CRIM 515 Project 3

Sydney Reuter

02 May 2024

Research Question

The purpose of this project is to examine the George Mason University Criminology, Law and Society (GMU CLS) faculty research. There are four main areas of study. First, to identify the most common topics by prevalence. Second, to identify which researchers write about what topics. Third, to examine how topics differ between researchers. And lastly, to decide what researcher I am most interested in and why, which will be based on personal interest in the topic.

Literature Review

This study aimed to examine how police officer experiences can offer insights into policing reforms. The researchers conducted 38 interviews with patrol officers in an unnamed police department in the US. Interview transcripts were closely read at least five times and then coded and classified into patterns. This data revealed five main insights. These were to focus more on the interaction between police officer characteristics and the specified reform, to utilize science to aid in creating performance criteria, to focus on police interactions (not just outcomes), to improve officer communication skills with research, and to understand values that guide discretion.

Willis, J., & Mastrofski, S. (2016). Improving policing by integrating craft and science: What can patrol officers teach us about good police work? Policing & Society.

Data

This project utilizes 98 research articles from GMU's CLS department. The articles are contained in a Google Drive folder and stored as a PDF. The folder includes articles from 1995 to 2023. It is not a complete list of research articles from the CLS department. The articles were chosen by Matthew D'anna and represent a wide range of various topics.

Methods

This project was created using R and R Markdown.

Data was loaded from a desktop folder and input into a data frame where each row was one article. Text data was formatted to character data, an ID number was added for each document, and text data was tokenized. The resulting output was a data table that included all words from all documents. Words were stemmed to remove common endings. Numbers and stop words, such as "as" and "the", were removed from the table. Finally, stemmed and unstemmed words were combined into one data table that included the file name, file ID, word, word count, and stemmed word.

Sentiment was attached to words using AFINN, which gave each word a score from -5 to 5. These words were input into a separate data table. Word counts were calculated for words assigned a sentiment (those in the AFINN data table) and all total words. The AFINN data set was filtered to include only the top 1% of words, and the total words data set was filtered to include only the top 0.01% of words.

Latent Dirichlet allocation (LDA) was utilized for topic modeling. Models including 4, 8, 10, 12, 14, 16, and 20 topics were created. Models with less than 10 topics were determined to not have enough detail, and models with 16 or more topics included topics that were meaningless or unsensible. Following this identification, models with 12, 13, 14, and 15 topics were tested. Models with 13-15 topics were ideal. A model with 15 topics was chosen. This model included more meaningful detail than the 13 and 14-topic models. All topics but one were considered meaningful within the 15-topic model. Perplexity and coherence scores were not examined, which is a limitation of this project. A more robust project should compare the perplexity and coherence scores of various models.

The output from LDA gave the top nine words and the proportion of those words appearing in a topic (*beta*). Top terms were then analyzed to produce generic titles for each topic.

The first research question was analyzed in three ways. First, a bar chart displaying word counts with sentiment was created. Second, a bar chart displaying all words was created. Third, the topic modeling output was utilized to create a bar graph faceted by topic, displaying the top 9 words per topic.

