AN ANALYSIS ON NETWORK AND TOPOLOGY OF LEGAL DOCUMENTS USING TEXT MINING AND GRAPH APPROACH

MR. SUPAWIT SOMSAKUL

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF ENGINEERING
(COMPUTER ENGINEERING)
FACULTY OF ENGINEERING
KING MONGKUT’S UNIVERSITY OF TECHNOLOGY THONBURI
2021

An Analysis on Network and Topology of Legal Documents Using Text Mining and Graph Approach

Mr. Supawit Somsakul B.Eng. (Computer Engineering)

A Thesis Submited in Partial Fulfillment of the Requirements for
the Degree of Master of Engineering (Computer Engineering)
Faculty of Engineering
King Mongkut's University of Technology Thonburi
2021

Thesis Committee

วาริน เชาวทัต.
.... Chairman of Thesis Committee
(Asst. Prof. Varin Chouvatut, Ph.D.)

.... Member and Thesis Advisor
(Asst. Prof. Santitham Prom-On, Ph.D.)

.... Member
(Asst. Prof. Kejkaew Thanasuan, Ph.D.)

Khajonpong Ak
.... Member
(Asst. Prof. Khajonpong Akkarajitsakul, Ph.D.)

.... Member
(Lect. Unchalisa Taetragool, Ph.D.)

Copyright reserved

Thesis Title	An Analysis on Network and Topology of Legal Documents Using Text Mining and Graph Approach
Thesis Credits	12
Candidate	Mr. Supawit Somsakul
Thesis Advisors	Asst. Prof. Dr. Santitham Prom-On
Program	Master of Engineering
Field of Study	Computer Engineering
Department	Computer Engineering
Faculty	Engineering
Academic Year	2021

Abstract

The number of documents in a dataset available on the Internet is increasing. However, the limitation of using textual information derived from documents in data analysis requires more computation time and resources as the data grow. An analysis of the documents in a dataset can be conducted using other types of features. This thesis presented a computational study on documents with references. The analysis chose Thai legal documents as a dataset. The data were collected from an information service system of the Supreme Court of Thailand's website. The study utilized text mining and network analysis to gain insight from the corpus. The analysis method included connected component analysis, graph clustering, and induced graphs. The study also built an information retrieval system from the network features obtained from the network analysis. The evaluation revealed an increase in the performance after incorporating the network features with textual features. This indicated the potential benefit of using network features in a real-world information retrieval system.

Keywords: Document Mining/ Graph Analysis/ Graph Clustering/ Information Retrieval System/ Network Science

หัวข้อวิทยานิพนธ์	การวิเคราะห์ความสัมพันธ์เชิงโครงข่ายของชุดเอกสารทางกฎหมายโดยวิธีการทำเหมืองข้อมูลและการวิเคราะห์กราฟ
หน่วยกิต	12
ผู้เขียน	นายศุภวิชญ์ สมสกุล
อาจารย์ที่ปรึกษา	ผศ. ดร. สันติธรรม พรหมอ่อน
หลักสูตร	วิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชา	วิศวกรรมคอมพิวเตอร์
ภาควิชา	วิศวกรรมคอมพิวเตอร์
คณะ	วิศวกรรมศาสตร์
ปีการศึกษา	2564

บทคัดย่อ

ด้วยจำนวนข้อมูลเอกสารที่สามารถนำไปใช้งานได้บนอินเตอร์เน็ตได้เพิ่มขึ้นอย่างรวดเร็ว ข้อจำกัดหลักของการใช้คุณลักษณะทางภาษาศาสตร์ในการวิเคราะห์นั้นคือขั้นตอนการวิเคราะห์จะใช้เวลาและทรัพยากรในการคำนวณมากขึ้นตามจำนวนข้อมูล การวิเคราะห์เอกสารนั้นสามารถทำได้โดยใช้คุณลักษณะรูปแบบอื่นๆ ในงานวิจัยนี้นำเสนอถึงการวิเคราะห์เชิงคำนวณกับข้อมูลเอกสารซึ่งมีการอ้างอิง โดยได้เลือกข้อมูลเอกสารทางกฎหมายมาใช้ในการวิเคราะห์ ซึ่งทำการเก็บข้อมูลจากเว็บไซต์ระบบสืบค้นคำพิพากษาศาลฎีกา ผู้วิจัยได้ทำการใช้ขั้นตอนวิธีทางการทำเหมืองเอกสารและการวิเคราะห์กราฟในการสกัดคุณลักษณะที่น่าสนใจของข้อมูล โดยการวิเคราะห์ได้ใช้ขั้นตอนการวิเคราะห์ส่วนประกอบที่เชื่อมถึงกัน การตัดแบ่งกราฟและการลดรูปกราฟ และยังได้สร้างระบบค้นคืนสารสนเทศซึ่งใช้ประโยชน์จากคุณลักษณะทางโครงข่ายที่ได้จากการวิเคราะห์ข้างต้น ในการประเมินผลพบว่าระบบค้นคืนสารสนเทศซึ่งสร้างมาจากคุณลักษณะทางโครงข่ายมีประสิทธิภาพในการค้นคืนเอกสารได้ดีขึ้น ซึ่งแสดงให้เห็นถึงความเป็นไปได้ในการใช้งานคุณลักษณะทางโครงข่ายกับระบบค้นคืนสารสนเทศในสถานการณ์จริง

คำสำคัญ: การตัดแบ่งกราฟ/ การทำเหมืองเอกสาร/ การวิเคราะห์กราฟ/ การวิเคราะห์โครงข่าย/ ระบบค้นคืนสารสนเทศ

CONTENTS

PAGE

ENGLISH ABSTRACT ii
THAI ABSTRACT iii
CONTENTS iv
LIST OF TABLES vi
LIST OF FIGURES vii
LIST OF EQUATIONS ix

CHAPTER

1. INTRODUCTION 1
    1.1 Statement of Purpose 1
    1.2 Objective 2
    1.3 Scopes 2
    1.4 Expected Benefits 3

2. LITERATURE REVIEW 4
    2.1 Related Works 4
    2.2 Theory 6

3. METHODOLOGY 16
    3.1 Proposed Methodology 16
    3.2 Dataset 17
    3.3 Data Preprocessing 17
    3.4 Exploratory Data Analysis 20
    3.5 Graph Representation 23
    3.5 Document Retrieval System 24
    3.6 Performance Evaluation Metrics 25

4. RESULTS 27
    4.1 Network Analysis Result 27
    4.2 Integration with Document Retrieval System 39

5. DISCUSSION AND SUGGESTION 41
    5.1 Conclusion 41
    5.1 Discussion 41
    5.2 Suggestion and Future Works 43
    5.3 Publication 43

CONTENTS (Cont’d)

PAGE

REFERENCES 44

CURRICULUM VITAE 47

LIST OF TABLES

TABLE **PAGE**

3.1 Settings of Document Retrieval System 24
4.1 Clustering metrics 33
4.2 Document Retrieval System Metrics 40

LIST OF FIGURES

FIGURE **PAGE**

2.1 An example of dataset used in PageRank calculation 4
2.2 A text mining workflow 6
2.3 Schematic representation of Bag-of-Words approach 8
2.4 An example of using Cosine similarity in finding similar documents 9
2.5 Steps in K-means clustering with K=2 10
2.6 A simple graph 11
2.7 Adjacency matrix representation of graph 12
2.8 Simple graph with different number of connected components 12
2.9 An example of induced graph 15
3.1 Block diagram of Proposed Methodology 16
3.2 A sample of retrieved documents and their information field 17
3.3 Raw Dataset 18
3.4 The process in isolating each document 18
3.5 Dataset after transformed into indicator variables 19
3.6 Dataset after tag documents with category 20
3.7 A sample of vectorized data 20
3.8 A distribution of number of legal articles referred of a judgement 21
3.9 Number of cases involved an article (top 10) 21
3.10 A distribution of case category (top 20) 22
3.11 A distribution of number of words in a judgement 22
3.12 Judgements that shared common legal article 23
3.13 Adjacency matrix 23
3.14 Experiment Setup of Document Retrieval System 24
3.15 An example of using Entropy in determining purity of group 25
3.16 An example of Modularity in measuring graph clustering performance 26
4.1 Network structure of the dataset 27
4.2 Distribution of number of nodes on each component 28
4.3 An example of outlying components that shared different legal articles 28
4.4 Main component 29
4.5 Eigenvalue 29
4.6 First 100 Eigenvalues 30
4.7 Cluster member distribution from Spectral Clustering using first *k* eigenvectors as a feature with K-means at different *k* 30
4.8 Cluster member distribution from Spectral Clustering using all eigenvectors as a feature with K-means at different *k* 31
4.9 Cluster member distribution from Louvain method 31
4.10 A network with colorized node according to their cluster 32

LIST OF FIGURES (Cont’d)

FIGURE **PAGE**

4.11 An induced network colorized according to their base cluster 33
4.12 Sample of documents in Cluster 0 34
4.13 Sample of documents in Cluster 1 34
4.14 Sample of documents in Cluster 2 35
4.15 Sample of documents in Cluster 3 35
4.16 Sample of documents in Cluster 4 36
4.17 Sample of documents in Cluster 5 36
4.18 Sample of documents in Cluster 6 36
4.19 Sample of documents in Cluster 7 37
4.20 Sample of documents in Cluster 8 37
4.21 Sample of documents in Cluster 9 38
4.22 Sample of documents in Cluster 10 38
4.23 Sample of documents in Cluster 11 39
4.24 Sample of documents in Cluster 12 39
5.1 A degree distribution of node in the network 41
5.2 A comparison of graph clustering method 42
5.3 An example of Hypergraph 42

LIST OF EQUATIONS

EQUATION **PAGE**

2.1 TF-IDF Term weighting 8
2.2 Cosine Similarity 9
2.3 Graph Laplacian 13
2.4 Modularity 14
3.1 Adjacency matrix 23
3.2 Entropy 25

CHAPTER 1 INTRODUCTION

1.1 Statement of Problem

In the past two decades, the Internet has become an important part of human's everyday life. A study shows that the number of people that has access to the World Wide Web has doubled every five years since 2000 [1]. Nowadays, there are over 4.66 billion active Internet users worldwide, which comprises up to 59.5% of the world population [2]. It is also expected that data generated per day is around 2.2 exbibytes (2.5 million terabytes) [3].

A text document is one category of data on the Internet that has dramatically increased with the expansion of the World Wide Web. It is expected that over 400 billion text-based communications occur per day, including blog posts, e-mails, instant message services, and news articles [4]. Those data can play an important role in addressing some business questions. For example, a company may gather the data that users talk about their products or services online and conduct research to identify the potential problem to their profit. They can also conduct marketing research to determine a potential new product that could address a current unsolved issue.

A common process to gain insight from textual information is called text mining. This consists of a process to retrieve documents, transform them into a feasible format for analytics, and use a natural language processing method to consider their internal structure and gather useful information.

However, a particular concern is the gigantic size of the data is slowing down the process of text mining. This is because the document needs to be transformed into a specific format for each analytical method beforehand. Thus, this step can take an exceptionally long period of time to process and/or heavily consumes computational resources.

To overcome the aforementioned limitation, a study on textual representation for data mining tried to extract meaningful information from a document in different aspects. As such, different types of documents may require specific ways to be processed. For example, schema rich documents like e-mail, a research paper, or program's source code require a certain step in the data extraction to parse textual data out of the structure. Some documents that have references to other documents may also have relationships between the documents that are meaningful and can be derived as document features.

There are many forms of documents with references that may be beneficial from analyzing the relationship in a corpus. Furthermore, a considerable amount of literature has been published on analyzing different types of documents with references. Gasevic et al. [5] proposed modeling the structure of research papers in the field of distance learning and online courses. One of the research topics in the study was to examine the network structure of documents through references. The authors were able to capture useful information, notably a "citation network" that was drawn from the citations in the papers. Using such network, the authors could obtain the most prominent work as well as how it affected other work in the related fields. In addition, the authors analyzed this network in the unit of people to locate the most influential researchers in this field of study.

A set of documents can be modeled in a network structure by considering the degree of similarity and relationship between the terms. Stanchev [6] [7] constructed a network structure of a news dataset by using term similarity drawn from WordNet called a similarity graph. The study reported higher precision and recall in using k-means clustering on the similarity graph than on simple keyword matching. The study suggested the potential benefits of using a network structure of documents to perform clustering.

In this current study, the researchers investigated one form of the most important documents with references in Thai society: legal documents. This type of documents has references to other documents based on the legal articles involved. For example, when a court rules on a case, the judges would have to refer to existing legal articles from the code of law related to the case. This study used the relationship to model a network structure and extract insights from the network.

The overall structure of the study took the form of five chapters. Chapter 1 introduces the problem, objective, and scope. Chapter 2 stages the literature survey with discussion on existing studies. Chapter 3 presents the research methodology, including data collection, transformation, network modeling techniques, and evaluation metrics. Chapter 4 presents the analysis of the results. Chapter 5 is the detailed conclusion and recommendations of this study.

1.2 Objectives

1. To study the network and topological structure of Thai legal documents.
2. To evaluate different methods in the network and topological analysis that would affect the performance of an information retrieval (IR) system.

1.3 Scopes

1. Dataset
    - The legal document dataset used in this study was the Supreme Court of Thailand’s judgments.
    - The dataset was obtained from an IR service of the Supreme Court of Thailand’s website [8].
    - The study considered only the essential part of the data.
    - There were no corrections or modifications on any text field of the data before the analysis.
2. Analysis method

    The analysis included the following methods.
    - Building an IR system
    - Document clustering
    - Graph representation of the network data
    - Connected component analysis

- Graph clustering
    - Spectral clustering
    - Louvain method for community detection
- Induced graph

1.4 Expected Benefits

1. The performance of the information retrieval system could be improved by integrating the network features into text features.
2. To discover the network and topological structure of Thai legal documents that may benefit further studies.

CHAPTER 2 LITERATURE REVIEW

2.1 Related Work

An effort to make computers understand the meaning of natural language has been made since the beginning of the Information Age. In 1957, Chomsky [9] characterized the concept of natural language processing (NLP) by stating that “To make a computer understand natural language, the sentence structure must have been changed.” This resulted in the work on natural language processing in the preceding decades by focusing on how to construct a set of grammatical rules that could be used for the transformation to natural language sentences. In the 1990s, there was a revolution with the introduction of computing intelligence with machine learning. As well as the growing volume of text data available for analysis, the World Wide Web appeared in a later decade.

At present, it is estimated that over 400 billion text-based communications occur per day [4]. With such gigantic size, the system that utilizes those data must be specifically designed to perform a certain task with that data. For example, to search for a particular document, a system called an information retrieval (IR) system is used. This is based on building an index of each document on a corpus (text dataset) and usually incorporates pure text features, which often lead to the process being dramatically slowed down for a very large corpus.

To overcome those aforementioned drawbacks, a number of studies on acquiring information from different aspects of the data have been conducted. This was because different types of documents may require a specific way of processing; for example, a document may contain a reference to others, or a document is under an explicit markup; such as, web pages [10].

A remarkable study of accompanying information beyond textual data was the PageRank [11]. The study developed an algorithm that incorporated the web page and links to other web pages. Figure 2.1 shows the structure of the dataset used in the study. The dataset consisted of roughly 150 million web pages and 1.7 billion links. Every page had a certain number of links to other pages as well as number of links to that page. The study constructed a graph and performed an infinite graph traversal and calculated the probability to arrive at each page. That probability was then used to rank the popularity of the web page. This study suggested that using information beyond textual data could be beneficial in some analyses. As a consequence, this led to a rise in the research about using links in the dataset as features in analysis.

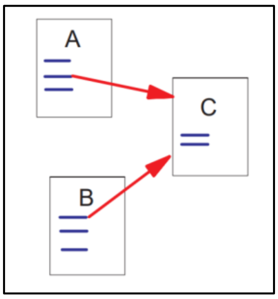


Figure 2.1 An example of a dataset used in a PageRank calculation [11].

As for improving the performance of the IR system, there has been an effort of incorporating several machine learning algorithms. van Rijsbergen [12] discussed the benefits of using a clustering algorithm to narrow down the search space of the IR system. The study discussed that because similar documents were grouped into the same cluster, the performance of the IR system was improved both in the retrieval performance (precision and recall) and runtime.

There is also some evidence that has suggested a document clustering algorithm could potentially benefit from the network features. Stanchev [6] [7] constructed a network of documents from a news article dataset. The relationship between the documents that defined the network in this study was obtained from WordNet, which is a collection of all the nouns, verbs, adverbs, and adjectives of the English language that are grouped into a set of synonyms. This provided semantic conceptual relationships of the words that could be used in modelling similarity. Hence, the network was called a similarity graph. The study used k-means clustering to group the documents based on the constructed similarity graph and simple keyword matching. According to the study, the approach with the similarity graph yielded a higher precision and recall compared to keyword matching.

Moreover, Ali and Melton [13] conducted a study that compared a semantic feature learning approach to clustering documents with the graph theory. The study built a semantic graph from a collection of documents and utilized a semantic reduction technique to maintain the primary topic and filter out the less important items. The study constructed a corpus graph representation of the dataset using a semantic graph from the dataset and ontology from Wordnet, and then performed clustering using the Louvain community detection algorithm. The results of the study reported that the graph approach produced lower entropy in each cluster compared to the latent semantic analysis approach. The study further suggested that the utilization of a network analysis algorithm in document clustering could possibly be useful.

Additionally, Yoo and Hu [14] performed a study on measuring the performance of clustering biomedical documents using several text features of a clustering algorithm; such as, k-means, bisecting k-means and CLUTO’s *v*cluster in comparison with the proposed Clustering Ontology-enriched Graph Representation. The study reported nearly twice an improvement in the performance of using the proposed method compared to a traditional method like k-means. This study also suggested the benefits of using network features in clustering.

There has been very little research about the computational study of Thai legal documents. Kowsrihawat and Vateekul [15] study utilized textual information from the verdict to build an IR system of a summarized verdict dataset. However, the study tended to focus on the textual features in building the IR system. However, the performance of the system could possibly be improved if the study utilized network features in the dataset as part of the information used in the system.

Previous studies involving document clustering with the accompanying network features were also mostly conducted using synthetic network features; for example, a feature from Wordnet or term similarity. This could lead to a decrease in the potential performance in some cases because synthetic network features could not guarantee a real-world or explainable relationship between each document. Moreover, only a few studies used a dataset with organic network features. Nevertheless, using this type of network features might improve the performance of the system because it would reflect a real-world relationship between the entities.

In this current study, the researchers divided the experiments into two parts: 1. Obtaining network features from the network analysis, and 2. building an IR system based on those features. The technical details of the technique used in each part are presented in the next section.

2.2 Theory

2.2.1 Text mining

Zanini and Dhawan [16] gave the definition of text mining as "A set of statistical and computer science techniques specifically developed to analyze text data." Text mining is also a broad definition of a collection of tasks required to achieve an insight from a text dataset. This also includes information retrieval, data mining, machine learning, statistics, and computational linguistics. In this current study, the researchers utilized the techniques in text mining. The following subsections provide a detailed explanation.

1. Text Mining Workflow

A generic process in text mining usually consists of the following steps [17]:

- Collecting text data from the source(s). These are usually available in various formats. For example: plain text, document file, or web pages.
- Preprocessing to remove the anomalies out of the data and transform them into an appropriate format for analysis.
- Analyze the text. The technique involved in this step would be different according to the goal of the process; for example, using a cosine similarity to find similar documents, or use a machine learning model to create a recommendation system.
- Store the extracted information in the knowledge base for further usage.

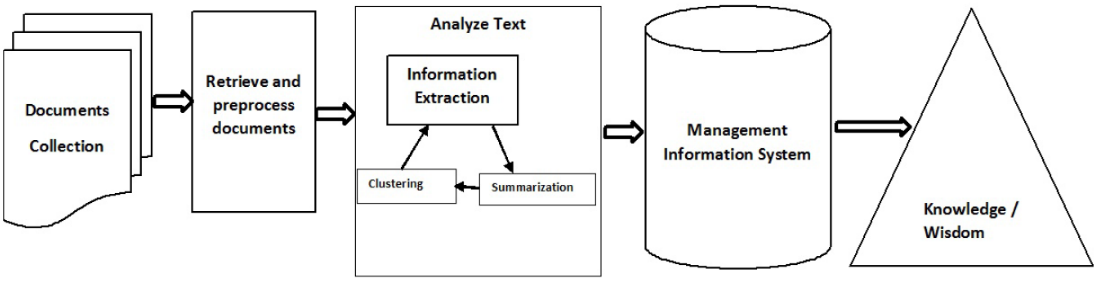


Figure 2.2 A text mining workflow [17]

The workflow could be varied according to the task and other unseen limitations.

2. Information Retrieval System

According to Manning et al. [10], an IR system is "Finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).” In general, an IR system often refers to a search engine for a particular dataset; for example, web pages, books, etc.

In addition, an IR system is designed to analyze, process, and store information sources in order to retrieve those that match a user's queries. Hence, the IR system usually consists of the following procedures [18]:

- Indexing:
  The system constructs a metadata or representation about each document to use in the search.
- Query formulation:
  The system creates a query representation of the user’s input.
- Searching:
  The system compares the user's query to the existing entity representation and returns the result.
- Feedback:
  Some systems have a feature called relevance feedback. It allows users to rate how relevant a result is to their query. The feedback data can then be utilized to improve the system's performance in the future.

2.2.2 Natural Language Processing

Bonaccorso [19] gave a definition of natural language processing as “A collection of machine learning techniques that allows the analysis of text documents, enables working with their internal structure, and results in word distribution.” Generally defined, this refers to a collection of algorithms and analytical techniques that make a computer capable of understanding natural language data. Several natural language processing techniques were implemented in this current study. The following subsections provide a detailed explanation of each technique.

1. Representation of Text in Vector Space

Because most machine learning and analytical methods can only operate on a numeric vector space, text documents must be transformed into a vector representation to enable the numeric operation. The mechanism of encoding text documents in a numeric vector space is called vectorization, which is a vital step toward natural language research [20].

The simplest and widely adopted technique of vectorization is called a bag-of-words. The basic assumption is that context, meaning, and similarity of documents are encoded within words, and the order of each single word in a document is unimportant. A common pipeline in the bag-of-words consists of the following steps.

- Tokenization:
  Each document is divided into the smallest meaningful units: a word, normally called a token.
- Stopwords removal:
  Some words may occur frequently and do not contribute to useful semantic information, so it is a good practice to remove those words.

- Normalization:
  Some words may occur in many forms but have the same meaning. In some cases. it is better to convert the variation back to the base form to reduce the dimensionality of the data.
- Building a vocabulary table:
  All tokens occurring in the dataset are used in constructing a vocabulary table.
- Vectorize text according to the vocabulary table:
  - Count Vectorizer:
    Representing a token considering how many times it appears in a document.
  - TF-IDF Vectorizer:
    The term frequency-inverse document frequency (TF-IDF) weighs the importance of the token by scaling down the weight of common terms and emphasizing rare terms that occur among multiple documents. This is displayed in the following equation.

$$tf \cdot idf(d,t) = \log\,(tf_{d,t}) \times \log\frac{N}{df_t} \tag{2.1}$$

Where:

$tf_{d,t}$	indicates the frequency of term t in document d
N	indicates the number of documents in dataset
df_t	indicates the number of documents that contains term t

Note that in this equation, $tf_{d,t}$ and df_t could not be zero, as the logarithm of zero and division by zero would be undefined. A common approach in mitigating the problem would be to add one to both terms to ensure that the logarithm and division by zero would not occur.

In this current study, the researchers used a Python's library for Thai natural language processing called PyThaiNLP [21] to perform the vectorization process. Tokenization was done by using its word tokenizer module with maximum matching algorithms. A Thai stopwords was also obtained by the library and used to filter out the noninformative token.

To emphasize only important features, as well as remove possible mistokenized tokens, the researchers only considered the tokens with the top 20,000 frequencies occurring in the dataset. Overall, the process of the bag-of-words approach could be visually described as shown in Figure 2.3.

