Investigating the Case of Jack the Ripper

With Data Mining

Audrey Crockett & T.S. Yeap

Syracuse University

Abstract

In 1888, a killer known only as Jack the Ripper murdered five women and sent wide spread panic throughout London. Jack the Ripper has yet to be identified. Through using data mining techniques (Kmeans Clustering, Decision Tree, and Support Vector Machine), we will compare the famous Jack the Ripper letters with writings and other forms of prose from known suspects and see if a prolific killer is among them.

Keywords: data mining, Jack the Ripper, machine learning

Investigating the Case of Jack the Ripper

With Data Mining

In 1888, an unknown killer caused fear and mayhem in the streets of London after five women were murdered (Ryder, S. P., Johno, & Schachner, T., 2013). The killer was known only to the public as Jack the Ripper. Jack the Ripper is one of the most famous unsolved mysteries of all time. This case has perplexed detectives and scholars alike for the past 130 years. The authorities of the time had unsophisticated techniques for collecting evidence and were never able to narrow in on one suspect (Ryder, S. P., Johno, & Schachner, T., 2013). Very little still exists that might be able to finally catch this age-old killer. Jack the Ripper often taunted the investigators of his (or possibly her) crimes through letters, and these letters still exist to this day. Through using data mining techniques, we will compare the famous Jack the Ripper letters with writings and other forms of prose from known suspects and see if a prolific killer is among them. We will be creating our own dataset using the Jack the Ripper letters by generating the frequency value of each word used in each primary source. We will repeat this procedure with primary source data from each suspect which may be testimony, written letters, or other interviews.

Methods

We plan to establish a comparison between the suspects and Jack the Ripper using data mining. Specifically, we will be using k-means clustering, decision trees, and support vector machine. Through visualizations generated within R, we will be able to see the resulting clusters and classifications.

Data Preprocessing

We began our data collection process by acquiring the texts from the original Jack the Ripper. Then, we converted the original Jack the Ripper letters into individual text files. Next, we

researched prominent suspects in the Jack the Ripper case. We had to rule out some suspects due to lack of accessible writings. The suspect pool for our experiment included six suspects; Joe Barnett, Lewis Carroll, Prince Albert, Carl Feigenbaum, Mary Pearcey, and Walter Richard Sickert (Ryder, S. P., Johno, & Schachner, T., 2013). All the suspects have at one time or another in the 130 years since Jack the Ripper slayings been implicated as the famous murderer. Once we identified the suspect pool, we then acquired writing and quotes by these suspects. Our data set includes writings, testimonies, or quotes from each of the six suspects. All suspect primary source documents were divided into individual text files. R was the employed to aid in the data wrangling of the text files. This required a few packages in R, “tidytext”, “readtext”, and “tidyverse”. Using “readtext”, text files can be read in and formatted. Then using “tidytext” and the “tidyverse”, this allowed for the manipulation of the data into word frequencies (Seigel, J. & Robinson, D., 2017). Once the data frame was set into a useable format, the data was transformed using the min/max transformation.

Data Mining Algorithms

We plan to establish a comparison between the suspects and Jack the Ripper using Cluster Analysis. Specifically, we will be using k-means clustering, classification analysis, and decision trees. We chose k-means clustering because the algorithm partitions objects that are alike into like clusters (Tan, P., Kumar, V. & Steinbach, M., 2018, pg. 496-515).. This can be especially useful when determining patterns in writing or language. We also implore the use of decision trees in attempt to classify the writings and determine if Jack the Ripper is among the suspects (Tan, P., Kumar, V. & Steinbach, M., 2018, pg. 150). Lastly, we used support vector machine for further classification. Through visualizations generated within R, we will be able to see the resulting clusters and classifications.

Results

Exploratory Analysis

Through the newly generated normalized word frequencies, we wanted to perform some initial analyses. Using data wrangling methods from the R package dpylr and aggregation, the data was formatted in a way that we could then create a word cloud (Seigel, J. & Robinson, D., 2017).

1: Word Cloud of Jack the Ripper's Most Frequently Used Words.

From the word cloud, the most frequently used word is “ha”. This makes sense with what we know about Jack the Ripper, who frequently like to taunt police over not being able to catch him (Ryder, S. P., Johno, & Schachner, T., 2013).

