Checkpoint 5: Natural Language Processing
Team: The Powerful Turtles

Introduction
In this checkpoint, we would like to study how the most common tokenized words for complainants against officers are different among different racial groups.

Relational Analytics Questions
In this checkpoint, we would like to address the following questions:

- *Extract common tokenized words for complaints against officers in each race based on the narrative text data, and use these words to analyze hints on potential police misconducts*

Colab link:
https://colab.research.google.com/drive/1OlMYhKKK9KrWALy0hqx8et-Obk4SfKdM?usp=sharing

NLP Preprocessing Step:
1. Install packages psycopg2-binary and nltk
2. Import dataset “narratives.csv”
3. Data cleaning, we removed the column names that are not in allegations and spaces
4. Integrate narratives information with race information of the complainants from “data_complainant.csv”
5. Tokenize the sentences using nltk
6. Clean the tokenizations and remove stopwords
7. Calculate the frequency of the word distribution
8. Compare the uniqueness of among groups

Results
1. **Overall Top 100 tokens**

Figure. 1 shows the word frequency of the common tokenized words in narratives from the complainants. The common words like “accused”, “officer”, “reporting”, “arrested” are normal to be shown in the results. It is interesting to find some special words. For example, the 4th place word “party” indicates that maybe most situations of misconduct occur during the parties, where people allege together and are easy to have conflicts. The same with the word “vehicles” and “cars”, showing that people and officers have some problems with traffic. “Ipra” stands for Illinois Park and Recreation Association,

which has been mentioned a lot by the complainants, where you can share information and resources, post questions, and use email to communicate directly with people who share your professional interests. It indicates that people want to use online resources as a way to report their problems. Other interesting words, such as “weapon”, “arrests”, “verbally”, “struck”, “threatened” and “handcuffed” showed the types of the officers’ misconduct. The only word related to the race is “white”. From the perspective of the complainants, the “white” refers to the officer’s race, which indicates that white officers are more likely to have misconducts.

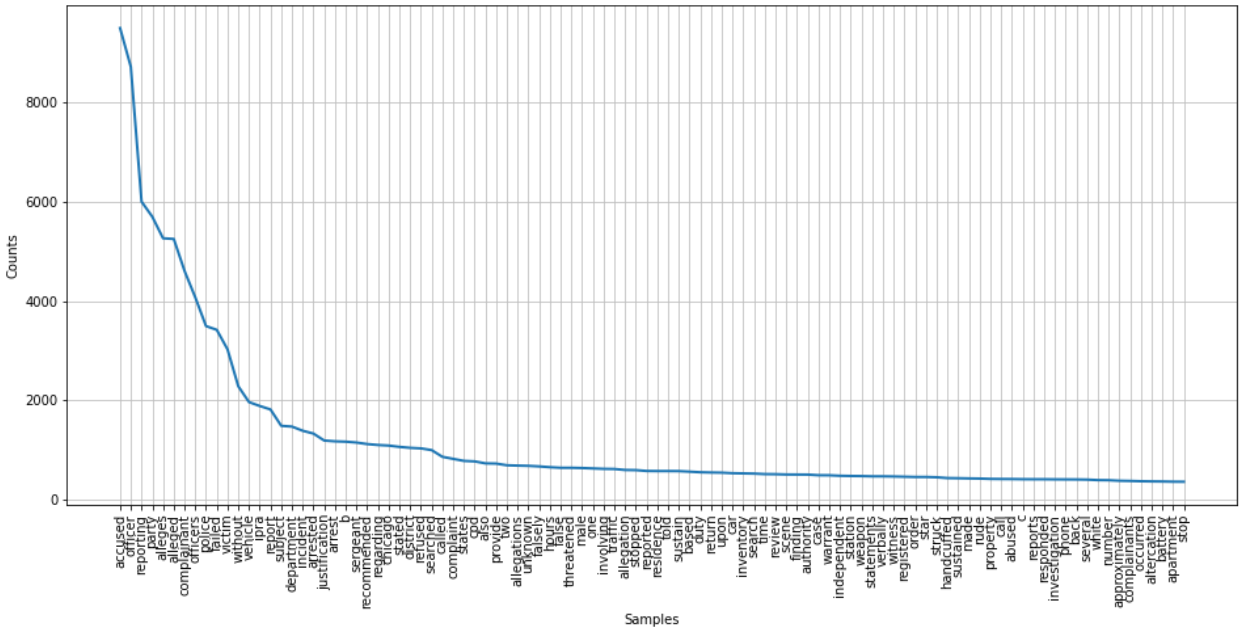

Figure 1. The plot of the word frequency counts from **all** complainants.

