Text Data Mining for Detective Stories

Yun-Hsiang (Ray) Chan

1. Abstract

Detective story is a subgenre of crime and mystery fiction in which an investigator conducts investigation on a crime, always a murder. These detective stories occurred since mid-nineteenth century and has remained extremely popular over years. In this paper, we collaborate with researchers from Department of English Literature, collecting a dataset with over 400 manual annotations and text of different detective stories. Combining traditional statistical analysis as well as various text data mining methods, we focus on investigating the situation of affective states of detective stories with different components, the evolution of components in detective stories, and the potential text features that might differentiates the stories. In conclusion, although we can't specify the true factors that differentiates the stories, we successfully find the difference of topics over decades and prove that the affective states varies substantially when stories are composed of different components.

2. Introduction

Detective story is a type of crime and mystery fiction remains popular for decades. Over the decades, many researchers are strongly interested in analyzing the detective stories from different aspects. English literature researchers always want to know the most important component that contributes to a good detective stories, the evolution of detective stories over years, and the situation of affective states in detective stories with different components.

Previously, researchers can only collect human-annotated dataset and conduct traditional statistical analysis such as hypothesis testing or regression analysis to do quantitative analysis. Recently, with the advancement of text data mining and natural language processing techniques, researchers start to utilize these novel techniques for finding hidden facts from the text data. In this paper, by combining traditional statistical analysis skills such as hypothesis testing and the latest text data mining techniques such as sentiment analysis, we try to tackle the evolution of detective stories and the situation of affective states in different detective stories. In the meanwhile, we also try to investigate the hidden text features that might differentiates the stories.

For the remainder of the paper:

- Section 3 is about the method, where 3.1 is about EDA method, 3.2 is about sentiment analysis method, 3.3 is about emotion analysis method, and 3.4 is about tf-idf feature extraction and clustering method.
- Section 4 and its subsection are about the conclusion we obtain from each method section.
- Section 5, the last section, is the conclusion and discussion of this research.

For the analysis about evolution of detective stories, although we try to investigate the question through different methods (either in the sentiment or emotion analysis part, or in the clustering part), it's hard to see any trend since the size of the dataset is small, and our main conclusions are from the EDA of annotated dataset.

For analysis about the situation of affective states, we reach our conclusion through sentiment analysis and emotion analysis.

And lastly, for analysis about the hidden text features that might differentiates the stories, we conduct the investigation by doing clustering on tf-idf features, while there's no discrete conclusion that we can reach at the current stage. The question can be furthered investigated by other researchers in future.

3. Methods

3.4 Clustering through tf-idf features

In this subsection, our goal is to assign clusters to detective stories based on their content. The main purpose is to dig out the potential text features that differentiates the stories, where our main expectation is that some stories are differentiating by their publication years.

There are 3 main steps for this method:

1. Calculate tf-idf scores for each word (excluding some rare and common words) as the representation vector for each detective story.

2. Implement K-means clustering algorithm for assigning the clusters.

3. Plot out the clustering result in a low-dimensional space, as a preliminary of next step.

These 3 steps will be dismantled into 6 subsections.

We will introduce the idea, the implementation details and the conclusion from all steps above.

3.4.1 Introduction to the idea of tf-idf

The tf-idf score, short for **term frequency-inverse document frequency**, is a numerical statistic that is intended to reflect how important a word is to a document in a collection of corpus. It's often used as a weighting factor in searches of information retrieval, text mining, and user modeling.

The tf–idf value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general.

The formula of tf-idf score of term t in document d is:

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

$f_{t,d}$ is the number of times that term t occurs in document occurs in document d

Intuition:

- If a term occurs less frequently among all documents while occurring many times in this specific document, it indicates that this term is valuable to this document.
- For example, if the word "doctor" appears 100 times in this story and it only appears in 2 different stories, the tf-idf score is 50. The other word, "knife", appears 10 times in this story and it appears in 10 different stories, the tf-idf score is 10.
- This indicates that "doctor" is a comparatively crucial word than "knife" to this story.

3.4.2 tf-idf scores as representation vector

After introducing the idea of tf-idf, it's time to introduce how can we use them as a representation vector.

Popular method in information retrieval

Since tf-idf scores for each word can represent how important the word to this story, by forming them as a vector, it could be a representation learned from the text of the story.

Implementation Details

To do that, for each document, we need to calculate the tf-idf score for every word in the document, and use these scores to form a vector.

To eliminate the negative influence of some rare words or common words, we set the threshold for adding the word to the representation vector.

- If a word appears in more than 50% of the documents or less than 80% of the documents, then the tf-idf score for this word in every document will be included in the representation vector.
- Otherwise, the tf-idf score for this word won't be included.

The reason to set threshold is to reduce the noise of common word like 'he' or 'she', and the noise of rare terms that might appear only in a specific type of stories (e.g. Holmes). The 50% and 80% are picked based on empirical experience.

We hope that by reducing these noises, the vector can capture the pattern of commonly occurred verb or noun so that we can detect the topic or writing habits of a particular era.

At the end, we have a vector with 425 dimensions for each detective story, with each dimension represent the tf-idf of the term. Each term must appear in more than 50% of documents, or less than 80% of documents.

3.4.3 K-means clustering and Dimensionality reduction by MDS.

As we have the representation vectors, we can assign clusters for each story based on the vectors.

The reason we do this step is to figure out why some stories are assigned to a specific cluster in the next part.

Is it because of the publication year? Or it is due to authors, topics, or something we never notice before.

Implementation Details

The entire process was done in python. Comments are provided in Appendix 1.

We implement K-means clustering to assign the cluster and draw the plots by:

1. Prepare the tf-idf matrix, where each row represents a detective story, and each column represent the tf-idf score for a unique word.
2. Calculate the distance between each row vectors by cosine similarity and form a matrix.
3. Choose the number of clusters, and apply the K-means algorithm on both tf-idf matrix and distance matrix.
4. Reduce the dimension of distance similarity to 2 by Multidimensional scaling (MDS). This is a technique to visualize the level of similarity between vectors.
5. Draw the plots with publication year and author code as a preliminary insight.