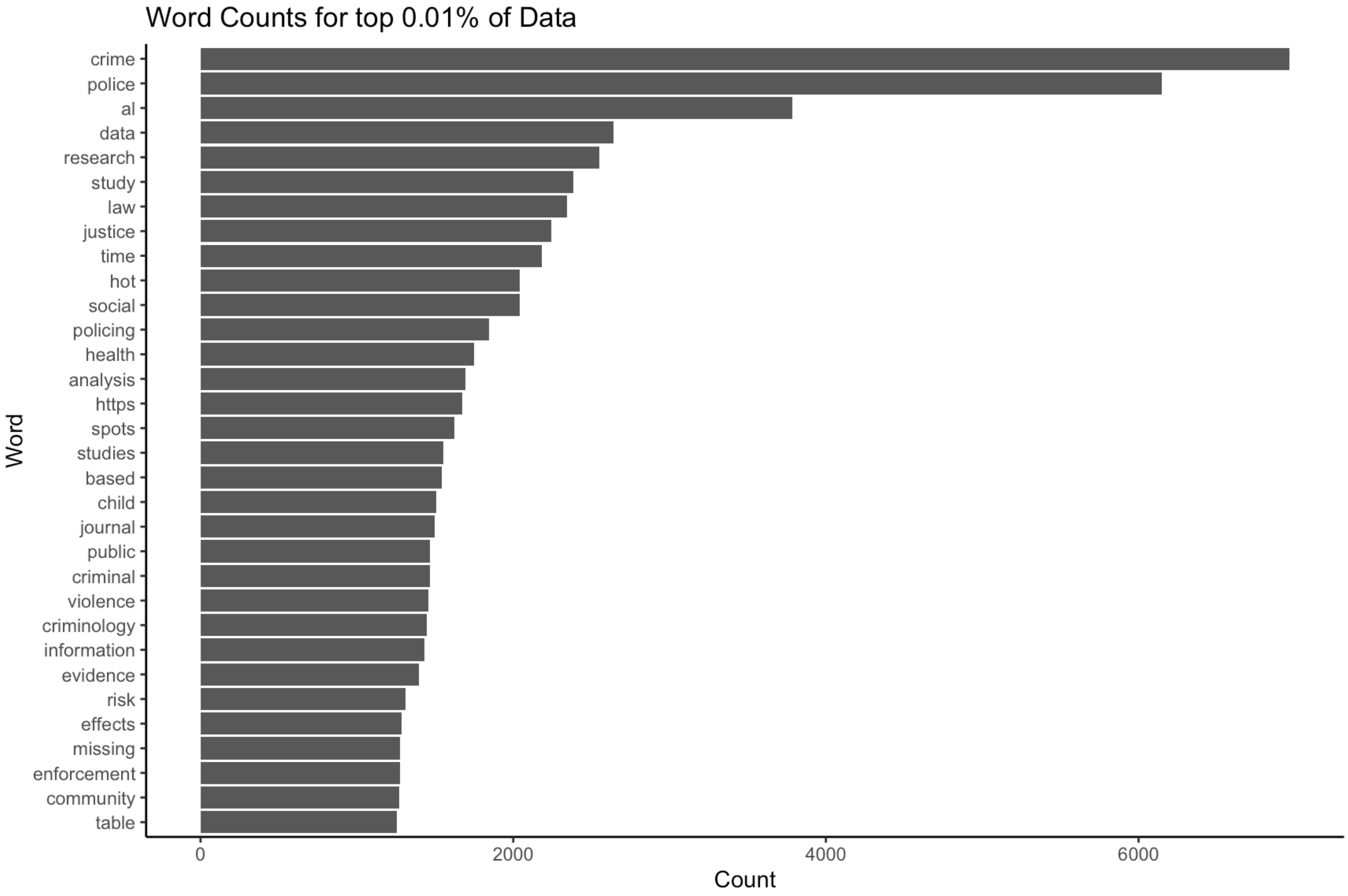
The second research question utilized LDA as well; however, instead of using beta in the output, gamma was used instead. This provided the list of documents and the proportion of topics within the document (*gamma*). A bar chart summarizing the proportion of topics (y) per document (x) was then created.

