Evolution of Components in 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. 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 various detective stories. Combining traditional statistical analysis as well as various text data mining methods, we mainly focus on the evolution of components in detective stories. In the meanwhile, we also investigate the condition of affective states in the detective stories by sentiment analysis and the potential factors that might differentiates the stories by clustering.

In conclusion, we find that the proportion of using Murder as crime increases over decades, the proportion of using Theft decreases over decades, the proportion of choosing Jealousy or Greed as motive increases over decades, and there is no obvious trend for the proportion of sentence with different sentiments and the text features over decades. As for other research questions, by doing hypothesis testings, we successfully reach the conclusion that the situation of affective states in the detective stories is not affected by the situation of female victim. Lastly, we suppose that some clusters of stories are differentiated by topics while some clusters are not, but the true factors need further investigation.

2. Introduction

Detective fiction has already been popular between readers for many years. Over the decades, many researchers are strongly interested in analyzing the detective stories from different aspects. English literature researchers always want to know the evolution of the components in the detective stories, including the content, the text feature, and the affective states in the stories. In addition, the situation of affective states in detective stories with different components and whether the detective stories can be differentiated by the text features are also some intersting questions that intrigue the researchers.

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 investigate three research questions:

1. How does detective stories evolve over years?
2. What is the condition of affective states (the proportion of positive, negative, and neutral sentiments) in the detective stories? In particular, are the stories with female victims having more sentences with negative sentiments?
3. What are potential factors that might differentiates the stories?

The evolution of detective stories is the main research question. However, even if we try to uncover this question through different methods (either in the sentiment analysis 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 the annotated dataset.

For the analysis of the sentiments, in addition to the evolution, we mainly investigate whether the gender of victim affects the proportion of sentences with negative sentiments in the stories.

And lastly, for the analysis about the hidden factors 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.

For the remainder of the paper, section 3 is about the method, where 3.1 is about types of EDA plots, and 3.2 is about sentiment analysis method, and 3.2 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.

3. Methods

3.1 EDA

For the EDA content, we investigate the evolutionary of detective stories mainly by box plot and line plot (time series plot). The dataset preprocessing details are in Appendix 1.

3.2 Sentiment Analysis

3.2.1 Methods to do Sentiment Analysis

The sentiment analysis in this report is lexcion-based. This is mainly because the dataset is unsupervised (i.e. we don't have lablelled sentence to train a machine learning model for doing classification). In this case, analyzing the sentiment through lexcion features is the most appropriate way.

The 8 steps of lexicon-based sentiment analysis are:

1. Perform sentence tokenization (i.e. separate a complete paragraph to different sentences) by nltk.
2. Clean each sentence by removing punctuation and digits.
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. Each lemma list receive a score. The score is restricted between -1 and 1. A score greater than 0.5 imples a positive sentence. A score less than -0.5 implies a negative sentence. A score between -0.5 and 0.5 implies a netural sentence.
8. Record the proportion of positive, negative and neutral sentences of each detective stories in each portion. Save the csv file.

The data files cleaning process includes finding the reveal border sentence, and the special cases encountered are listed in Appendix 2, and the details of implementation steps are presented in Appendix 3.

3.2.2 Methods to Investigate Research Questions

There are two research questions: the evolution of detective stories and the sentiment variation by gender of victims).

In the evolution analysis part, we are mainly doing EDA of time series plots to show the results. As for the other research question, we conduct the investigation by two sample t-test. Specifically, we divide the stories into two groups, with female victims and without female victims, and conduct two sample t-test on different sentiment statistics (e.g. the proportion of sentences with negative sentiments in the first portion).

3.3 Clustering through tf-idf features

The main goal of clustering is to uncover the potential text features that differentiate the detective stories. The main expectation is that some stories are differentiating because of the publication year.

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 can be dismantled into 4 subsections.

3.3.1 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

The intuition behind this calcuation is that, 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 more crucial word than "knife" to this story.

3.3.2 tf-idf scores as representation vector

Since tf-idf scores for each word can represent how important the word to this story, by forming them as a vector, it becomes a representation learned from the text of the story. This is a popular method to create representation for documents in information retrieval. To do that, we calculate the tf-idf score for every word in each document, and use these scores to form a vector.

In order 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 the threshold is to reduce the noise from common words such as ‘he’ or ‘she’ and the noise from rare terms that might appear only in a specific type of stories (e.g. Holmes). The 50% and 80% are picked based on empirical experiments.

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.

3.3.3 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.

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.

Comments pf the implementation are provided in Appendix 4.

Choice of Methods

K-means clustering and MDS are both popular methods in machine learning.

K-means clustering is a method of vector quantization that aims to partition n observations into k clusters, where 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).

Clustering is one the most popular unsupervised methods for pattern recognition. And the reason to do clustering is to differentiate the stories by text feature vectors so that we can find out the hidden reason behind cluster.

In terms of clustering methods, we also try the hierarchical clustering. 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.