Briefly Introduction to the methods

- K-means clustering is a method of vector quantization, that aims to partition n observations into k clusters, wehre each observation belongs to the cluster with the nearest mean.
- MDS is a means of visualizing the level of similarity of individual cases of a dataset. It's used to translate "information about the pairwise distances among a set of n objects" into a configuration of n points mapped into an abstract Cartesian space (2-dimensional space).

Why doing clustering on both tf-idf matrix and similarity distance matrix?

- The similarity distance matrix was originally computed for visualization. However, we find something interesting when doing the clustering.
- The clustering result will be mostly the same when applying k-means clustering on tf-idf and distance matrices. But those tiny differences raise our interest.
- We believe both type of matrix can capture some special patterns that the other one can't capture. This is why we keep the result of clustering on both matrices.

Why clustering and why K-means?

- The reason to do clustering is to differentiate the stories by text feature vectors so that we can find out the reason behind cluster. Clustering is one the most popular unsupervised methods for pattern recognition. This is the reason for choosing it.
- In terms of clustering methods, we also try the hierarchical clustering method. The hierarchical clustering method will produce a similar clustering result for distance matrix, like the one we can obtain by K-means clustering, while it produces an inconsistent result for tf-idf matrix clustering. Therefore, to compare the clusters produced by two matrices at the same time, we believe K-means is more suitable in this case.

4.4 Text Feature Clustering Results

4.4.1 Clustering EDA results

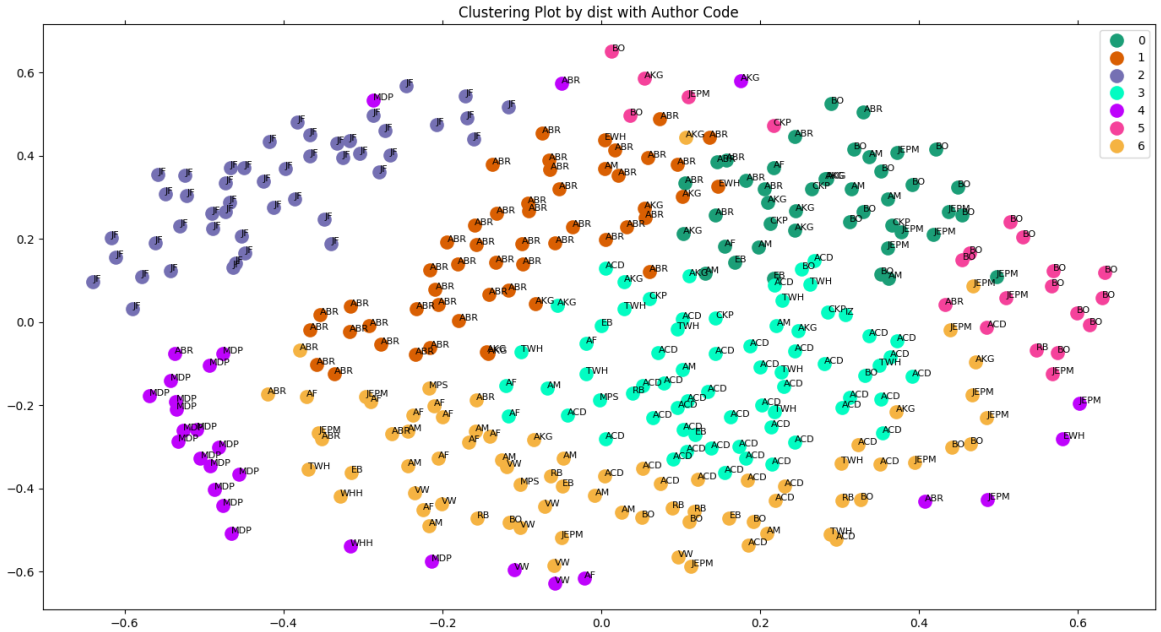
The two plots are clustering result on tf-idf matrix or distance matrix with author code.

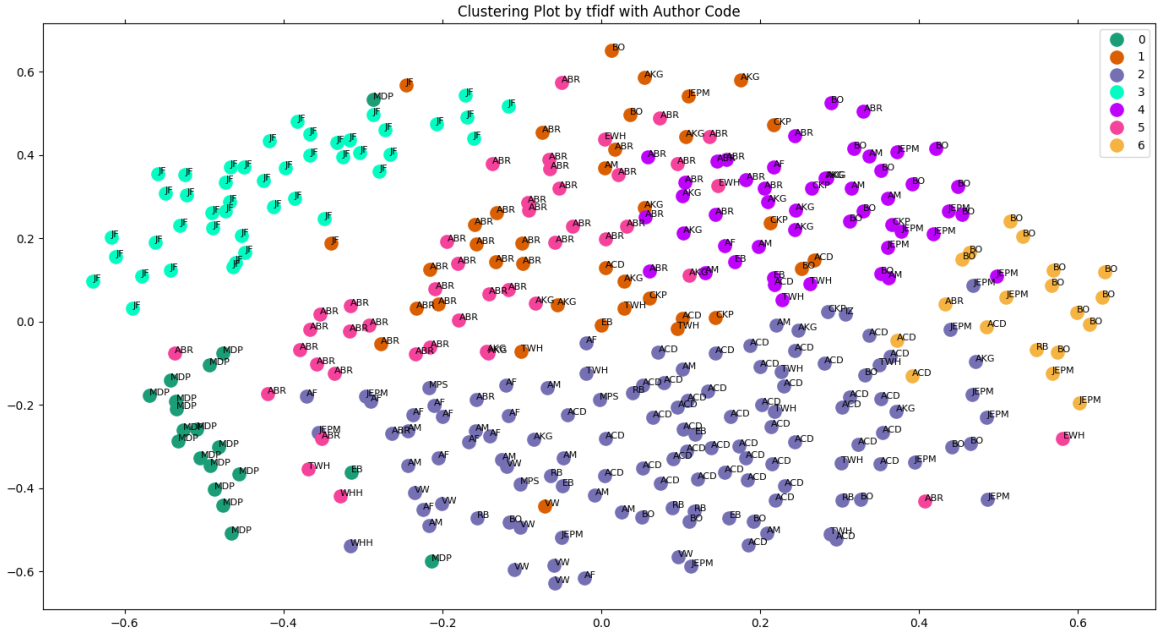
In these plots, each point represents one story, and each color represents the cluster.