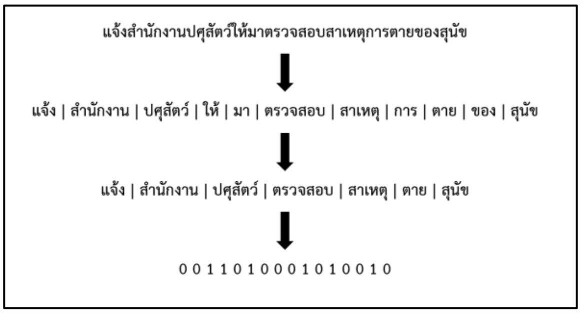

แจ้งสำนักงานปศุสัตว์ให้มาตรวจสอบสาเหตุการตายของสุนัข
แจ้ง | สำนักงาน | ปศุสัตว์ | ให้ | มา | ตรวจสอบ | สาเหตุ | การ | ตาย | ของ | สุนัข
แจ้ง | สำนักงาน | ปศุสัตว์ | ตรวจสอบ | สาเหตุ | ตาย | สุนัข
0 0 1 1 0 1 0 0 0 1 0 1 0 0 1 0

Figure 2.3 Schematic representation of Bag-of-Words approach

After the vectorization process, the text documents were in a numeric vector representation and could be used in machine learning or other analytical processes.

2. Cosine Similarity

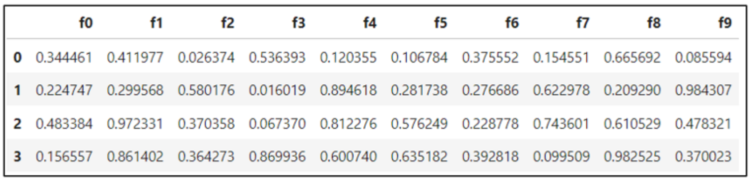
Cosine similarity is the measurement of the similarity between two non-zero vectors that would utilize the cosine of the angle between them [22]. A cosine ranges from -1 to 1. The value closer to 1 would indicate that the vector is going in the same direction and thus similar. The value closer to -1 would indicate that the vector is going in the opposite direction, whereas the value closer to 0 would indicate that the vector pair is orthogonal. In IR, cosine similarity ranges between 0 and 1 because the term frequencies cannot be negative. A cosine similarity can be defined by the following equation.

$$sim(\vec{A},\vec{B}) = \frac{\vec{A}\cdot\vec{B}}{\|\vec{A}\|\|\vec{B}\|} \tag{2.2}$$

Where:

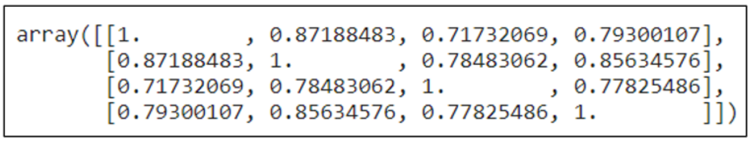
$\vec{A}\cdot\vec{B}$	indicates Euclidian dot product between vector $\vec{A}$ and $\vec{B}$
$\|\vec{A}\|$ and $\|\vec{B}\|$	indicates the norm of vector $\vec{A}$ and $\vec{B}$ respectively

Figure 2.4 shows the example of using cosine similarity in finding similar documents. If each document in the collection could be vectorized using the technique discussed in the previous section, then a cosine similarity of each pair of documents could be computed to determine the similarity between them. If it was necessary to find the document that was the most similar to document 0. From the calculation, document 1 was the most similar because document 0 had the highest similarity score with document 1 (excluding self).



	f0	f1	f2	f3	f4	f5	f6	f7	f8	f9
0	0.344461	0.411977	0.026374	0.536393	0.120355	0.106784	0.375552	0.154551	0.665692	0.085594
1	0.224747	0.299568	0.580176	0.016019	0.894618	0.281738	0.276686	0.622978	0.209290	0.984307
2	0.483384	0.972331	0.370358	0.067370	0.812276	0.576249	0.228778	0.743601	0.610529	0.478321
3	0.156557	0.861402	0.364273	0.869936	0.600740	0.635182	0.392818	0.099509	0.982525	0.370023

(a)



```
array([[1.        , 0.87188483, 0.71732069, 0.79300107],
       [0.87188483, 1.        , 0.78483062, 0.85634576],
       [0.71732069, 0.78483062, 1.        , 0.77825486],
       [0.79300107, 0.85634576, 0.77825486, 1.        ]])
```

(b)

Figure 2.4 An example of using Cosine similarity in finding similar documents
(a) Vectorized documents
(b) Pairwise Cosine similarity

By utilizing cosine similarity, the researchers could quantify the similarity between the documents and use it as the indicator on retrieving a document that was similar to the user's queries or other documents of interest.

3. Document Clustering

According to Manning et al. [23], clustering is an algorithm that groups a set of documents or entities for considering the subsets or clusters, where a cluster is coherent and contains similar entities internally and clearly different from other clusters. Clustering is the most common form of unsupervised learning tasks. In this current study, the researchers incorporated well-known forms of clustering called k-means clustering.

K-means clustering divides data into k groups based on their centroids. A centroid is the average value of the observations on the vector space in the cluster. K-means clustering is an iterative process that can be divided into two steps. The first step is to assign every observation to the clusters based on its nearest centroid. The second step is to recompute the value of the centroids based on recently assigned observations. The process is repeated until it reaches a stable state or the stop condition is met [24]. The algorithm is described as follows:

Algorithm 1: K-means clustering

```
Required: set of observation: X, number of clusters: k
Initialize: a random set of centroids $m_1, m_2, m_3, \dots, m_k$
While: stop condition is not met
    Assign each observation X to the cluster of nearest centroid
    Recalculate centroids based on mean of observation assigned to each cluster
Return: Cluster label for each observation
```

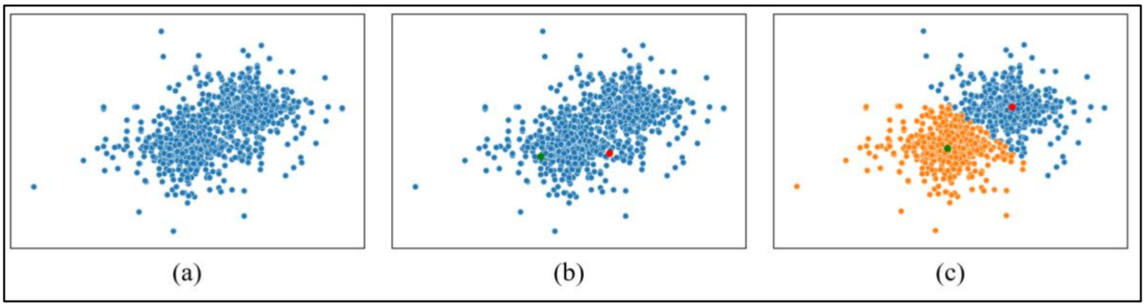

(a)
(b)
(c)

Figure 2.5 Steps in K-means clustering with K=2
(a) Initial dataset
(b) A randomly assigned centroids in the first iteration
(c) Centroids after assigning/updating for several iterations

In an ordinary k-means clustering, a stop condition would be often used to avoid further computation that would not improve the clustering. The condition would usually involve a tolerance limit ε, which would be a small number that would be used to compare the difference in the centroids of two consecutive iterations. If the difference was more than the tolerance limit, the computation could proceed to the next iteration. Otherwise, it could be stated that the cluster would be converged, and no further computation would be required. In some cases, a stop condition could only be the number of the iteration limit.

In an information retrieval system, adopting document clustering resulted in better performance in terms of the runtime and precision/recall. This was because similar documents could be assigned to the same cluster and tend to be relevant to the user's queries.

In this study, the researchers employed k-means clustering on the connectivity matrix of the dataset, which described the connection between the entities. As a result, the dataset was clustered based on how each entity was connected.

2.2.3 Network Science and Network Analysis

Barabasi [25] defined a network as "A collection of system's components (nodes or vertices) and a direct interaction between them (edges or links)". According to this definition, a complex system that could be represented as a form of interaction between the entities could be modeled as a network. For example, a collaboration in academic research, E-mail, phone calls, or links in web pages.

The term "network science" can be broadly defined as a study of network representation of real-world phenomena that would lead to an insight or predictive model [26]. In general, this would refer to the study of the interaction between the entities in the network and how each entity would affect other items. According to Börner et al. [27], an approach in network science can be divided into two categories: 1. network analysis and 2. network modeling. Network analysis aims for the generation of a descriptive model that would explain the system, while network modeling focuses on designing the process model that could reproduce and predict the later state of the system.

The researchers used network analysis techniques in this current study. The following sections are a detailed description of each technique.

1. Graph representation

A common form of representing interaction between entities is called a graph. An arbitrary graph is defined by a pair of sets $G = (V, E)$ where V is non-empty set of entities called nodes or vertices and E is a set of pairs of nodes called edges or links [27]. Figure 2.6 depicted a simple graph.

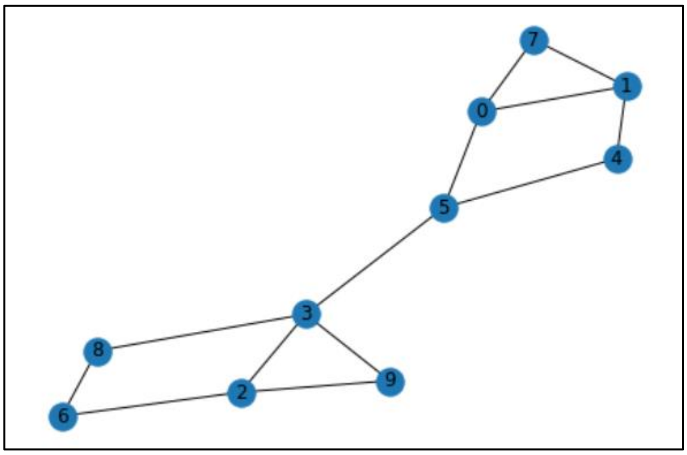
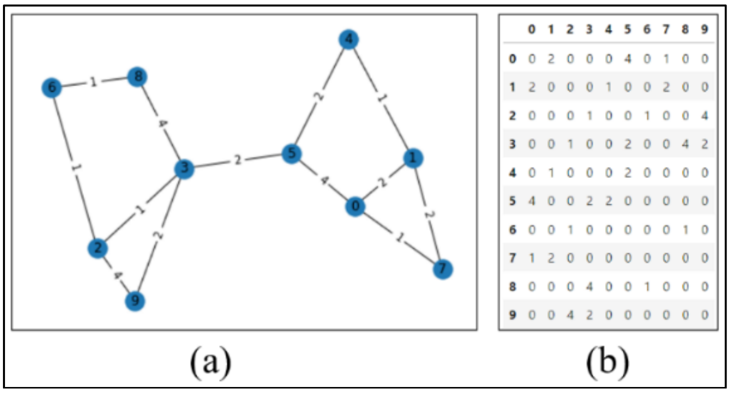


Figure 2.6 A simple graph contains 9 vertices and 12 edges

A graph can be weighted or unweighted. A weighted graph is a graph where the edges have been given a numerical weight, which could be used to indicate the importance of the link or connection. Moreover, a graph can be either directed or undirected; for example, if a particular vertex i has a link to vertex j but not vice versa. An edge $(i,\ j)$ is called a directed edge and would be represented in a graph as a line with an arrowhead that would indicate the direction. In this current study, the researchers modeled the network using a weighted undirected graph.

In this study, the researcher modeled the network using weighted undirected graph.

From a mathematical perspective, it would be more convenient to define a graph using an adjacency matrix, which is a square matrix of size N that each element x_{ij} in the matrix would represent the connection between the vertices i and j. In a weighted graph representation, x_{ij} would be the weight of the edge connecting the vertices i and j, whereas in an undirected graph, the adjacency matrix would be symmetric (Figure 2.7).



	0	1	2	3	4	5	6	7	8	9
0	0	2	0	0	0	4	0	1	0	0
1	2	0	0	0	1	0	0	2	0	0
2	0	0	0	1	0	0	1	0	0	4
3	0	0	1	0	0	2	0	0	4	2
4	0	1	0	0	0	2	0	0	0	0
5	4	0	0	2	2	0	0	0	0	0
6	0	0	1	0	0	0	0	0	1	0
7	1	2	0	0	0	0	0	0	0	0
8	0	0	0	4	0	0	1	0	0	0
9	0	0	4	2	0	0	0	0	0	0

Figure 2.7 Adjacency matrix representation of graph
(a) A simple weighted undirected graph
(b) Adjacency matrix of graph in (a)

2. Connected Component

In an arbitrary undirected graph, a connected component would be the number of maximal subgraphs that any pair of vertices could be traversed by a path [28]. A connected component analysis would enable a finding of the already existing clusters in the graph according to how each vertices would be connected. Figure 2.8 illustrates a graph with a different number of connected components.

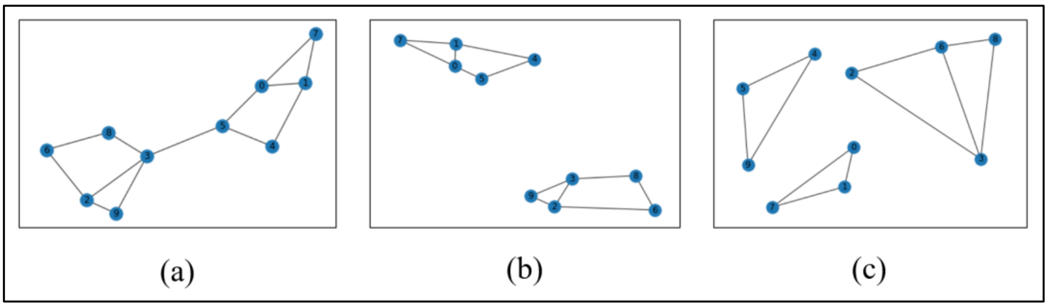


Figure 2.8 Simple graph with different number of connected components
(a) 1 component
(b) 2 components
(c) 3 components

In this study, the researchers used a connected component analysis as an initial analysis of the dataset. This enabled finding the existing cluster as well as locate the entities that were not connected to other entities.

3. Graph Clustering

To inspect the hidden components in the graph that could not be obtained using a connected component analysis, a graph clustering analysis was employed to study how each entity formed a connection in the graph. In this current study, the researchers used two graph clustering methods: 1. Spectral clustering and 2. Louvain method for community detection. A detailed explanation on each method is as follows:

- Spectral Clustering

  This clustering technique is used to identify the clusters in a graph based on density. The method applies a spectrum (Eigendecomposition) to the graph Laplacian to perform the dimensionality reduction and use the spectrum to do the clustering process [29].

  A graph Laplacian is defined by the subtraction of degree matrix and adjacency matrix.

  $$L = D - W \tag{2.3}$$

  Where:

  L indicates a graph Laplacian
  D indicates a diagonal degree matrix
  W indicates an adjacency matrix

  Eigendecomposition is then applied to the graph Laplacian. The resulting eigenvalue is then sorted ascendingly. If the smallest eigenvalue is 0, this would reflect that the graph would have one connected component. If not, then the graph would have more than one component. In this study, the researchers ensured that the graph that was used in the clustering comprised only one component by taking the connected component analysis first. This only used a certain component in the graph clustering.

  The corresponding eigenvector to the smallest eigenvalue would be a constant vector. The eigenvector corresponding to the eigenvalue after the first one would be the one that spectral clustering would be considered. The first non-constant eigenvector is called a Fiedler's eigenvector. This vector could be used to determine the cluster of the data points by looking at the sign of the components. The element with a positive value would be in one cluster, and the negative value would be in another.

  If taking two eigenvectors that corresponded to the two smallest and non-zero eigenvalues into consideration, the cluster points would become three clusters. The element with both eigenvectors that would be positive would be in cluster one. On the other hand, the element with only one positive value would be in cluster two, and the element with both negative values would be in cluster three. From this procedure, the first k eigenvector could be chosen to cluster the data into k clusters according to the value of the element of each eigenvector.

In this study, the researchers used the eigenvector that corresponded to the non-zero eigenvalue as a feature of k-means clustering. Using this method, the vertices could be clustered in the graph to become an arbitrary k cluster based on the connection between each vertex.

- Louvain method for community detection
  Blondel et al. [30] developed an algorithm in an unfolding hidden community in a large network structure called a Louvain method. The algorithm was a heuristic method that was based on modularity optimization. The modularity of the graph is defined as follows [31]:

$$Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta\left(c_i, c_j\right) \tag{2.4}$$

Where:

Q	indicates modularity of network
A_{ij}	indicates the edge weight between node i and j
$k_i,\ k_j$	indicates sum of the edge's weight attached to node i and j
m	indicates sum of all edge's weight in the graph
c_i, c_j	indicates cluster or community of node i and j respectively
δ	indicates a function that yields1 if $i = j$, 0 otherwise

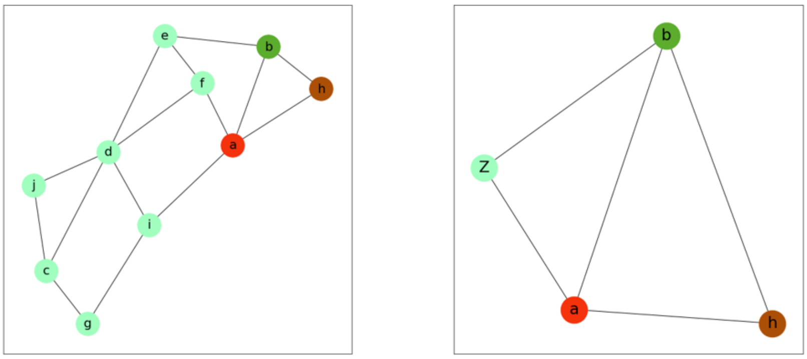
Modularity is a measure of the strength of in-cluster connections in comparison to overall network connections. The value range of modularity would be [-0.5, 1] with -0.5 indicating that a graph would be fully connected and could not be separated by any cut strategy. The value 1 would indicate that the connectivity within a cluster would be significantly stronger than between the clusters.

The Louvain method aimed to optimize this value by first assigning each node to its own cluster, then merging into the group of nodes that resulted in the greatest improvement in modularity, followed by merging iteratively until no such merger that could improve the modularity could occur.

In this study, the researchers chose a graph clustering method that focused on the different aspects of the connection. The spectral clustering focused on the connectivity of each node, and the Louvain method tried to maximize the modularity.

4. Induced Graph

According to Diestel [32], given an arbitrary graph $G = (V, E)$, an induced graph or induced subgraph is a graph composed of a subset of vertices of a graph G and edges that would have both endpoints in those vertices called S. Any two pairs of vertices in S would be adjacent in this induced graph if and only if they were adjacent in G. Figure 2.9 shows example of graph and its induced form.



(a) (b)

Figure 2.9 An example of induced graph
(a) Original graph
(b) Induced graph

In Figure 2.9, the original graph consisted of nine vertices. The induced graph of this graph was constructed by partially setting the vertices. In this example, the vertices were divided into $\{\{a\},\{b\},\{h\},\{c,d,e,g,g,i,j\}\}$. The induced vertex Z comprised $\{c,\ d,\ e,\ g,\ g,\ i,\ j\}$ as well as the edges that had both termini in those vertices. The edges (a,Z) and (b,Z) were from the definition that a vertices pair would be adjacent in the induced graph if they were adjacent in the original graph. In this case, a would be adjacent to Z because a would be adjacent to f and i in the original graph, as well as b would be adjacent to e, thus producing the edge (b,Z).

In this study, an induced graph was used for analyzing the structure of the cluster from the graph clustering process. Each cluster could be induced to a set of vertices that represented the clusters. As such, the connectivity of each cluster could be observed as whether they were fully connected, or there were special relationships between them.

In the next section, we present the methodology and overall process before the analysis. As well as performance measuring metrics.

CHAPTER 3 METHODOLOGY

This section presents the methodology of the study, dataset, data preprocessing, and performance evaluation metrics.

3.1 Proposed Methodology

As stated in Section 1.2, the purpose of this study was to examine the effect of different methods of the network analysis and build a document retrieval system based on that network.

As such, an experiment was organized into two parts to meet the intended objective. Figure 3.1 displays the proposed methodology.

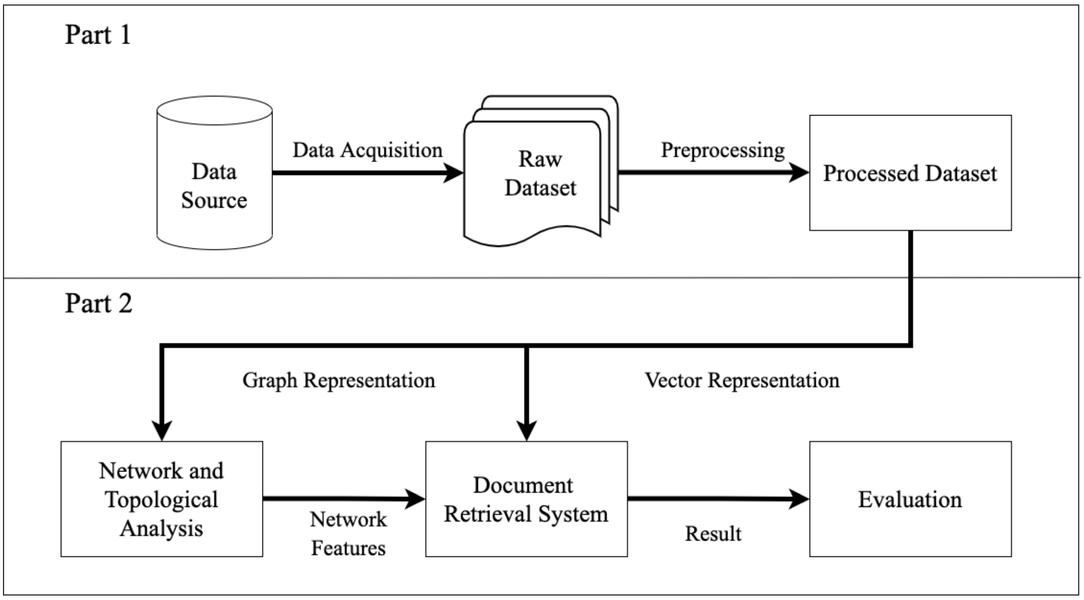


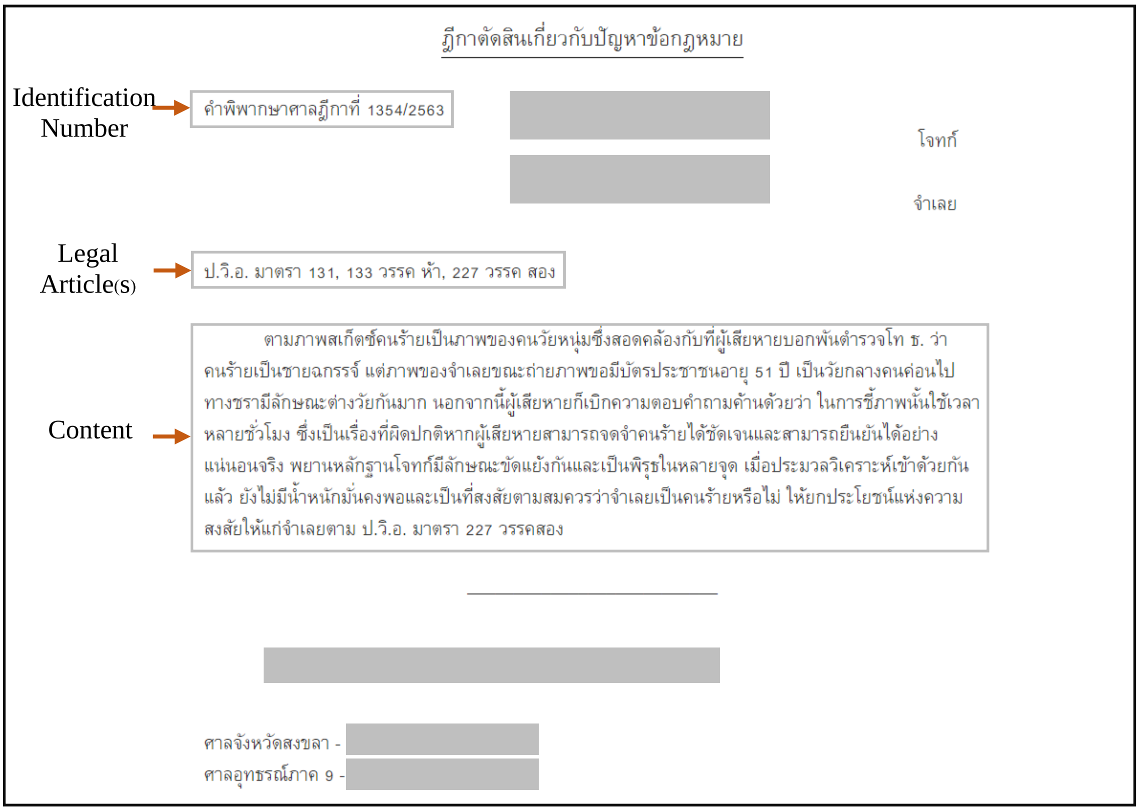
Figure 3.1 Block diagram of Proposed Methodology

From Figure 3.1, the first part of the study was to collect and preprocess the dataset. In this part, the data were gathered from the respective sources and transformed into the appropriate format for each type of analysis. For the network analysis, the data were transformed into an adjacency matrix for the graph representation as mentioned in Section 2.2.3. For the document retrieval system, a text vectorizer was used to transform the data into the vector space as stated in Section 2.2.2.