2. The difference in Top 100 tokens for each race compared against overall Top 100 tokens

While we attempt to analyze the differences for these lists, it is important to note that we are humans, and while our ability to connect the dots is powerful, it is also very prone to bias. There is no good way to accurately interpret meaning just from the bag of words, so the analysis below may (probably even 100%) be biased, and are only suggestions. Table 1 shows the complaint counts from each race. Table 2 shows the top 100 token words from complaints of each race that are not in the overall top 100 token list, and the top 100 token words from each race are shown in the Appendix.

- Black

What we first notice is that because black complainants filed the most complaints, most of the tokens in their summary are in the overall top 100 tokens, making their list short. But this also makes this list more significant.

From the list, several words such as “door”, “home”, “son” seem to suggest that many of the complaints happen around the complainant’s home when the complaint is black, unlike the other two races. Some other words like “domestic” and “female” could suggest that many allegations were filed while the police officers were solving a domestic case involving the complainant’s family, at their home. The word “offender” could be pointing to the offending police officer, but it could also support the previous possibility as it’s not a common top 100 token. The word “uniformed” is on the list, which is interesting considering most cops on duty wear uniforms. A guess would be that the word is actually “not uniformed” while “not” is filtered out as a stop/conjunction word (and even if it’s not, it’d be a separate token anyways), and the officers were not wearing uniforms and were “unprofessional”, also a word from the list.

Another word that stands out is “black”, considering for the other races, the word for their respective races are not on the list and not in the common top 100. It’s even stranger that “black” is also on the list for Asian/pacific islander complainants. It seems to suggest that a person being black, be it an officer or complainant, seems to be important, more important than any other race.

- White

For white complainants, “firearm” seems to stand out, however, it is also found in the lists for Hispanic and Asian/P.I. However, “discharged” is on the list too. It is unlikely that so many incidents revolve around discharging for other sources, such as hospitals, and so we can assume with relative certainty that it is a firearm that discharged. Similarly, “engaged” should be interpreted not in a wedding sense, but as engaging in a fight of some sort, including gunfights. “Mediation” supports this that someone, perhaps the officer, tried to mediate in the fight.

- Hispanic

From this list, several keywords such as “driver” “license” stand out, and they suggest that Hispanic complainants are often involved in vehicle incidents with the traffic police. A lot of words such as “grabbed” “pushed” suggests that a lot of physical contacts occurred during the incidents and made them file the complaint.

- Asian/Pacific Islander

The first word that stands out is “abdelmajeid”. What is an “abdelmajeid”? We looked it up and apparently it’s the name of a police officer, whose race is Asian/Pacific Islander. We do note that there are only 60 complaints from Asian/Pacific Islander

complainants, but it seems among these complaints, the officer's name is very often mentioned, occurring more often even than common words for other races such as "day". Other than that, the complaint count may be too small to draw any useful conclusion.

- Native American/Alaskan Native

There are only 13 complaints from complainants of these races, but that means that these words carry a large weight from each of the 13 complaints. It seems these cases involve a lot of drug/narcotics selling, and sex.

Table 1. Number of complaints for each complainant race

Complainant Race	**Complain Counts**	**Percentage**
Black	3667	61.27%
White	1541	25.75%
Hispanic	704	11.76%
Asian/Pacific Islander	60	1.00%
Native American/Alaskan Native	13	0.22%
Total	5985	100.00%

Table 2. The list of top 100 frequent tokens for each race that are not in the overall top 100 tokens