3.3.4 Analyze the differentiation through Fisher’s Exact Test

At the end of the clustering, we also conduct the Fisher’s Exact Test to check whether some stories are divided into different clusters due to topic differences.

The purpose of testing is to verify whether the two clusters are having similar topic. We test the topic between different clusters by comparing the overall distribution of crime and topic of two clusters. As for the choice of testing method, instead of choosing the popular Chi-square test, we do Fishers’ Exact test because it’s more useful when the sample size is large, and our clusters are assigned from the small-size dataset.

Table 1: Number of publications in different decade

decade	count
1891-1900	66
1901-1910	195
1911-1927	90

There are also other possible ways to analyze the differentiation. For example, it's possible 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 word level. 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. These influential words (top-10) for each cluster in tf-idf matrix clustering can be found in Appendix 5.

4 Results

4.1 EDA Results

4.1.1 Distribution of Publication Year

The distribution of publication year is in Figure 1.

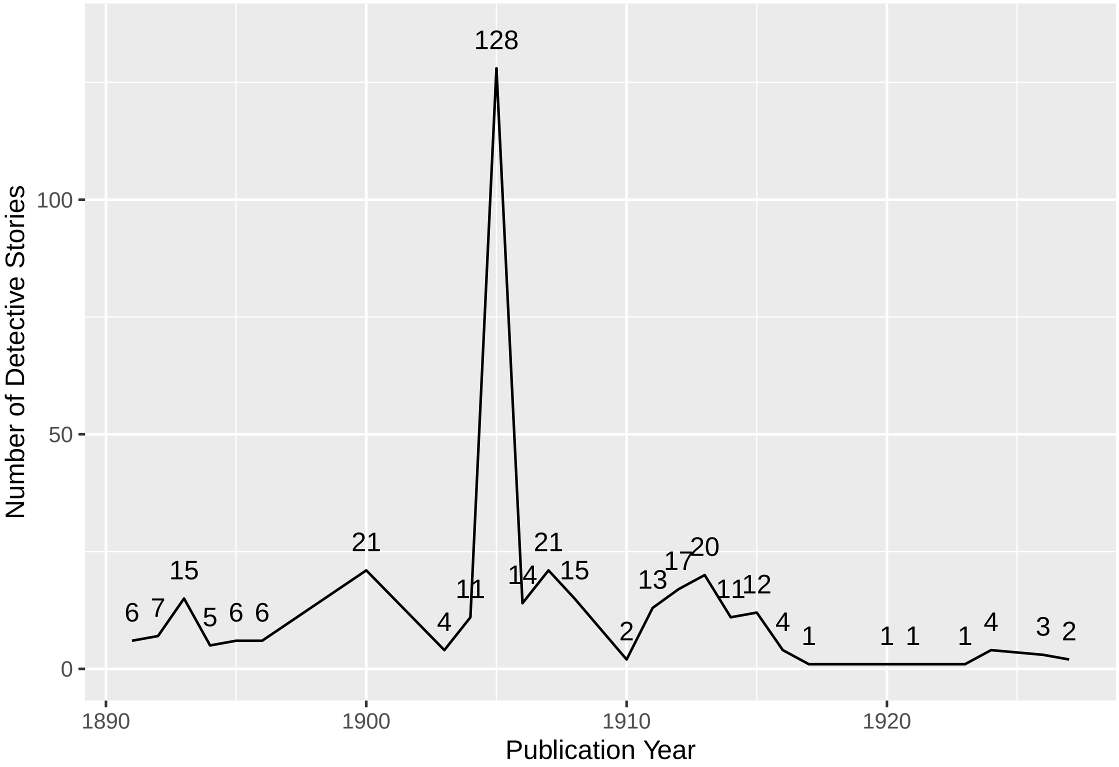


Figure 1: distribution of the publication years of stories

From the plot, we can see that the stories annotated in this dataset were mainly published at 1905. Since 1915, the number of stories were decreasing drastically.

Also, the numbers of publications in each decade are concluded in Table 1.

4.1.2 Trend Analysis for Proportion of different components

There are some interesting trends of components in stories that we can found. The two key components presented in this report are crime types and motive types.

Trend of Crime Types

There are 13 types of crime concluded in the annotated dataset. Among all these crimes, assault, fraud, murder and theft are the most popular crime written in the stories.

We calculate the proportion of these crimes in each year, group them by decade, and use box plot to visualize the results. No obvious trend has been seen from the plots of assault, fraud and theft. The final plot for murder is in Figure 2.