The second part of the study was to perform the analysis on the processed data. The analysis method used in this study included a graph analysis and construction of the document retrieval system. The results of each analysis were then evaluated using evaluation metrics.

3.2 Dataset

There were many types of legal documents that could be modeled and analyzed using the network analysis. In this current study, the researchers focused on the Supreme Court of Thailand’s judgments. The data were available on an information retrieval service of the Supreme Court of Thailand’s website [8], which contained 129,402 records of judgments from 1920-2020. However, for this study, only the data for the period of 2010-2019, which consisted of 7,090 documents were utilized. The data were obtained by conducting a web crawler via Python’s programming language. Figure 3.2 shows the sample of the documents obtained and their information field that were used in the study.


Identification Number
Legal Article(s)
Content

ฎีกาตัดสินเกี่ยวกับปัญหาข้อกฎหมาย

คำพิพากษาศาลฎีกาที่ 1354/2563

[redacted] โจทก์

[redacted] จำเลย

ป.วิ.อ. มาตรา 131, 133 วรรค ห้า, 227 วรรค สอง

ตามภาพสเก็ตซ์คนร้ายเป็นภาพของคนวัยหนุ่มซึ่งสอดคล้องกับที่ผู้เสียหายบอกพันตำรวจโท ธ. ว่าคนร้ายเป็นชายฉกรรจ์ แต่ภาพของจำเลยขณะถ่ายภาพขอมีบัตรประชาชนอายุ 51 ปี เป็นวัยกลางคนค่อนไปทางชรามีลักษณะต่างวัยกันมาก นอกจากนี้ผู้เสียหายก็เบิกความตอบคำถามค้านด้วยว่า ในการชี้ภาพนั้นใช้เวลาหลายชั่วโมง ซึ่งเป็นเรื่องที่ผิดปกติหากผู้เสียหายสามารถจดจำคนร้ายได้ชัดเจนและสามารถยืนยันได้อย่างแน่นอนจริง พยานหลักฐานโจทก์มีลักษณะขัดแย้งกันและเป็นพิรุธในหลายจุด เมื่อประมวลวิเคราะห์เข้าด้วยกันแล้ว ยังไม่มีน้ำหนักมั่นคงพอและเป็นที่สงสัยตามสมควรว่าจำเลยเป็นคนร้ายหรือไม่ ให้ยกประโยชน์แห่งความสงสัยให้แก่จำเลยตาม ป.วิ.อ. มาตรา 227 วรรคสอง

[redacted]

ศาลจังหวัดสงขลา - [redacted]
ศาลอุทธรณ์ภาค 9 - [redacted]

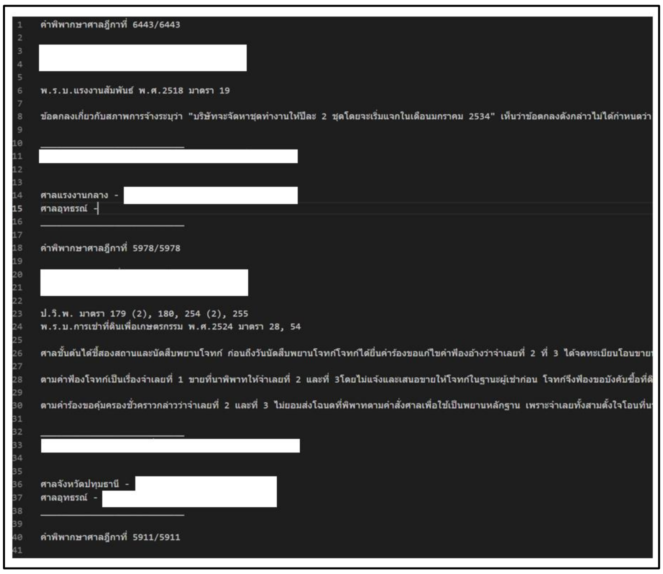
Figure 3.2 A sample of retrieved documents and their information field

3.3 Data Preprocessing

The data used in the study were obtained in a raw text format that consisted of all the crawled documents in a single file. Preprocessing was needed to isolate one document from the others, as well as separate the information field inside the document. The process comprised four steps as follows.

1. Isolate each documents

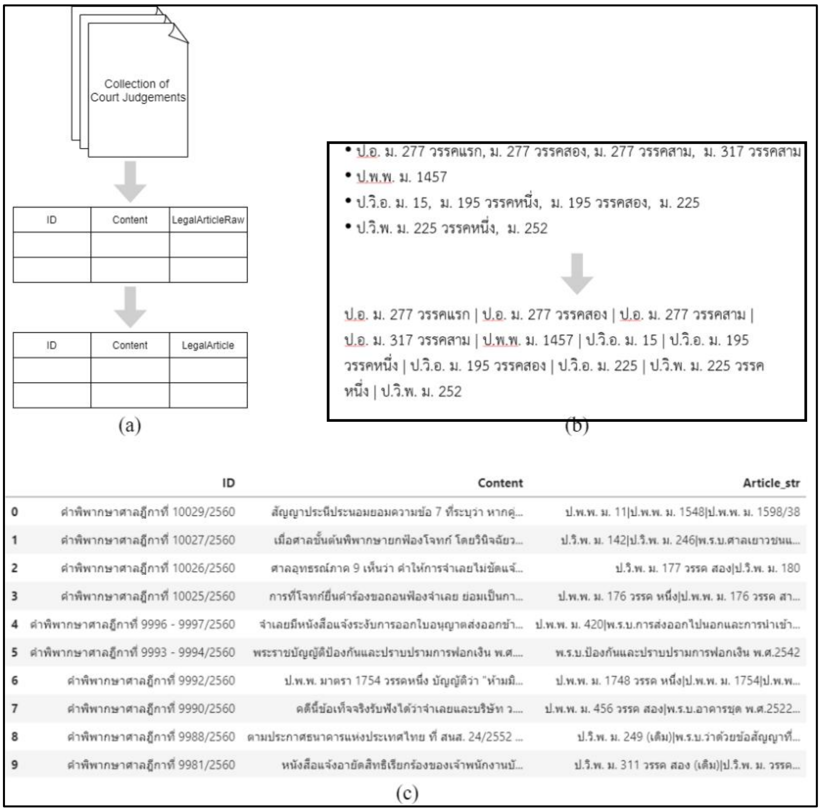
In this step, a raw text result from the crawler was fed into a Python script that separated each document and extracted only the information that would be used in the study of each document into a row in a tabular file. Figure 3.3 shows the sample of raw dataset.



```
คำพิพากษาศาลฎีกาที่ 6443/6443

[redacted]

พ.ร.บ.แรงงานสัมพันธ์ พ.ศ.2518 มาตรา 19

ข้อตกลงเกี่ยวกับสภาพการจ้างระบุว่า "บริษัทจะจัดหาชุดทำงานให้ปีละ 2 ชุดโดยจะเริ่มแจกในเดือนมกราคม 2534" เห็นว่าข้อตกลงดังกล่าวไม่ได้กำหนดว่า

______________________________
[redacted]


ศาลแรงงานกลาง - [redacted]
ศาลอุทธรณ์ -
______________________________

คำพิพากษาศาลฎีกาที่ 5978/5978

[redacted]

ป.วิ.พ. มาตรา 179 (2), 180, 254 (2), 255
พ.ร.บ.การเช่าที่ดินเพื่อเกษตรกรรม พ.ศ.2524 มาตรา 28, 54

ศาลชั้นต้นได้ชี้สองสถานและนัดสืบพยานโจทก์ ก่อนถึงวันนัดสืบพยานโจทก์โจทก์ได้ยื่นคำร้องขอแก้ไขคำฟ้องอ้างว่าจำเลยที่ 2 ที่ 3 ได้จดทะเบียนโอนขาย

ตามคำฟ้องโจทก์เป็นเรื่องจำเลยที่ 1 ขายที่นาพิพาทให้จำเลยที่ 2 และที่ 3โดยไม่แจ้งและเสนอขายให้โจทก์ในฐานะผู้เช่าก่อน โจทก์จึงฟ้องขอบังคับซื้อที่ดิ

ตามคำร้องขอคุ้มครองชั่วคราวกล่าวว่าจำเลยที่ 2 และที่ 3 ไม่ยอมส่งโฉนดที่พิพาทตามคำสั่งศาลเพื่อใช้เป็นพยานหลักฐาน เพราะจำเลยทั้งสามตั้งใจโอนที่น

______________________________
[redacted]


ศาลจังหวัดปทุมธานี - [redacted]
ศาลอุทธรณ์ -
______________________________

คำพิพากษาศาลฎีกาที่ 5911/5911

```

Figure 3.3 Raw Dataset

The field "Legal Articles" on the raw dataset was in a combined and shortened format. If a legal article came from the same code of law, it would be written together under the same topic. A transformation was required in order to divide each legal article into the proper name for use in the network analysis. Figure 3.4 (a-c) depicted the process and result.



- ป.อ. ม. 277 วรรคแรก, ม. 277 วรรคสอง, ม. 277 วรรคสาม, ม. 317 วรรคสาม
- ป.พ.พ. ม. 1457
- ป.วิ.อ. ม. 15, ม. 195 วรรคหนึ่ง, ม. 195 วรรคสอง, ม. 225
- ป.วิ.พ. ม. 225 วรรคหนึ่ง, ม. 252

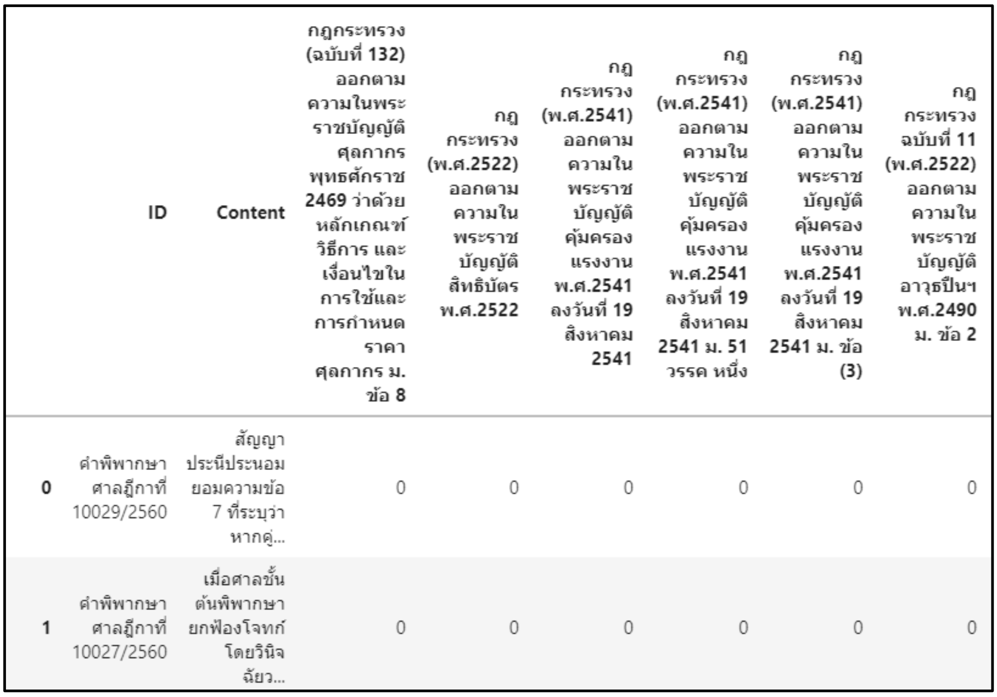
ป.อ. ม. 277 วรรคแรก | ป.อ. ม. 277 วรรคสอง | ป.อ. ม. 277 วรรคสาม | ป.อ. ม. 317 วรรคสาม | ป.พ.พ. ม. 1457 | ป.วิ.อ. ม. 15 | ป.วิ.อ. ม. 195 วรรคหนึ่ง | ป.วิ.อ. ม. 195 วรรคสอง | ป.วิ.อ. ม. 225 | ป.วิ.พ. ม. 225 วรรคหนึ่ง | ป.วิ.พ. ม. 252

	ID	Content	Article_str
0	คำพิพากษาศาลฎีกาที่ 10029/2560	สัญญาประนีประนอมยอมความข้อ 7 ที่ระบุว่า หากคู่...	ป.พ.พ. ม. 11|ป.พ.พ. ม. 1548|ป.พ.พ. ม. 1598/38
1	คำพิพากษาศาลฎีกาที่ 10027/2560	เมื่อศาลชั้นต้นพิพากษายกฟ้องโจทก์ โดยวินิจฉัยว...	ป.วิ.พ. ม. 142|ป.วิ.พ. ม. 246|พ.ร.บ.ศาลเยาวชนแ...
2	คำพิพากษาศาลฎีกาที่ 10026/2560	ศาลอุทธรณ์ภาค 9 เห็นว่า คำให้การจำเลยไม่ชัดแจ้...	ป.วิ.พ. ม. 177 วรรค สอง|ป.วิ.พ. ม. 180
3	คำพิพากษาศาลฎีกาที่ 10025/2560	การที่โจทก์ยื่นคำร้องขอถอนฟ้องจำเลย ย่อมเป็นกา...	ป.พ.พ. ม. 176 วรรค หนึ่ง|ป.พ.พ. ม. 176 วรรค สา...
4	คำพิพากษาศาลฎีกาที่ 9996 - 9997/2560	จำเลยมีหนังสือแจ้งระงับการออกใบอนุญาตส่งออกข้า...	ป.พ.พ. ม. 420|พ.ร.บ.การส่งออกไปนอกและการนำเข้า...
5	คำพิพากษาศาลฎีกาที่ 9993 - 9994/2560	พระราชบัญญัติป้องกันและปราบปรามการฟอกเงิน พ.ศ....	พ.ร.บ.ป้องกันและปราบปรามการฟอกเงิน พ.ศ.2542
6	คำพิพากษาศาลฎีกาที่ 9992/2560	ป.พ.พ. มาตรา 1754 วรรคหนึ่ง บัญญัติว่า "ห้ามมิ...	ป.พ.พ. ม. 1748 วรรค หนึ่ง|ป.พ.พ. ม. 1754|ป.พ.พ...
7	คำพิพากษาศาลฎีกาที่ 9990/2560	คดีนี้ข้อเท็จจริงรับฟังได้ว่าจำเลยและบริษัท ว....	ป.พ.พ. ม. 456 วรรค สอง|พ.ร.บ.อาคารชุด พ.ศ.2522...
8	คำพิพากษาศาลฎีกาที่ 9988/2560	ตามประกาศธนาคารแห่งประเทศไทย ที่ สนส. 24/2552 ...	ป.วิ.พ. ม. 249 (เดิม)|พ.ร.บ.ว่าด้วยข้อสัญญาที่...
9	คำพิพากษาศาลฎีกาที่ 9981/2560	หนังสือแจ้งอายัดสิทธิเรียกร้องของเจ้าพนักงานบั...	ป.วิ.พ. ม. 311 วรรค สอง (เดิม)|ป.วิ.พ. ม. วรรค...

Figure 3.4 The process in isolating each document
(a) Overall process
(b) Legal article field transformation
(c) Result

2. Transform Legal Article field into indicator variable

In statistics, an indicator variable is a variable with only two possible values that indicates the presence or absence of a property in each entry [33]. In this analysis, an indicator variable described which legal articles that each judgment was involved. These properties could be later used to determine the connection between the common judgments by the legal article. In graph analysis, this representation is called an incidence matrix, which shows the relationship between the vertices and edges. This matrix could be transformed into an adjacency matrix by multiplying its transpose, which was mainly used in this analysis. Figure 3.5 shows the result of the transformation.



	ID	Content	กฎกระทรวง (ฉบับที่ 132) ออกตามความในพระราชบัญญัติศุลกากร พุทธศักราช 2469 ว่าด้วยหลักเกณฑ์วิธีการ และเงื่อนไขในการใช้และการกำหนดราคาศุลกากร ม. ข้อ 8	กฎกระทรวง (พ.ศ.2522) ออกตามความในพระราชบัญญัติสิทธิบัตร พ.ศ.2522	กฎกระทรวง (พ.ศ.2541) ออกตามความในพระราชบัญญัติคุ้มครองแรงงาน พ.ศ.2541 ลงวันที่ 19 สิงหาคม 2541	กฎกระทรวง (พ.ศ.2541) ออกตามความในพระราชบัญญัติคุ้มครองแรงงาน พ.ศ.2541 ลงวันที่ 19 สิงหาคม 2541 ม. 51 วรรค หนึ่ง	กฎกระทรวง (พ.ศ.2541) ออกตามความในพระราชบัญญัติคุ้มครองแรงงาน พ.ศ.2541 ลงวันที่ 19 สิงหาคม 2541 ม. ข้อ (3)	กฎกระทรวง ฉบับที่ 11 (พ.ศ.2522) ออกตามความในพระราชบัญญัติอาวุธปืนฯ พ.ศ.2490 ม. ข้อ 2
0	คำพิพากษาศาลฎีกาที่ 10029/2560	สัญญาประนีประนอมยอมความข้อ 7 ที่ระบุว่าหากคู่...	0	0	0	0	0	0
1	คำพิพากษาศาลฎีกาที่ 10027/2560	เมื่อศาลชั้นต้นพิพากษายกฟ้องโจทก์โดยวินิจฉัยว...	0	0	0	0	0	0

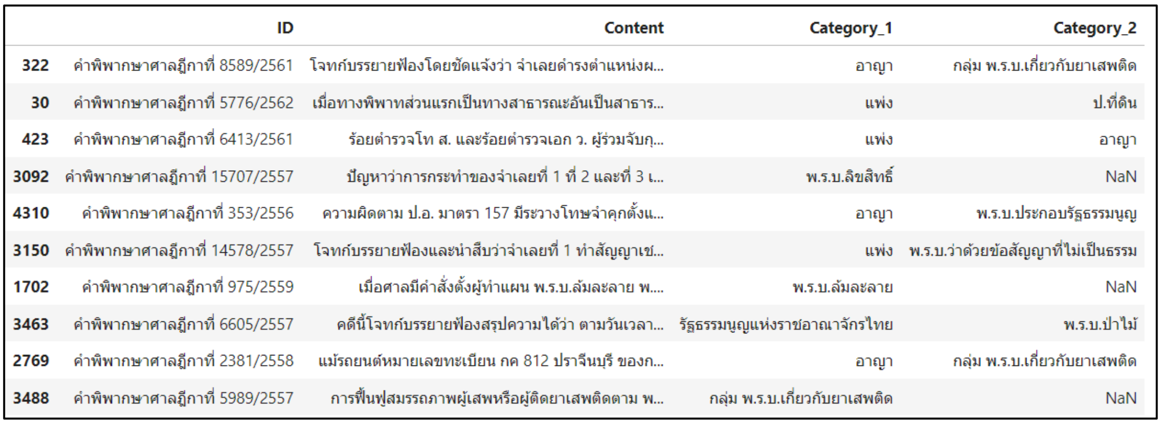
Figure 3.5 Dataset after transformed into indicator variables

3. Category tagging of documents

To enables the use of some metrics in the study, each document was assigned with a predetermined label called a “category”. In this current study, the category of the documents was determined by the types of legal articles involved with the case; for example, if the case involved a legal article from the “Civil and Commercial Code”, its category would be annotated as “civil”.

Furthermore, a particular case could involve a legal article from more than one code of law; for example, a case may involve an article from the “Civil and Commercial Code” and “Code of Revenue”. To address this issue, a “subcategory” was added to allow a document to be in more than one category.

The annotation process was done by running a script that considered the amount of legal articles of each type that was involved in the case, and this applied a category tag accordingly. Figure 3.6 shows the sample of a dataset with a category.

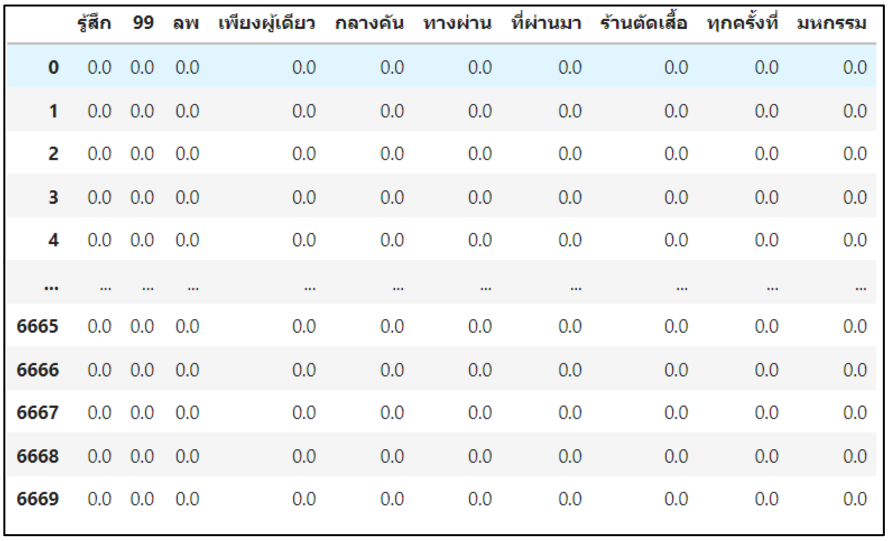


	ID	Content	Category_1	Category_2
322	คำพิพากษาศาลฎีกาที่ 8589/2561	โจทก์บรรยายฟ้องโดยชัดแจ้งว่า จำเลยดำรงตำแหน่งผ...	อาญา	กลุ่ม พ.ร.บ.เกี่ยวกับยาเสพติด
30	คำพิพากษาศาลฎีกาที่ 5776/2562	เมื่อทางพิพาทส่วนแรกเป็นทางสาธารณะอันเป็นสาธาร...	แพ่ง	ป.ที่ดิน
423	คำพิพากษาศาลฎีกาที่ 6413/2561	ร้อยตำรวจโท ส. และร้อยตำรวจเอก ว. ผู้ร่วมจับกุ...	แพ่ง	อาญา
3092	คำพิพากษาศาลฎีกาที่ 15707/2557	ปัญหาว่าการกระทำของจำเลยที่ 1 ที่ 2 และที่ 3 เ...	พ.ร.บ.ลิขสิทธิ์	NaN
4310	คำพิพากษาศาลฎีกาที่ 353/2556	ความผิดตาม ป.อ. มาตรา 157 มีระวางโทษจำคุกตั้งแ...	อาญา	พ.ร.บ.ประกอบรัฐธรรมนูญ
3150	คำพิพากษาศาลฎีกาที่ 14578/2557	โจทก์บรรยายฟ้องและนำสืบว่าจำเลยที่ 1 ทำสัญญาเช...	แพ่ง	พ.ร.บ.ว่าด้วยข้อสัญญาที่ไม่เป็นธรรม
1702	คำพิพากษาศาลฎีกาที่ 975/2559	เมื่อศาลมีคำสั่งตั้งผู้ทำแผน พ.ร.บ.ล้มละลาย พ....	พ.ร.บ.ล้มละลาย	NaN
3463	คำพิพากษาศาลฎีกาที่ 6605/2557	คดีนี้โจทก์บรรยายฟ้องสรุปความได้ว่า ตามวันเวลา...	รัฐธรรมนูญแห่งราชอาณาจักรไทย	พ.ร.บ.ป่าไม้
2769	คำพิพากษาศาลฎีกาที่ 2381/2558	แม้รถยนต์หมายเลขทะเบียน กค 812 ปราจีนบุรี ของก...	อาญา	กลุ่ม พ.ร.บ.เกี่ยวกับยาเสพติด
3488	คำพิพากษาศาลฎีกาที่ 5989/2557	การฟื้นฟูสมรรถภาพผู้เสพหรือผู้ติดยาเสพติดตาม พ...	กลุ่ม พ.ร.บ.เกี่ยวกับยาเสพติด	NaN

Figure 3.6 Dataset after tagging documents with category

4. Vectorizing Text

To enable certain analytical methods on the documents, text data were transformed into the vector representation. In this study, a cosine similarity was utilized to measure the similarity of a pair of documents. This method required transforming the text into the vector space. The study used a TF-IDF Vectorizer as discussed in Section 2.2.2 to transform the text data into the vector space. Figure 3.7 shows a sample of vectorized data.