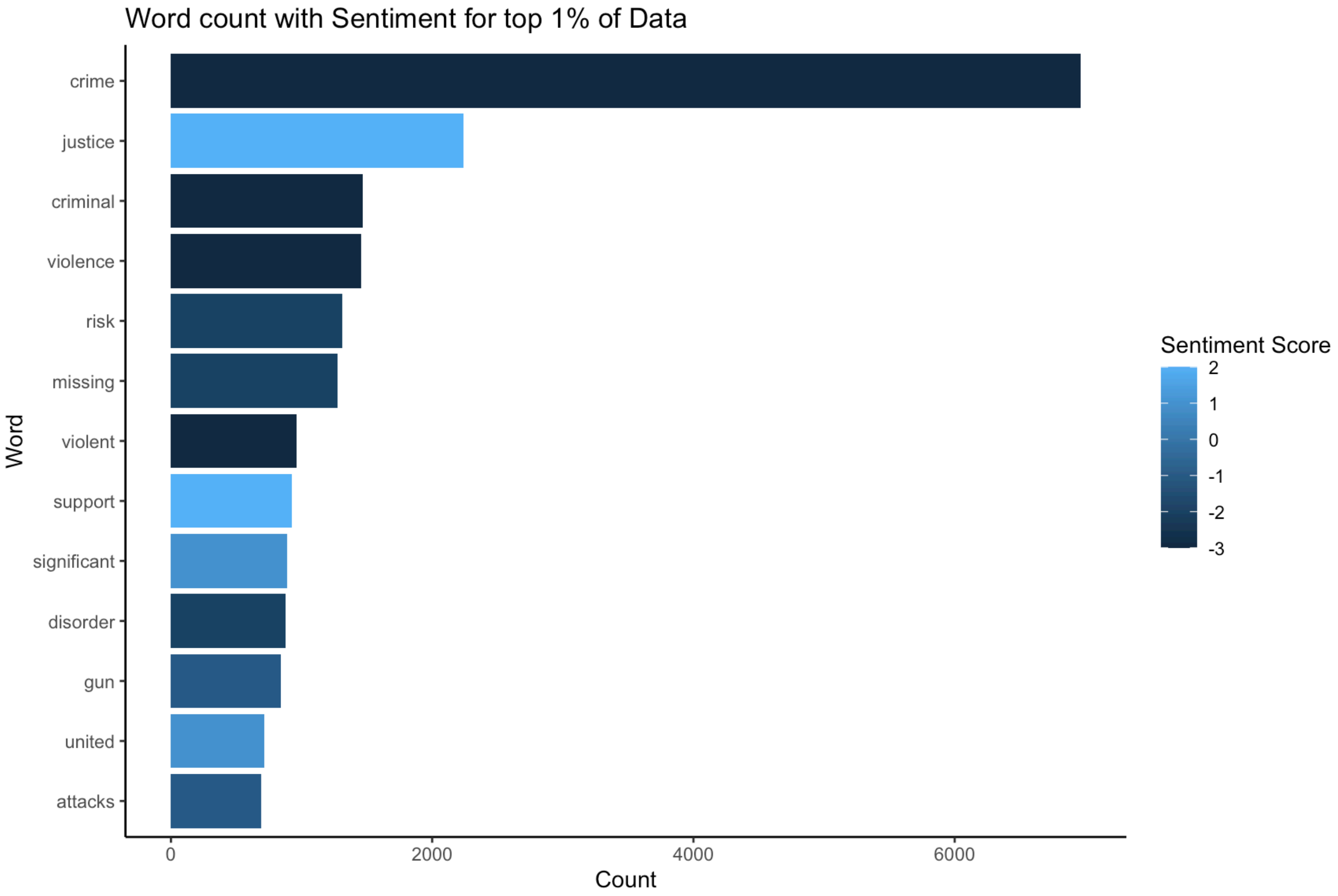
Research question three used two graphs. The first was the same topic modeling graph from research question one. The second graph was a collection of networking charts. Topic titles were examined to identify two topics that were similar to each other. Topics 3 and 11 were used. The gamma data frame was then utilized to identify the top documents for each of these topics (Weisburd 2015 and Wu Koper Lum 2022). Two network charts, one per document, were then created.

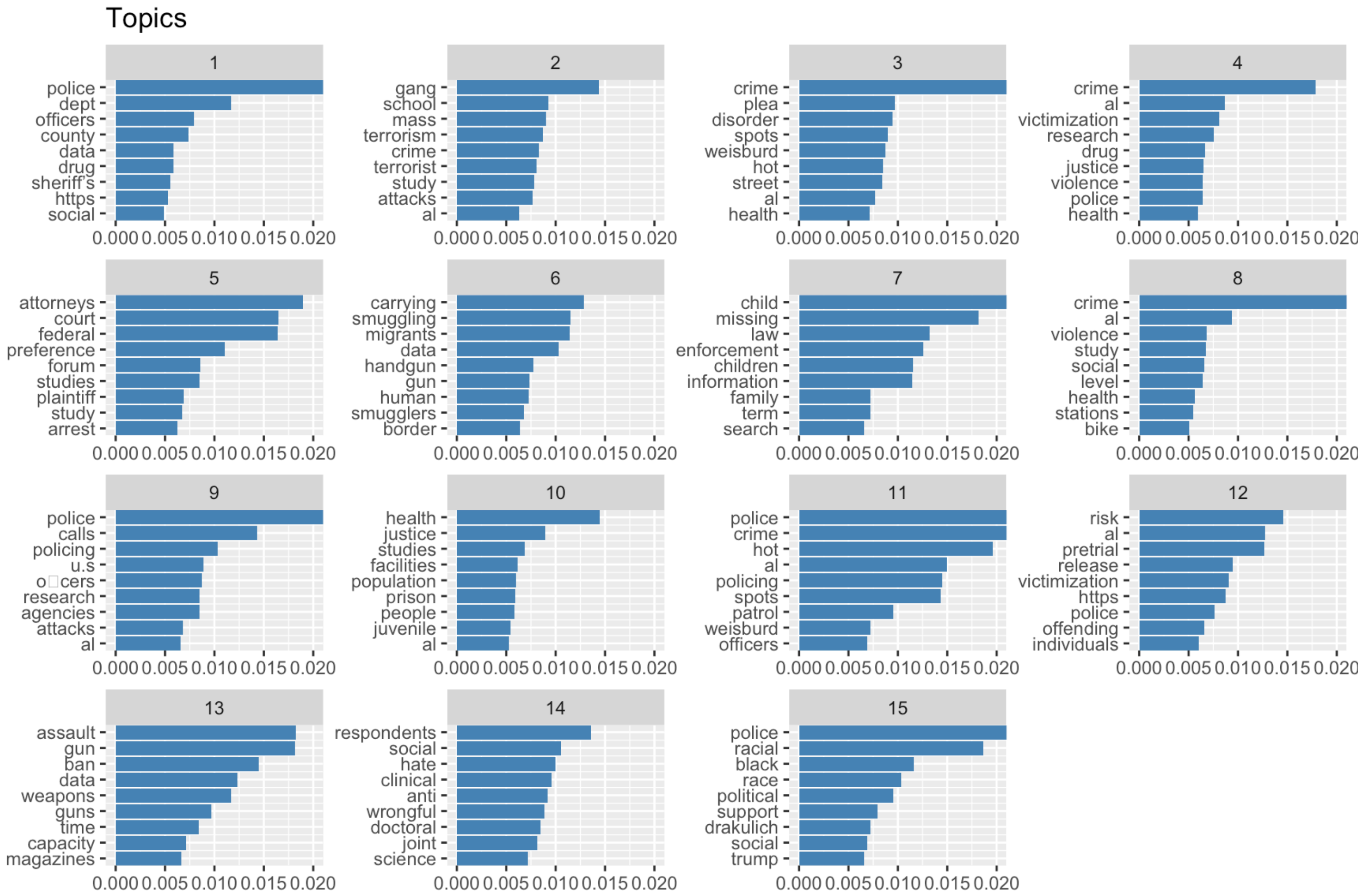
Research question four did not involve any technical methods. First, I identified which topics I found most interesting. No specific criteria were utilized to determine interest, only personal preference was used. After finding these topics, I identified which topics I am least familiar with, and therefore the most interested in. The gamma data table (from research question two) was then utilized to identify the author who wrote about that topic the most often. This author was then identified as the most interesting author.