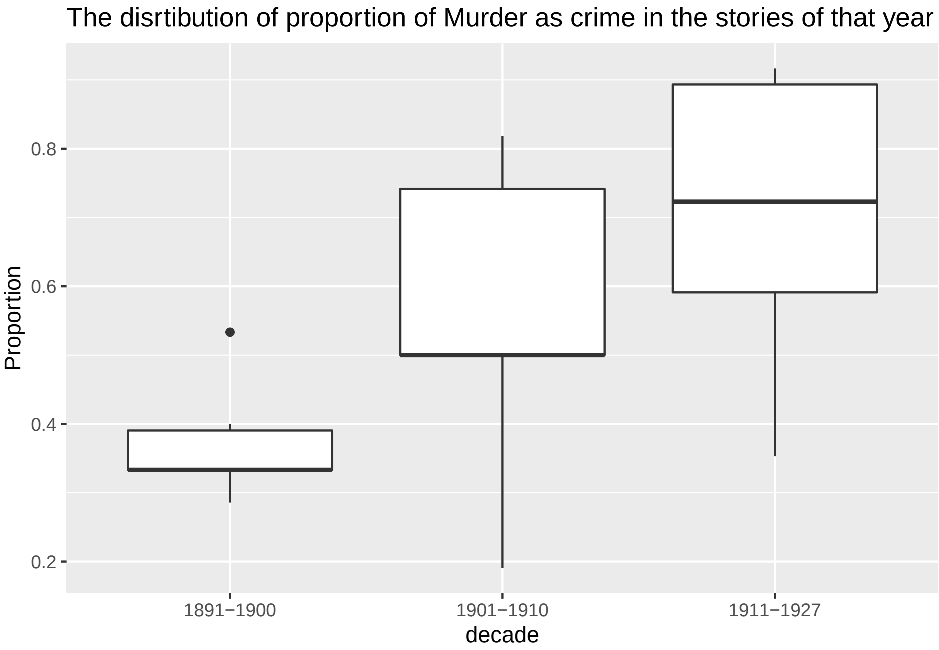


Figure 2: distribution of proportion of crime type murder in different decade

From Figure 2, we can mostly believe that the proportion of using Murder as crime increases over decades.

Trend of Motive Types

There are also a variety of motive types concluded by the annotated dataset.

We calculate the proportion of these types in each year, group them by decade, and use boxplot to visualize the results. The final plots for Jealousy and Greed are in Figure 3 and Figure 4, and no obvious trend is seen in plots of other motive types.

From Figure 3 and Figure 4, we can conclude that the proportion of choosing Jealousy or Greed as motive increases over decades.

4.2 Sentiment Analysis Results

4.2.1 Trend Analysis for Proportion of sentences with different sentiments

This subsection tries to figure out the trend of proportions of sentences with different sentiments change over years.