Note:

- Empirically, we select $num\_clusters = 7$ for both distance matrix and tf-idf matrix clustering.
- The x and y axis in the plot have no actual meaning. They are only calculated to measure the distance between different stories.
- The position of each story in the plot are the same because they are both calculated on the distance matrix, while the color will be different because of clustering on different matrices (tf-idf or distance).
- The number of cluster will be slightly different by orders. This is because the clustering are done separately.

We first look at the clustering plot by distance matrix and tf-idf matrix with author code information.





From these two plots, we can know that although we really hope that the clustering of stories can rely on different eras, **it's more obvious that stories from the same author are more likely to be assigned to one cluster.**

For example, the stories from JF (Jacques Furtelle) are likely to be assigned to one cluster (cluster 2 in plot 1 and cluster 3 in plot 2). It's reasonable that they are classified in one cluster when the writing habits of the author is unique. Moreover, the topics of these stories are similar (most of them are from the series of detective stories, thinking machine).

Other clusters based on authors are stories from ABR (Arthur B. Reeve) or MDP (Melville Davisson Post) by distance matrix clustering.

Interesting Facts There are still some clusters are assigned with other reasons, and these reasons are not publication years.

Note: the plots with publication year information are provided in the Appendix 2. There are also comments for justifying my claim.

Fact A: Cluster 1 and 5 in tf-idf matrix clustering One interesting fact is that, while stories from ABR (Arthur B. Reeve) are assigned to the same cluster by distance matrix clustering, they are divided into different clusters by tf-idf matrix clustering. And from empirical experiments, if we choose *n_clusters = 6* (now it's 7), then these stories won't be divided into different clusters by tf-idf matrix clustering.

This indicates that there are some crucial text features in tf-idf matrix that could be used to distinguish between cluster 1 and 5 in plot 2, while these features can't be captured by similarity distances.

Fact B: Cluster 3 and 6 in distance matrix clustering The other interesting fact is that, while cluster 2 from tf-idf matrix clustering (plot 2) looks like a combination of all remaining stories (this is even more obvious when *n_clusters = 3* or *4*), the similarity distance clustering method (plot 1) can assign different clusters for them (cluster 3 and cluster 6).

This indicates that there are some crucial features in similarity distance matrix that could be used to distinguish between cluster 3 and 6 (plot 1), while these features can't be captured by tf-idf features.

2.6 Next Steps

Therefore, in the next few sections, we will investigate:

1. What is differentiating cluster 3 and cluster 6 in plot 1 (similarity distance matrix clustering)?
2. What is differentiating cluster 1 and cluster 5 in plot 2 (tf-idf matrix clustering)?

Implementation

We plan to conduct several hypothesis testings to see whether these clusters are assigned due to similar topics.

If there's no explicit reason from the annotated dataset, then the authenticate logic behind the classification should be left for the experts of detective stories, our collaborators.

Other choices and limitations.

- It's also viable to look at the most influential column values (i.e. the tf-idf scores for some words) contributes to each cluster after tf-idf matrix clustering, and try to find out the reason from the word.
- However, these influential tf-idf scores are often noisy. There might exist some patterns when scanning them altogether, but these patterns are hard to be observed individually.
- Therefore, as we already have annotation of each story, we instead take the annotations for investigation.
- You can also find the most influential words (top 10) to each cluster for tf-idf matrix clustering in Appendix 3.
- Similarity distance clustering do not have these words, since the dimension is changed when calculating the matrix from tf-idf matrix and therefore the dimensions don't have actual meanings.

III. Statistical Anlaysis

In this section, we will mainly focus on hypothesis testing. We want to testify whether the implicit logic behind the grouping. For two the groups of clusters, are they assigned because they have similar topics or written in different eras?

3.1 Fisher's Exact Test of Independence

For examining the topics of two groups (cluster 3 / 6 from similarity distance matrix clustering, and cluster 1 / 5 from tf-idf matrix clustering), we can will specifically focus on the crime type and motive in the stories. These two components are the most representative annotations in our dataset.

Let's first look at the two tables below.

Crime.Types	Cluster.3	Cluster.6	Cluster.1	Cluster.5
Murder	43	36	24	33
Suspected.murder	10	7	1	4
Theft	30	42	19	14
Fraud	33	32	17	19
Blackmail	19	5	5	4
Bribery	0	3	0	0
Assault	29	21	14	17
Forgery	3	15	5	8
Kidnapping	13	4	4	5
Mischief	11	8	7	8
Breaking.and.entering	8	13	8	6
Trafficking	0	1	0	2
Illegal.gambling	0	1	1	0
NA.	1	2	1	1