	รู้สึก	99	ลพ	เพียงผู้เดียว	กลางคัน	ทางผ่าน	ที่ผ่านมา	ร้านตัดเสื้อ	ทุกครั้งที่	มหกรรม
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...
6665	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6666	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6667	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6668	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6669	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 3.7 A sample of vectorized data

3.4 Exploratory Data Analysis

In this section, the initial exploratory data analysis is presented to gain some basic insight from the data before the in-depth analysis.

The dataset used in this study consisted of 7,090 judgments, in which each judgment was tagged with a legal article involved with the case. In total, there were 5,545 articles mentioned. The legal articles came from 225 codes of law, including the Civil and Commercial Code, Criminal Code, Revenue Code, and other acts and decrees.

An individual judgment may involve one or more legal articles. In this dataset, there were up to 21 legal articles referred in a single judgment. The distribution of the number of articles were positively skewed with an average of 3.16 and median of 3 referred articles. Figure 3.8 shows the distribution of the number of articles referred in a judgment.

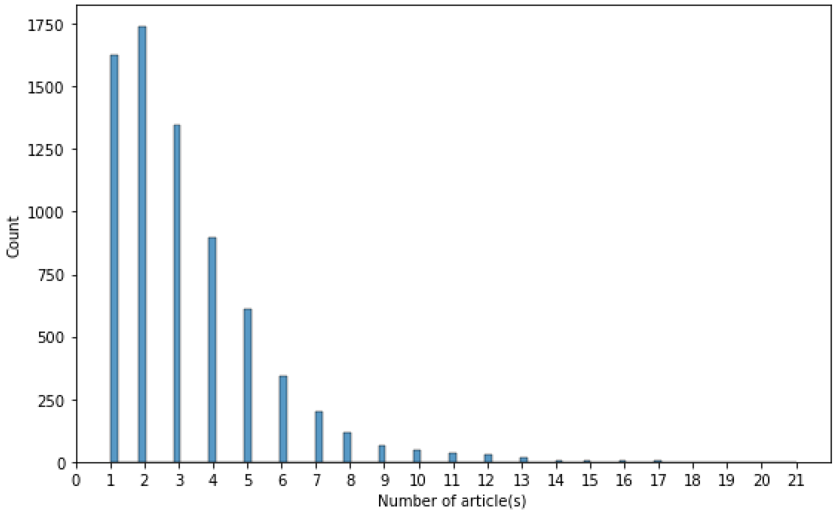


Figure 3.8 A distribution of number of legal articles referred of a judgement

On the other hand, an inspection on each legal article revealed that there were some articles that were involved in more than 200 cases. The average cases per article was 4.04, and median was one because there were 3,064 articles involved in only one case with the top article being "Criminal Procedural Code Article 225". The article involved permitting the using of legislation on the procedures and judgment of the Court of Appeal in the Supreme Court of Thailand. This was followed by "Criminal Procedural Code Article 195 Section 2", which would allow the courts and appellant to refer to the legal article that was not raised in previous legal proceedings if and only if the case involved the peace and order of the nation, or a procedural code was misapplied. Figure 3.9 shows the top 10 articles that were referred.

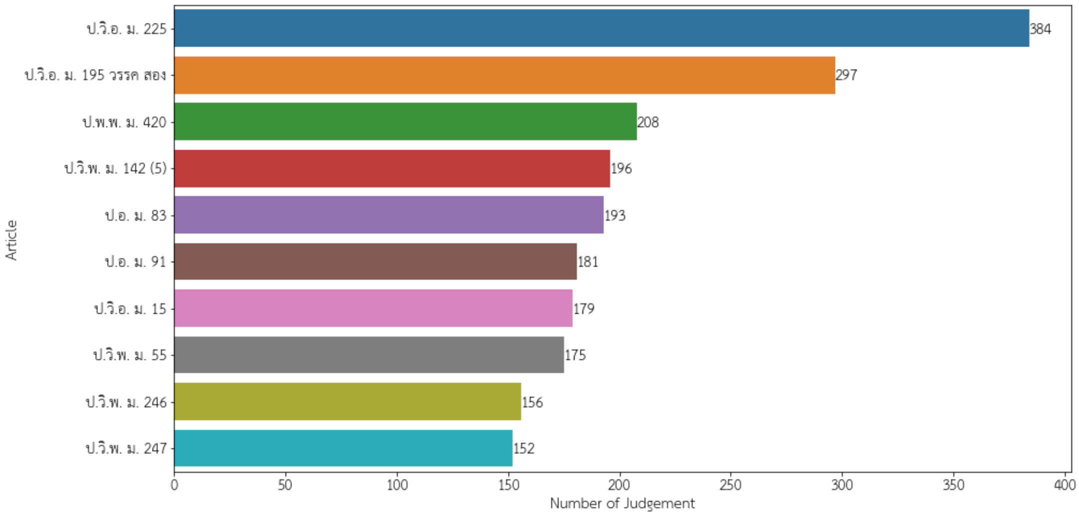


Figure 3.9 Number of cases involved an article (top 10)

After tagging the judgment with the categories as stated in Section 3.3.3, the number of times could be counted where the case mainly involved a certain code of law. The inspection showed that the most common category of a case was the “Civil and Commercial Code”, which used approximately 42% of the dataset. This was followed by the “Criminal Code” (28%), “Acts and Codes on Drugs” (4%), and Revenue Code (2%), respectively. The remaining categories consisted of up to 24% of the dataset. Figure 3.10 shows the number of the top 20 case categories.

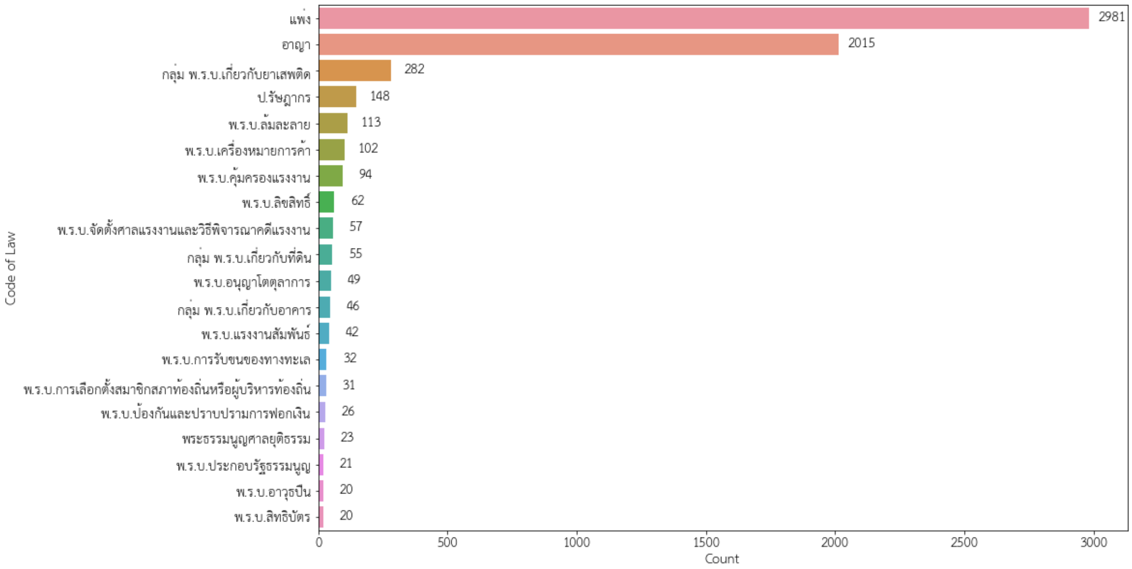


Figure 3.10 A distribution of case category (top 20)

On the aspect of the text data, tokenization was performed on the content of each judgment and the number of words or tokens that occurred were counted. The results showed that a single judgment contained between 26 to 4,256 words with the mean at 306 and median at 255 words, respectively. Figure 3.11 shows the distribution of the number of words in a judgement.

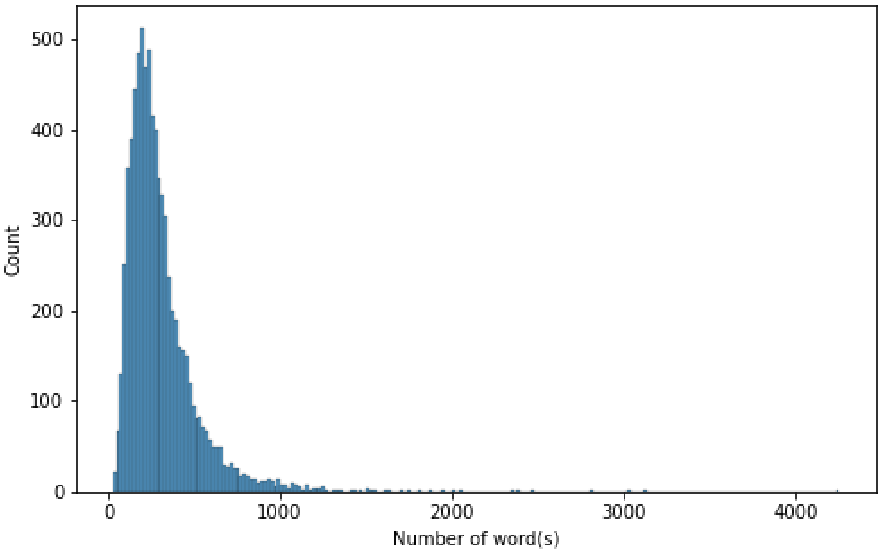
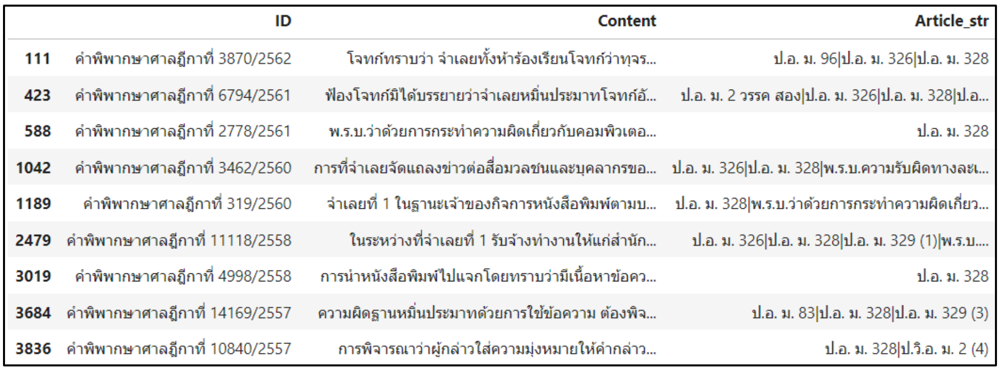


Figure 3.11 A distribution of number of words in a judgement

3.5 Graph Representation

As discussed in Section 2.2.3, each entry was defined, as to how it was connected in a graph using a common shared property. In this study, the common property was the legal article. If a pair of judgments shared a common legal article, thus they were related. Figure 3.12 shows an example of the judgments of a shared common legal article from the Criminal Code Article 328.



	ID	Content	Article_str
111	คำพิพากษาศาลฎีกาที่ 3870/2562	โจทก์ทราบว่า จำเลยทั้งห้าร้องเรียนโจทก์ว่าทุจร...	ป.อ. ม. 96|ป.อ. ม. 326|ป.อ. ม. 328
423	คำพิพากษาศาลฎีกาที่ 6794/2561	ฟ้องโจทก์มิได้บรรยายว่าจำเลยหมิ่นประมาทโจทก์อั...	ป.อ. ม. 2 วรรค สอง|ป.อ. ม. 326|ป.อ. ม. 328|ป.อ...
588	คำพิพากษาศาลฎีกาที่ 2778/2561	พ.ร.บ.ว่าด้วยการกระทำความผิดเกี่ยวกับคอมพิวเตอ...	ป.อ. ม. 328
1042	คำพิพากษาศาลฎีกาที่ 3462/2560	การที่จำเลยจัดแถลงข่าวต่อสื่อมวลชนและบุคลากรขอ...	ป.อ. ม. 326|ป.อ. ม. 328|พ.ร.บ.ความรับผิดทางละเ...
1189	คำพิพากษาศาลฎีกาที่ 319/2560	จำเลยที่ 1 ในฐานะเจ้าของกิจการหนังสือพิมพ์ตามบ...	ป.อ. ม. 328|พ.ร.บ.ว่าด้วยการกระทำความผิดเกี่ยว...
2479	คำพิพากษาศาลฎีกาที่ 11118/2558	ในระหว่างที่จำเลยที่ 1 รับจ้างทำงานให้แก่สำนัก...	ป.อ. ม. 326|ป.อ. ม. 328|ป.อ. ม. 329 (1)|พ.ร.บ....
3019	คำพิพากษาศาลฎีกาที่ 4998/2558	การนำหนังสือพิมพ์ไปแจกโดยทราบว่ามีเนื้อหาข้อคว...	ป.อ. ม. 328
3684	คำพิพากษาศาลฎีกาที่ 14169/2557	ความผิดฐานหมิ่นประมาทด้วยการใช้ข้อความ ต้องพิจ...	ป.อ. ม. 83|ป.อ. ม. 328|ป.อ. ม. 329 (3)
3836	คำพิพากษาศาลฎีกาที่ 10840/2557	การพิจารณาว่าผู้กล่าวใส่ความมุ่งหมายให้คำกล่าว...	ป.อ. ม. 328|ป.วิ.อ. ม. 2 (4)

Figure 3.12 Judgements that shared common legal article

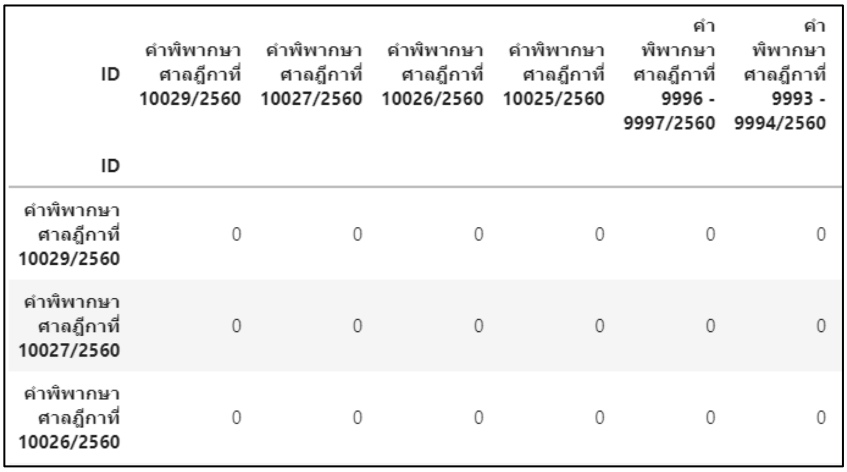
Hence, the data could be used in the form of an incidence matrix from Section 3.3 to construct an adjacency matrix. The adjacency matrix would be constructed based on how each entity shared a common legal article. In this study, the construction of the adjacency matrix is defined as follows:

$$Adj(M) = M \cdot M^T \tag{3.1}$$

Where:

$Adj(M)$	indicates adjacency matrix constructed from M
M	indicates the initial matrix
M^T	indicates the transpose of M

The results were used in the network analysis. Figure 3.13 depicts the adjacency matrix used in the study.



ID	คำพิพากษาศาลฎีกาที่ 10029/2560	คำพิพากษาศาลฎีกาที่ 10027/2560	คำพิพากษาศาลฎีกาที่ 10026/2560	คำพิพากษาศาลฎีกาที่ 10025/2560	คำพิพากษาศาลฎีกาที่ 9996 - 9997/2560	คำพิพากษาศาลฎีกาที่ 9993 - 9994/2560
ID						
คำพิพากษาศาลฎีกาที่ 10029/2560	0	0	0	0	0	0
คำพิพากษาศาลฎีกาที่ 10027/2560	0	0	0	0	0	0
คำพิพากษาศาลฎีกาที่ 10026/2560	0	0	0	0	0	0

Figure 3.13 Adjacency matrix

3.6 Document Retrieval System

As mentioned in Section 2.2.1, a document retrieval system could be built on any corpus of the text data. In this study, a document retrieval system was built using different approaches, including using the network features obtained from the network and topological analysis. The detailed setting on each system is presented in Table 3.1.

Table 3.1 Settings of Document Retrieval System

	Name	Scheme
1	Text Features	TF-IDF Vectorizer + Cosine Similarity
2	Network Connectivity	Network Neighbor Expansion
3	Network Clustering + Text Features	In-Cluster + Text Features

In system one, a standard text mining approach was used as discussed in Section 2.2.1, which consisted of representing the text in the vector space using the TF-IDF vectorizer. This was followed by using the cosine similarity to precompute the similarity between each document.

A network of documents was formed in systems two and three based on how each document was connected. Then, that network was utilized to find the documents that were similar to the one under consideration. In system two, a neighbor expansion strategy was employed by obtaining the neighbor of the node of interest. Then, the neighbors of the neighbor were obtained till the number of retrieved documents reached the limit.

For system three, a graph clustering algorithm was performed for the network. Then, all the nodes that were in the same cluster as the node were obtained till achieving the result. If the number of retrieved documents was not met, the approach of system one was used to fulfill the rest.

The approach in measuring the performance of the system consisted of the process of obtaining 100 most related documents from a randomly selected document and compute the performance evaluation metrices. The process was repeated 1,000 times for each system, and the average performance score was computed for each system. The average score from each system was then used to compare the performance of each setup. Figure 3.14 shows the process of calculating the metrics of each setup.

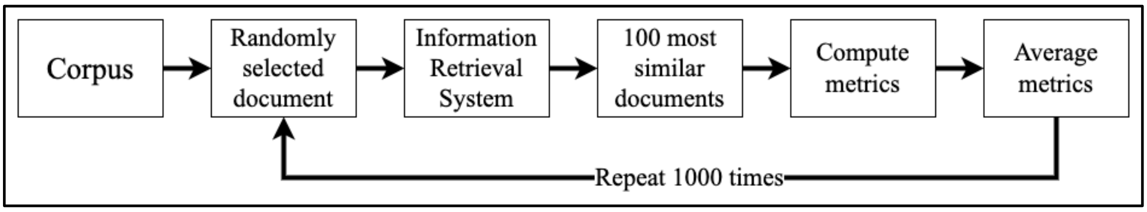

Corpus
Randomly selected document
Information Retrieval System
100 most similar documents
Compute metrics
Average metrics
Repeat 1000 times

Figure 3.14 Experiment Setup of Document Retrieval System

The concept of this experiment was to measure the performance of the system in retrieving documents that were similar to any given document when the features used in each system were different. In this study, the similarity was measured by the purity of the category of the retrieved documents. The metrics used in the study was presented in next section.

3.7 Performance Evaluation Metrics

To measure the performance of the different methods of analysis; namely, the graph clustering analysis, which could be measured by how relevant each member of a cluster was compared to outside the cluster, the following evaluation metrics were used in the study to determine the performance of each method.

1. Entropy

Entropy is a measurement of the information or uncertainty of the possible outcome in a random variable using the probability of an independent outcome that could occur [34]. Entropy is defined as follows

$$H(X) = -\sum_{i} P_x(x_i)\log_b(P_x(x_i)) \tag{3.2}$$

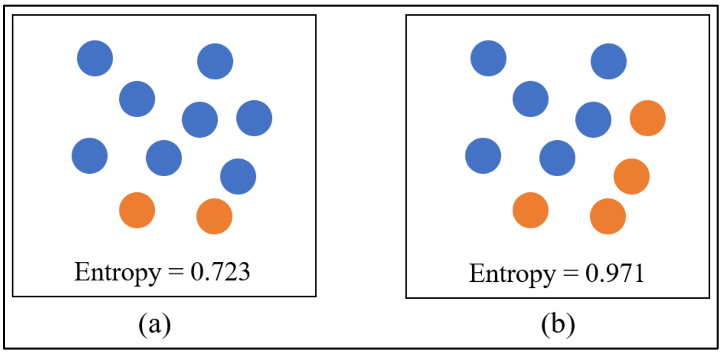
Where:

$H(X)$ indicates entropy of observation X
b indicates numbers of possible outcome that can be found in X
$P_x(x_i)$ indicates probability of finding a member with outcome i in X

This definition of entropy allowed for its use as a measurement of impurity in the data. If a group or cluster consisted of only one possible outcome, the entropy would be zero, which would indicate that the group was pure. In contrast, if a group or cluster had more than one possible outcome and the chance of finding each outcome was equal, the entropy would be 1, and the group would be impure.

For example, if a group of 10 contained eight members with label “A” and two members with label “B”, the entropy of this group would be equal to 0.723. On the other hand, if the group had six members of “A” and four members of “B”, the entropy would be equal to 0.971.

By using this metric, it could be stated that the first setting was purer than the latter. Figure 3.15 depicted this example.



(a) (b)

Figure 3.15 An example of using Entropy in determining purity of group
(a) A group with 8 members of “A” and 2 members of “B”
(b) A group with 6 members of “A” and 4 members of “B”

In addition to measuring the clustering algorithm performance, entropy would be used in measuring the performance of the document retrieval system as well.