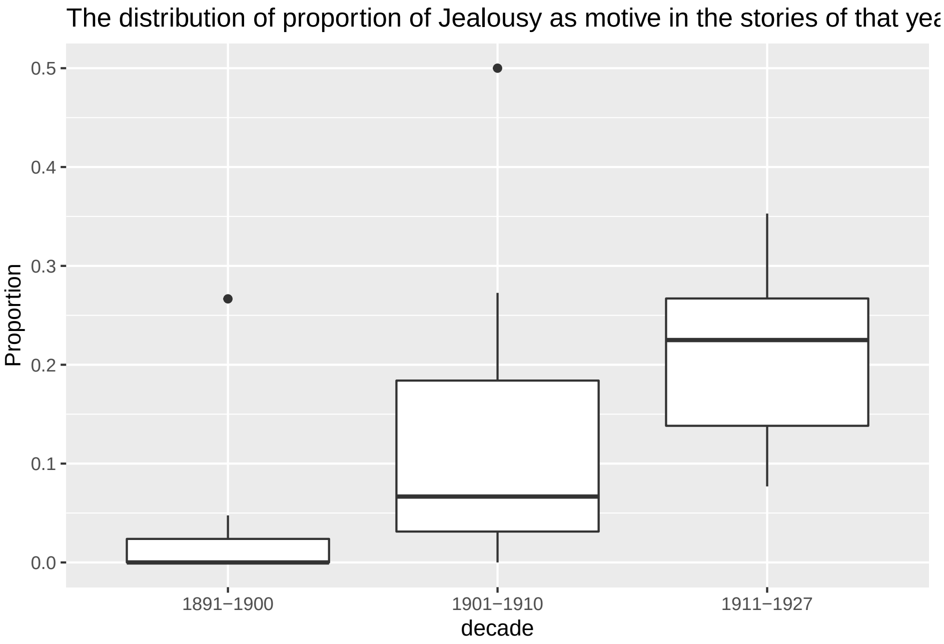


Figure 3: distribution of proportion of motive type Jealousy in different decade

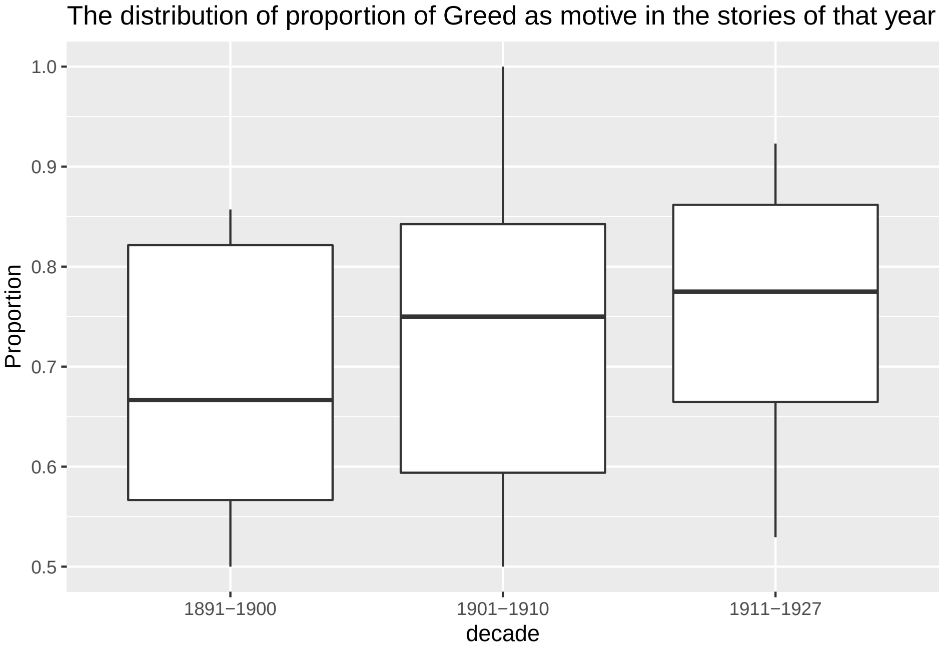


Figure 4: distribution of proportion of motive type Greed in different decade

Table 2: two sample t-test result

Experiment.Result	First.Portion	Second.Portion
t	0.43	1.56
df	319.05	326.09
p-value	0.67	0.1195

Figure 5 and Figure 6 are the line plot of proportion of positive, negative, and neutral sentences in the first and second portion changes over year.

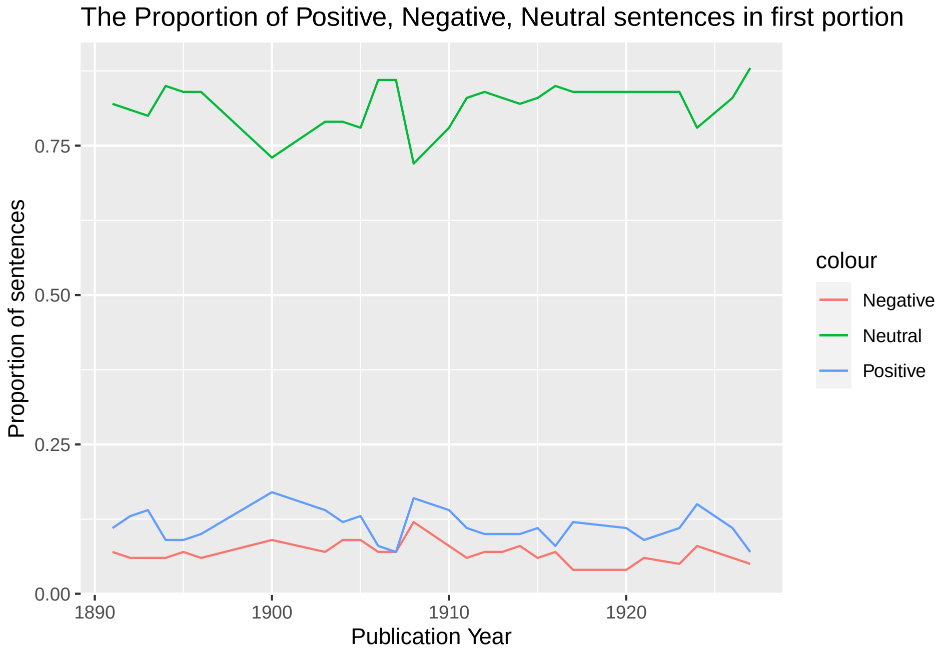


Figure 5: The proportion of positive, negative, and Neutral sentences in first portion over years

From Figure 5 and Figure 6, it's obvious that there is no explicit trend.

However, we can also find some interesting fact if we plot these sentiments together by portions.

From Figure 7, we can know that there are always more sentences with negative sentiments in the second portion (reveal phase). This is reasonable because the reveal phase may include more sentences that people show pity to the victims or show anger to the culprits.

4.2.2 Sentiments Variation by Gender in stories

In this section, we conduct two sample t-test on the proportion of sentences with negative sentiments in the first and second portion. And t-statistic, degree of freedom and p-value are presented in Table 2.

From Table 2, we can see that both p-value are greater than 0.05, which indicates that both tests fail to reject the null hypothesis that the stories with and without female victims are having different mean of proportion of sentences with negative sentiments.