Findings

What are the common topics and themes?




Word count with Sentiment for top 1% of Data
Word
crime
justice
criminal
violence
risk
missing
violent
support
significant
disorder
gun
united
attacks
0
2000
4000
6000
Count
Sentiment Score
2
1
0
-1
-2
-3

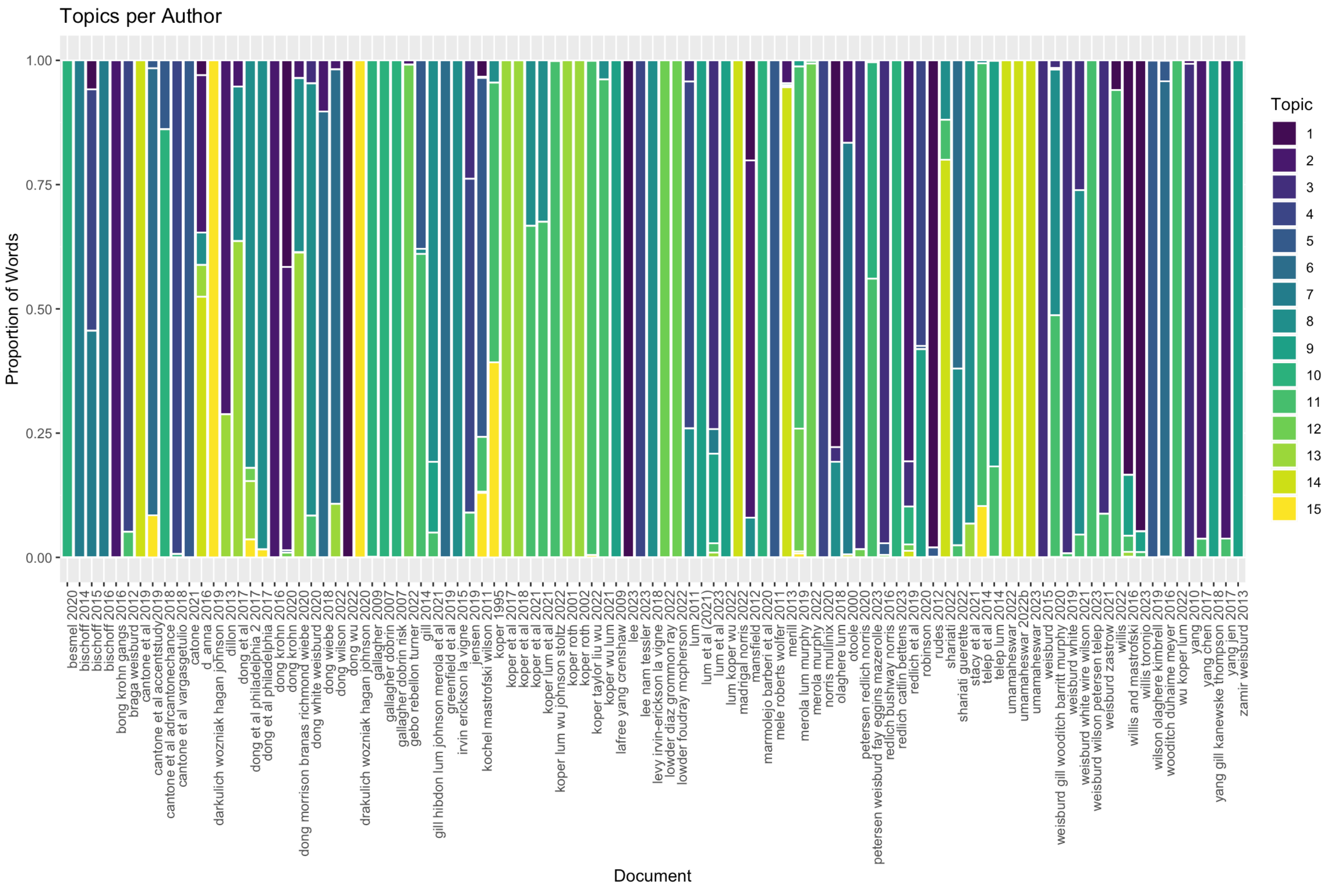

Topics
1
police
dept
officers
county
data
drug
sheriff's
https
social
2
gang
school
mass
terrorism
crime
terrorist
study
attacks
al
3
crime
plea
disorder
spots
weisburd
hot
street
al
health
4
crime
al
victimization
research
drug
justice
violence
police
health
5
attorneys
court
federal
preference
forum
studies
plaintiff
study
arrest
6
carrying
smuggling
migrants
data
handgun
gun
human
smugglers
border
7
child
missing
law
enforcement
children
information
family
term
search
8
crime
al
violence
study
social
level
health
stations
bike
9
police
calls
policing
u.s
o□cers
research
agencies
attacks
al
10
health
justice
studies
facilities
population
prison
people
juvenile
al
11
police
crime
hot
al
policing
spots
patrol
weisburd
officers
12
risk
al
pretrial
release
victimization
https
police
offending
individuals
13
assault
gun
ban
data
weapons
guns
time
capacity
magazines
14
respondents
social
hate
clinical
anti
wrongful
doctoral
joint
science
15
police
racial
black
race
political
support
drakulich
social
trump
0.000
0.005
0.010
0.015
0.020

According to the bar chart above (Word Count for top 0.01% of Data), the most common words are crime, police, data, research, study, law, and justice (“al” was not counted as it is not meaningful). This indicates that common themes are those relating to crime, police, scientific studies (research, data, study), and law and policy (law, justice).

The bar chart for words with sentiment indicates similar themes. The top words with sentiment are crime, justice, criminal, violence, risk, and missing. These indicate similar themes to the other bar chart (crime, law and policy), but include other themes such as violent crimes and missing people. Additionally, five of these words have a sentiment rating of -2 or less. Only justice has a positive sentiment rating. This indicates that common themes in the articles, such as crime and violence, can be perceived as negative.