Black	White	Hispanic	Asian/Pacific Islander	Native American/ Alaskan Native
door	day	day	properly	selling
son	member	door	named	drugs
uniformed	provided	used	april	card
get	firearm	license	december	conduct
unprofessional	members	times	members	wallet
black	properly	left	location	assaulted
directed	agreed	firearm	abdelmajeid	could
offender	medical	members	incidents	audio
took	engaged	witnesses	cab	finger
supervisor	accept	driver	date	april
home	suspension	may	brought	supposed
female	mediation	personal	day	entered
file	inattentive	grabbed	involved	maybe
domestic	discharged	pushed	discredit	lied
	unit	detective	barz	dick

	city detective information december	directed domestic	uniformed aziz detective body secure arrived punched suspension marriage black firearms face possibly civilian times dates unprofessional medical recorded fist illegally michael watch	allow sword ni make rd date oes thorough af suck sexually ref questioned compl people coming arrived narcotics implied park cell black stuck informed get possibly telephoned video easy id masseur threat removed unprofessional sim telling spoke sold vagina receiving apri asked contained pertaining knew

3. Ranking order for the common words in Top 100 tokens

We were then interested in seeing whether the common words among all racial groups have similarities in their rankings. Table 3 shows the words that are in the top 100 token lists for all races, and they are presented in descending order by frequency.

There are 27 words that are on the top 100 token lists for all races. These words can be grouped into three groups.

The first group is ranked 1-10. In this group, almost all words in the list are also the top 1-10 words in the top 100 list for their corresponding race (the only exception is “failed” for Native Americans which is rank #13 in the overall list). For all groups, the top two most frequent words are “accused” and “officer,” which we can infer from the text that the complainants complain about officers accusing them. This indicates that the officers tend to use words or attitudes that make the complainants feel offended, and the officers should be more careful about the message they convey.

In the second (rank 11-20) and third (rank 21-27) groups, the words show more randomized orders, and they are present on the top 100 token list away from each other. However, from the two lists, we can find interesting features (for this analysis, we exclude Native American/Alaskan Native groups because of the low sample size and many words from group 3 occur only once or twice).

For example, the word “sergeant” is in group 3 for Black, but it is in group 2 for the other races. The words “arrest” and “arrested” are in group 2 for Black, but they are in group 3 for the other races, which indicates that black people are faced more with the arrest problem than other races. The word “time” and “unknown” are in group 2 for Asian/Pacific Islander, but they are in group 3 for the other races, which probably means that Asian people cared more about the details of the description. The word “victim” is in group 2 for Asian/Pacific Islander, but it is in group 1 for the other races. Other races claimed themselves as victims more than Asian people.

It is so interesting to extract all of these common words from the narratives, which will be useful to see the different features of race groups and as a reference to do future research study.

Table 3. Common words from the top 100 tokens from each race in their ranked orders

Group	Black	White	Hispanic	Asian/Pacific Islander	Native American/ Alaskan Native
1 (rank 1-10)	accused officer reporting party alleges alleged officers complainant	officer accused alleged reporting complainant party failed alleges	officer accused alleged complainant reporting party alleges victim	accused officer reporting party complainant alleged alleges failed	accused officer alleges complainant reporting alleged party without

	failed victim	officers victim	failed officers	officers without	justification failed
2 (rank 11-20)	without vehicle report arrest arrested justification stated incident refused regarding	report incident sergeant regarding vehicle district without complaint stated refused	without report vehicle incident sergeant district regarding refused called stated	unknown report incident victim sergeant vehicle called refused time regarding	stated victim officers arrested vehicle refused provide district report sergeant
3 (rank 21-27)	called district sergeant provide complaint unknown time	justification arrest called arrested provide unknown time	justification arrest complaint unknown arrested provide time	district stated arrest arrested provide justification complaint	incident unknown complaint arrest called regarding time