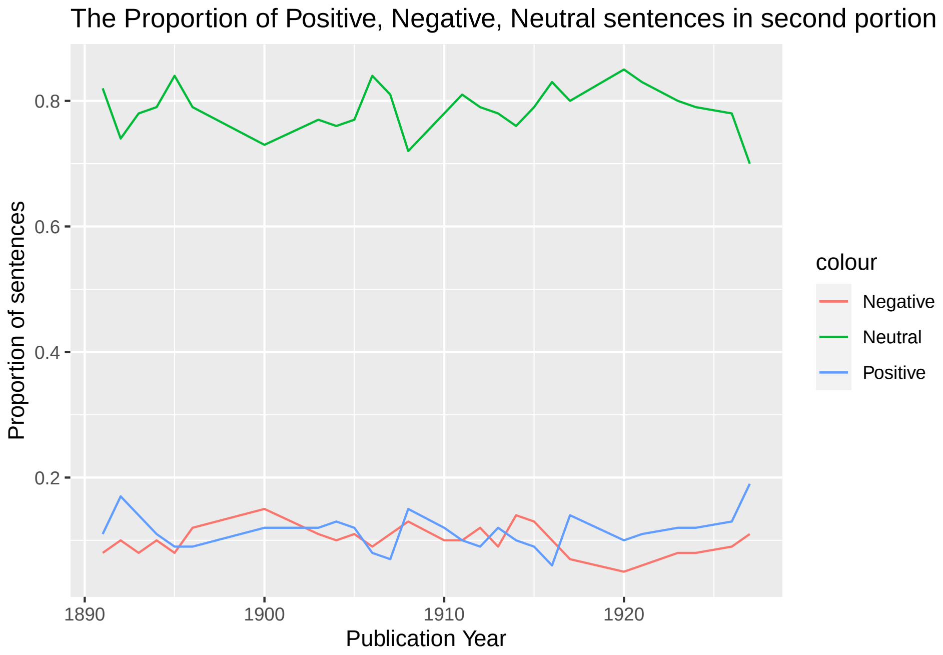


Figure 6: The proportion of positive, negative, and Neutral sentences in second portion over years

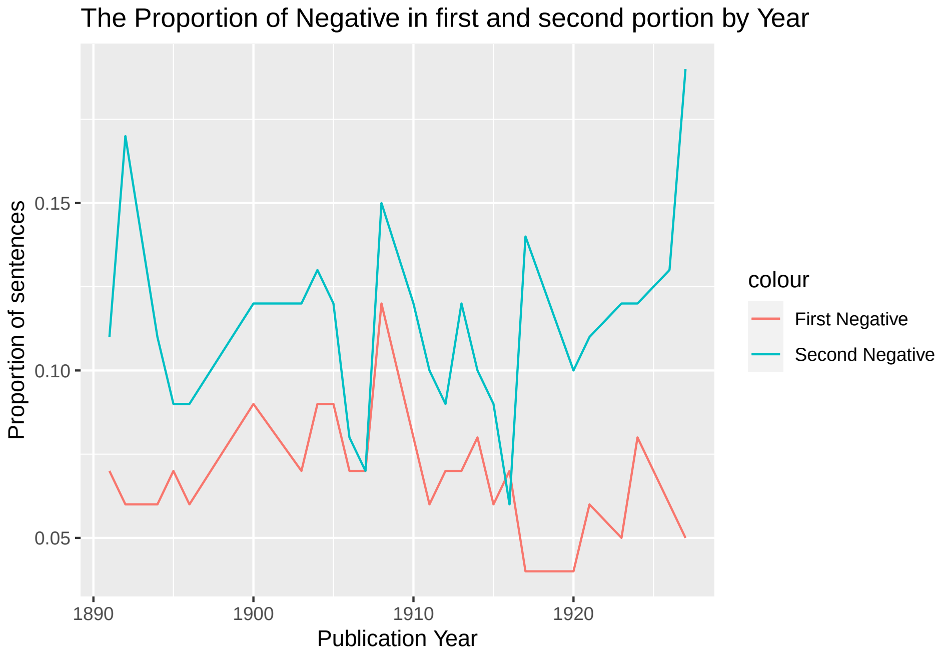


Figure 7: The proportion of sentences with negative sentiment in different year

4.3 Text Feature Clustering Results

4.3.1 Clustering EDA results

The two plots, Figure 8 and Figure 9, 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.