Motive.Types	Cluster.3	Cluster.6	Cluster.1	Cluster.5
Greed	47	52	23	29
Revenge	23	14	8	7
Jealousy	11	3	7	9
Love	8	3	6	4
Ideology	2	12	4	7
Pride	17	10	4	6
Duress	0	0	0	0
Crime.for.crime.s.sake	3	2	3	0
Paranoia	3	2	2	0
Self.defence	7	3	1	0
Insanity	0	1	2	3
NA.	2	3	0	2

```
cluster_crime <- left_join(dist_cluster_data, crime_data, by = 'story_code') %>%
  filter(label == 3 | label == 6)
cluster_crime %>% arrange(publication_year) %>% select(story_name, label, publication_year, author_code
```

```
##                                                 story_name label
## 1                                   A Case of Identity     3
## 2                                 The Five Orange Pips     6
## 3                                The Red-Headed League     6
```

```
## 4                                          The Boscombe Valley Mystery     3
## 5                                        The Man with the Twisted Lip     3
## 6                                     The Adventure of the Speckled Band     3
## 7                                     The Adventure of the Beryl Coronet     3
## 8                                    The Adventure of the Blue Carbuncle     6
## 9                                    The Adventure of the Copper Beeches     3
## 10                                   The Adventure of the Noble Bachelor     3
## 11                                                       Silver Blaze     3
## 12                                                 The Musgrave Ritual     3
## 13                                                The Resident Patient     6
## 14                                               The Greek Interpreter     3
## 15                                                  The Reigate Puzzle     6
## 16                                                   The Final Problem     3
## 17                                                    The Naval Treaty     3
## 18                                                The "Gloria Scott"     3
## 19                                                   The Crooked Man     3
## 20                                                   The Yellow Face     3
## 21                                            The Stock-Broker's Clerk     6
## 22                                                   The Cardboard Box     3
## 23                                                The Redhill Sisterhood     3
## 24                                              The Affair of the Tortoise     6
## 25                                              The Loss of Sammy Crockett     6
## 26                                           The Case of the Dixon Torpedo     6
## 27                                                 The Case of Mr. Foggatt     3
## 28                                           The Case of the Missing Hand     6
## 29                                               The Nicobar Billion Case     6
## 30                                               The Ivy Cottage Mystery     3
## 31                                            The Case of Laker, Absconded     3
## 32                                          The Case of the Lost Foreigner     6
## 33                                 The Case of the "Flitterbat Lancers"     6
## 34                                                   The Forged Cheque     6
## 35                                                  The Big Loan Fraud     6
## 36                                                   Tracing a Traitor     6
## 37                                              The Clue of the Silver Jug     6
## 38                                                   An Unsolved Problem     6
## 39  The Gold-Seeker's Strange Fate: An Astounding Romance of Real Life     6
## 40                                                  With a Passing Glory     6
## 41                                       The Mystery of the Gravel Pits Farm     6
## 42                                                        The New Tenant     6
## 43                                                 The Norwood Builder     3
## 44                                                     The Dancing Men     3
## 45                                     The Adventure of the Empty House     3
## 46                                                 The Solitary Cyclist     3
## 47                                              The Lisson Grove Mystery     3
## 48                                                   The Three Students     6
## 49                                                 The Golden Pince-Nez     3
## 50                                                    The Six Napoleons     3
## 51                                                    The Priory School     3
## 52                                                          Black Peter     3
## 53                                              Charles Augustus Milverton     3
## 54                               The Adventure of the Missing Three-Quarter     3
## 55                                        The Adventure of the Second Stain     3
## 56                                        The Adventure of the Abbey Grange     3
## 57                                             The Liberation of Wyoming Ed     6
```

```
## 58                                        Peter Crane's Cigars     6
## 59                                      Missing: Page Thirteen     3
## 60                                        The Aluminium Dagger     3
## 61                                     Midnight in Beauchamp Row     6
## 62                                       A Case of Premeditation     3
## 63                                    A Message from the Deep Sea     6
## 64                                          Cheating the Gallows     3
## 65                                  The Mystery of the Steel Room     6
## 66                                      The Stanway Cameo Mystery     6
## 67                                         Percival Bland's Proxy     3
## 68                                     The Absent-Minded Coterie     6
## 69                    The Mystery of the Five Hundred Diamonds     6
## 70                                         The Riddle of the 5.28     3
## 71                          The Doctor, His Wife, and the Clock     3
## 72                                The Clue of the Silver Spoons     6
## 73                                       The Grotto Spectre     6
## 74                                       The Moabite Cipher     6
## 75                                      The Stolen Necklace     6
## 76                                    The Caliph's Daughter     3
## 77                                   The Quinton Jewel Affair     6
## 78                                           Room Number 3     3
## 79                             Lord Chizelrigg's Missing Fortune     6
## 80                                  The Riddle Of The Sacred Son     3
## 81                                      The Missing Mortgagee     6
## 82                                        The Ayrsham Mystery     6
## 83                                        The Haunted Jarvee     6
## 84                                The Riddle Of The Rainbow Pearl     3
## 85                                       The Race of Orven     3
## 86                                  The Problem Of The Red Crawl     3
## 87                                     The Case of Oscar Brodski     6
## 88                                              The Old Lag     6
## 89                                                  The Hog     6
## 90                                       The Wizard's Belt     3
## 91                                  The Murder at Troyte's Hill     3
## 92                                       A Memorable Night     3
## 93                                         The Blue Sequin     6
## 94                                The Riddle Of The Ninth Finger     3
## 95                                                The Thief     6
## 96                                                 The S.S.     6
## 97                                  The Riddle Of The Siva Stones     3
## 98                                  The Man with the Nailed Shoes     6
## 99                                         How the Bank Was Saved     6
## 100                                       The Lion's Smile     6
## 101                                 Sir Gilbert Murrell's Picture     6
## 102                               The Mystery of the Boat Express     6
## 103                The Tragedy on the London and Mid-northern     6
## 104                                       The Tremarn Case     6
## 105                       The Affair of the Corridor Express     6
## 106                                       The Divided House     3
## 107                                     The Echo of a Mutiny     6
## 108                                     The Mandarin's Pearl     6
## 109                           The Staircase at the Heart's Delight     6
## 110                              The Black Bag Left on a Door-Step     3
## 111                                   The Ghost with the Club-Foot     3
```

```
## 112                                The Stone of the Edmundsbury Monks     6
## 113                                             The Dublin Mystery     6
## 114                                       The Regent's Park Murder     6
## 115                                     The Bruce-Partington Plans     3
## 116                                The Robbery in Phillimore Terrace     6
## 117                                           The Liverpool Mystery     6
## 118                                The Fenchurch Street Mystery     3
## 119                                   The De Genneville Peerage     6
## 120                                             Wisteria Lodge     3
## 121                                          The Devil's Foot     3
## 122                                       The Terror in the Air     6
## 123                                             The Steel Door     6
## 124                                             The Red Circle     3
## 125                           (A Case Of) Spontaneous Combustion     6
## 126                                             The Black Hand     6
## 127                                 The Master Counterfeiter     6
## 128                                    The Coin of Dionysius     6
## 129                          The Last Exploit of Harry the Actor     6
## 130                                        The Dying Detective     6
## 131                                 The Tilling Shaw Mystery     3
## 132                          The Knight’s Cross Signal Problem     6
## 133                                The Game Played in the Dark     3
## 134                                             His Last Bow     6
## 135                                       The Three Gables     3
## 136                                      The Mazarin Stone     6
## 137                                       The Creeping Man     3
## 138                                  The Illustrious Client     3
## 139                                     The Three Garridebs     3
## 140                                   The Retired Colourman     3
## 141                                    The Blanched Soldier     3
## 142                                         The Lion's Mane     3
## 143                                   Shoscombe Old Place     6
##     publication_year author_code
## 1               1891         ACD
## 2               1891         ACD
## 3               1891         ACD
## 4               1891         ACD
## 5               1891         ACD
## 6               1892         ACD
## 7               1892         ACD
## 8               1892         ACD
## 9               1892         ACD
## 10              1892         ACD
## 11              1892         ACD
## 12              1893         ACD
## 13              1893         ACD
## 14              1893         ACD
## 15              1893         ACD
## 16              1893         ACD
## 17              1893         ACD
## 18              1893         ACD
## 19              1893         ACD
## 20              1893         ACD
## 21              1893         ACD
```

```
## 22              1893         ACD
## 23              1893         CKP
## 24              1894          AM
## 25              1894          AM
## 26              1894          AM
## 27              1894          AM
## 28              1895          AM
## 29              1895          AM
## 30              1895          AM
## 31              1895          AM
## 32              1895          AM
## 33              1896          AM
## 34              1900        JEPM
## 35              1900        JEPM
## 36              1900        JEPM
## 37              1900        JEPM
## 38              1900        JEPM
## 39              1900        JEPM
## 40              1900        JEPM
## 41              1900        JEPM
## 42              1900        JEPM
## 43              1903         ACD
## 44              1903         ACD
## 45              1903         ACD
## 46              1903         ACD
## 47              1904          BO
## 48              1904         ACD
## 49              1904         ACD
## 50              1904         ACD
## 51              1904         ACD
## 52              1904         ACD
## 53              1904         ACD
## 54              1904         ACD
## 55              1904         ACD
## 56              1904         ACD
## 57              1905          RB
## 58              1905          VW
## 59              1905         AKG
## 60              1905          AF
## 61              1905         AKG
## 62              1905          AF
## 63              1905          AF
## 64              1905          IZ
## 65              1905         TWH
## 66              1905          AM
## 67              1905          AF
## 68              1905          RB
## 69              1905          RB
## 70              1905         TWH
## 71              1905         AKG
## 72              1905          RB
## 73              1905         AKG
## 74              1905          AF
## 75              1905          VW
```

```
## 76              1905         TWH
## 77              1905          AM
## 78              1905         AKG
## 79              1905          RB
## 80              1905         TWH
## 81              1905          AF
## 82              1905          BO
## 83              1905         WHH
## 84              1905         TWH
## 85              1905         MPS
## 86              1905         TWH
## 87              1905          AF
## 88              1905          AF
## 89              1905         TWH
## 90              1905         TWH
## 91              1905         CKP
## 92              1905         AKG
## 93              1905          AF
## 94              1905         TWH
## 95              1905         AKG
## 96              1905         MPS
## 97              1905         TWH
## 98              1905          AF
## 99              1905          VW
## 100             1905         TWH
## 101             1905          VW
## 102             1905          VW
## 103             1905          VW
## 104             1905          BO
## 105             1905          VW
## 106             1905         TWH
## 107             1905          AF
## 108             1905          AF
## 109             1905         AKG
## 110             1905         CKP
## 111             1905          RB
## 112             1905         MPS
## 113             1908          BO
## 114             1908          BO
## 115             1908         ACD
## 116             1908          BO
## 117             1908          BO
## 118             1908          BO
## 119             1908          BO
## 120             1908         ACD
## 121             1910         ACD
## 122             1911         ABR
## 123             1911         ABR
## 124             1911         ACD
## 125             1911         ABR
## 126             1911         ABR
## 127             1912         ABR
## 128             1913          EB
## 129             1913          EB
```

```
## 130              1913         ACD
## 131              1913          EB
## 132              1913          EB
## 133              1913          EB
## 134              1917         ACD
## 135              1920         ACD
## 136              1921         ACD
## 137              1923         ACD
## 138              1924         ACD
## 139              1924         ACD
## 140              1926         ACD
## 141              1926         ACD
## 142              1926         ACD
## 143              1927         ACD
```

