:, i].nlargest(2).index.tolist()
  # print(neighbors)
  G.add_edges_from([(i,j) for i in neighbors for j in neighbors if i > j])
```

Let's look at the length of outstanding components we have.

```
# nx.connected_components(G)
[len(c) for c in sorted(nx.connected_components(G), key=len, reverse=True) if len(c) >= 15]
```

```
[58, 40, 35, 26, 23, 22, 22, 21, 21, 21, 18, 18, 16, 16, 15]
```

Finally, we calculated the average income for each cluster as opposed to the mean income for the whole population.

```
train_ori_map = {i: train_ori.iloc[i, 1] for i in range(train_ori.shape[0])}
```

```
avg_income = []

for c in sorted(nx.connected_components(G), key=len, reverse=True):
  if len(c) >= 15:
    avg_income.append(np.mean(list(map(lambda x: train_ori_map[x], c))))
```

```
avg_income
```

```
[50464.18965517241,
 49322.6,
 45312.4,
 50771.42307692308,
 41052.608695652176,
 44056.13636363636,
 33738.36363636364,
 42120.619047619046,
 47985.380952380954,
 46466.666666666664,
 46018.444444444445,
 49241.38888888889,
 39037.5625,
 34008.4375,
 52417.13333333333]
```

Mean income for all texts.

```
np.mean(np.array(train_ori.iloc[:,1]))
```

```
45572.10407239819
```

Conclusion

From the above analysis of embedding of the texts, we found that similar texts cluster have outstanding average income than the whole population. For example, several clusters with over 15 data points have a lower income of 30,000 while the average income is about 45,000. We can conclude that different economic areas have different complaint texts with each other, which means a different type of policing in these areas.

Question 3

Question 3: Is there any bias in the complaint report? In other words, is the complaint report narrative different from the incident summary?

Analytics

In this question, we planned to group the data into high-income and low-income areas. According to our previous finding that high-income areas have more complaints toward the police officers compared with the low-income areas, our hypothesis was that high-income areas have more bias in complaints. The bias degree can be further used to measure if the complained over-policing has really happened in the area. And this question can help to answer question 1.

Discussion

To test our hypothesis, we planned to embed the text by using BERT and calculate the text-similarity by using distance matrices. Although the incident reports have some sort of biases, the texts are still profiling the incident in an objective fashion. The complaint report, on the other hand, profiled the subjective description from the complainant. Unfortunately, after querying the data, we found only 17 complaint reports. The data sample is too insufficient to be analyzed.