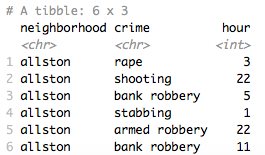
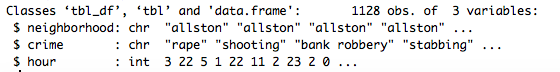
**1. Scraping Boston Crime Data**

1. Create a tibble called crime.data which contains *all* crimes in *all* neighborhoods (i.e., each row represents a crime), and has three columns: crime, hour, and neighborhood



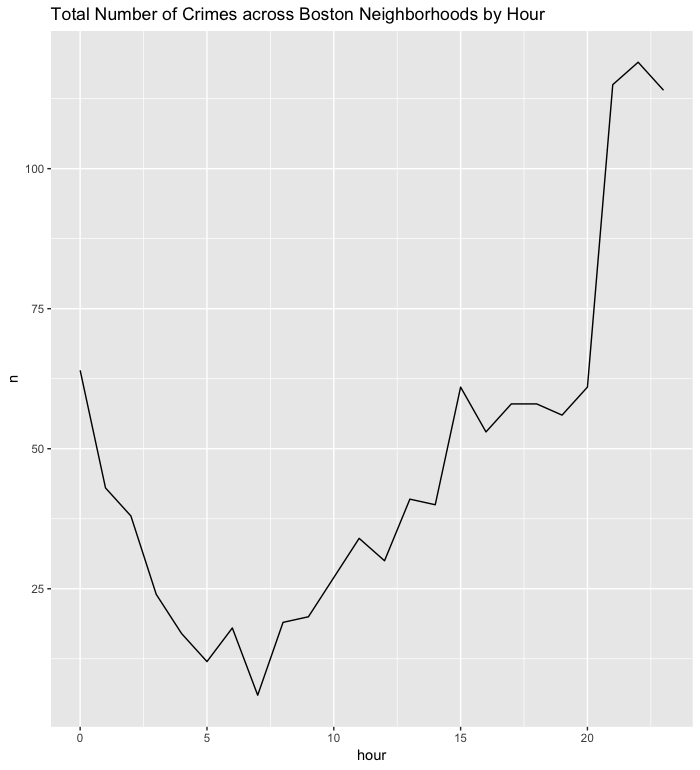


1. What are the five most common crime types (aggregated across neighborhoods and hours), and how many of each such crime occurred?

**The five most common crime types (aggregated across neighborhoods and hours) are:**

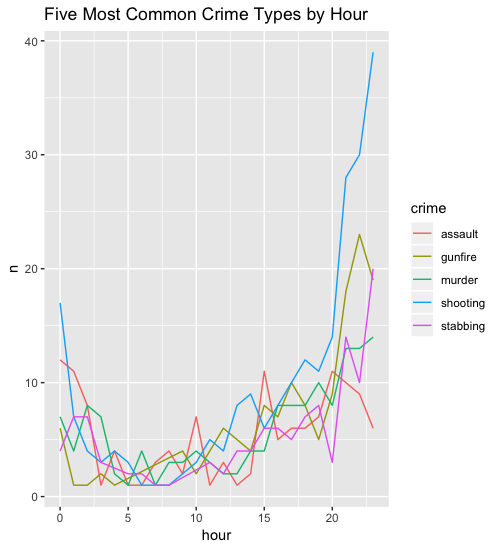
* **Shooting 227 occurrences**
* **Gunfire 143 occurrences**
* **Murder 141 occurrences**
* **Assault 132 occurrences**
* **Stabbing 119 occurrences**

1. Make a plot of the total number of crimes (aggregated across neighborhoods and crime types) by hour. Write a few sentences about the pattern you observe.



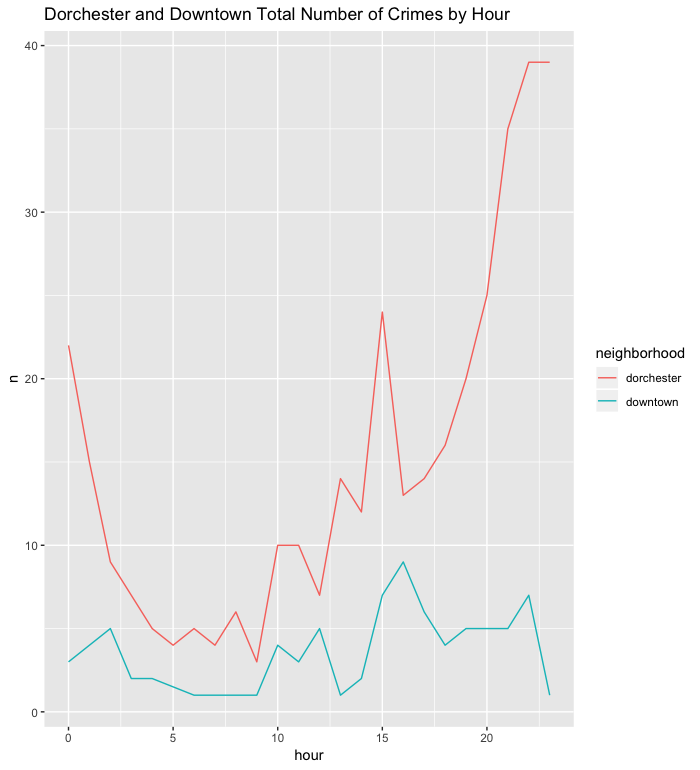
**This line plot (see Figure 1) allows us to observe a variety of patterns. The most obvious is that crimes tend to be committed during later hours, such 9 pm (21), 10 pm (22), 11 pm (23), and 12 am (0). Crimes tend to be less frequently committed during typical work/daytime hours, such as 5 am (5), 6 am (6), 7 am (7) and so on.**