Appendix. Word count frequency plot for each race

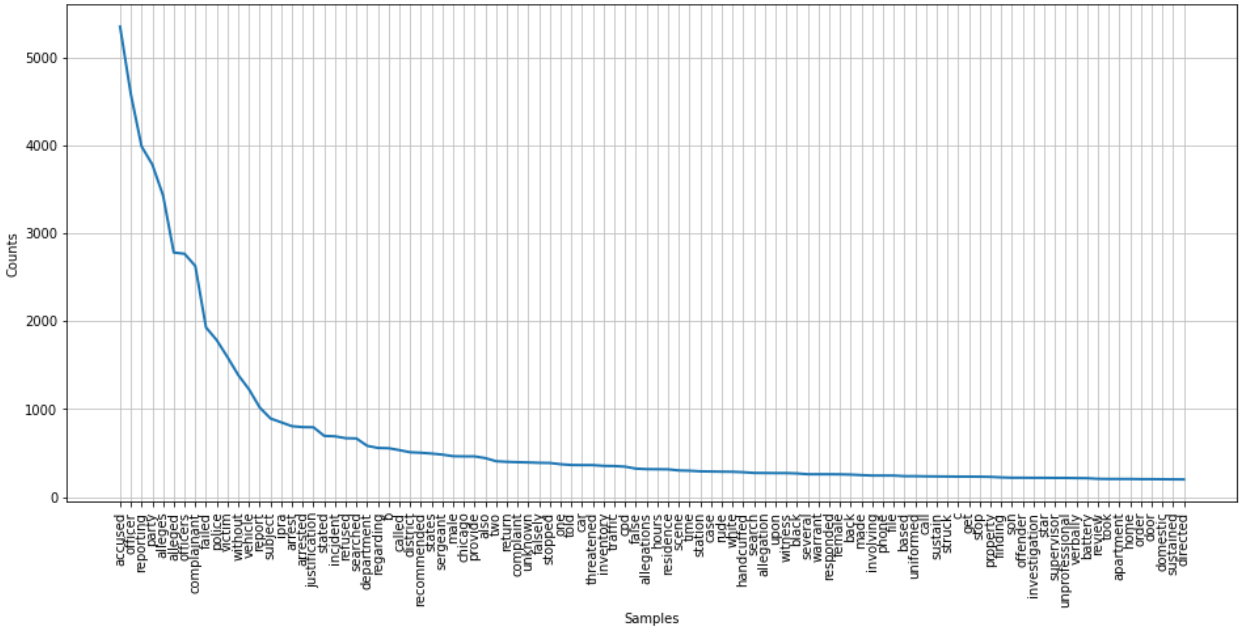

Figure A1. The plot of the word frequency counts from **black** complainants.

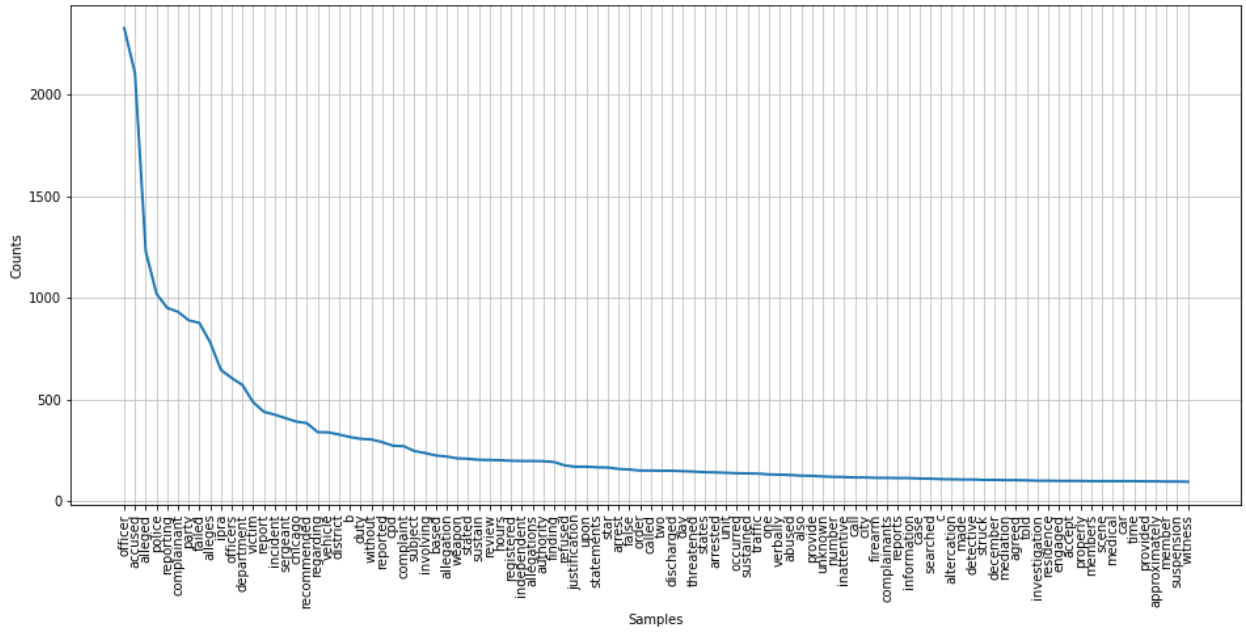

Figure A2. The plot of the word frequency counts from **white** complainants.