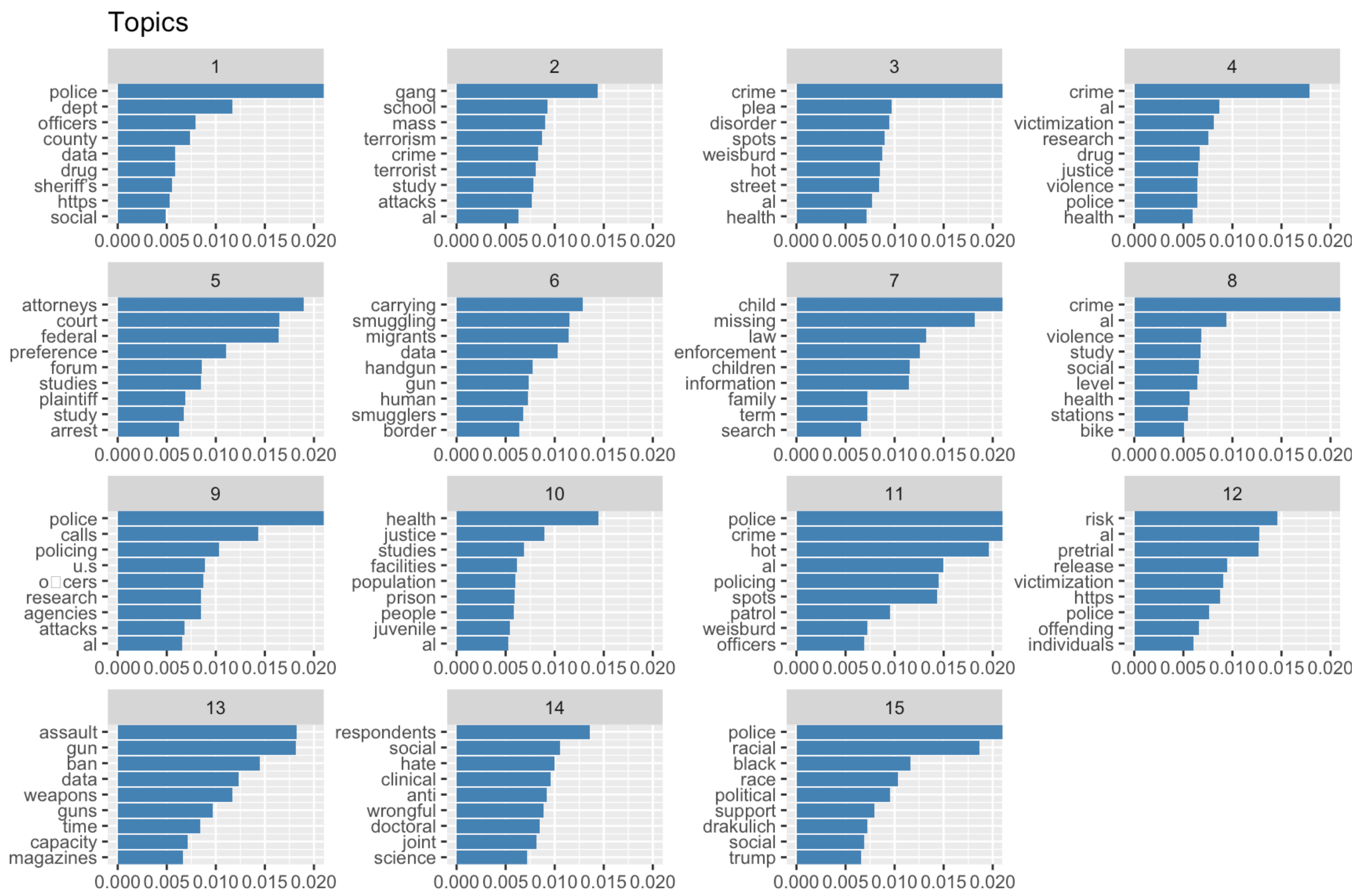
The bar chart for topics includes 15 different topics and the top nine words per topic. Topics were given a generalized title. The topic titles, in order of 1 to 15, are police administration, gang-related crime, crime hot spots, drug victimization, legal, smuggling and border security, missing children, violent crime, police calls for service, incarceration justice, hot spot policing, pretrial, gun assaults and bans, scientific studies, and racial/political policing. “Scientific studies” was deemed an unmeaningful topic, as the top words appeared to relate more to the methods section of articles and less to the actual article topic. The identified topics are similar to the themes described above but are more specific. For example, instead of policing in general, there are multiple topics covering different types and tactics of policing (such as hot spot policing).

Who writes about what?