2. Modularity

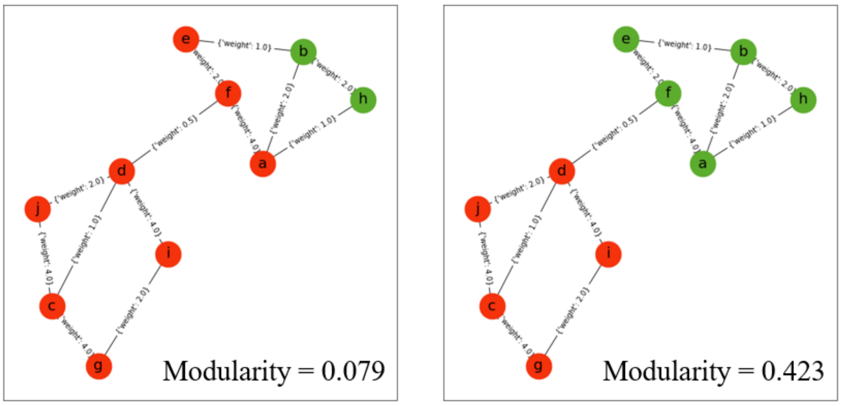
Modularity is a measurement of strength of in-cluster connections compared to the overall network connections. Modularity ranges between [-0.5, 1] where -0.5 is when a graph would be fully connected and could not be separated, and 1 being the connection within the cluster would have much more prominence than between the clusters. For any undirected graph, modularity was defined as follows [31].

$$Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j) \tag{2.4}$$

Where:

Q	indicates modularity of network
A_{ij}	indicates the edge weight between node i and j
$k_i,\ k_j$	indicates sum of the edge's weight attached to node i and j respectively
m	indicates sum of all edge's weight in the graph
c_i, c_j	indicates cluster or community of node i and j respectively
δ	indicates a function that yields 1 if $i = j$, 0 otherwise

Figure 3.16 illustrates the example of modularity in measuring the performance of a graph clustering algorithm. Figure 3.16 (a) is an example of bad clustering. However, it could be clearly observed that the connection within cluster one ($\{b,\ h{:}\,2\}$) was not significantly stronger than the connection between the clusters ($\{a,\ b{:}\,2\}, \{a,\ h{:}\,1\}, \{b, e{:}\,1\}$). In contrast, Figure 3.16 (b) shows the clustering that yielded a stronger in-cluster connection compared to between the clusters connection ($\{d,\ f{:}\,0.5\}$).



(a) (b)

Figure 3.16 An example of Modularity in measuring graph clustering performance
(a) A graph clustering with $\{\{b, h\}, \{a, c, d, e, f, g, i, j\}\}$
(b) A graph clustering with $\{\{a, b, e, f, h\}, \{c, d, g, i, j\}\}$

CHAPTER 4 RESULTS

In this section, the results of the study are presented, including the interpretation and detailed explanation on each method of analysis.

4.1 Network Analysis Result

Using techniques in network and topological analysis as discussed in Section 2.2.3, the results of the study were as follows.

1. Connected component analysis

After obtaining the adjacency matrix from Section 3.5, the network could be visualized using a Python's library called NetworkX [35]. Figure 4.1 shows the overall network topology of the dataset.

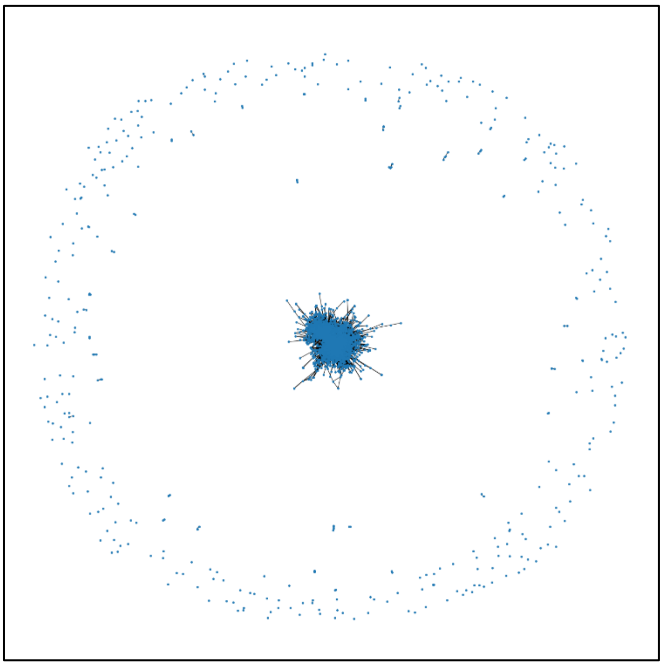


Figure 4.1 Network structure of the dataset

From the topology, a densely connected cluster was observed in the middle of the network, while on the outskirts of the graph, a small number of nodes were isolated or only loosely connected. The inspection showed that this network consisted of 355 connected components. However, 94% (6,670 of 7,090) of the nodes were in a single component, henceforth referred to as the main component. Nevertheless, 6% (420 of 7,090) of the nodes that were not in the main component were loosely connected or isolated. Figure 4.2 shows the distribution of the number of nodes in each component.

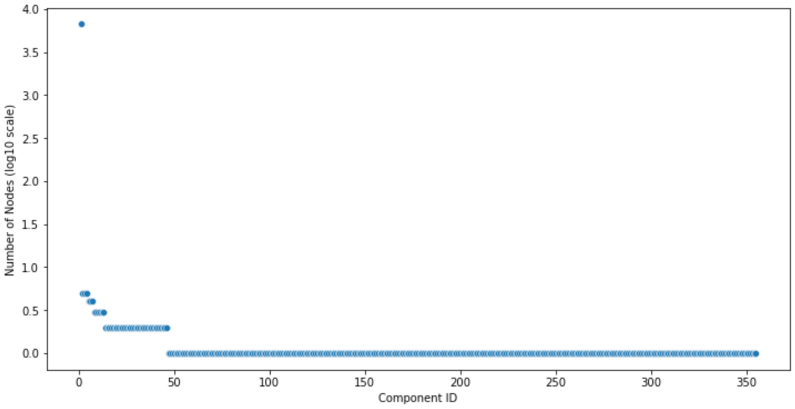
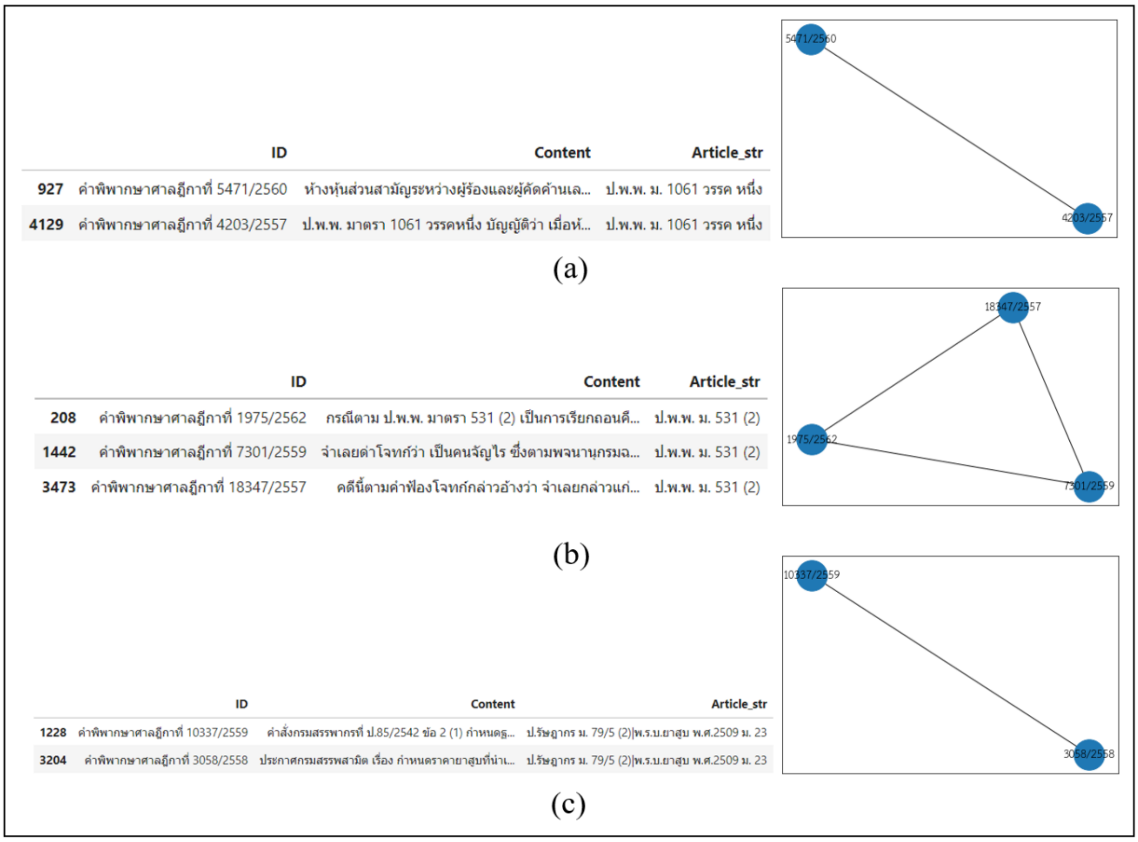


Figure 4.2 Distribution of number of nodes on each component

From Figure 4.2, it could be observed that only 13% (47 of 355) of the components contained more than one node, while the remaining components only consisted of a single node. This indicated that some legal articles had no relationship with other items through a common court judgment.

The inspection also showed that the outlying components were an extremely specific case; for example, the judgment on the case about pensions for retired government officers, which occurred only once across the dataset and had no relationship to the other cases. Figure 4.3 shows an example of such components.



	ID	Content	Article_str
927	คำพิพากษาศาลฎีกาที่ 5471/2560	ห้างหุ้นส่วนสามัญระหว่างผู้ร้องและผู้คัดค้านเล...	ป.พ.พ. ม. 1061 วรรค หนึ่ง
4129	คำพิพากษาศาลฎีกาที่ 4203/2557	ป.พ.พ. มาตรา 1061 วรรคหนึ่ง บัญญัติว่า เมื่อห้...	ป.พ.พ. ม. 1061 วรรค หนึ่ง

(a)

	ID	Content	Article_str
208	คำพิพากษาศาลฎีกาที่ 1975/2562	กรณีตาม ป.พ.พ. มาตรา 531 (2) เป็นการเรียกถอนคื...	ป.พ.พ. ม. 531 (2)
1442	คำพิพากษาศาลฎีกาที่ 7301/2559	จำเลยด่าโจทก์ว่า เป็นคนจัญไร ซึ่งตามพจนานุกรมฉ...	ป.พ.พ. ม. 531 (2)
3473	คำพิพากษาศาลฎีกาที่ 18347/2557	คดีนี้ตามคำฟ้องโจทก์กล่าวอ้างว่า จำเลยกล่าวแก่...	ป.พ.พ. ม. 531 (2)

(b)

<table>
<tr><th></th><th>ID</th><th>Content</th><th>Article_str</th></tr>
<tr><td><b>1228</b></td><td>คำพิพากษาศาลฎีกาที่ 10337/2559</td><td>คำสั่งกรมสรรพากรที่ ป.85/2542 ข้อ 2 (1) กำหนดฐ...</td><td>ป.รัษฎากร ม. 79/5 (2)|พ.ร.บ.ยาสูบ พ.ศ.2509 ม. 23</td></tr>
<tr><td><b>3204</b></td><td>คำพิพากษาศาลฎีกาที่ 3058/2558</td><td>ประกาศกรมสรรพสามิต เรื่อง กำหนดราคายาสูบที่นำเ...</td><td>ป.รัษฎากร ม. 79/5 (2)|พ.ร.บ.ยาสูบ พ.ศ.2509 ม. 23</td></tr>
</table>

(c)

Figure 4.3 An example of outlying components that shared different legal articles
(a) Component with Civil and Commercial Code article 1061 section 1
(b) Component with Civil and Commercial Code article 531 (2)
(c) Component with Revenue Code article 79/5 (2)

According to the findings, the study mainly considered the main component in the further analysis. Figure 4.4 shows the main component, which contained 6,670 nodes that were linked with 383,098 edges.

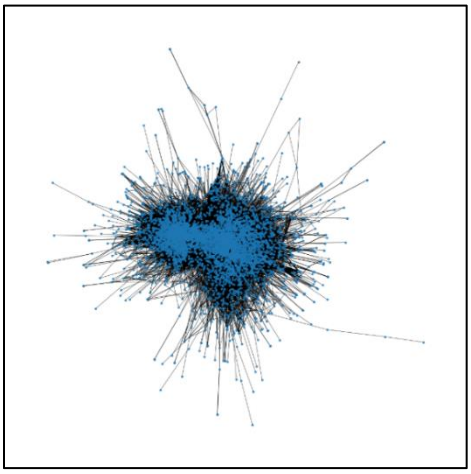


Figure 4.4 Main component

2. Graph clustering

The graph clustering algorithm was used to find hidden clusters of documents based on how they were referenced in the network. As discussed in Section 2.2.3, the algorithm used in this study consisted of spectral clustering and the Louvain method.

- Spectral Clustering

  In the spectral clustering process, first the graph Laplacian would be computed by subtracting the degree matrix with the adjacency matrix. From there, it would be possible to obtain the spectrum of a graph Laplacian by using Eigendecomposition. The result eigenvector and eigenvalue were used in the later analysis.

  To determine an appropriate number of clusters in spectral clustering, the differences between each eigenvalue after being sorted ascendingly were taken into consideration. This is called a spectral gap. The position of the biggest spectral gap would be the number of clusters. Figure 4.5 shows the position and value of the eigenvalue from the decomposition.

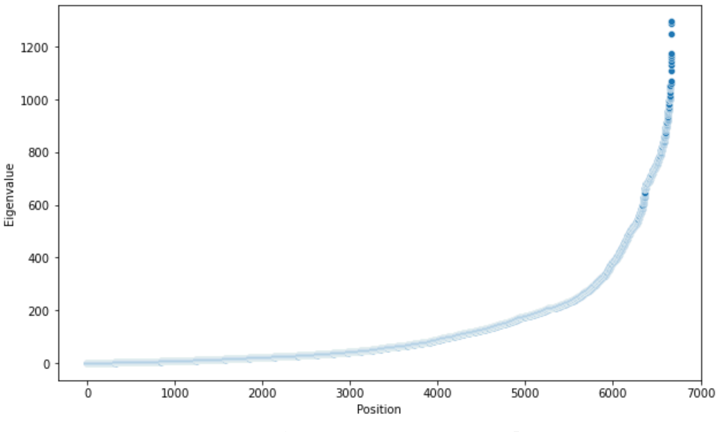


Figure 4.5 Eigenvalue

Figure 4.5 reveals that the greatest spectral gap was at positions 6,669-6,670. This indicated that each member belonged to its own cluster, which provided no useful information for the study. From the findings, a truncation approach was utilized to only analyze the top N positions of the eigenvalues. The first 100 eigenvalues are shown in Figure 4.6.

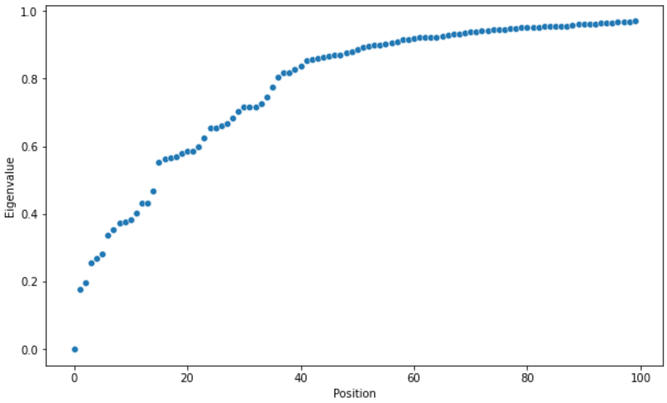
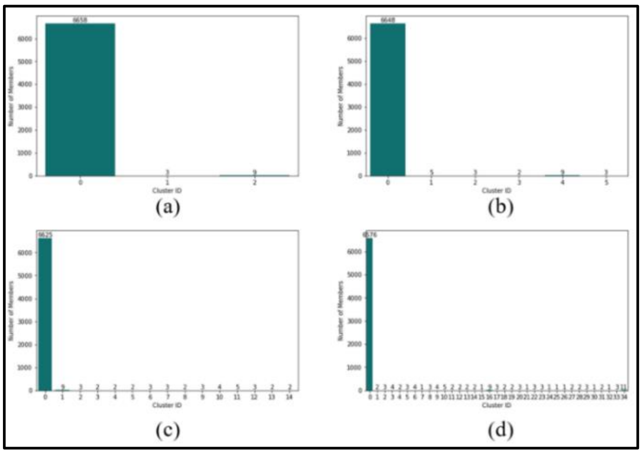


Figure 4.6 First 100 Eigenvalues

Figure 4.6 shows that the greatest spectral gap occurred at positions 3, 6, 15, and 35. These numbers were then used as the number of clusters in a simple clustering algorithm; such as, k-means. In this study, two methods were used for selecting the features for the clustering algorithm. The first method used only the first *k* eigenvectors as a feature. Figure 4.7 depicts the outcome of the k-means clustering in terms of the cluster member distribution.



(a) (b) (c) (d)

Figure 4.7 Cluster member distribution from Spectral Clustering using first *k* eigenvectors as a feature with K-means at different *k*.

(a) K=3
(b) K=6
(c) K=15
(d) K=35

The result from clustering with a stated approach yielded an unsatisfactory result. The presence of an over dominated cluster could be observed, which the cluster took over 98% of the nodes leaving other clusters with only a few tens of nodes.

The second approach used all the eigenvectors as a feature for the clustering algorithm. Figure 4.8 shows the clustering result of such process.

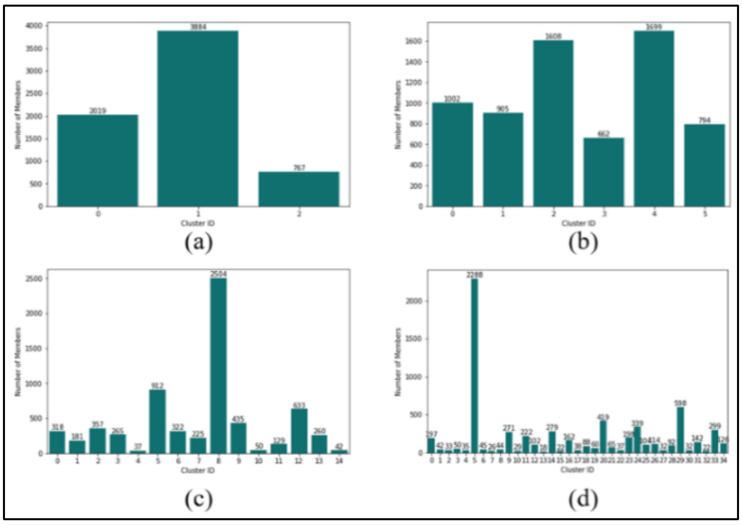


Figure 4.8 Cluster member distribution from Spectral Clustering using all eigenvectors as a feature with K-means at different *k*
(a) K=3
(b) K=6
(c) K=15
(d) K=35

From both diagrams, the over dominated cluster could be observed even though the effect of such cluster was lessened in the second approach. This was due to the behavior of the spectral clustering algorithm which will be discussed in later section.

- Louvain method for community detection
  As discussed in Section 2.2.3, the Louvain method tried to cluster the graph in which the modularity was optimized. This study used a Python's community detection library called python-louvain [36]. The algorithm discovered 13 communities. Each communities' member distribution was depicted in Figure 4.9.

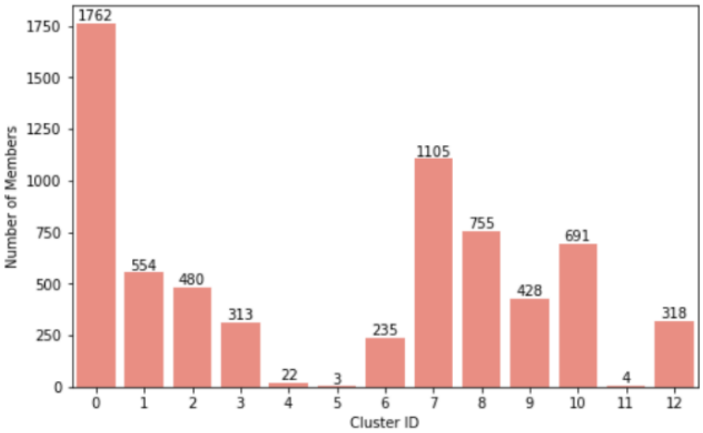


Figure 4.9 Cluster member distribution from Louvain method

From Figure 4.9, it could be observed that there was more balance in the cluster member distribution compared to the initial spectral clustering although the cluster with only a few members was still present.

3. Cluster Interpretation in Computational Aspect

In this topic, an interpretation of the cluster obtained from Section 4.2 in the aspect of topology, connectivity, and purity was presented.

- Cluster induction

  In Figure 4.10, each node was colorized on the graph based on the cluster to which it belonged using the visualization of the main cluster.

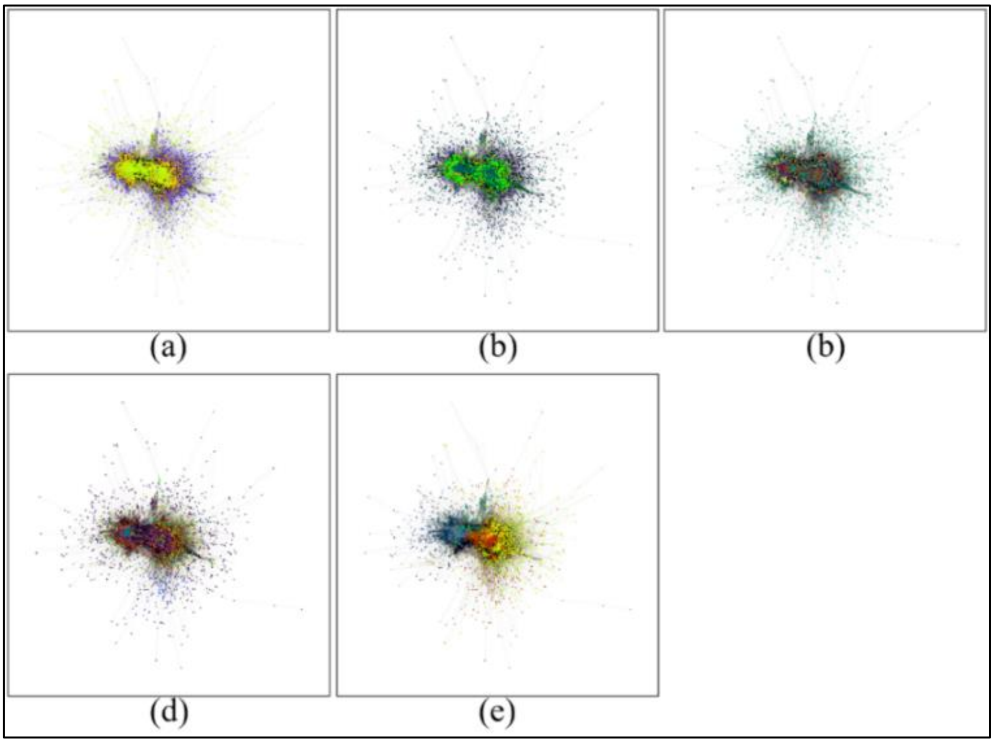


Figure 4.10 A network with colorized node according to their cluster
(a) Spectral Clustering K=3
(b) Spectral Clustering K=6
(c) Spectral Clustering K=15
(d) Spectral Clustering K=35
(e) Louvain method

To visualize the connection between the clusters, each cluster was induced into a single node. Figure 4.11 depicted the results.

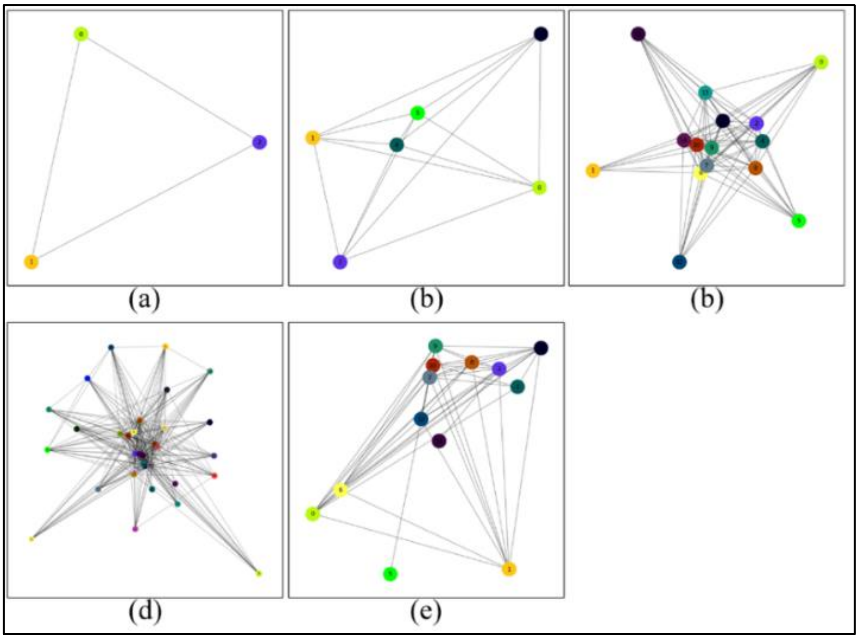


Figure 4.11 An induced network colorized according to their base cluster
(a) Spectral Clustering K=3
(b) Spectral Clustering K=6
(c) Spectral Clustering K=15
(d) Spectral Clustering K=35
(e) Louvain method

From Figure 4.11, a dense to fully connected graph was observed for spectral clustering. This suggested that some clusters may have been impure and inseparable. In the Louvain method, some clusters were only connected to a few neighbor clusters. In addition, some clusters served as network hubs; as such, this indicated the possibility of separation.