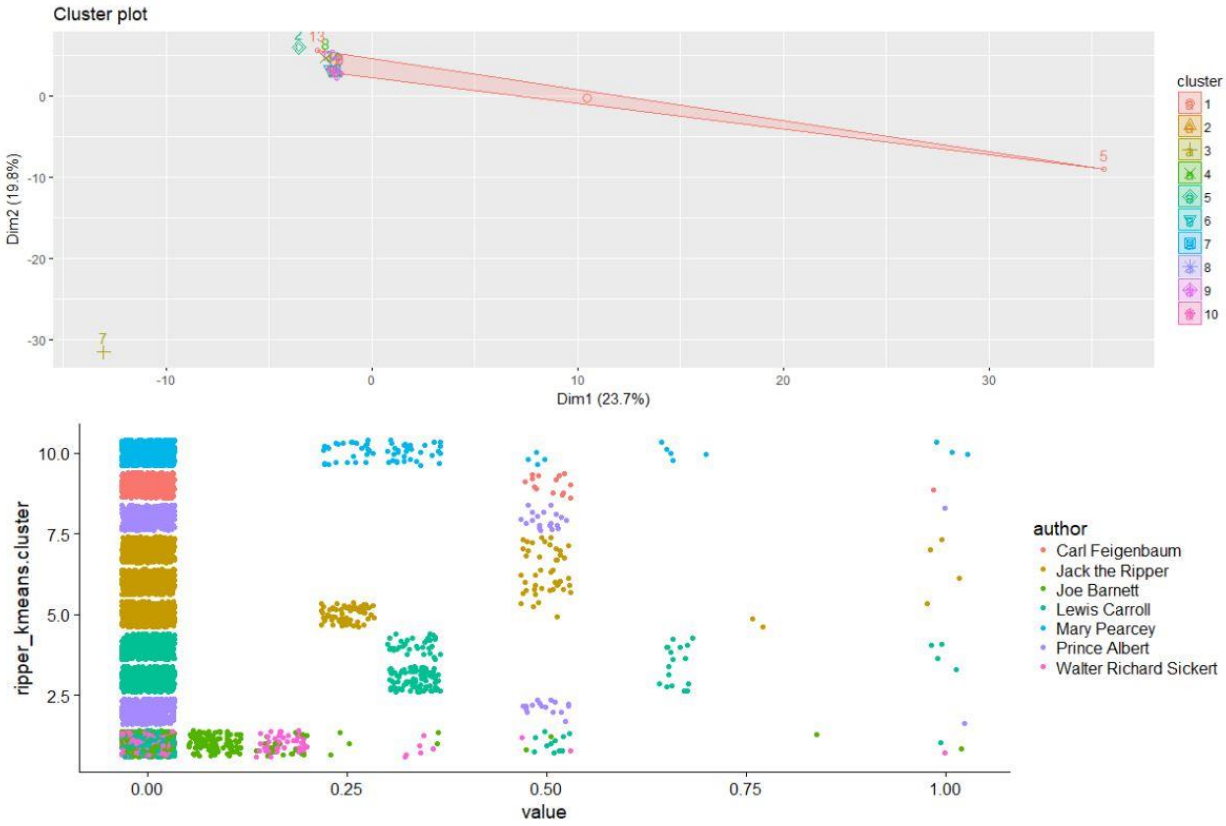
Kmeans Clustering

We chose K means clustering because K means an algorithm often used for text mining due to its ability to manage unstructured data (Tan, P., Kumar, V. & Steinbach, M., 2018, pg. 496-515). K means is also a great algorithm to use when exploring your data (Tan, P., Kumar, V. & Steinbach, M., 2018, pg. 496-515). For our K means model, 10 centriods were modeled. We found that 10 centriods provided the most distinct clusters.

```
K-means clustering with 10 clusters of sizes 3, 1, 1, 1, 1, 1, 1, 1, 1, 2

Cluster means:
      a.m abhorred abjure absorbing  afraid     age     ago  agreed aldershot  alike   alive allowed anoying answer
1 0.0000   0.0000 0.0000       0.0 0.02778 0.05556 0.05556 0.02778       0.0 0.0000 0.02778  0.0000     0.0 0.0000
  apologising appreciated apron arisen    army arranged arrived artistic ate avarice bachelor     bad baffled    bag
1         0.0          0.0  0.00 0.0000 0.02778  0.02778     0.0   0.0000 0.0  0.0000   0.0000 0.08333  0.0000 0.0000
  bags   bear beer belief belong bethnal bishopsgate    bit   black bloodthirsty bloody blotting    body   bold    born
1    0 0.0000 0.00 0.0000 0.0000 0.02778     0.02778 0.1111 0.05556       0.0000    0.0  0.05556 0.02778 0.0000 0.02778
  boss bottle  brave break. brothers   brush buckled   burn called calvinists  can.t canvas canvases capsules cardiff
1 0.00   0.00 0.0000  0.000  0.02778 0.05556    0.00 0.0000 0.0000     0.0000 0.0000 0.3333   0.1111      0.0 0.08333
  careful carnarvonshire carrying catch caught  chance christian christian.s christians clause clever clip   coat
1     0.0        0.02778   0.1667  0.00   0.00 0.05556    0.0000      0.0000     0.0000 0.0000   0.00 0.00 0.1111
  codding collier  colour commercial  common concludes confess continue contrary control conveyed coroner    court
1     0.0 0.02778 0.05556    0.05556 0.02778       0.0  0.0000      0.0   0.0000     0.0   0.0000 0.02778 0.02778
   cousin covered covering cruelty curse custom    cut d.être   dark darkest   darks  davies   davis   dawn     day
1 0.02778 0.05556  0.05556  0.0000  0.00 0.0000 0.0000 0.0000 0.0000  0.0000 0.05556 0.05556 0.02778 0.0000 0.05556
     days    deal dear dearest   death deceased decline  demon desire despised    die    died diet directly disagreeable
1 0.05556 0.05556 0.00  0.0000 0.05556  0.05556  0.0000 0.0000    0.0   0.0000 0.0000 0.02778  0.0   0.0000      0.05556
  disastrous disease dispute doctor  don.t dont double downfall   drank   drink   drunk     dry  drying   duty    ear
1     0.0000     0.0       0   0.00 0.0000 0.00    0.0   0.0000 0.02778 0.02778 0.02778 0.05556 0.05556 0.0000 0.1944
    ears    egg   eggs embarrass employed employers england evening event experience explosion express extreme    eyes
1 0.0000 0.0000 0.0000   0.05556   0.0000    0.0000   0.000 0.02778   0.0    0.05556   0.02778 0.02778  0.0000 0.02778
  falls father.s favour    fear feeling feelings  female fierce fifty finish   fire    fish  fit fits  fix flemming