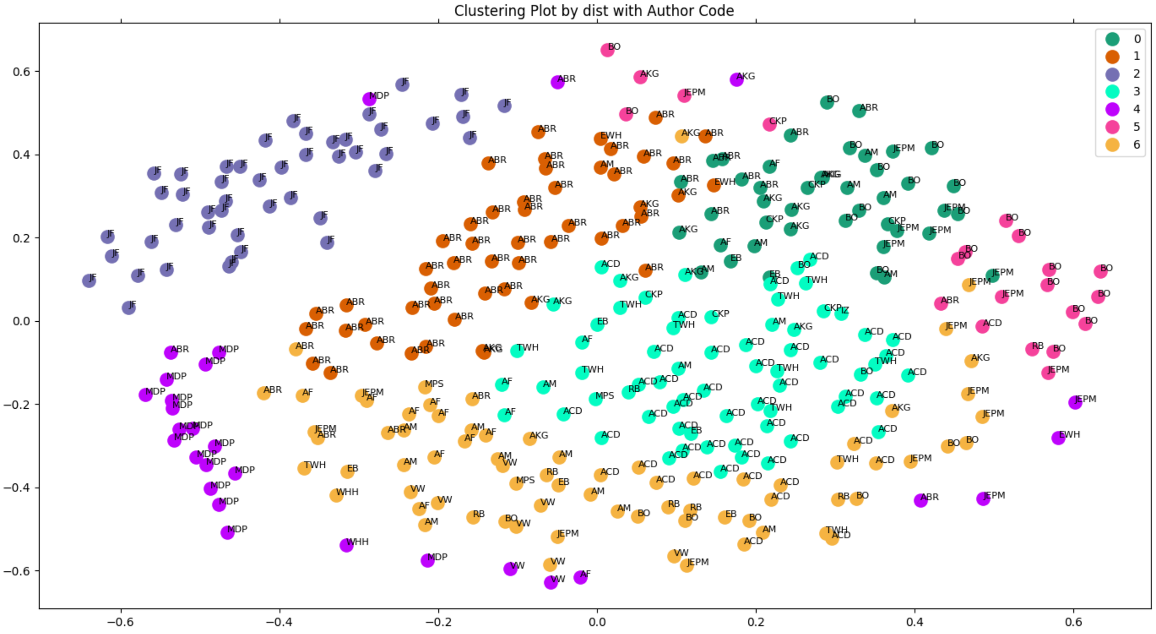


Figure 8: Clustering plot by distance matrix with Author Code

From Figure 8 and Figure 9 (clustering plot by distance matrix and tf-idf matrix with author code information), we can see 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).

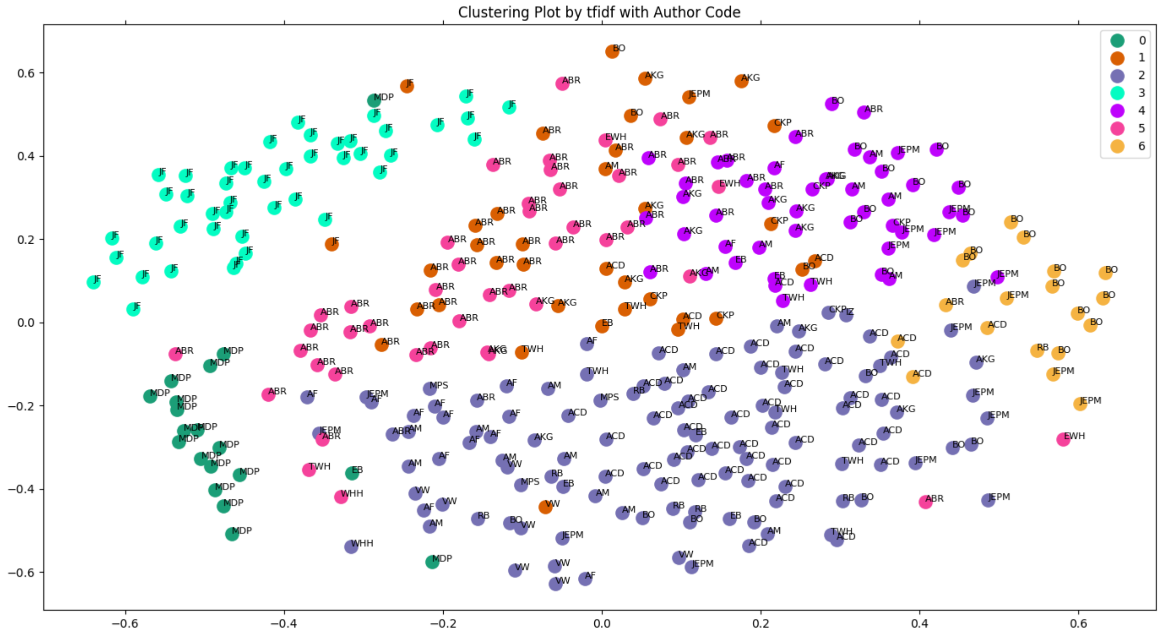


Figure 9: Clustering plot by tf-idf matrix with Author Code

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. The plots with publication year information are provided in the Appendix 6. There are also comments for justifying why the reasons are not publication years.

Fact A: Cluster 3 and 6 in distance matrix clustering (Figure 8)

One 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.

Fact B: Cluster 1 and 5 in tf-idf matrix clustering (Figure 9)

The other 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

Table 3: Number of crime types in selected clusters

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

between cluster 1 and 5 in plot 2, while these features can't be captured by similarity distances.

4.3.2 Does topic differentiate the cluster?

There are two questions come up here:

1. What is differentiating cluster 1 and cluster 5 in Figure 8 (tf-idf matrix clustering)?
2. What is differentiating cluster 3 and cluster 6 in Figure 9 (similarity distance matrix clustering)?

In particular, since we have annotated information about topics (crimes, motives, etc.), we plan to conduct several hypothesis testings (Fisher's Exact Test) to see whether these clusters are assigned due to similar topics.