From these two tables, we can know that for either group, the two clusters are not coming from a particular crime type or motive type.

Therefore, we would like to use Fisher exact test of independence to check whether the distributions of crime and motive types from these two clusters in each group are independent.

The p-value for the test of each group are below.

Group	p.value.for.crime.distribution.test	p.value.for.motive.distribution.test
Cluster 3 and 6 from dist clustering	0.001	0.02
Cluster 1 and 5 from tf-idf clustering	0.870	0.50

From the p-value above, we can conclude that:

1. Overall, stores in cluster 3 and 6 from distance clustering are more likely to have different topics.

- The p-values for the test on crime and motive type distributions of cluster 3 and 6 are smaller than 0.05.
- It's a strong evidence against the null hypothesis that these two groups are from the same distribution (topic).

2. Overall, stories in cluster 1 and 5 from tf-idf matrix clustering are more likely to have similar topics.

- The p-values for the test on crime and motive type distributions of cluster 1 and 5 are larger than 0.05.
- We fail to reject the null hypothesis that these two groups are from the same distribution (topic).

Choice of testing methods

Why choosing fisher's exact test

- We want to see whether they two clusters are having similar topic.
- It's hard to define topic by the annotations we have. However, we can look at the overall distribution of crime and topic as an overview of topics in the cluster.

Why not choosing chi-square test?

- The reason for not doing chi-square test is simple. Chi-square test is more useful when the sample size is large, while here these clusters are specifically assigned from our the small-size sample.
- Therefore, fisher's exact test is more suitable under this case.