1   0.0  0.02778 0.0000 0.02778  0.0000   0.0000 0.02778 0.0000 0.000 0.1111 0.0000 0.02778 0.00 0.00 0.00  0.05556
    fond   foot  force forcing fortnight  found  france  frank friday fried friend friendly friendship fulness funny
1 0.02778 0.1667 0.0000  0.0000   0.02778 0.1944 0.05556 0.0000 0.0000   0.0  0.000  0.02778     0.000  0.0000  0.00
  future  gained games     gas  gauger     gay gentleman  george ginger giving glete glue gospel grand   grays   green
1 0.0000 0.05556  0.00 0.02778 0.02778 0.02778   0.02778 0.05556   0.00   0.00   0.0 0.00 0.0000  0.00 0.02778 0.02778
     grey ground  guided  ha   habit  habits half hands   head hear hearing  heart hearted  heavy   held hell   henry
1 0.05556 0.0000 0.05556 0.0 0.05556 0.02778  0.0  0.00 0.0000 0.00  0.0000 0.0000   0.000 0.0000 0.0000  0.0 0.02778
  highway   hold    home honour   hope horrible    hour  hours   house husband.s   idea identify ignorant imagined
1 0.02778 0.3333 0.02778 0.0000 0.0000   0.0000 0.08333 0.0000 0.05556   0.02778 0.0000  0.02778   0.0000   0.0000
  impression improving incarnate inconveniences indefinite individual induces infirmary inflicting ink     inn ironworks
1    0.05556   0.05556    0.0000            0.0    0.05556    0.02778     0.0   0.02778     0.0000 0.0 0.02778   0.02778
  irving jack jacky.s jeanette    job    john joke jolly  joseph    joy keeping  kelly kidne kill knif knife.s laborer
1 0.0000 0.00     0.0  0.02778 0.1667 0.02778 0.00  0.00 0.02778 0.0000 0.05556 0.1111   0.0  0.0  0.0    0.00 0.02778
  labour lady ladys    laid    lane laughed lawfully    lay   leah leather   left letter    life limerick limited line
1 0.0000 0.00  0.00 0.05556 0.02778    0.00  0.02778 0.0000 0.0000    0.00 0.1667 0.0000 0.02778  0.02778     0.0  0.0
   lines linseed     lit    live  lived  living lodging lodgings  london   lose   loss   love loving luck lusk    m.e
1 0.0000 0.05556 0.05556 0.02778 0.2778 0.02778 0.02778  0.02778 0.08333 0.0000 0.0000 0.0000 0.0000  0.0    0 0.0000
   maiden  man.s manifests   marie  market married martyr mason.s materials  meant millers mind minutes mishter mistake
1 0.02778 0.0000       0.0 0.02778 0.05556  0.1111 0.0000 0.02778   0.02778 0.0000 0.02778 0.00   0.000     0.0  0.0000
   model moment monday  months morganstone morning    move  moving murders mutilate  nature needless nice   night nise
1 0.1111 0.0000 0.0000 0.08333     0.05556  0.0000 0.05556 0.05556 0.02778     0.0 0.05556   0.0000 0.00 0.05556  0.0
   note o.clock obedient objected occasions october officers    oil omitted outrage outsider   p.s pain  paint   paper
1 0.0000  0.0000   0.0000  0.02778   0.02778 0.05556      0.00 0.1667  0.0000  0.0000   0.0000 0.000 0.00 0.1667 0.05556
  parents passion   past    pay pennington percep perfectly permitting  person  picked piece plasterer      pm police
1 0.02778       0 0.0000 0.0000    0.02778    0.0       0.0    0.05556 0.02778 0.02778   0.0   0.02778 0.02778    0.0
   porter portia portpool positive post praises prasarved   pray preparations prepared presence presently prompting