However, these hypothesis testing are superficial and can't represent the authenticate logic behind the classification. The logic should be left for the experts of detective stories, our collaborators.

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 will specifically focus on the crime type and motive in the stories. These two components are the most representative annotations in our dataset.

The numbers of crime types and motive types in selected clusters are in Table 3 and 4.

From the tables, we can see 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 in Table 5.

From the p-value in Table 5, we can conclude that:

Table 4: Number of motive types in selected clusters

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

Table 5: p-value for crime and motive distribution test

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

1. Overall, stories 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).

If the clusters are from similar distribution, it indicates that there might be some commonly appeared word (e.g. police) in these particular topics. And 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.

5 Conclusions and Discussions

5.1 Conclusion

In this research, we investigate the three research questions: How does detective stories evolve over years? What is the condition of affective states (the proportion of positive, negative, and neutral sentiments) in the detective stories? What are potential factors that might differentiates the stories?

From the EDA, we can know that the proportion of using Murder as crime increases over decades, and the proportion of choosing Jealousy or Greed as motive increases over decades.

From the sentiment analysis, we can know that there is no obvious trend of sentiment components in the stories and the proportions of sentences with negative sentiments for the stories with or without female victims are similar. However, the proportion of sentences with negative sentiments in the second portion (reveal phase) is always higher than the proportion of sentences with negative sentiments in the first portion (investigation phase).

From the tf-idf feature extraction, clustering, and fisher's exact test, we can know that some stories may be divided into different groups due to different topics, while some stories are divided into different groups with mysterious reasons that we can't explain now.

5.2 Dicussions & Limitations

There are three main limitations for this report are:

The first limitation is the size of dataset. The size of the annotated dataset itself is not large, not the imbalanced problem is severe. The number of stories in most year are only 10-20, while there are more than 100 stories in 1905.

The second limitation is the assumptions of test. The assumptions of Fisher's Exact Test can't be fully satisfied, which implies that the result can only be viewed as a simple preliminary insight. The section is still kept in the report for reminding others to do more in future.

The third limitation is the lack of domain knowledge. Further investigation in clustering part is too difficult to statistician without domain (detective stories) knowledge.

There are many future work that can be further investigated. For example, it's important to know why some types of crime or motives are being used increasingly over decades. Moreover, the result from tf-idf feature extraction and clustering is still mysterious. For the cluster 3 and 6 having different topic, what are the implicit logic behind the assignment of cluster 3 and 6? And for the cluster 1 and 5, what are the explicit language features in those specific topics that exactly differentiate the clusters. All these questions worth more investigation from both the collaborators and statisticians.

In future, we believe topic modeling might be a good point to further trace the hidden factors that affect evolution and the result of clustering. Also, emotion analysis can be an extended research for analyzing the condition of affective states (sentences with different sentiments) in the detective stories.

Appendix 1: Prepartion before EDA

In the EDA section, we mainly prepare the dataset by:

- Add the publication year information for our investigation.

Our research interest includes the evolution of detective stories, while the original dataset doesn't contain the publication year information. Therefore, we add the publication year information for each story.

- Create two new variables "decade" and "decade_level".

The number of publications is small in each year. Therefore, we create two variables related to decade with 3 level ('decade': 1891-1900, 1901-1910, 1911-1928, 'decade-level': 1, 2, 3). We utilize them to create plots for investigating the trend in decades.

decade	decade_level
1891-1900	1
1901-1910	2
1911-1927	3

- Change the publication year of all stories > 1916 to 1916.

This is because for publication years later than 1916, only few stories published each year. To reduce the bias of that, we summarize and analyze them together as one year.

Appendix 2: Cleaning the text data files

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.

Appendix 3: Text Data Preprocessing Details

In general, people preprocess the text data before text data mining, since the raw text data may be noisy and interfere the conclusion. A common preprocessing procedure includes: sentence tokenization, word tokenization, stop words removal, punctuation and digits removal, and pos tagging.

The sentence tokenization is about splitting a complete paragraph into sentences, and word tokenization is about splitting the sentence into words. The reason for doing tokenization is to convert the unstructured text data into structured data so that the analysis can be performed in a better manner.

The punctuation, digits and stop words are removed because they are noisy when doing sentiment analysis. As for lemmatization and pos tagging, as the lexicon-based sentiment analysis model analyzes the sentences from linguistic perspective, these additional process can increase the efficiency of the model.

The python packagas we use are "nltk" and "vaderSentiment" in Python.

Table 6: 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

Appendix 4: Data Preparation for tf-idf feature extractions and clustering

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).