1. Restrict to the five most common crime types, and plot the total number of crimes (aggregated across neighborhoods) for each crime type by hour (i.e., your plot should have five lines). Write a few sentences about the pattern you observe.



**From the line graph (see Figure 2), we see that shootings are by far the most frequent crime type from 8pm (20) until 11pm (23). At 12 am (0) shootings remain the most frequent crime type, however they decrease in frequency in typical working and commuting hours. We can also see that gunfire is correlated with shootings, as the lines that represent these crimes tend to increase and decrease at similar points. Assaults appear to be the most frequent crime during typical working and commuting hours.**

1. Restrict to just the neighborhoods of Dorchester and Downtown, and plot the total number of crimes (aggregated across crime types (include all crime types, not just the top five)) for each of the two neighborhoods by hour (i.e., your plot should have two lines). Write a few sentences about the pattern you observe.



**From the above plot, we can see that no matter the hour, Dorchester tends to have the higher total number of crimes compared to Downtown. For both Dorchester and Downtown, there appears to be a spike in crime at around 3 pm (15), this then sharply declines for Dorchester and slightly declines for Downtown. The total number of crimes then steadily increases for Dorchester from 4 pm (16) to 11 pm (23). Following a slight spike in crime at around 9 or 10 pm (21/22), crime steadily declines Downtown at 11 pm (23).**

**2. Characterizing Appeals Courts**

1. Import the data in the file recent\_opinions.tsv as a tibble called ‘appeals.data’. Add an ‘opinion\_id’ column to appeals.data, which gives each row a unique id.

A close up of a newspaper

Description automatically generated

Load tidytext’s stop\_words tibble (using the command data(stop\_words)), and add the words in custom\_words.txt to it to create a custom dictionary of stop words (you can set the lexicon as “custom” for these words).

A screenshot of a cell phone

Description automatically generated

1. You will now build a simple bag-of-words classifier using the top 100 words (that are not stop words) in the corpus.
2. Unnest the tokens in appeals.data and remove custom stop words. What are the 10 most common words in the entire corpus, and in each of the two circuits (not including stop words)?

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Top 10 common words in the entire corpus** | | | **Top 10 common words in the 5th circuit** | | | **Top 10 common words in the 9th circuit** | | |
| 1 | district | 101273 | 1 | district | 50846 | 1 | district | 50427 |
| 2 | evidence | 52583 | 2 | evidence | 22772 | 2 | evidence | 29811 |
| 3 | review | 38205 | 3 | judgment | 17173 | 3 | review | 25933 |
| 4 | claim | 34757 | 4 | motion | 16548 | 4 | decision | 21505 |
| 5 | law | 33089 | 5 | appeal | 16380 | 5 | claim | 19637 |
| 6 | motion | 32865 | 6 | claim | 15120 | 6 | law | 18561 |
| 7 | judgment | 30653 | 7 | sentence | 14932 | 7 | motion | 16317 |
| 8 | appeal | 29901 | 8 | law | 14528 | 8 | petition | 15838 |
| 9 | decision | 29775 | 9 | claims | 12816 | 9 | federal | 14448 |
| 10 | claims | 27139 | 10 | review | 12272 | 10 | claims | 14323 |

1. Build a document-term tibble, where each row represents an opinion (there should be 16389 rows). There should be 102 columns: the circuit (code "fifth" as 1, and "ninth" as 0), the opinion\_id, and the number of occurrences of the 100 most common words in the corpus (that are not stop words).

A close up of text on a white background

Description automatically generated

Randomly shuffle the rows of this tibble.

A screenshot of a cell phone

Description automatically generated

Split into a 50% training set and 50% test set.

A close up of a device

Description automatically generated

A close up of a logo

Description automatically generated

1. Fit a logistic regression model on the training set that predicts the circuit as a function of all other predictors.

A screenshot of text

Description automatically generated

If you got warning messages, what do they say?