1 0.02778 0.0000  0.02778  0.02778 0.00  0.0000       0.0 0.0000      0.05556  0.05556  0.02778    0.0000   0.02778
  proper proportions prostitute   ps public quantity quickly quit raison ratcliffe     raw    read real realise  reason
1 0.1667     0.05556    0.02778 0.00 0.0000      0.0 0.02778 0.00 0.0000   0.02778 0.05556 0.02778 0.00  0.0000 0.02778
   red reduce regularly rejoice religion religious  remain remarks  remove  reply  reside respect returned revolting
1 0.00 0.0000       0.0  0.0000   0.0000    0.0000 0.02778     0.0 0.05556 0.0000 0.02778     0.0  0.02778    0.0000
  ripper ripping    road roche   round roused   rowed  safe saturday saucy   save saved  scene seldom   sell send
1    0.00    0.00 0.02778   0.0 0.05556 0.0000 0.02778 0.000  0.02778   0.0 0.0000  0.00 0.0000 0.0000 0.0000 0.00
  sentence sentiment separated sept  series servant session shakespeare shakespeare.s shant sharp  shock shocked shortly
1    0.0000    0.0000   0.02778 0.00 0.05556  0.0000 0.05556      0.0000       0.0000  0.00  0.00 0.0000  0.0000 0.02778
  shylock sign signed simply singular    sir  sister sister.s    size  sleep slight   sober sor   soul spirit
1       0  0.0    0.0 0.0000      0.0 0.0000 0.02778  0.02778 0.05556 0.0000      0 0.05556 0.0 0.0000 0.0000
  spitalfields   spoke spoken squeal squealed   start stating station    stay stepney stopped   stove straight street
1      0.02778 0.02778 0.0000   0.00     0.0 0.05556 0.05556     0.0 0.02778 0.02778       0 0.05556     0.00 0.1667
  strike strikes strong stronger stuff suddenly suffered suggestion sunday superb supposing supreme surely  sword
1 0.0000  0.0000 0.0000   0.0000  0.00   0.0000      0.0    0.0000 0.0000 0.0000          0  0.0000 0.0000 0.0000
  sympathies sympathy   takes  talk tendency   terms thick thinking thousand threatened throbs thursday  tight     till
1     0.0000   0.0000 0.05556 0.000   0.0000 0.05556  0.00    0.000  0.0000     0.0000  0.000  0.05556 0.1667 0.02778
     time   times tip tiresome   told tomorrow  tones tother touches    town track trade traveller  treat  trial triumph
1 0.05556 0.08333 0.0      0.0 0.3333      0.0 0.1111    0.0  0.0000 0.02778  0.00  0.00   0.02778 0.0000 0.0000  0.0000
  trouble troubles TRUE. trust tuesday   turps   ugly unable understood varnish vendors victim   visit   wales    wash
1     0.0   0.0000 0.000 0.000  0.0000 0.05556 0.1667    0.0     0.0000 0.05556  0.0000 0.0000 0.02778 0.05556 0.05556
  wasnt wate wealth  weary wednesday    west whil  whilst   white whores wine wishing   woman won.t wondered wont   wood
1  0.00  0.0 0.0000 0.0000    0.0000 0.02778  0.0 0.02778 0.05556   0.00  0.0  0.0000 0.05556 0.000   0.1667 0.00 0.1667
    words  worse wriggled  write  wrong
1 0.0000 0.0000   0.1667 0.0000 0.0000
 [ reached getOption("max.print") -- omitted 9 rows ]

Clustering vector:
 [1]  9  5  6  7  1  1  3  4 10 10  8  2  1

Within cluster sum of squares by cluster:
 [1] 7.435 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 3.597
 (between_SS / total_SS =  85.4 %)

Available components:
```