- Cluster purity and modularity

  As discussed in Section 3.6, the purity of a cluster obtained from each algorithm was measured using entropy. In addition, modularity was used to assess the strength of the cluster's connections. Table 4.1 displays the outcome.

Table 4.1 Clustering metrics

	Algorithm				
Measure	Spectral K=3	Spectral K=6	Spectral K=15	Spectral K=35	Louvain
Number of clusters	3	6	15	35	13
Average entropy	0.503	0.530	0.542	0.612	**0.456**
Modularity	0.062	0.108	0.089	0.087	**0.569**

According to Table 4.1, the Louvain method had the best performance in terms of document category differentiation, as well as structural strength optimization. This suggested that using the Louvain method with this dataset in the further analysis of integrating it with the document retrieval system would provide benefits to the system more than using spectral clustering.

4. Cluster Interpretation in Legal Aspect

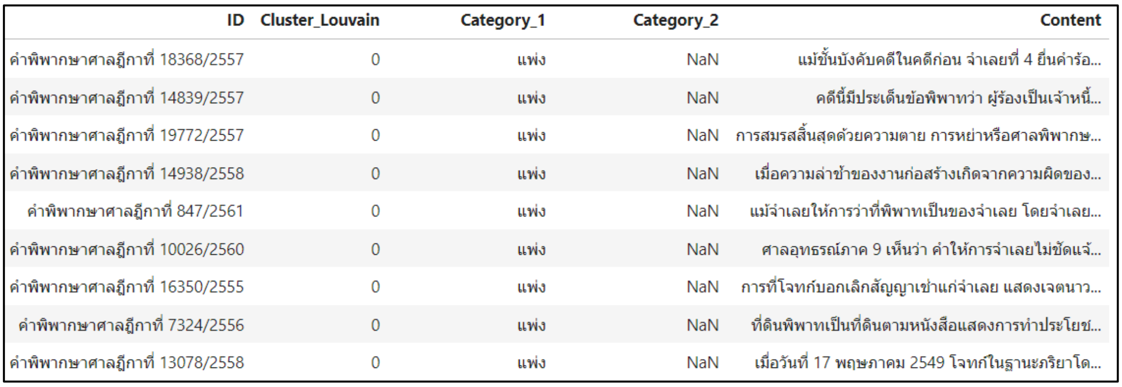
Apart from the interpretation in the computational aspect, an interpretation in the legal aspect allowed the inspection to have consistency of each cluster in terms of the legal document category, as well as to define the “topic” of each cluster. This interpretation was based on the cluster obtained from the Louvain method. In each cluster, the number of cluster members and a fraction of the top three legal article categories were provided.

- Cluster 0

  Members: 1762

  Dominant Category: Civil and Commercial Code (0.68), Bankruptcy Act (0.07), Arbitration Act (0.03)

  This cluster comprised the common Civil and Commercial cases, which often occurred in Thailand, such as, the verdict on a contract, marriage, property, and inheritance. This cluster could be interpreted as the representation of “popular” Civil and Commercial cases that often occurred in Thailand.



ID	Cluster_Louvain	Category_1	Category_2	Content
คำพิพากษาศาลฎีกาที่ 18368/2557	0	แพ่ง	NaN	แม้ชั้นบังคับคดีในคดีก่อน จำเลยที่ 4 ยื่นคำร้อ...
คำพิพากษาศาลฎีกาที่ 14839/2557	0	แพ่ง	NaN	คดีนี้มีประเด็นข้อพิพาทว่า ผู้ร้องเป็นเจ้าหนี้...
คำพิพากษาศาลฎีกาที่ 19772/2557	0	แพ่ง	NaN	การสมรสสิ้นสุดด้วยความตาย การหย่าหรือศาลพิพากษ...
คำพิพากษาศาลฎีกาที่ 14938/2558	0	แพ่ง	NaN	เมื่อความล่าช้าของงานก่อสร้างเกิดจากความผิดของ...
คำพิพากษาศาลฎีกาที่ 847/2561	0	แพ่ง	NaN	แม้จำเลยให้การว่าที่พิพาทเป็นของจำเลย โดยจำเลย...
คำพิพากษาศาลฎีกาที่ 10026/2560	0	แพ่ง	NaN	ศาลอุทธรณ์ภาค 9 เห็นว่า คำให้การจำเลยไม่ชัดแจ้...
คำพิพากษาศาลฎีกาที่ 16350/2555	0	แพ่ง	NaN	การที่โจทก์บอกเลิกสัญญาเช่าแก่จำเลย แสดงเจตนาว...
คำพิพากษาศาลฎีกาที่ 7324/2556	0	แพ่ง	NaN	ที่ดินพิพาทเป็นที่ดินตามหนังสือแสดงการทำประโยช...
คำพิพากษาศาลฎีกาที่ 13078/2558	0	แพ่ง	NaN	เมื่อวันที่ 17 พฤษภาคม 2549 โจทก์ในฐานะภริยาโด...

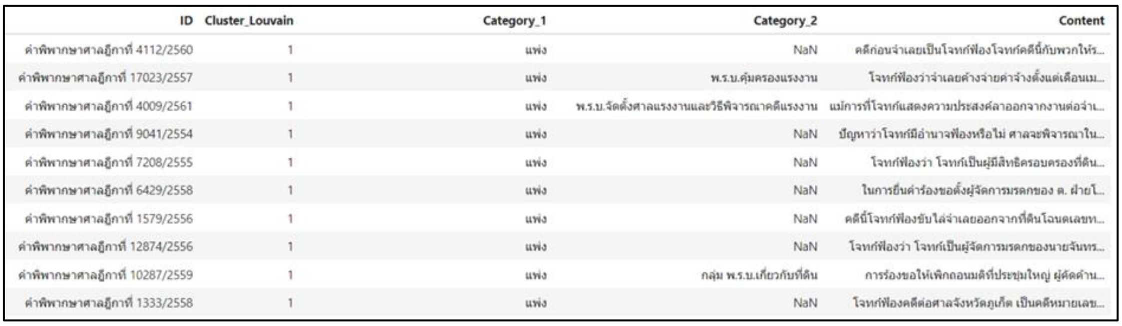
Figure 4.12 Sample of documents in Cluster 0

- Cluster 1

  Members: 554

  Dominant Category: Civil and Commercial Code (0.74), Act on Establishment of Labor Court and Labor Court Procedure (0.05), The Law for the Organization of Court of Justice (0.02)

  The cluster was composed of mainly Civil and Commercial cases comparable to cluster 0. However, the case in this cluster contained a mixture of more complex codes of law; such as, the Act on the Labor Court, Land and Settlement Code, and others.



ID	Cluster_Louvain	Category_1	Category_2	Content
คำพิพากษาศาลฎีกาที่ 4112/2560	1	แพ่ง	NaN	คดีก่อนจำเลยเป็นโจทก์ฟ้องโจทก์คดีนี้กับพวกให้ร...
คำพิพากษาศาลฎีกาที่ 17023/2557	1	แพ่ง	พ.ร.บ.คุ้มครองแรงงาน	โจทก์ฟ้องว่าจำเลยค้างจ่ายค่าจ้างตั้งแต่เดือนเม...
คำพิพากษาศาลฎีกาที่ 4009/2561	1	แพ่ง	พ.ร.บ.จัดตั้งศาลแรงงานและวิธีพิจารณาคดีแรงงาน	แม้การที่โจทก์แสดงความประสงค์ลาออกจากงานต่อจำเ...
คำพิพากษาศาลฎีกาที่ 9041/2554	1	แพ่ง	NaN	ปัญหาว่าโจทก์มีอำนาจฟ้องหรือไม่ ศาลจะพิจารณาใน...
คำพิพากษาศาลฎีกาที่ 7208/2555	1	แพ่ง	NaN	โจทก์ฟ้องว่า โจทก์เป็นผู้มีสิทธิครอบครองที่ดิน...
คำพิพากษาศาลฎีกาที่ 6429/2558	1	แพ่ง	NaN	ในการยื่นคำร้องขอตั้งผู้จัดการมรดกของ ต. ฝ่ายโ...
คำพิพากษาศาลฎีกาที่ 1579/2556	1	แพ่ง	NaN	คดีนี้โจทก์ฟ้องขับไล่จำเลยออกจากที่ดินโฉนดเลขท...
คำพิพากษาศาลฎีกาที่ 12874/2556	1	แพ่ง	NaN	โจทก์ฟ้องว่า โจทก์เป็นผู้จัดการมรดกของนายจันทร...
คำพิพากษาศาลฎีกาที่ 10287/2559	1	แพ่ง	กลุ่ม พ.ร.บ.เกี่ยวกับที่ดิน	การร้องขอให้เพิกถอนมติที่ประชุมใหญ่ ผู้คัดค้าน...
คำพิพากษาศาลฎีกาที่ 1333/2558	1	แพ่ง	NaN	โจทก์ฟ้องคดีต่อศาลจังหวัดภูเก็ต เป็นคดีหมายเลข...

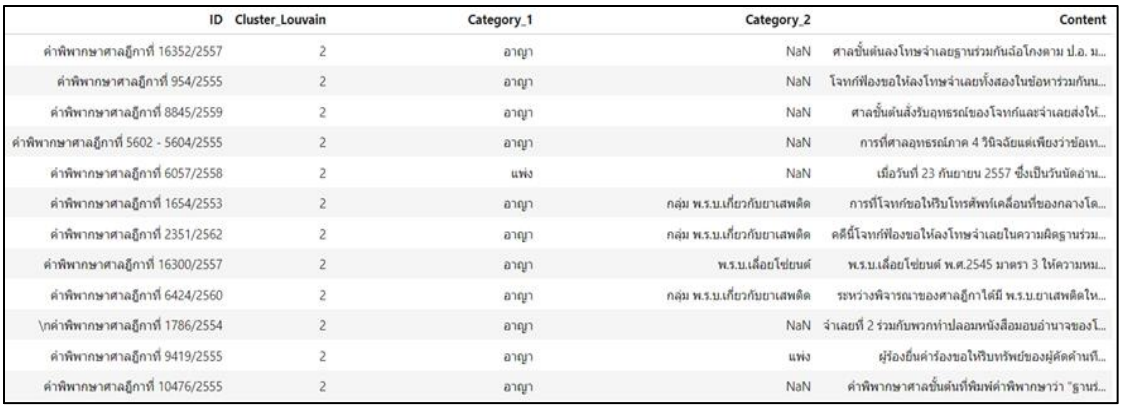
Figure 4.13 Sample of documents in Cluster 1

- Cluster 2

  Members: 480

  Dominant Category: Criminal Code (0.65), Acts and Codes on Drugs (0.08), Civil and Commercial Code (0.06)

  Documents assigned to this cluster were mainly involved with the Criminal Code with an emphasis on cases that involved drugs.



ID	Cluster_Louvain	Category_1	Category_2	Content
คำพิพากษาศาลฎีกาที่ 16352/2557	2	อาญา	NaN	ศาลชั้นต้นลงโทษจำเลยฐานร่วมกันฉ้อโกงตาม ป.อ. ม...
คำพิพากษาศาลฎีกาที่ 954/2555	2	อาญา	NaN	โจทก์ฟ้องขอให้ลงโทษจำเลยทั้งสองในข้อหาร่วมกันน...
คำพิพากษาศาลฎีกาที่ 8845/2559	2	อาญา	NaN	ศาลชั้นต้นสั่งรับอุทธรณ์ของโจทก์และจำเลยส่งให้...
คำพิพากษาศาลฎีกาที่ 5602 - 5604/2555	2	อาญา	NaN	การที่ศาลอุทธรณ์ภาค 4 วินิจฉัยแต่เพียงว่าข้อเท...
คำพิพากษาศาลฎีกาที่ 6057/2558	2	แพ่ง	NaN	เมื่อวันที่ 23 กันยายน 2557 ซึ่งเป็นวันนัดอ่าน...
คำพิพากษาศาลฎีกาที่ 1654/2553	2	อาญา	กลุ่ม พ.ร.บ.เกี่ยวกับยาเสพติด	การที่โจทก์ขอให้ริบโทรศัพท์เคลื่อนที่ของกลางโด...
คำพิพากษาศาลฎีกาที่ 2351/2562	2	อาญา	กลุ่ม พ.ร.บ.เกี่ยวกับยาเสพติด	คดีนี้โจทก์ฟ้องขอให้ลงโทษจำเลยในความผิดฐานร่วม...
คำพิพากษาศาลฎีกาที่ 16300/2557	2	อาญา	พ.ร.บ.เลื่อยโซ่ยนต์	พ.ร.บ.เลื่อยโซ่ยนต์ พ.ศ.2545 มาตรา 3 ให้ความหม...
คำพิพากษาศาลฎีกาที่ 6424/2560	2	อาญา	กลุ่ม พ.ร.บ.เกี่ยวกับยาเสพติด	ระหว่างพิจารณาของศาลฎีกาได้มี พ.ร.บ.ยาเสพติดให...
\nคำพิพากษาศาลฎีกาที่ 1786/2554	2	อาญา	NaN	จำเลยที่ 2 ร่วมกับพวกทำปลอมหนังสือมอบอำนาจของโ...
คำพิพากษาศาลฎีกาที่ 9419/2555	2	อาญา	แพ่ง	ผู้ร้องยื่นคำร้องขอให้ริบทรัพย์ของผู้คัดค้านที...
คำพิพากษาศาลฎีกาที่ 10476/2555	2	อาญา	NaN	คำพิพากษาศาลชั้นต้นที่พิมพ์คำพิพากษาว่า "ฐานร...

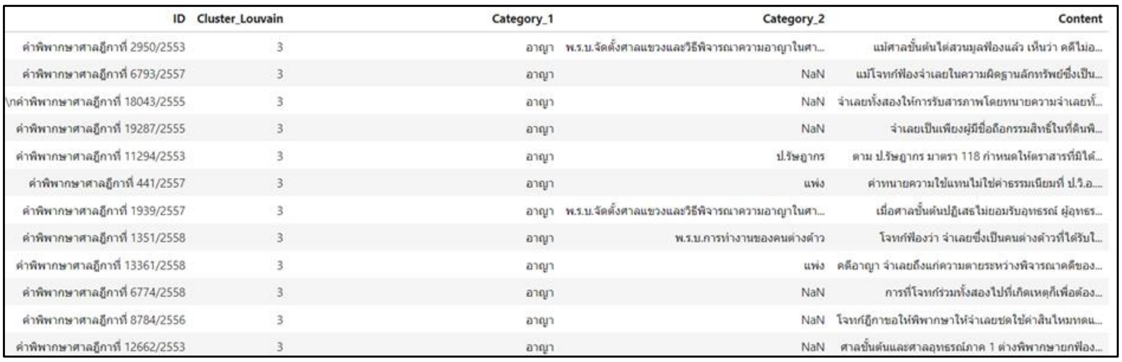
Figure 4.14 Sample of documents in Cluster 2

- Cluster 3

  Members: 313

  Dominant Category: Criminal Code (0.64), Civil and Commercial Code (0.12), Act on Establishment of District Court and Criminal Procedure in District Court (0.07)

  Documents in this cluster were mainly involved in the Criminal Code similar to Cluster 2. However, this cluster emphasized the responsibility of the district court on criminal cases.



ID	Cluster_Louvain	Category_1	Category_2	Content
คำพิพากษาศาลฎีกาที่ 2950/2553	3	อาญา	พ.ร.บ.จัดตั้งศาลแขวงและวิธีพิจารณาความอาญาในศา...	แม้ศาลชั้นต้นได้สวนมูลฟ้องแล้ว เห็นว่า คดีไม่อ...
คำพิพากษาศาลฎีกาที่ 6793/2557	3	อาญา	NaN	แม้โจทก์ฟ้องจำเลยในความผิดฐานลักทรัพย์ซึ่งเป็น...
\nคำพิพากษาศาลฎีกาที่ 18043/2555	3	อาญา	NaN	จำเลยทั้งสองให้การรับสารภาพโดยทนายความจำเลยทั้...
คำพิพากษาศาลฎีกาที่ 19287/2555	3	อาญา	NaN	จำเลยเป็นเพียงผู้มีชื่อถือกรรมสิทธิ์ในที่ดินพิ...
คำพิพากษาศาลฎีกาที่ 11294/2553	3	อาญา	ป.รัษฎากร	ตาม ป.รัษฎากร มาตรา 118 กำหนดให้ตราสารที่มิได้...
คำพิพากษาศาลฎีกาที่ 441/2557	3	อาญา	แพ่ง	ค่าทนายความใช้แทนไม่ใช่ค่าธรรมเนียมที่ ป.วิ.อ....
คำพิพากษาศาลฎีกาที่ 1939/2557	3	อาญา	พ.ร.บ.จัดตั้งศาลแขวงและวิธีพิจารณาความอาญาในศา...	เมื่อศาลชั้นต้นปฏิเสธไม่ยอมรับอุทธรณ์ ผู้อุทธร...
คำพิพากษาศาลฎีกาที่ 1351/2558	3	อาญา	พ.ร.บ.การทำงานของคนต่างด้าว	โจทก์ฟ้องว่า จำเลยซึ่งเป็นคนต่างด้าวที่ได้รับใ...
คำพิพากษาศาลฎีกาที่ 13361/2558	3	อาญา	แพ่ง	คดีอาญา จำเลยถึงแก่ความตายระหว่างพิจารณาคดีของ...
คำพิพากษาศาลฎีกาที่ 6774/2558	3	อาญา	NaN	การที่โจทก์ร่วมทั้งสองไปที่เกิดเหตุก็เพื่อต้อง...
คำพิพากษาศาลฎีกาที่ 8784/2556	3	อาญา	NaN	โจทก์ฎีกาขอให้พิพากษาให้จำเลยชดใช้ค่าสินไหมทดแ...
คำพิพากษาศาลฎีกาที่ 12662/2553	3	อาญา	NaN	ศาลชั้นต้นและศาลอุทธรณ์ภาค 1 ต่างพิพากษายกฟ้อง...

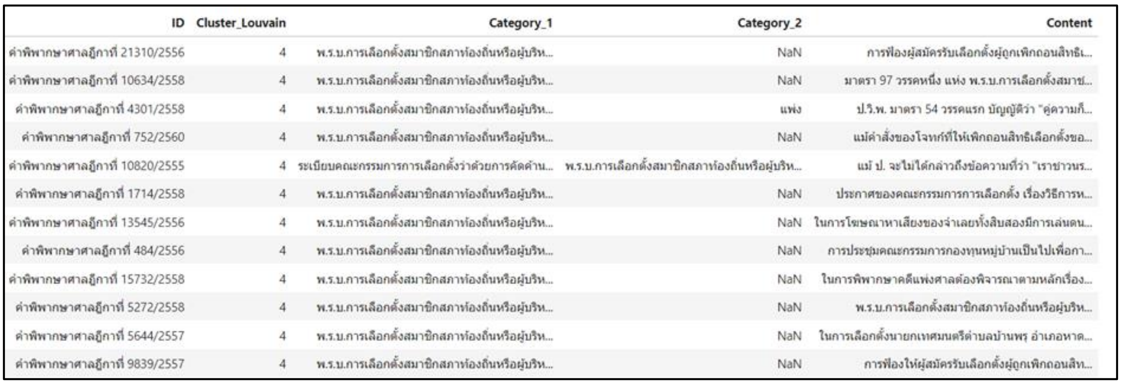
Figure 4.15 Sample of documents in Cluster 3

- Cluster 4

  Members: 22

  Dominant Category: Act on the Election of Members of the Local Administrative Organization (0.81), Organic Act on Election Commission (0.07), Administrative Procedure Act (0.04)

  Documents in this cluster were involved with the election of local administrative organizations. The involved legal articles were very specific and only had a few relationships to other categories.



ID	Cluster_Louvain	Category_1	Category_2	Content
คำพิพากษาศาลฎีกาที่ 21310/2556	4	พ.ร.บ.การเลือกตั้งสมาชิกสภาท้องถิ่นหรือผู้บริห...	NaN	การฟ้องผู้สมัครรับเลือกตั้งผู้ถูกเพิกถอนสิทธิเ...
คำพิพากษาศาลฎีกาที่ 10634/2558	4	พ.ร.บ.การเลือกตั้งสมาชิกสภาท้องถิ่นหรือผู้บริห...	NaN	มาตรา 97 วรรคหนึ่ง แห่ง พ.ร.บ.การเลือกตั้งสมาช...
คำพิพากษาศาลฎีกาที่ 4301/2558	4	พ.ร.บ.การเลือกตั้งสมาชิกสภาท้องถิ่นหรือผู้บริห...	แพ่ง	ป.วิ.พ. มาตรา 54 วรรคแรก บัญญัติว่า "คู่ความก็...
คำพิพากษาศาลฎีกาที่ 752/2560	4	พ.ร.บ.การเลือกตั้งสมาชิกสภาท้องถิ่นหรือผู้บริห...	NaN	แม้คำสั่งของโจทก์ที่ให้เพิกถอนสิทธิเลือกตั้งขอ...
คำพิพากษาศาลฎีกาที่ 10820/2555	4	ระเบียบคณะกรรมการการเลือกตั้งว่าด้วยการคัดค้าน...	พ.ร.บ.การเลือกตั้งสมาชิกสภาท้องถิ่นหรือผู้บริห...	แม้ ป. จะไม่ได้กล่าวถึงข้อความที่ว่า "เราชาวนร...
คำพิพากษาศาลฎีกาที่ 1714/2558	4	พ.ร.บ.การเลือกตั้งสมาชิกสภาท้องถิ่นหรือผู้บริห...	NaN	ประกาศของคณะกรรมการการเลือกตั้ง เรื่องวิธีการห...
คำพิพากษาศาลฎีกาที่ 13545/2556	4	พ.ร.บ.การเลือกตั้งสมาชิกสภาท้องถิ่นหรือผู้บริห...	NaN	ในการโฆษณาหาเสียงของจำเลยทั้งสิบสองมีการเล่นดน...
คำพิพากษาศาลฎีกาที่ 484/2556	4	พ.ร.บ.การเลือกตั้งสมาชิกสภาท้องถิ่นหรือผู้บริห...	NaN	การประชุมคณะกรรมการกองทุนหมู่บ้านเป็นไปเพื่อกา...
คำพิพากษาศาลฎีกาที่ 15732/2558	4	พ.ร.บ.การเลือกตั้งสมาชิกสภาท้องถิ่นหรือผู้บริห...	NaN	ในการพิพากษาคดีแพ่งศาลต้องพิจารณาตามหลักเรื่อง...
คำพิพากษาศาลฎีกาที่ 5272/2558	4	พ.ร.บ.การเลือกตั้งสมาชิกสภาท้องถิ่นหรือผู้บริห...	NaN	พ.ร.บ.การเลือกตั้งสมาชิกสภาท้องถิ่นหรือผู้บริห...
คำพิพากษาศาลฎีกาที่ 5644/2557	4	พ.ร.บ.การเลือกตั้งสมาชิกสภาท้องถิ่นหรือผู้บริห...	NaN	ในการเลือกตั้งนายกเทศมนตรีตำบลบ้านพรุ อำเภอหาด...
คำพิพากษาศาลฎีกาที่ 9839/2557	4	พ.ร.บ.การเลือกตั้งสมาชิกสภาท้องถิ่นหรือผู้บริห...	NaN	การฟ้องให้ผู้สมัครรับเลือกตั้งผู้ถูกเพิกถอนสิท...