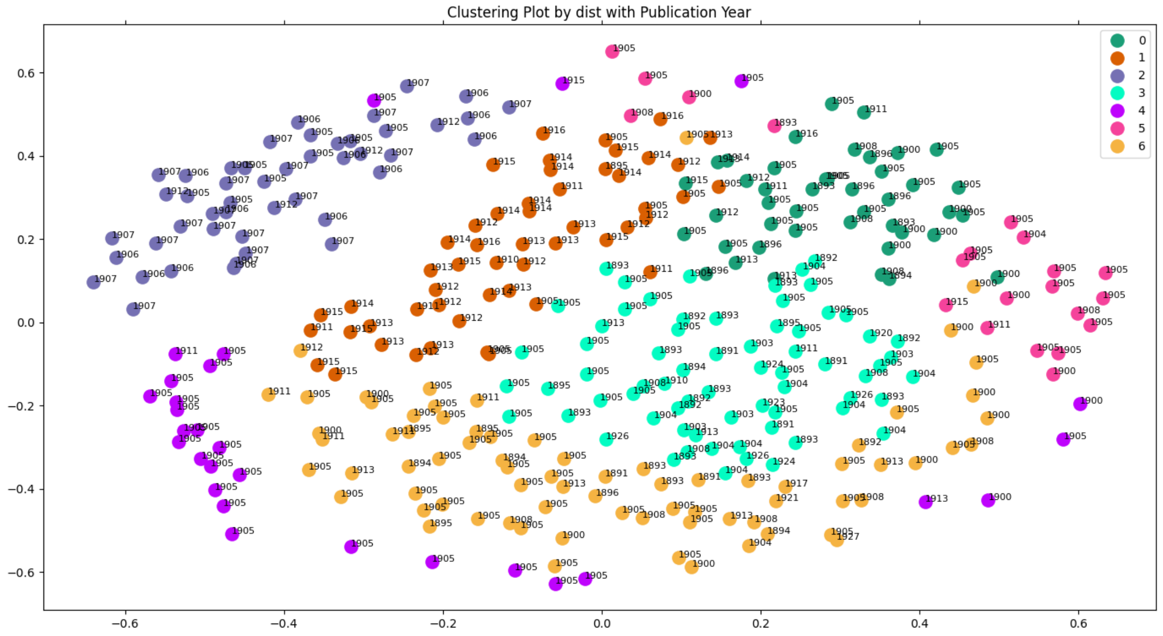
The preprocessing includes the text data normalization and partitions. The partition-by-reveal-border process was initially designed for doing sentiment analysis, which was mentioned by the collaborators. But since the target for this task is vague, and unsupervised sentiment analysis is harder than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

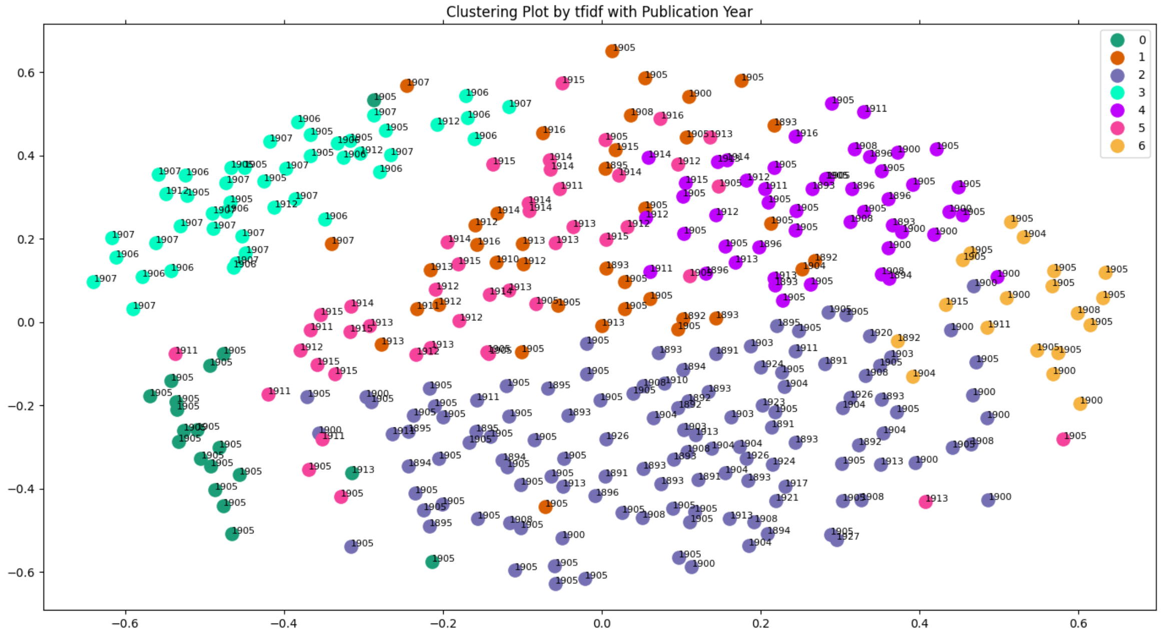
You can do ‘python {path}/data_pacakage/main.py’ in the terminal / bash to run the process of doing preprocessing, tf-idf counting, clustering and plotting. The path inside the python program might be changed, depending on the directory root that you run the program.

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 5: Top words contributes to tf-idf matrix clustering

Appendix 6: Clustering Plots with Publication Year information





Why we exclude publication 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.