**Warning messages we got are as follows:**

**1: glm.fit: algorithm did not converge**

**2: glm.fit: fitted probabilities numerically 0 or 1 occurred**

Compute the AUC of your model on the test set. Explain why your result is strange and which predictor is causing the strange result.

**The auc score computed is 99.78825. This score seems weird as it is a very high value. This model has a high value of intercept estimate due to the scale differences among the predictors. Among diverse predictors, “opinion\_ID ” is suggested to cause this strange result since it has a different range of values from 1 to 16389. The other predictors, however were measured by the word occurrences. Thus, we decided to remove the opinion\_ID predictor.**

1. Drop the predictor referred to in part c) above, and refit your logistic regression model on the training set.

A screenshot of text

Description automatically generated

What is your new AUC on the test set?

**The auc score newly computed was 95.8521.**

What are the five smallest and five largest coefficients in this model? Give a precise interpretation of the largest model coefficient.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Top 5 smallest coefficients in the model** | | | **Top 5 largest coefficients in the model** | | |
| 1 | disposition | -3.091804867 | 1 | argues | 0.205772469 |
| 2 | memorandum | -1.227435427 | 2 | filed | 0.169885861 |
| 3 | california | -0.888091213 | 3 | `2009` | 0.128133206 |
| 4 | precedent | -0.650205278 | 4 | denied | 0.110231721 |
| 5 | denying | -0.235946669 | 5 | error | 0.109504797 |

The largest coefficient is argues which has a value of 0.2058.One unit increase in the occurrence of the word filed is associated with the increase in the log odds of the opinion being related to the fifth circuit court of appeals by 0.2058 units.

1. Repeat sub-parts a), b), and d) of part B) above, but this time, consider the top 100 bigrams instead of individual words.
2. Unnest the tokens and remove custom stop words for bigram

A close up of text on a white background

Description automatically generated

Consider top 100 most common bigrams.

A screenshot of text

Description automatically generated

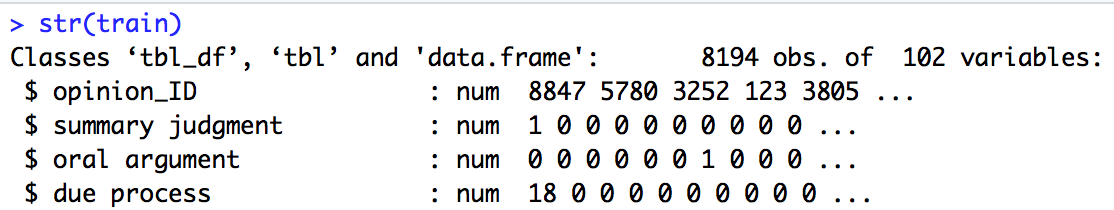
1. Build a document-term tibble for bigram where each row represents an opinion (there should be 16389 rows) and 102 columns.

A screenshot of a social media post

Description automatically generated

Randomly shuffle the rows of this tibble.

A screenshot of a cell phone

Description automatically generated

Split into a 50% training set and 50% test set.



A close up of a keyboard

Description automatically generated

1. Fit a logistic regression model on the training set that predicts the circuit as a function of all other predictors except the dropped predictor

A screenshot of a cell phone

Description automatically generated

Compute the AUC of your model on the test set.

**The auc score computed was 94.36046.**

What are the five smallest and five largest coefficients in this model? Give a precise interpretation of the largest model coefficient.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Top 5 smallest coefficients in the model** | | | **Top 5 largest coefficients in the model** | | |
| 1 | unanimously concludes | -19.60457207 | 1 | official product | 5.087793163 |
| 2 | displayed page | -5.495254327 | 2 | plea conviction | 2.255205897 |
| 3 | panel unanimously | -3.685613562 | 3 | texas law | 1.857622362 |
| 4 | jurisdiction pursuant | -2.690939683 | 4 | california 386 | 1.669904346 |
| 5 | california law | -2.337095680 | 5 | counsel's motion | 1.563643423 |

The largest coefficient is official product which has a value of 5.0878. One unit increase in the occurrence of the bigram “official product” is associated with the increase in the log odds of the opinion being related to the fifth circuit court of appeals by 5.0878 units.