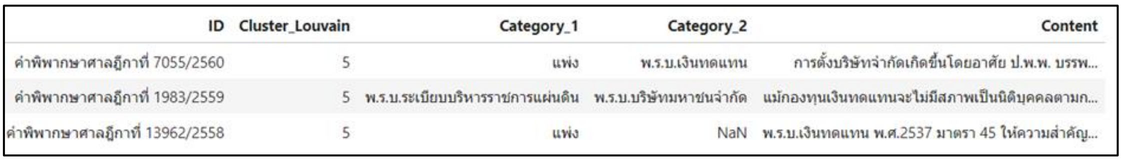
Figure 4.16 Sample of documents in Cluster 4

- Cluster 5

  Members: 3

  Dominant Category: Civil and Commercial Code (0.4), Public Limited Company Act (0.2), Administrative Procedure Act (0.2)

  This cluster involved very specific legal articles from the Civil and Commercial Code, which was about the mergers and acquisitions of a public company limited. This article only appeared twice in the dataset.



ID	Cluster_Louvain	Category_1	Category_2	Content
คำพิพากษาศาลฎีกาที่ 7055/2560	5	แพ่ง	พ.ร.บ.เงินทดแทน	การตั้งบริษัทจำกัดเกิดขึ้นโดยอาศัย ป.พ.พ. บรรพ...
คำพิพากษาศาลฎีกาที่ 1983/2559	5	พ.ร.บ.ระเบียบบริหารราชการแผ่นดิน	พ.ร.บ.บริษัทมหาชนจำกัด	แม้กองทุนเงินทดแทนจะไม่มีสภาพเป็นนิติบุคคลตามก...
คำพิพากษาศาลฎีกาที่ 13962/2558	5	แพ่ง	NaN	พ.ร.บ.เงินทดแทน พ.ศ.2537 มาตรา 45 ให้ความสำคัญ...

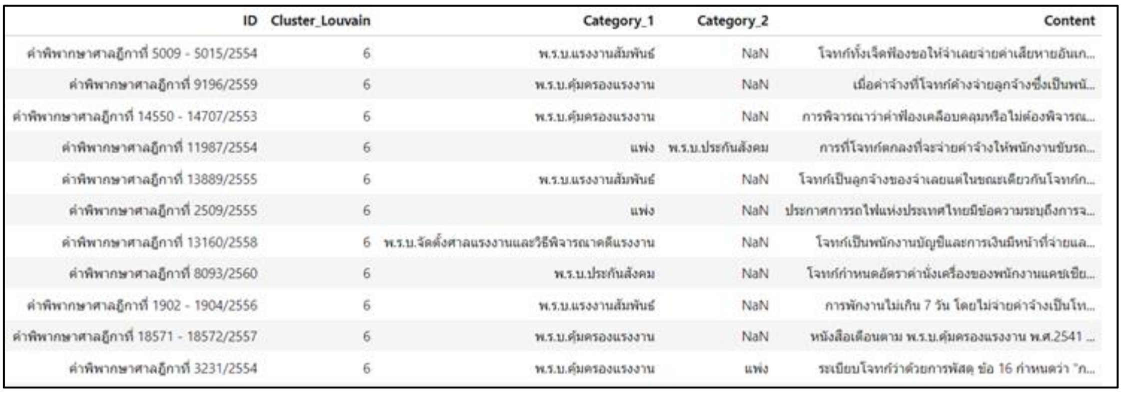
Figure 4.17 Sample of documents in Cluster 5

- Cluster 6

  Members: 235

  Dominant Category: Labor Protection Act (0.34), Civil and Commercial Code (0.22), Act on Establishment of Labor Court and Labor Court Procedure (0.18)

  This cluster comprised the case that was involved with the labor law. For example: the Labor Protection, Development, and Relations Act. This cluster could be interpreted as the cluster about the labor laws.



ID	Cluster_Louvain	Category_1	Category_2	Content
คำพิพากษาศาลฎีกาที่ 5009 - 5015/2554	6	พ.ร.บ.แรงงานสัมพันธ์	NaN	โจทก์ทั้งเจ็ดฟ้องขอให้จำเลยจ่ายค่าเสียหายอันเก...
คำพิพากษาศาลฎีกาที่ 9196/2559	6	พ.ร.บ.คุ้มครองแรงงาน	NaN	เมื่อค่าจ้างที่โจทก์ค้างจ่ายลูกจ้างซึ่งเป็นพนั...
คำพิพากษาศาลฎีกาที่ 14550 - 14707/2553	6	พ.ร.บ.คุ้มครองแรงงาน	NaN	การพิจารณาว่าคำฟ้องเคลือบคลุมหรือไม่ต้องพิจารณ...
คำพิพากษาศาลฎีกาที่ 11987/2554	6	แพ่ง	พ.ร.บ.ประกันสังคม	การที่โจทก์ตกลงที่จะจ่ายค่าจ้างให้พนักงานขับรถ...
คำพิพากษาศาลฎีกาที่ 13889/2555	6	พ.ร.บ.แรงงานสัมพันธ์	NaN	โจทก์เป็นลูกจ้างของจำเลยแต่ในขณะเดียวกันโจทก์ก...
คำพิพากษาศาลฎีกาที่ 2509/2555	6	แพ่ง	NaN	ประกาศการรถไฟแห่งประเทศไทยมีข้อความระบุถึงการจ...
คำพิพากษาศาลฎีกาที่ 13160/2558	6	พ.ร.บ.จัดตั้งศาลแรงงานและวิธีพิจารณาคดีแรงงาน	NaN	โจทก์เป็นพนักงานบัญชีและการเงินมีหน้าที่จ่ายแล...
คำพิพากษาศาลฎีกาที่ 8093/2560	6	พ.ร.บ.ประกันสังคม	NaN	โจทก์กำหนดอัตราค่านั่งเครื่องของพนักงานแคชเชีย...
คำพิพากษาศาลฎีกาที่ 1902 - 1904/2556	6	พ.ร.บ.แรงงานสัมพันธ์	NaN	การพักงานไม่เกิน 7 วัน โดยไม่จ่ายค่าจ้างเป็นโท...
คำพิพากษาศาลฎีกาที่ 18571 - 18572/2557	6	พ.ร.บ.คุ้มครองแรงงาน	NaN	หนังสือเตือนตาม พ.ร.บ.คุ้มครองแรงงาน พ.ศ.2541 ...
คำพิพากษาศาลฎีกาที่ 3231/2554	6	พ.ร.บ.คุ้มครองแรงงาน	แพ่ง	ระเบียบโจทก์ว่าด้วยการพัสดุ ข้อ 16 กำหนดว่า "ก...

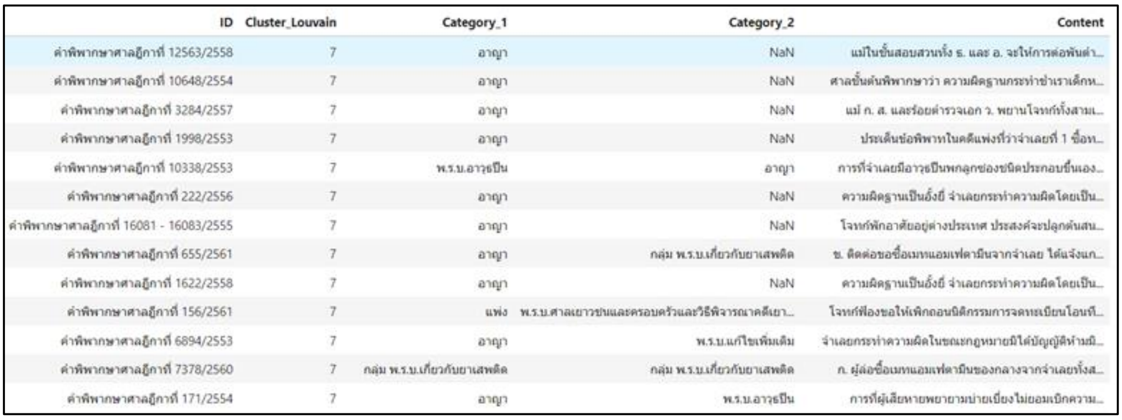
Figure 4.18 Sample of documents in Cluster 6

- Cluster 7

  Members: 1105

  Dominant Category: Criminal Code (0.65), Acts and Codes on Drugs (0.09), Civil and Commercial Code (0.03), Firearms Act (0.03)

  This cluster was similar to Cluster 2, but involved more variety in the code of law. For example: Radio and Telecommunication Act, Medical Professional Act and Copyright Act.



ID	Cluster_Louvain	Category_1	Category_2	Content
คำพิพากษาศาลฎีกาที่ 12563/2558	7	อาญา	NaN	แม้ในชั้นสอบสวนทั้ง ธ. และ อ. จะให้การต่อพันตำ…
คำพิพากษาศาลฎีกาที่ 10648/2554	7	อาญา	NaN	ศาลชั้นต้นพิพากษาว่า ความผิดฐานกระทำชำเราเด็กห…
คำพิพากษาศาลฎีกาที่ 3284/2557	7	อาญา	NaN	แม้ ก. ส. และร้อยตำรวจเอก ว. พยานโจทก์ทั้งสามเ…
คำพิพากษาศาลฎีกาที่ 1998/2553	7	อาญา	NaN	ประเด็นข้อพิพาทในคดีแพ่งที่ว่าจำเลยที่ 1 ซื้อท…
คำพิพากษาศาลฎีกาที่ 10338/2553	7	พ.ร.บ.อาวุธปืน	อาญา	การที่จำเลยมีอาวุธปืนพกลูกซองชนิดประกอบขึ้นเอง…
คำพิพากษาศาลฎีกาที่ 222/2556	7	อาญา	NaN	ความผิดฐานเป็นอั้งยี่ จำเลยกระทำความผิดโดยเป็น…
คำพิพากษาศาลฎีกาที่ 16081 - 16083/2555	7	อาญา	NaN	โจทก์พักอาศัยอยู่ต่างประเทศ ประสงค์จะปลูกต้นสน…
คำพิพากษาศาลฎีกาที่ 655/2561	7	อาญา	กลุ่ม พ.ร.บ.เกี่ยวกับยาเสพติด	ช. ติดต่อขอซื้อเมทแอมเฟตามีนจากจำเลย ได้แจ้งนก…
คำพิพากษาศาลฎีกาที่ 1622/2558	7	อาญา	NaN	ความผิดฐานเป็นอั้งยี่ จำเลยกระทำความผิดโดยเป็น…
คำพิพากษาศาลฎีกาที่ 156/2561	7	แพ่ง	พ.ร.บ.ศาลเยาวชนและครอบครัวและวิธีพิจารณาคดีเยา…	โจทก์ฟ้องขอให้เพิกถอนนิติกรรมการจดทะเบียนโอนที…
คำพิพากษาศาลฎีกาที่ 6894/2553	7	อาญา	พ.ร.บ.แก้ไขเพิ่มเติม	จำเลยกระทำความผิดในขณะกฎหมายมิได้บัญญัติห้ามมิ…
คำพิพากษาศาลฎีกาที่ 7378/2560	7	กลุ่ม พ.ร.บ.เกี่ยวกับยาเสพติด	กลุ่ม พ.ร.บ.เกี่ยวกับยาเสพติด	ก. ผู้ล่อซื้อเมทแอมเฟตามีนของกลางจากจำเลยทั้งส…
คำพิพากษาศาลฎีกาที่ 171/2554	7	อาญา	พ.ร.บ.อาวุธปืน	การที่ผู้เสียหายพยายามบ่ายเบี่ยงไม่ยอมเบิกความ…

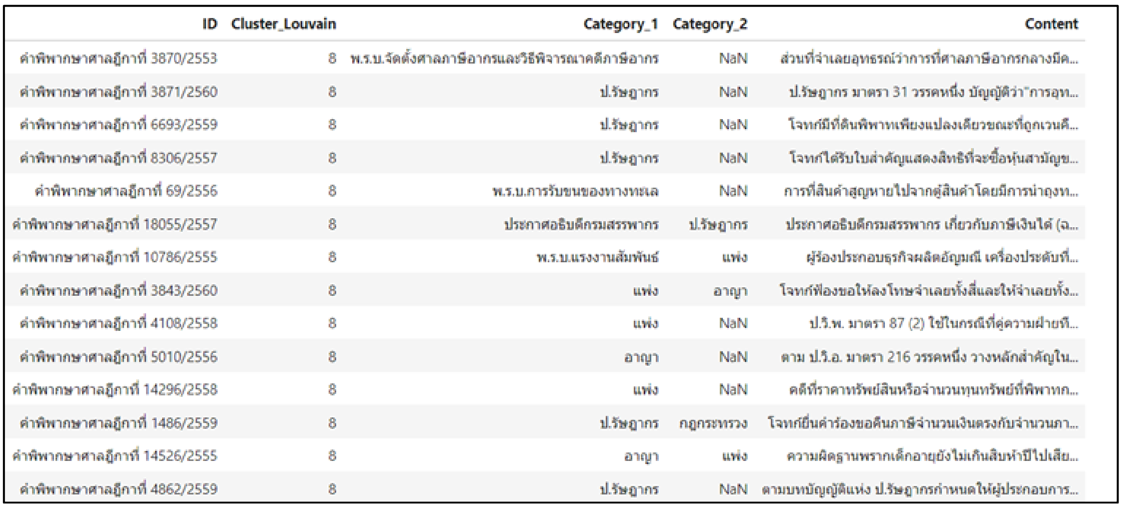
Figure 4.19 Sample of documents in Cluster 7

- Cluster 8

  Members: 755

  Dominant Category: Civil and Commercial Code (0.45), Criminal Code (0.20), Revenue Code (0.13)

  Documents in this cluster involved a legal article from the Revenue Code. It could be interpreted that this cluster represented cases involving revenue and taxes.



ID	Cluster_Louvain	Category_1	Category_2	Content
คำพิพากษาศาลฎีกาที่ 3870/2553	8	พ.ร.บ.จัดตั้งศาลภาษีอากรและวิธีพิจารณาคดีภาษีอากร	NaN	ส่วนที่จำเลยอุทธรณ์ว่าการที่ศาลภาษีอากรกลางมีค…
คำพิพากษาศาลฎีกาที่ 3871/2560	8	ป.รัษฎากร	NaN	ป.รัษฎากร มาตรา 31 วรรคหนึ่ง บัญญัติว่า"การอุท…
คำพิพากษาศาลฎีกาที่ 6693/2559	8	ป.รัษฎากร	NaN	โจทก์มีที่ดินพิพาทเพียงแปลงเดียวขณะที่ถูกเวนคื…
คำพิพากษาศาลฎีกาที่ 8306/2557	8	ป.รัษฎากร	NaN	โจทก์ได้รับใบสำคัญแสดงสิทธิที่จะซื้อหุ้นสามัญข…
คำพิพากษาศาลฎีกาที่ 69/2556	8	พ.ร.บ.การรับขนของทางทะเล	NaN	การที่สินค้าสูญหายไปจากตู้สินค้าโดยมีการนำถุงท…
คำพิพากษาศาลฎีกาที่ 18055/2557	8	ประกาศอธิบดีกรมสรรพากร	ป.รัษฎากร	ประกาศอธิบดีกรมสรรพากร เกี่ยวกับภาษีเงินได้ (ฉ…
คำพิพากษาศาลฎีกาที่ 10786/2555	8	พ.ร.บ.แรงงานสัมพันธ์	แพ่ง	ผู้ร้องประกอบธุรกิจผลิตอัญมณี เครื่องประดับที่…
คำพิพากษาศาลฎีกาที่ 3843/2560	8	แพ่ง	อาญา	โจทก์ฟ้องขอให้ลงโทษจำเลยทั้งสี่และให้จำเลยทั้ง…
คำพิพากษาศาลฎีกาที่ 4108/2558	8	แพ่ง	NaN	ป.วิ.พ. มาตรา 87 (2) ใช้ในกรณีที่คู่ความฝ่ายที่…
คำพิพากษาศาลฎีกาที่ 5010/2556	8	อาญา	NaN	ตาม ป.วิ.อ. มาตรา 216 วรรคหนึ่ง วางหลักสำคัญใน…
คำพิพากษาศาลฎีกาที่ 14296/2558	8	แพ่ง	NaN	คดีที่ราคาทรัพย์สินหรือจำนวนทุนทรัพย์ที่พิพาทก…
คำพิพากษาศาลฎีกาที่ 1486/2559	8	ป.รัษฎากร	กฎกระทรวง	โจทก์ยื่นคำร้องขอคืนภาษีจำนวนเงินตรงกับจำนวนภา…
คำพิพากษาศาลฎีกาที่ 14526/2555	8	อาญา	แพ่ง	ความผิดฐานพรากเด็กอายุยังไม่เกินสิบห้าปีไปเสีย…
คำพิพากษาศาลฎีกาที่ 4862/2559	8	ป.รัษฎากร	NaN	ตามบทบัญญัติแห่ง ป.รัษฎากรกำหนดให้ผู้ประกอบการ…

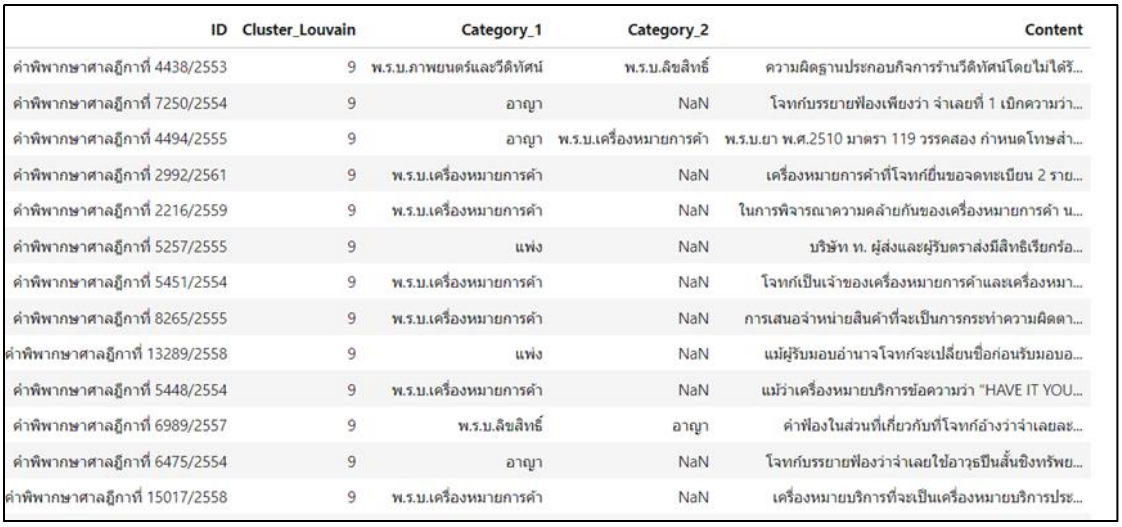
Figure 4.20 Sample of documents in Cluster 8

- Cluster 9

  Members: 428

  Dominant Category: Criminal Code (0.37), Trademark Act (0.16), Copyright Act (0.11)

  The legal articles involved in this cluster were mainly from the Trademarks Act. Thus, it could be concluded that this cluster represented cases about intellectual property.



ID	Cluster_Louvain	Category_1	Category_2	Content
คำพิพากษาศาลฎีกาที่ 4438/2553	9	พ.ร.บ.ภาพยนตร์และวีดิทัศน์	พ.ร.บ.ลิขสิทธิ์	ความผิดฐานประกอบกิจการร้านวีดิทัศน์โดยไม่ได้รั...
คำพิพากษาศาลฎีกาที่ 7250/2554	9	อาญา	NaN	โจทก์บรรยายฟ้องเพียงว่า จำเลยที่ 1 เบิกความว่า...
คำพิพากษาศาลฎีกาที่ 4494/2555	9	อาญา	พ.ร.บ.เครื่องหมายการค้า	พ.ร.บ.ยา พ.ศ.2510 มาตรา 119 วรรคสอง กำหนดโทษสำ...
คำพิพากษาศาลฎีกาที่ 2992/2561	9	พ.ร.บ.เครื่องหมายการค้า	NaN	เครื่องหมายการค้าที่โจทก์ยื่นขอจดทะเบียน 2 ราย...
คำพิพากษาศาลฎีกาที่ 2216/2559	9	พ.ร.บ.เครื่องหมายการค้า	NaN	ในการพิจารณาความคล้ายกันของเครื่องหมายการค้า น...
คำพิพากษาศาลฎีกาที่ 5257/2555	9	แพ่ง	NaN	บริษัท ท. ผู้ส่งและผู้รับตราส่งมีสิทธิเรียกร้อ...
คำพิพากษาศาลฎีกาที่ 5451/2554	9	พ.ร.บ.เครื่องหมายการค้า	NaN	โจทก์เป็นเจ้าของเครื่องหมายการค้าและเครื่องหมา...
คำพิพากษาศาลฎีกาที่ 8265/2555	9	พ.ร.บ.เครื่องหมายการค้า	NaN	การเสนอจำหน่ายสินค้าที่จะเป็นการกระทำความผิดตา...
คำพิพากษาศาลฎีกาที่ 13289/2558	9	แพ่ง	NaN	แม้ผู้รับมอบอำนาจโจทก์จะเปลี่ยนชื่อก่อนรับมอบอ...
คำพิพากษาศาลฎีกาที่ 5448/2554	9	พ.ร.บ.เครื่องหมายการค้า	NaN	แม้ว่าเครื่องหมายบริการข้อความว่า "HAVE IT YOU...
คำพิพากษาศาลฎีกาที่ 6989/2557	9	พ.ร.บ.ลิขสิทธิ์	อาญา	คำฟ้องในส่วนที่เกี่ยวกับที่โจทก์อ้างว่าจำเลยละ...
คำพิพากษาศาลฎีกาที่ 6475/2554	9	อาญา	NaN	โจทก์บรรยายฟ้องว่าจำเลยใช้อาวุธปืนสั้นชิงทรัพย...
คำพิพากษาศาลฎีกาที่ 15017/2558	9	พ.ร.บ.เครื่องหมายการค้า	NaN	เครื่องหมายบริการที่จะเป็นเครื่องหมายบริการประ...

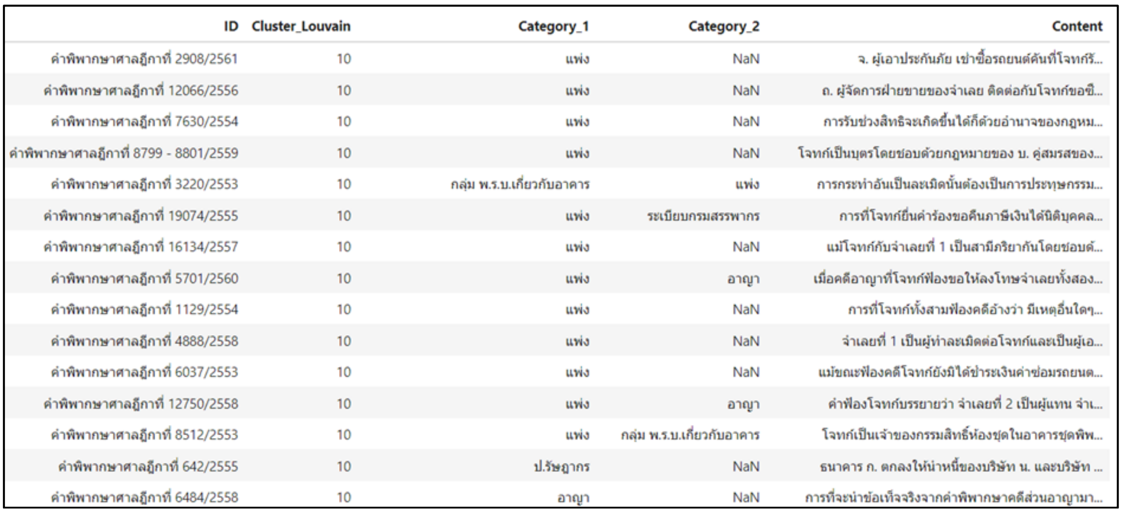
Figure 4.21 Sample of documents in Cluster 9

- Cluster 10

  Members: 691

  Dominant Category: Civil and Commercial Code (0.66), Criminal Code (0.08), Codes and Acts on Land and Settlement (0.03)

  This cluster involved mainly the Civil and Commercial Code, but there was an ambiguity in the subcategory of the Code. Hence, it could not be concluded which category of the Civil and Commercial Code that this cluster represented.