2 R Kmeans Output

The last part of the output shows the goodness of fit of the mode. This calculated by using the between sum of squares and total sum of squares, which was high at 85.4 %.



3 Kmeans Cluster Analysis

From the cluster analysis, Jack the Ripper stands on his own. He has three distinct clusters, which is interesting as portion of our data set is derived from three Jack the Ripper letters. Scholars have posed that the Jack the Ripper Letters were not written by the same person; however, the cluster analysis is clear the same person did likely write all three letters. Further analysis is needed in this area. While we found some interesting exploratory results, the clusters do not implicate or suggest that any of our suspects have writing similarities with Jack the Ripper.

Decision Tree

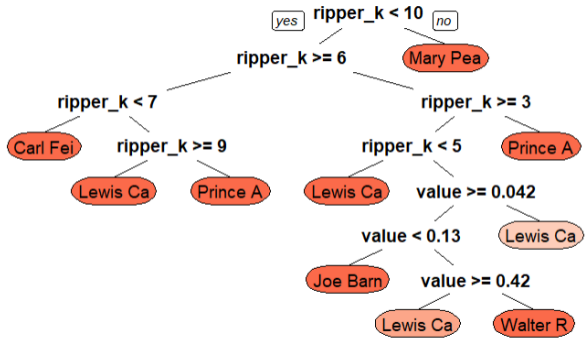
We chose decision tree, because decision trees can determine classifications in a more straightforward manner than other classification algorithms (Tan, P., Kumar, V. & Steinbach, M., 2018, pg. 150). The accuracy of the model was 91.1%.

```
CART

5230 samples
   3 predictor
   6 classes: 'Carl Feigenbaum', 'Joe Barnett', 'Lewis Carroll', 'Mary Pearcey', 'Prince Albert', 'Walter Richard Sickert'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 4706, 4708, 4705, 4707, 4707, 4707, ...
Resampling results across tuning parameters:

  cp       Accuracy  Kappa
  0.00000  0.9109    0.8861
  0.01504  0.8955    0.8693
  0.03008  0.8955    0.8693
  0.04511  0.8955    0.8693
  0.06015  0.8955    0.8693
  0.07519  0.8000    0.7500
  0.09023  0.8000    0.7500
  0.10526  0.8000    0.7500
  0.12030  0.8000    0.7500
  0.13534  0.8000    0.7500
  0.15038  0.5000    0.3151
  0.16541  0.5000    0.3151
  0.18045  0.5000    0.3151
  0.19549  0.5000    0.3151
  0.21053  0.5000    0.3151
  0.22556  0.5000    0.3151
  0.24060  0.5000    0.3151
  0.25564  0.5000    0.3151
  0.27068  0.5000    0.3151
  0.28571  0.3731    0.1152

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.
```