In-depth Interpretation of test result

- If the clusters are from similar distribution, it indicates that there might be some commonly appeared word (e.g. police) in these particular topics.
- If the clusters are from different distribution, it indicates that there might exist some other reasons (other linguistic features) that push the clustering algorithm to divide them.
- Either result could become the next research question for our collaborators.

3.2 Next Steps for the collaborators

The analysis is ended in this part, since further investigation is too difficult without the domain knowledege of detective stories.

Here are some questions that could be investigated further by our collabortors:

1. The two clusters have different topics, then what are the implicit logic behind the assignment of cluster 3 and 6?

What elements (either components of stories or linguistic features) lead to the special representation vectors that could be classified into different groups?

2. What are the explicit language features in those specific topics from cluster 1 and 5?

Topic is a broad definition here. Can we define them in a more specific way? What are the key factors lead to the final assignment?

IV. Explanation of Doing an Unconventional Analysis & Limitations

In my opinion, the main component of this analysis relies on the idea of text analysis and the choices of method to conduct the analysis (tf-idf, clustering, dimensionality reduction for visualization).

These methods (e.g. clustering) include some statistical component, while it's not the traditional testing or modeling expected in a general statistical analysis. This is mainly because of the characteristic of text data that I want to analyze.

I choose this topic as a challenge of learning new things and conduct insightful analysis from the text of detective stories, rather than the annotated dataset.

The design and choice of hypothesis test in the analysis part is not completely reasonable. The test itself and some assumptions might not be suitable for the research question. This is mainly because of the unconventional dataset we have and the difficult question that we want to investigate.

However, it's too hard to find the pattern and analyze further without the insight and domain knowledge from our collaborators. The only choice to keep the analysis part is to include these tests.

Therefore, even if the tests are not absolutely reasonable, I still keep them in the report and structure the report as the one that is originally expected in STA490.

I still believe that to obtain a more meaningful text data analysis, it's better to communicate and cooperate with our collaborators for in-depth understanding of the stories (those clusters) rather than doing the testing like this analysis.

I hope my decisions wouldn't be criticized and hurt the grade of this project.

V. Sentiment Analysis

In this section, we mainly focus on another research question: whether the sentiments in the first and second portion of detective stories change over years.

5.1 Data Cleaning

We first clean the data by unicode normalization and manual check of punctuation, and eventually find the reveal border sentences recorded in the input_form.csv. The issues encountered are provided below:

1. There are some annotations that don't match the stories (i.e. the reveal border sentence can't be found successfully). These stories are: 'OMIC04', 'ASH09', 'OMIC03', 'CKS53', 'PVDS41', 'CBSH05'. The problem for each of them are provided in project log.
2. There are some stories without an explicit reveal border sentence provided in the input_form.csv. These stories are: 'ASH01', 'CBSH10', 'LMSY03', 'MC03', 'GPM01', 'OSH05', 'TSOTR15', 'TEV02_02', 'TEV02_01'.
3. There are some stories provided but not annotated in the input_form.csv. These stories are These stories are: 'TEV02': 'TEV02 - Cassie Côté(1).txt', 'TCD03': 'TCD03 - Wen W..txt', 'TCD02': 'TCD02 - Wen W..txt', 'TCD01': 'TCD01 - Wen W..txt'.
4. There are some stories annotated but aren't provided as the plain texts. These stories are: 'CKS21'.

Apart from these problematic annotations or stories, we get the first and second (investigation & reveal) portions of each detective story annotated in the input_form.csv.

5.2 Lexicon-based Sentiment Analysis

For the sentiment analysis, we utilize the 'nltk' and 'vaderSentiment' package in python. The steps are:

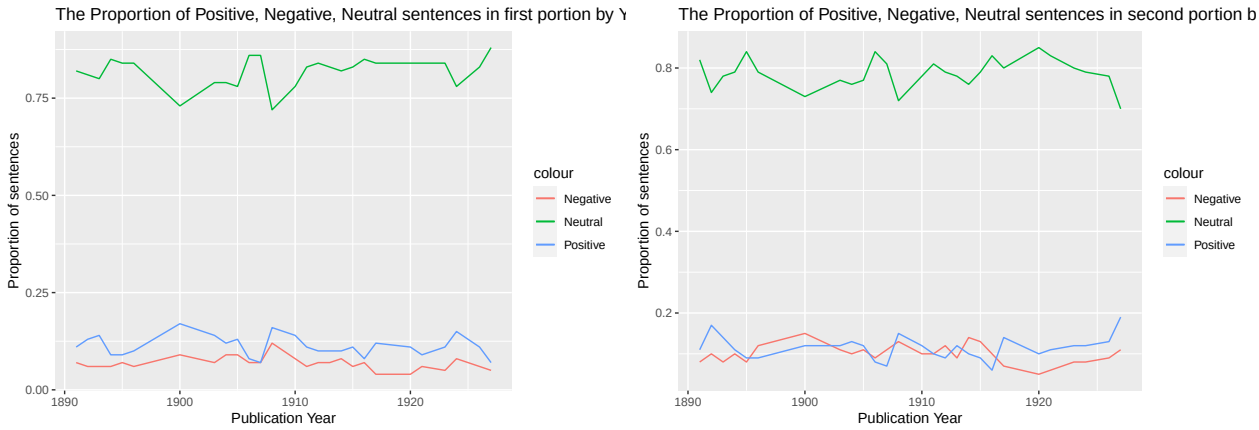
1. Perform sentence tokenization (i.e. separate a complete paragraph to different sentences) by nltk.
2. Clean each sentence by removing punctuation and number.
3. Perform word tokenization (i.e. separate a complete sentence to different words) on each sentence by nltk.
4. Do pos tagging (i.e. categorizing words in a text in correspondence with a particular part of speech) by nltk.
5. Remove stop words (i.e. the most common words in any language) by nltk.
6. Do lemmatization (i.e. grouping together the inflected forms of a word so they can be analysed as a single item. For example, make 'am', 'is', 'are' -> 'be') by ntlk.
7. Perform sentiment analysis on the lemma list of each sentence by vaderSentiment. The package provides the analysis for each lemma list and give a score.