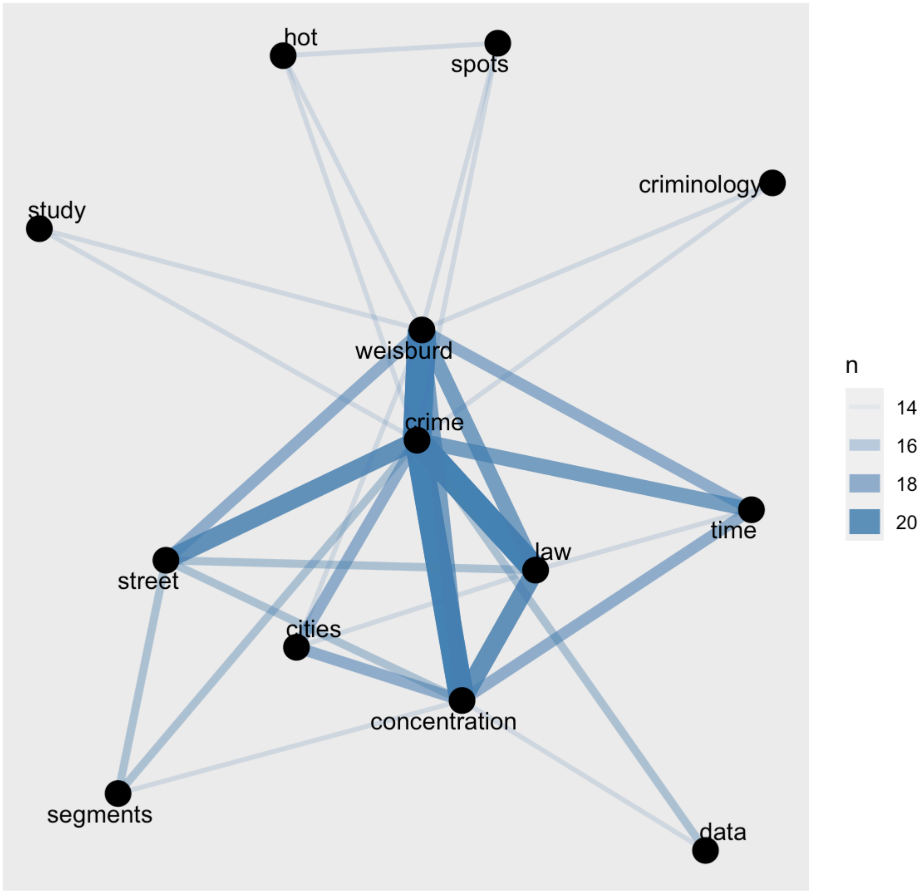


The bar graph above shows the topics present in each document. Most documents include multiple topics. However, there are some documents that only include specific topics. Documents that share topics typically are written by the same author; therefore, assumptions that this author write primarily about that topic, given the data, may be generated. Some examples of this are Gallager (topic 10), Dong (topics 1 and 6), Koper (topics 11 and 13), Lowder (topic 12), and Bischoff (topic 7). From this it is concluded that these authors write about incarceration justice, police administration and border security, hot spot policing and gun assaults, pretrial, and missing children, respectively.

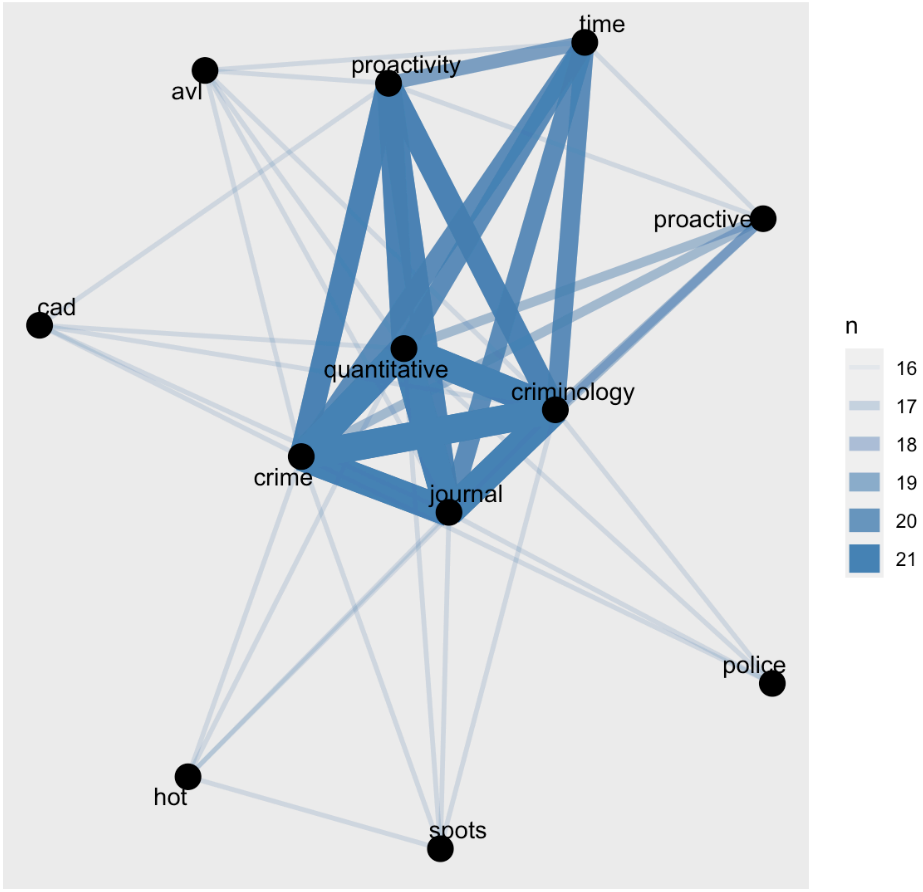
What differs by researcher?



Word Pairs for Weisburd 2015 (Topic 3)



Word Pairs for Wu Koper Lum 2022 (Topic 11)



The bar chart of topics shows how top words differ by topic. Many topics' top word is police or crime (specifically topics 1, 3, 4, 8, 9, 11, and 15). Topics 5, 6, 7, 8, 13, and 14 do not contain the words police or crime in their top words. These topics are also fairly unique, as the words included in them are not frequently included in other topics (the only exceptions to this is (1) topics 6 and 13 which both include the word gun and other variations of this word, and (2) topic 8 which includes the word violence, found in other topics).