4 R Decision Tree Output

```
prediction             Jack the Ripper
  Carl Feigenbaum                    0
  Joe Barnett                        0
  Lewis Carroll                    523
  Mary Pearcey                       0
  Prince Albert                   1046
  Walter Richard Sickert             0
```

5 Decision Tree Prediction



6 Decision Tree Visualization

The model predicts Prince Albert as Jack the Ripper 66.7% of the time and Lewis Carroll as Jack the Ripper the other 33.3% of the time. The model does not implicate any of the other 4 suspects.

Support Vector Machine (SVM)

We decided to experiment with Support Vector Machine (SVM) because it was widely used in Federalist Paper authorship attribution. There are several advantages of SVM. For example, it can handle both linear and non-linear methods. Besides, it can take all the features as inputs. In other words, one can add as many words as possible to train an SVM model. Unlike other algorithms, one needs to pick the characteristic words before train an SVM model (Diederich, 2003, pg. 113).

In this project, we used Radial Basis Function (RBF). This function separates the features to a higher dimension with chosen a priori. In this dimension, several hyperplanes will be built to classify the features (Diederich, 2003, pg. 113). We first created a train dataset by extracting the suspects into a dataset. Then we trained the RBF SVM model by setting the control parameters as repeatedcv for 3 times in resampling for optimization purposes.

```
Support Vector Machines with Radial Basis Function Kernel

 10 samples
522 predictors
  6 classes: 'Carl Feigenbaum', 'Joe Barnett', 'Lewis Carroll', 'Mary Pearcey', 'Prince Albert', 'Walter Richard
Sickert'

Pre-processing: centered (522), scaled (522)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 9, 9, 9, 9, 9, 9, ...
Resampling results across tuning parameters:

  sigma  C    Accuracy  Kappa
  0.0    0.0  NaN       NaN
  0.0    0.1  0.4         0
  0.0    0.2  0.4         0
  0.0    0.3  0.4         0
  0.0    0.4  0.4         0
  0.0    0.5  0.4         0
  0.0    0.6  0.4         0
  0.0    0.7  0.4         0
  0.0    0.8  0.4         0
  1.0    0.1  0.4         0
  1.0    0.2  0.4         0
  1.0    0.3  0.4         0
  1.0    0.4  0.4         0
  1.0    0.5  0.4         0
  1.0    0.6  0.4         0
  1.0    0.7  0.4         0
  1.0    0.8  0.4         0
  1.0    0.9  0.4         0
  1.0    1.0  0.4         0

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were sigma = 1 and C = 0.1.
[1] Lewis Carroll Lewis Carroll Lewis Carroll
Levels: Carl Feigenbaum Joe Barnett Lewis Carroll Mary Pearcey Prince Albert Walter Richard Sickert
```

7 R SVM Output

Running the parameters above, it is costly. The model takes the longest time to run compared to Kmeans and Decision Tree models. Furthermore, the accuracy is only 40%. It is not as great. The model predicts Lewis Carroll to be Jack the Ripper, which was also mentioned in the Decision Tree model.