- The score is restricted between -1 and 1.
- A score greater than or equal to 0.5 implies a positive sentence.
- A score less than or equal to -0.5 implies a negative sentence.
- Otherwise, it implies a neutral sentence.

8. Record the proportion of positive, negative and neutral sentences of each detective stories in each portion. Save the csv file.

5.3 Trend Analysis

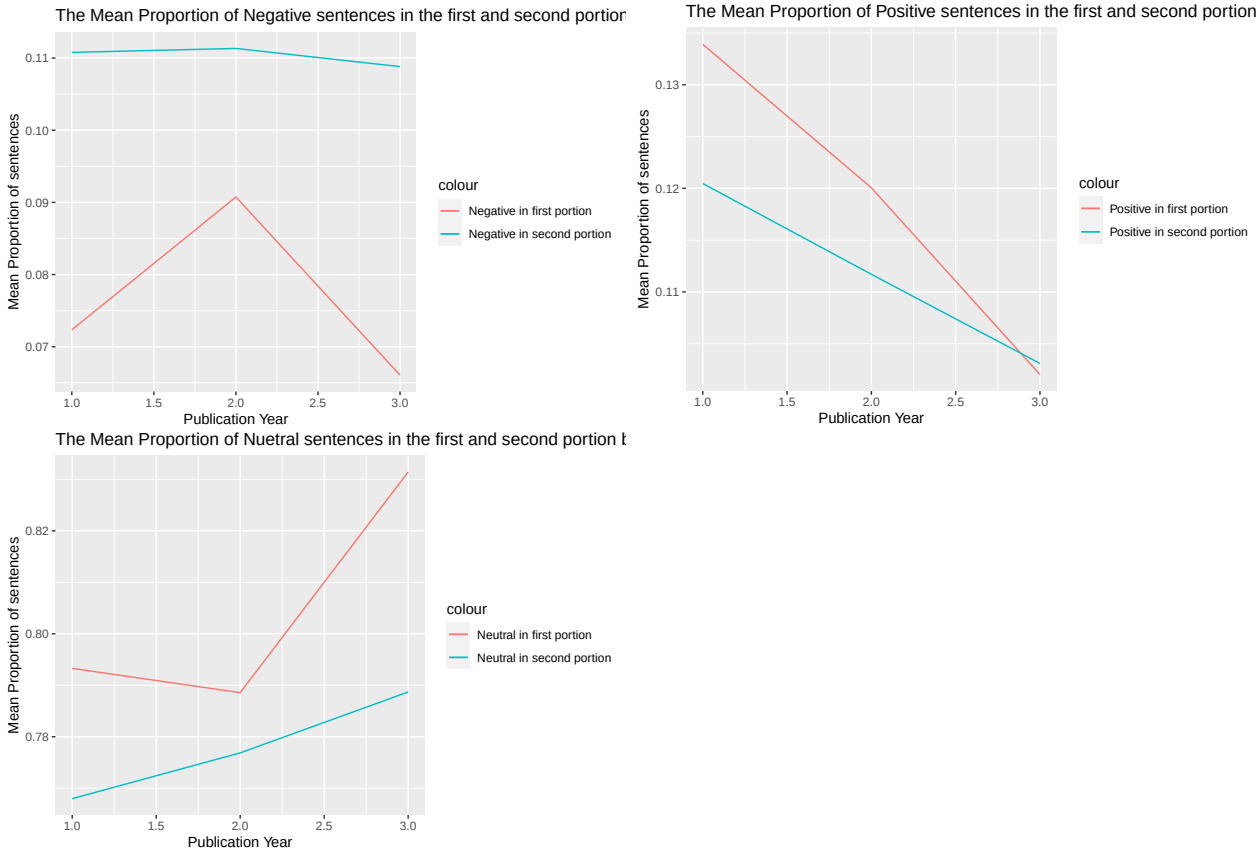
Now it's time to look at the trend of these proportions. These two plots are the line plot of proportion of positive, negative, and neutral sentences in the first and second portion changes over year. From these two plots, it's obvious that there is no explicit trend.



To compare the difference of each sentiment between first and second proportion, we group the publication year by decade as follow:

decade	decade_level	count
1891-1900	1	64
1901-1910	2	188
1911-1927	3	84

And the plots of mean proportion of different sentiments in different decade are plotted below.



From these three plots, we can know that the trend is still not obvious between each decade. Therefore, we can conclude that the sentiments in different portions of decective stories didn't change a lot over years.

VI. XGBoost for Choosing Readers' Favorite Topics

In this section, we utilize XGBoost Classifier for choosing the readers' (annotators') favorite topics (components of stories).

6.1 Introduction to XGBoost

XGBoost (Extreme Gradient Boosting) is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework.

The reason to choose this method is that decision tree based algorithms are considered the best type of algorithms for handling small-to-medium tabular data, and XGBoost is considered the best one among them in the Kaggle competition. It outperforms most on the model and trains very fast.

6.2. Construct XGBoost Classifiers for Readers' Recommendation

The labels for training XGBoost is "Would you recommend this story to a friend", which is a binary classification problem.

To examine the influence of each components of the stories, we train 3 different XGBoost Classifiers for motive types, crime types and other potential categorical variables in the input_form.csv dataset.

We didn't construct one classifier for all these variables, because the connections between these three factors is not obvious and we want to focus on the internal influences from each of these three factors itself.