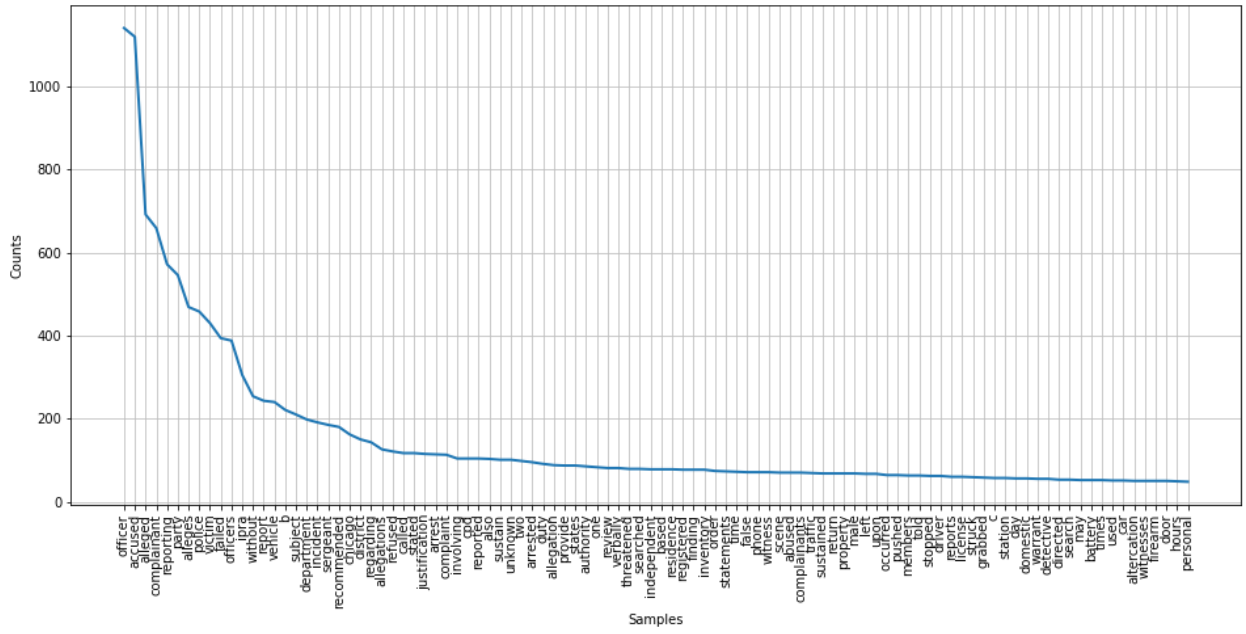

Figure A3. The plot of the word frequency counts from **Hispanic** complainants.