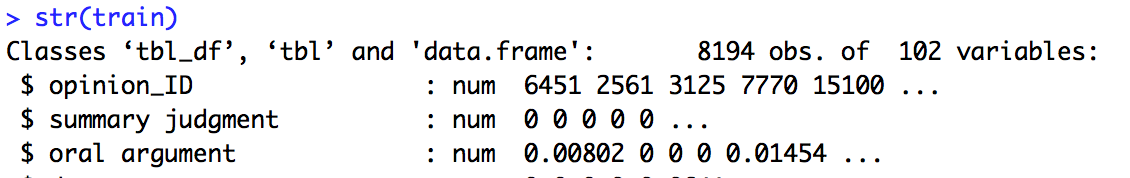
1. Repeat sub-parts b) and d) of part B) above, considering the top 100 bigrams and using the tf-idf value for each of the top 100 bigrams. Compute the tf-idf value for each bigram using the entire corpus of data, not just the training set. ​
2. Build a document-term tibble for bigram using the tf-idf value. Compute the tf-idf value for each bigram using the entire corpus of data, not just the training set.

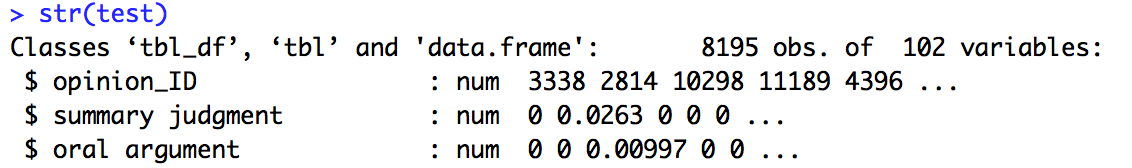
A screenshot of a social media post

Description automatically generated

Randomly shuffle the rows of this tibble.

A screenshot of a cell phone

Description automatically generated

Split into a 50% training set and 50% test set.





1. Fit a logistic regression model on the training set that predicts the circuit as a function of all other predictors except the dropped predictor.

~~A screenshot of a cell phone

Description automatically generated~~

Compute the AUC of your model on the test set.

**The auc score computed was 94.1268.**

What are the five smallest and five largest coefficients in this model? Give a precise interpretation of the largest model coefficient.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Top 5 smallest coefficients in the model** | | | **Top 5 largest coefficients in the model** | | |
| 1 | panel unanimously | -2841.1639818 | 1 | texas law | 213.2303189 |
| 2 | forest service | -1661.1658851 | 2 | displayed page | 151.4988467 |
| 3 | unanimously concludes | -583.3810397 | 3 | plea conviction | 57.2444520 |
| 4 | california law | -563.9656235 | 4 | pro se | 46.2362180 |
| 5 | official product | -253.5393737 | 5 | pleaded guilty | 39.7319830 |

The largest coefficient is official product which has a value of 213.2303. One unit increase in the occurrence of the bigram “texas law” is associated with the increase in the log odds of the opinion being related to the fifth circuit court of appeals by 213.2303 units.

1. Suppose you wanted to apply the model you fit in part D) to a single new opinion. Think through how you would do this (write a few sentences about your thoughts). Does part D) actually make sense as a strategy to build a classifier? If not, what is one way you could still use tf-idf values to build a classifier?
2. Generate all trigrams, making sure you remove trigrams that contain a stopword as either of the three words in the trigram.

A close up of text on a white background

Description automatically generated

Examine, for each circuit, the top 10 trigrams (by frequency in the corpus) that contain the word “supreme.”

|  |  |  |  |
| --- | --- | --- | --- |
| **Top 10 trigrams including “supreme” in the fifth circuit** | | **Top 10 trigrams including “supreme” in the ninth circuit** | |
| Rank | Trigram (occurrences) | Rank | Trigram (occurrences) |
| 1 | retroactively applicable supreme (37) | 1 | banc decision supreme (7) |
| 2 | controlling texas supreme (5) | 2 | unreasonably apply supreme (4) |
| 3 | controlling louisiana supreme (4) | 3 | montana supreme court.see (3) |
| 4 | controlling mississippi supreme (4) | 3 | rhode island supreme (3) |
| 5 | unreasonably applied supreme (4) | 5 | 2001 california supreme (2) |
| 6 | 2006 recent supreme (2) | 5 | 2005 california supreme (2) |
| 7 | fn9 relevant supreme (2) | 5 | 2008 california supreme (2) |
| 8 | recent texas supreme (2) | 5 | appeal pending supreme (2) |
| 9 | supreme beef processors (2) | 5 | dearly established supreme (2) |
| 10 | supreme courtpage 13 (2) | 5 | decision contradicts supreme (2) |

Write a few sentences about what you see (e.g., what are the different contexts in which the 5th vs. 9th circuit opinions are mentioning “supreme”?).

**The differences between the fifth circuit and the ninth circuit seemed heavily based on the differences in geography of supreme courts mentioned. For example, the ninth circuit has top trigrams such as 2001, 2005, and 2008 California supreme while the fifth circuit has top trigrams such as controlling supreme located in Texas, Louisiana, and Mississippi. In addition, two circuits have different trigrams in the role of supreme which appeared in retroactively applicable supreme in the fifth circuit while the ninth circuit has banc decision supreme and unreasonably apply supreme the most.**