6.3

Appendix 1: Data Preparing Process

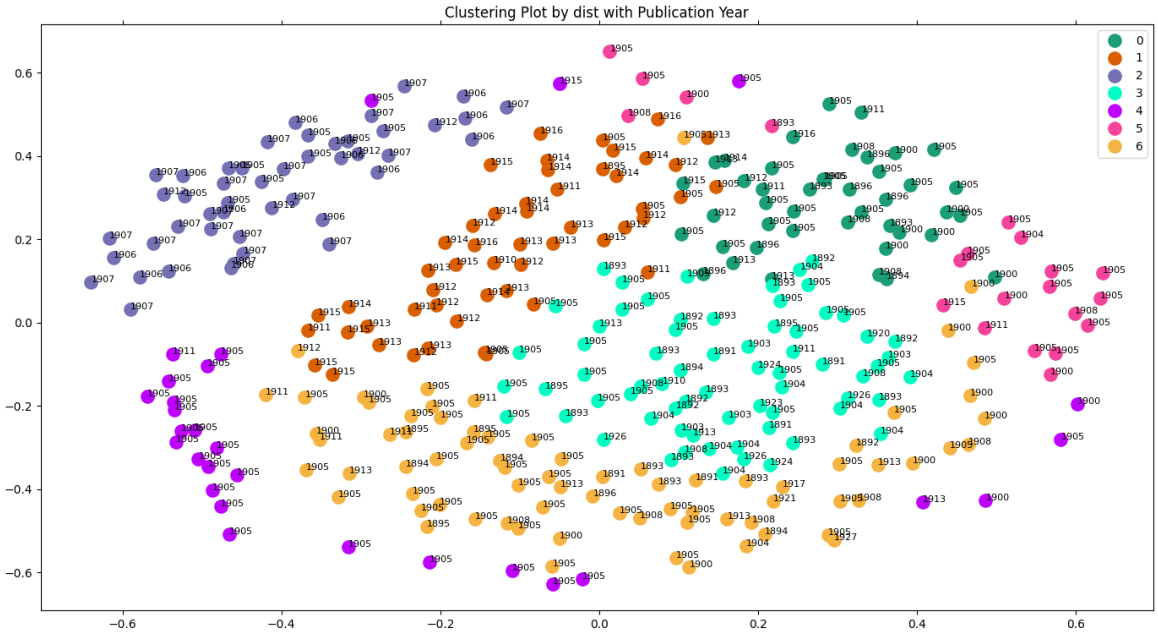
To deal with the text data, we utilize python instead of R. I upload all python code, R code, produced csv and plots in github repository (https://github.com/rachan1637/detective).

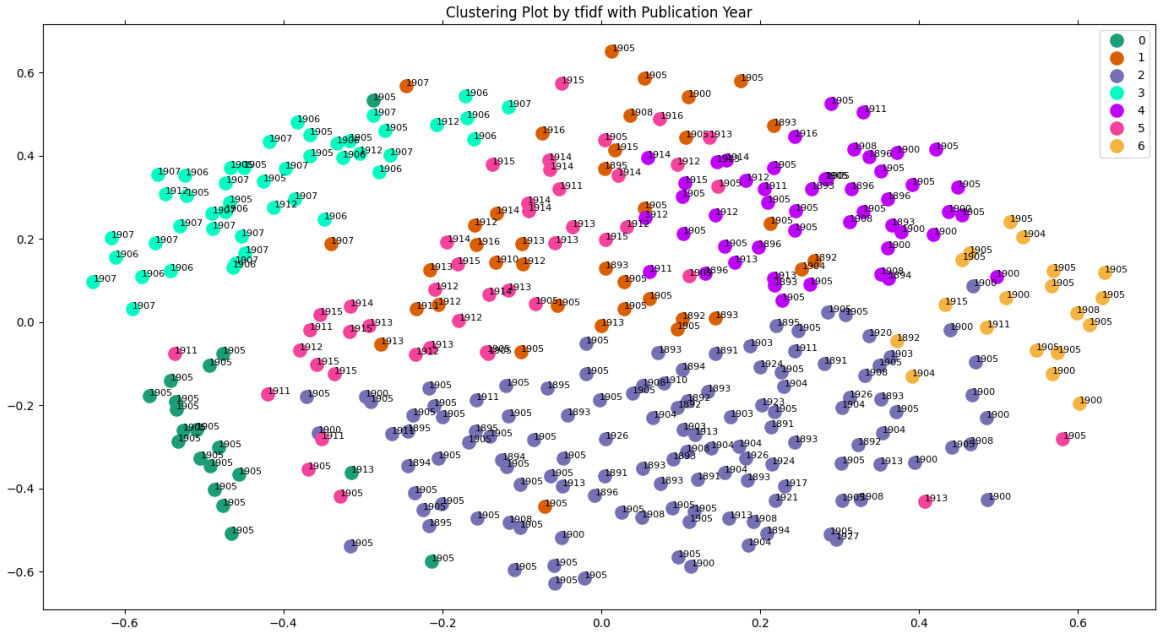
The preprocessing includes the text data normalization and partitions. The partition-by-reveal-border process was initially designed for doing sentiment analysis, which was mentioned by the collaborators. But since the target for this task is vague, and unsupervised sentiment analysis is harder than expected, we give up the plan eventually.

You can do 'python {path}/data_pacakage/main.py' in the terminal / bash to run the process of doing preprocessing, tf-idf counting, clustering and plotting. The path inside the python program might be changed, depending on the directory root that you run the program.

I was too busy in these few weeks. As the grading rubric doesn't ask for a cleaned python code, I didn't add much annotations and clean my python code for your convenience. However, I am willing to do that if the TA, the instructor or the collaborators want to trace my code in future.

Appendix 2: Clustering Plots with Publication Year information





Why we exclude pubilcation year as a reason of clustering:

- We know that the collection of the dataset is biased, only the detective stories that could be found on the Internet are collected.
- The reason that the clusters seem to be assigned due to the year is that many stories written by the same author were published in adjacent year, and all stories we have in some particular years are from only one or two authors. This conclusion can be proved by the plot with author information above.
- Therefore. The publication year should not be considered as the primary reason of clustering.

Appendix 3: Top words contributes to tf-idf matrix clustering

Table 5: Top-10 influential words for each cluster

Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
replied	miss	sir	thinking	mrs	doctor	lady
cried	father	friend	miss	woman	new	miss
big	woman	police	detective	lady	mrs	woman
dead	girl	london	didn	wife	quickly	girl
father	mrs	window	woman	miss	miss	dear
girl	lady	round	mrs	girl	replied	london
woman	window	paper	replied	police	remarked	police
sir	herself	street	girl	herself	air	sir
fingers	replied	father	clock	replied	body	wife
chair	paper	business	money	money	blood	herself