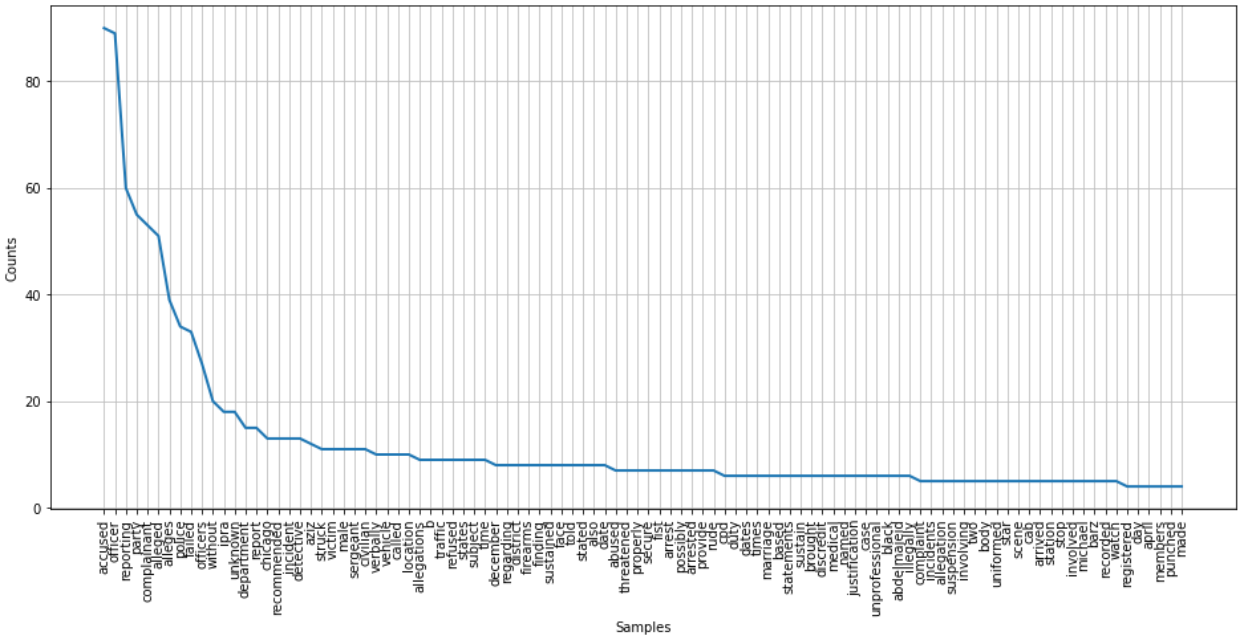

Figure A4. The plot of the word frequency counts from **Asian/Pacific Islander** complainants.

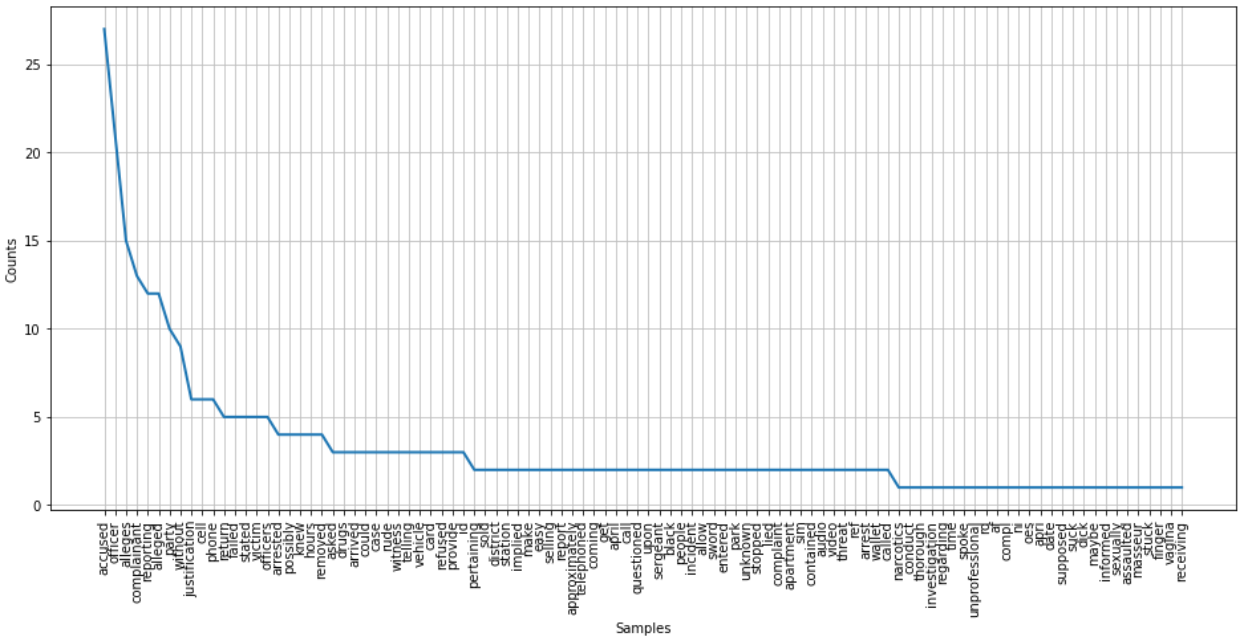

Figure A5. The plot of the word frequency counts from **Native American/Alaskan Native** complainants.