Discussion

K means analysis is quick with easy to read results and did not show that any suspects clustered with Jack the Ripper. K means was able to identify three distinct clusters with in the data that did belong to Jack the Ripper. Decision Tree has a longer runtime with higher accuracy and implicates Prince Albert as Jack the Ripper 67% of the time with Writer Lewis Carroll 33% of the time. SVM had the longest runtime with 40% accuracy and predicted Lewis Carroll to be Jack the Ripper 100% of the time. Could Lewis Carroll or Prince Albert be Jack the Ripper? Lewis Carroll was identified as a Jack the Ripper suspect due to his odd life style. In 1996, a book on him as a Jack the Ripper suspect was written and was based on pseudoscientific methods, such as fitting odd anagrams from Carroll's writings (Ryder, S. P., Johno, & Schachner,

T., 2013). At the time, there was not much physical or circumstantial evidence to suggest that Lewis Carroll was in fact Jack the Ripper. Prince Albert Victor has throughout the years been implicated as Jack the Ripper or been associated with Jack the Ripper through some royal conspiracy. It is known that Prince Albert Victor did enjoy the company of prostitutes and all five Ripper victims were prostitutes (Trayer, D., 2016). At the time there was not circumstantial or physical evidence to link him to Jack the Ripper (Ryder, S. P., Johno, & Schachner, T., 2013).

Since all the models did not come to a consensus, it is difficult to say beyond a reasonable doubt who Jack the Ripper really was. This is because we had not much time and resources. It was challenging to find the letters of the suspects. There were 30 suspects at the time. However, we only managed to locate the letters of 6 out of 30 suspects.

Conclusion

Did we finally solve the mystery of Jack the Ripper after so many years? Based on the models we built, Kmeans had 85.41% accuracy but it did not find any suspect clusters who overlapped with Jack the Ripper. Decision Tree had the highest accuracy, 91%. The model predicted Prince Albert as the most likely match to be Jack the Ripper by 67% but 67% still leaves room for reasonable doubt. Lastly, the SVM model had the lowest accuracy (40%) but it predicted Lewis Carroll as Jack the Ripper 100% of the time. Further researcher is needed. The acquisition of more primary source documents would improve the efficacy of the results. Until there is further research, the jury is still out on this 130-year-old mystery.

References

Bennett-Smith, M. (2013, September 25). Jack The Ripper Mystery Solved? Retrieved from https://www.huffingtonpost.com/2013/09/24/jack-the-ripper-solved-investigation-german-sailor_n_3981837.html

Diederich, J. "Authorship Attribution with Support Vector Machines", *Applied Intelligence* 19, pg. 109–123, 2003

Lin, Y. (2018). Week 4 Cluster Analysis [html]. Retrieved from https://drive.google.com/drive/folders/1ZBGiZr8BEL_d4Gubr1R7iFWMY_RgPl5E?usp=sharing

Lin, Y. (2018). Week 5 Decision Tree [html]. Retrieved from https://drive.google.com/drive/folders/1ZBGiZr8BEL_d4Gubr1R7iFWMY_RgPl5E?usp=sharing

Lin, Y. (2018). Week 8 KNN, Ensemble, SVM [html]. Retrieved from https://drive.google.com/drive/folders/1ZBGiZr8BEL_d4Gubr1R7iFWMY_RgPl5E?usp=sharing

Ryder, S. P., Johno, & Schachner, T. (2013). *Casebook: Jack the Ripper.* Retrieved from https://www.casebook.org/

Silge, J., & Robinson, D. (2017). Text mining with R: A tidy approach. Beijing: OReilly. https://www.tidytextmining.com/

Sutherland, J., & Gundry, D. (2004). The Project Gutenberg eBook of *The Life and Letters of Lewis Carroll (REV. C. L. Dogson)*. Retrieved from https://www.gutenberg.org/files/11483/11483-h/11483-h.htm

Tan, P., Steinbach, M., Karpatne, A., & Kumar, V. (2018). Introduction to data mining. New York: Pearson.

The Proceedings of the Old Bailey: MARY ELEANOR PEARCEY. (1890). Retrieved from https://www.oldbaileyonline.org/browse.jsp?div=t18901124-43

Trayner, D. (2016, February 25). "Does this prove Jack the Ripper was member of Royal Family?" *Daily Star*. Retrieved from https://www.dailystar.co.uk/news/latest-news/497089/jack-the-ripper-prince-albert-victor-Duke-Clarence-Avondale-evidence-letters-gonorrhoea

Upstone, R., & Hackney, S. (2012, May 01). A Marengo c.1903–4 by Walter Richard Sickert. Retrieved from https://www.tate.org.uk/art/research-publications/camden-town-group/walter-richard-sickert-a-marengo-r1136451