Topics that have their top word as police (1, 9, 11, and 15) are relatively similar. These topics tend to focus on police departments, agencies, calls for service, styles of policing (specifically hot spot policing), and political policing. However, topics that have their top word as crime (3, 4, and 8) are not incredibly similar. These topics focus on crime hot spots, victimization, and violence.

Topics that stand out as extremely unique include 2, 5, 6, 7, 13, and 15. These topics' top three words are rarely found in other topics. For example, topic 5 has attorneys, court, and federal as its top three words. These words are not found in other topics.

Topics that are mildly unique include 10, 12, and 14. These topics' top three words are found in other topics but are found at a lower proportion. For example, topic 10 has health, justice, and studies as its top three words. These words are found in other topics, but are not as frequently used (health is the ninth word in topic 3 and 4, and studies/study is the eighth word in topic 5).

Topics that are similar to each other are 1, 3, 4, 8, 9, and 11. As previously discussed, topics 1, 9, and 11 are all relating to policing activity. Topics 3, 4, and 8 all relate heavily to crime, as crime is their top word. Although these topics differ in terms of where they are focused on crime (as can be seen in the less frequently used words in each topic), their over-arching theme is general crime, which is substantially different than other, more unique topics. This warrants them being grouped together.

Topics 3 and 11 were grouped together as each deal with hot spot policing. The documents that had the highest proportion of these topics were Weisburd 2015 for topic 3 (*gamma* = 9.999628e-01) and Wu Koper Lum 2022 for topic 11 (*gamma* = 9.999619e-01). Two network analysis charts are displayed above to examine how these two authors, who write about similar topics, differ.

Weisburd 2015 has strong connections between crime, concentration, and law. These three words are moderately connected to cities, time, and street. Weaker connections include hot-spots, time-law, law-street, data-crime, data-concentration, criminology-crime, and study-crime. "Weisburd" has a frequent word but was not considered in this analysis as it is not meaningful.

Wu Koper Lum 2022 has strong connections between criminology, crime, quantitative, journal, and proactivity. These words are moderately connected to crime and loosely connected to proactive. Weaker connections included hot-spots, police-crime, police-proactivity, cad-police, cad-quantitative, and proactive-time.

Both documents have strong or moderate connections with crime and time, and a weak connection between hot and spots. Both articles also have connections relating to data (quantitative and data). However, Wu Koper Lum 2022 has stronger data-related connections than Weisburd 2015 does. Weisburd 2015 has more connections relating to locations (cities, street). Wu Koper Lum 2022 has more connections relating to policing styles (police, proactive, proactivity, cad).

Who is the most interesting researcher?

Topics I find interesting include 6, 7, and 13. These topics are related to smuggling and border security, missing children, and gun assaults and bans. The topic I know the least about would be smuggling, and therefore I would be most interested in reading about this topic. The documents that have the highest proportion of topic 6 are Greenfield et al. 2019 (*gamma* = 9.999843e-01) and Wooditch Duhaime Meyer 2016 (*gamma* = 9.561176e-01). Since Greenfield has a higher proportion of topic 6, I would consider Greenfield to be the most interesting researcher.

So What

This project has demonstrated that text mining can quickly, efficiently, and effectively summarize and visualize the contents of multiple text documents. This process identifies topics, key words, relationships between words, and other characteristics that may be important (such as author or document title). This process can be applied to other forms of text documents and adjusted to better suit the document type in the same way it was adjusted to include authors per article. Use of text mining can be very useful to federal, state, and local law enforcement agencies.

On a federal level, text mining can be useful in analyzing intelligence. Instead of individually parsing through text documents, analysts should use text mining to identify main trends and themes within the intelligence. On a state and local level, text mining can be used for intelligence analysis as well. However, these departments would also benefit from implementing text mining on a smaller scale.

Departments can use text mining in report writing to identify the most common themes and topics in reports. This can reveal problem areas regarding crime patterns in the community that may not be revealed with traditional summary statistics. For example, larceny reports may include information about the type and style of objects stolen, which would not be recorded in summary data. Text mining may reveal this pattern of objects that are stolen. Traditionally, departments rely on crime analysts and detectives to make these connections; however, utilizing text mining alongside these experts can make this a more effective process. Text mining can identify patterns that these experts may not yet see, and the experts can manually verify that these patterns do exist.

A successful integration of text mining on a state and local scale would be crime analysts and detectives utilizing the output and visuals of text mining to identify trends in crime reports that may otherwise be overlooked. These trends should then be utilized to aid in investigation.