ID	Cluster_Louvain	Category_1	Category_2	Content
คำพิพากษาศาลฎีกาที่ 2908/2561	10	แพ่ง	NaN	จ. ผู้เอาประกันภัย เช่าซื้อรถยนต์คันที่โจทก์รั...
คำพิพากษาศาลฎีกาที่ 12066/2556	10	แพ่ง	NaN	ถ. ผู้จัดการฝ่ายขายของจำเลย ติดต่อกับโจทก์ขอซื...
คำพิพากษาศาลฎีกาที่ 7630/2554	10	แพ่ง	NaN	การรับช่วงสิทธิจะเกิดขึ้นได้ก็ด้วยอำนาจของกฎหม...
คำพิพากษาศาลฎีกาที่ 8799 - 8801/2559	10	แพ่ง	NaN	โจทก์เป็นบุตรโดยชอบด้วยกฎหมายของ บ. คู่สมรสของ...
คำพิพากษาศาลฎีกาที่ 3220/2553	10	กลุ่ม พ.ร.บ.เกี่ยวกับอาคาร	แพ่ง	การกระทำอันเป็นละเมิดนั้นต้องเป็นการประทุษกรรม...
คำพิพากษาศาลฎีกาที่ 19074/2555	10	แพ่ง	ระเบียบกรมสรรพากร	การที่โจทก์ยื่นคำร้องขอคืนภาษีเงินได้นิติบุคคล...
คำพิพากษาศาลฎีกาที่ 16134/2557	10	แพ่ง	NaN	แม้โจทก์กับจำเลยที่ 1 เป็นสามีภริยากันโดยชอบด้...
คำพิพากษาศาลฎีกาที่ 5701/2560	10	แพ่ง	อาญา	เมื่อคดีอาญาที่โจทก์ฟ้องขอให้ลงโทษจำเลยทั้งสอง...
คำพิพากษาศาลฎีกาที่ 1129/2554	10	แพ่ง	NaN	การที่โจทก์ทั้งสามฟ้องคดีอ้างว่า มีเหตุอื่นใดๆ...
คำพิพากษาศาลฎีกาที่ 4888/2558	10	แพ่ง	NaN	จำเลยที่ 1 เป็นผู้ทำละเมิดต่อโจทก์และเป็นผู้เอ...
คำพิพากษาศาลฎีกาที่ 6037/2553	10	แพ่ง	NaN	แม้ขณะฟ้องคดีโจทก์ยังมิได้ชำระเงินค่าซ่อมรถยนต...
คำพิพากษาศาลฎีกาที่ 12750/2558	10	แพ่ง	อาญา	คำฟ้องโจทก์บรรยายว่า จำเลยที่ 2 เป็นผู้แทน จำเ...
คำพิพากษาศาลฎีกาที่ 8512/2553	10	แพ่ง	กลุ่ม พ.ร.บ.เกี่ยวกับอาคาร	โจทก์เป็นเจ้าของกรรมสิทธิ์ห้องชุดในอาคารชุดพิพ...
คำพิพากษาศาลฎีกาที่ 642/2555	10	ป.รัษฎากร	NaN	ธนาคาร ก. ตกลงให้นำหนี้ของบริษัท น. และบริษัท ...
คำพิพากษาศาลฎีกาที่ 6484/2558	10	อาญา	NaN	การที่จะนำข้อเท็จจริงจากคำพิพากษาคดีส่วนอาญามา...

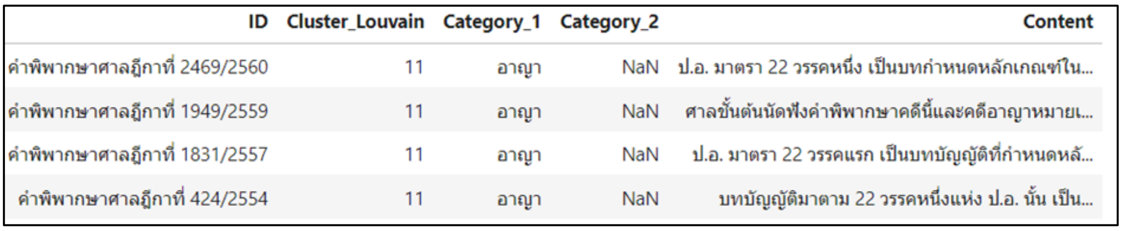
Figure 4.22 Sample of documents in Cluster 10

- Cluster 11

  Members: 4

  Dominant Category: Criminal Code (1)

  This cluster consisted of four cases that involved only one legal article: Article 22 of the Criminal code. No other legal articles were involved, so they were grouped together.



ID	Cluster_Louvain	Category_1	Category_2	Content
คำพิพากษาศาลฎีกาที่ 2469/2560	11	อาญา	NaN	ป.อ. มาตรา 22 วรรคหนึ่ง เป็นบทกำหนดหลักเกณฑ์ใน...
คำพิพากษาศาลฎีกาที่ 1949/2559	11	อาญา	NaN	ศาลชั้นต้นนัดฟังคำพิพากษาคดีนี้และคดีอาญาหมายเ...
คำพิพากษาศาลฎีกาที่ 1831/2557	11	อาญา	NaN	ป.อ. มาตรา 22 วรรคแรก เป็นบทบัญญัติที่กำหนดหลั...
คำพิพากษาศาลฎีกาที่ 424/2554	11	อาญา	NaN	บทบัญญัติมาตาม 22 วรรคหนึ่งแห่ง ป.อ. นั้น เป็น...

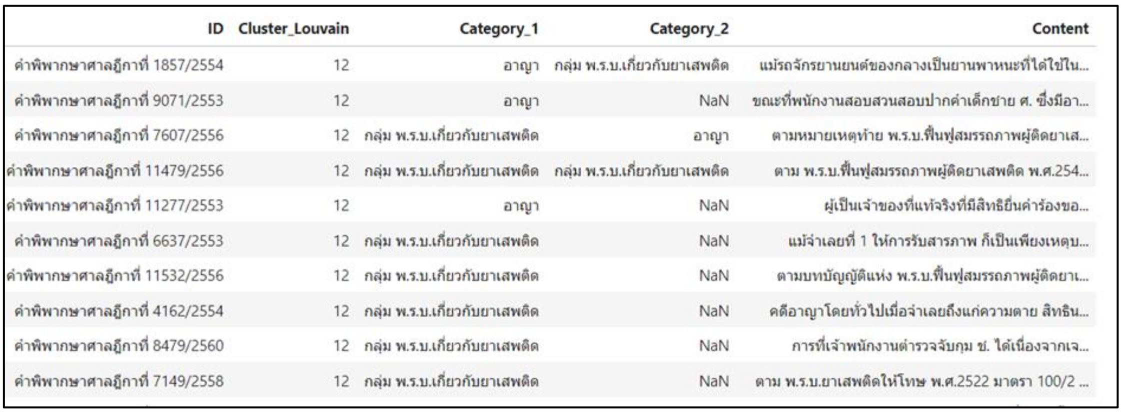
Figure 4.23 Sample of documents in Cluster 11

- Cluster 12

  Members: 318

  Dominant Category: Codes and Acts on Drugs (0.52), Criminal Code (0.33), Road Traffic Act (0.04)

  The cluster emphasized cases with the drugs trade and uses. The cluster was comparable with Cluster 2, which also had a high fraction of Codes and Acts on Drugs.



ID	Cluster_Louvain	Category_1	Category_2	Content
คำพิพากษาศาลฎีกาที่ 1857/2554	12	อาญา	กลุ่ม พ.ร.บ.เกี่ยวกับยาเสพติด	แม้รถจักรยานยนต์ของกลางเป็นยานพาหนะที่ได้ใช้ใน...
คำพิพากษาศาลฎีกาที่ 9071/2553	12	อาญา	NaN	ขณะที่พนักงานสอบสวนสอบปากคำเด็กชาย ศ. ซึ่งมีอา...
คำพิพากษาศาลฎีกาที่ 7607/2556	12	กลุ่ม พ.ร.บ.เกี่ยวกับยาเสพติด	อาญา	ตามหมายเหตุท้าย พ.ร.บ.ฟื้นฟูสมรรถภาพผู้ติดยาเส...
คำพิพากษาศาลฎีกาที่ 11479/2556	12	กลุ่ม พ.ร.บ.เกี่ยวกับยาเสพติด	กลุ่ม พ.ร.บ.เกี่ยวกับยาเสพติด	ตาม พ.ร.บ.ฟื้นฟูสมรรถภาพผู้ติดยาเสพติด พ.ศ.254...
คำพิพากษาศาลฎีกาที่ 11277/2553	12	อาญา	NaN	ผู้เป็นเจ้าของที่แท้จริงที่มีสิทธิยื่นคำร้องขอ...
คำพิพากษาศาลฎีกาที่ 6637/2553	12	กลุ่ม พ.ร.บ.เกี่ยวกับยาเสพติด	NaN	แม้จำเลยที่ 1 ให้การรับสารภาพ ก็เป็นเพียงเหตุบ...
คำพิพากษาศาลฎีกาที่ 11532/2556	12	กลุ่ม พ.ร.บ.เกี่ยวกับยาเสพติด	NaN	ตามบทบัญญัติแห่ง พ.ร.บ.ฟื้นฟูสมรรถภาพผู้ติดยาเ...
คำพิพากษาศาลฎีกาที่ 4162/2554	12	กลุ่ม พ.ร.บ.เกี่ยวกับยาเสพติด	NaN	คดีอาญาโดยทั่วไปเมื่อจำเลยถึงแก่ความตาย สิทธิน...
คำพิพากษาศาลฎีกาที่ 8479/2560	12	กลุ่ม พ.ร.บ.เกี่ยวกับยาเสพติด	NaN	การที่เจ้าพนักงานตำรวจจับกุม ช. ได้เนื่องจากเจ...
คำพิพากษาศาลฎีกาที่ 7149/2558	12	กลุ่ม พ.ร.บ.เกี่ยวกับยาเสพติด	NaN	ตาม พ.ร.บ.ยาเสพติดให้โทษ พ.ศ.2522 มาตรา 100/2 ...

Figure 4.24 Sample of documents in Cluster 12

4.2 Integration with Document Retrieval System

After acquiring the network information for a corpus from the previous section, this information could be used to create a document retrieval system. As discussed in Section 3.5, three document retrieval systems were built: one based on the text features and two network-based systems.

An entropy metric, as stated in Section 3.6.1, was utilized to assess the performance. One thousand documents were selected at random from the corpus and each system was used to find the 100 most similar documents to the one that was selected. Then, the entropy for each result set was computed based on the category to which it belonged. For system three, the cluster features were incorporated from the Louvain method. The experiment was conducted on a personal computer with the following specifications: AMD Ryzen 5 5600X CPU, 32 GB DDR4 Ram. The results are presented in Table 4.2.

Table 4.2 Document Retrieval System Metrics

ID	Name	Entropy Average	Entropy std.	Runtime (S)
1	Text Features	0.526	0.077	5.13
2	Network Connectivity	0.492	0.073	7.26
3	Network Clustering + Text Features	0.472	0.091	9.09

As a result of incorporating the network features into the document retrieval system, an increase in the performance of retrieving a document that was in the same category of the initial document could be observed. However, this came at the cost of additional computation time. This also confirmed the expected benefits in Section 1.4 that using network features in a document retrieval system would improve the performance of the system.

In the next section, we present a conclusion of the study, discussion on the result as well as a suggestion that may benefits future works.

CHAPTER 5 CONCLUSION AND SUGGESTION

5.1 Conclusion

This study focused on analyzing the topological structure of the Supreme Court of Thailand's judgment dataset, as well as evaluated the effects of using different methods of analysis to investigate the performance of the information retrieval (IR) system. The study acquired the dataset from the source and preprocessed it into a suitable format for each type of analysis. Then, an IR system was constructed using different features, including text features and network features from multiple approaches. The performance of the system was then measured and compared to find the most effective settings according to the study.

From the study, the performance improvement in the document retrieval system after integrating document clustering and network features was observed. The results concurred with a previous study on the topic. This study also confirmed that using organic network features that were already present in the dataset were also capable of improving the performance similar to a previous study on synthetic network features.

5.2 Discussion

In this study, two different graph clustering algorithms that were spectral clustering and the Louvain method were examined. The results, however, yielded a dissimilarity in both considered metrics (entropy and modularity). From the cluster induction analysis, it could be inferred that spectral clustering was not suitable with a densely connected graph. This was because this network consisted of 6,670 nodes, but the average shortest path length was only 2.97. Additionally, the network's median degree of connectivity was 50. This signified that 50% of the node's neighbors were 50 or more. As a result, this network was densely connected.

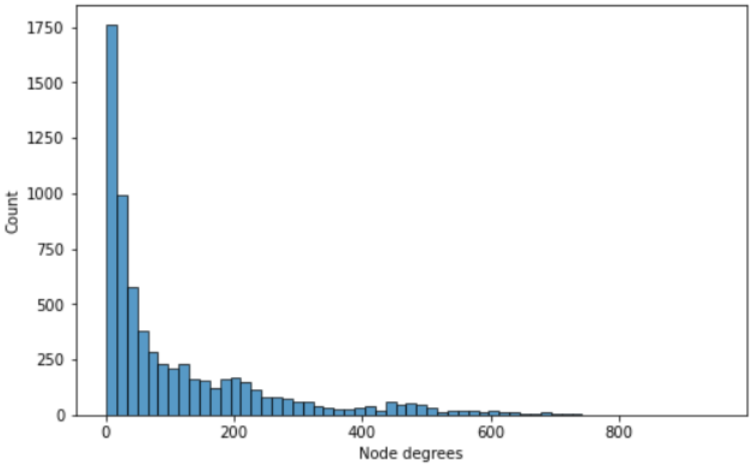
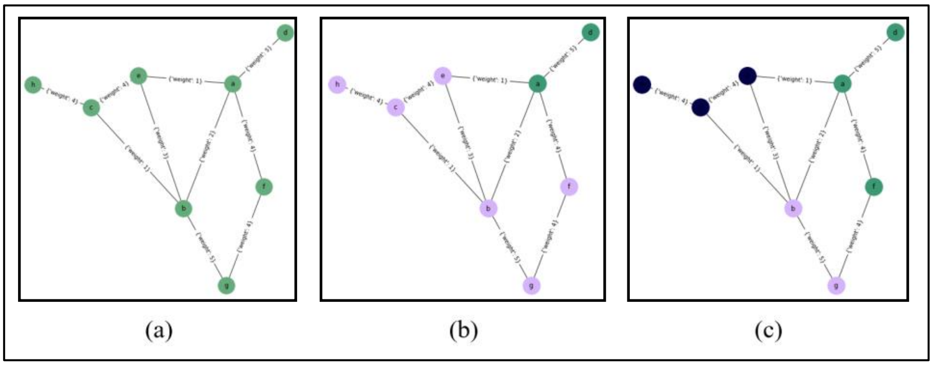


Figure 5.1 A degree distribution of node in the network

Another topic was the behavior of spectral clustering. It was observed that clustering always yielded an over dominated cluster when using a lower k as pictured in Figure 4.7 and 4.8. This was the result of the behavior of spectral clustering called minimal cut that in some cases resulted in a non-optimal solution. Figure 5.3 illustrates the comparison of using spectral clustering and the Louvain method on a simple graph.



(a) (b) (c)

Figure 5.2 A comparison of different graph clustering methods
(a) Original graph
(b) Spectral Clustering with K=2
(c) Louvain method

From Figure 5.3, it could be seen that spectral clustering chose to cut the edges *ab*(2), *ae*(1) and *af*(4) rather than cut *ae*(1), *be*(3) and *bc*(1), which had less total weight. On the other hand, the Louvain method was not affected by this behavior because it used modularity optimization.

There was also a study that raised a concern about the performance of the Louvain method. Traag et al. [37] discussed that the Louvain method could yield badly connected or in the worst case, disconnected communities. The study also proposed an algorithm that could address this problem called the Leiden method, which ascribed that it had better performance in both uncovering the communities and the runtime.

Finally, this study forced the network into an ordinary graph in the representation of each judgment in the dataset. In a real-world setting, however, each legal article could function as a "hyperedge" by connecting more than two nodes. This resulted in making a "hypergraph" as the suitable representation of this network. A hypergraph analysis could thus be a viable method for extracting information from this corpus.

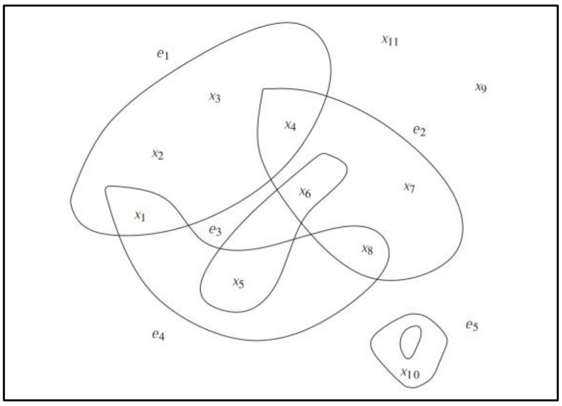


Figure 5.3 An example of Hypergraph with 11 vertices and 5 edges. A hyperedges (e.g. e_1, e_2) connects to more than 2 vertices. [38]

5.3 Suggestion and Future Works

Potential research that could be developed from this study could include a comparison with other graph clustering methods to investigate the behavior and performance of different algorithms in document clustering. A comparison could also be extended to compare a graph-based approach with text feature-based clustering to examine the effect of using different techniques of feature extraction.

As described in the work, this study modeled the network structure by using court judgments-court judgment relationship through referenced legal articles. This was undertaken by taking different aspects, so the dataset could be modeled into a legal article-legal article relationship through the court judgments. This enabled an inspection of the connection between legal articles that were co-referenced within the court judgments.

Ultimately, one could take an approach from a hypergraph to analyze the network and topological structure of this dataset, as the hypergraph maybe more suitable in its representation according to the legal article that would act as a hyperedge.

5.4 Publication

This work has been published in the 2020 1st International Conference on Big Data Analytics and Practices (IBDAP) with the title of “On the Network and Topological Analyses of Legal Documents using Text Mining Approach”. Full manuscript can be found at https://doi.org/10.1109/IBDAP50342.2020.9245615.

REFERENCES

1. Roser, M., Ritchie, H., and Ortiz-Ospina, E., 2015, **Internet**, [Online], Available: https://ourworldindata.org/internet [2021, October 12].

2. Johnson, J., 2021, **Global Digital Population as of January 2021**, [Online], Available: https://www.statista.com/statistics/617136/digital-population-worldwide/ [2021, October 12].

3. Ramalingam, D., 2014, "Analysis on Big Data over the Years", **International Journal of Scientific and Research Publications**, Vol. 4, No. 1, pp. 1-7.

4. Desjardins, J., 2019, **How Much Data Is Generated Each Day?**, [Online], Available: https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/ [2021, October 13].

5. Gasevic, D., Kovanovic, V., Joksimovic, S., and Siemens, G., 2014, "Where Is Research on Massive Open Online Courses Headed? A Data Analysis of the Mooc Research Initiative", **The International Review of Research in Open and Distributed Learning**, Vol. 15, No. 5, pp. 134-176.

6. Stanchev, L.,2014. "Creating a Similarity Graph from Wordnet". **4th International Conference on Web Intelligence, Mining and Semantics**, 2 - 4 June, Thessaloniki, Greece, pp., 1–11.

7. Stanchev, L.,2016. "Semantic Document Clustering Using a Similarity Graph". **IEEE Tenth International Conference on Semantic Computing**, 4-6 Febuary, California, USA, pp., 1-8.

8. Information Technology and Communication Center of Supreme Court of Thailand, 2020, ระบบสืบค้นคำพิพากษา คำสั่งคำร้องและคำวินิจฉัยศาลฎีกา, [Online], Available: http://deka.supremecourt.or.th [2020, September 10].

9. Chomsky, N., 1957, **Syntactic Structures**, Mouton, Oxford, England, Pages.

10. Manning, C.D., Raghavan, P., and Schütze, H., 2009, "Boolean Retrieval", In **An Introduction to Information Retrieval**, Cambridge University Press, Cambridge, pp 1-2.

11. Page, L., Brin, S., Motwani, R., and Winograd, T., 1999, The Pagerank Citation Ranking: Bringing Order to the Web. Stanford InfoLab, Vol., pp.

12. Van Rijsbergen, C.J., 1975, "Search Strategies", In **Information Retrieval**, Butterworths, London, pp 68-77.

13. Ali, I. and Melton, A.,2018. "Semantic-Based Text Document Clustering Using Cognitive Semantic Learning and Graph Theory". **IEEE 12th International Conference on Semantic Computing**, 31 January - 2 Febuary, California, USA, pp., 243-247.

14. Yoo, I. and Hu, X.,2006. "Clustering Ontology-Enriched Graph Representation for Biomedical Documents Based on Scale-Free Network Theory". **3rd International IEEE Conference Intelligent Systems**, 4-6 September, London, UK, pp., 851-858.

15. Kowsrihawat, K. and Vateekul, P.,2015. "An Information Extraction Framework for Legal Documents: A Case Study of Thai Supreme Court Verdicts". **2015 12th International Joint Conference on Computer Science and Software Engineering (JCSSE)**, 22-24 July 2015, pp., 275-280.

16. Zanini, N. and Dhawan, V., 2015, **Text Mining: An Introduction to Theory and Some Applications**, [Online], Available: https://www.cambridgeassessment.org.uk/Images/466185-text-mining-an-introduction-to-theory-and-some-applications-.pdf

17. Talib, R., Kashif, M., Ayesha, S., and Fatima, F., 2016, "Text Mining: Techniques, Applications and Issues", **International Journal of Advanced Computer Science and Applications**, Vol. 7, No. 11, pp. 414-418.

18. Soergel, D., 2004, **Information Retrieval**, [Online], Available: https://www.researchgate.net/publication/308874065_Information_Retrieval [2021, October 19].

19. Bonaccorso, G., 2017, "Introduction to Natural Language Processing", In **Machine Learning Algorithms**, Packt, Birmingham, pp 242-260.

20. Bengfort, B., Bilbro, R., and Ojeda, T., 2018, "Machine Learning on Text", In **Applied Text Analysis with Python**, O'Reilly Media, California, pp 41-80.

21. Phatthiyaphaibun, W., Chaovavanich, K., Polpanumas, C., Suriyawongkul, A., and Lowphansirikul, L., 2016, Pythainlp: Thai Natural Language Processing in Python. Zenodo, Vol., pp.

22. Dangeti, P., 2017, "Cosine Similarity", In **Statistics for Machine Learning**, Packt, Birmingham, pp 281-282.

23. Manning, C.D., Raghavan, P., and Schütze, H., 2009, "Flat Clustering", In **An Introduction to Information Retrieval**, Cambridge University Press, Cambridge, pp 349-372.

24. Bonaccorso, G., 2017, "K-Means", In **Machine Learning Algorithms**, Packt, Birmingham, pp 183-198.

25. Barabási, A.-L., 2015, "Networks and Graphs", In **Network Science**, Cambridge University Press, Cambridge, pp 32-35.

26. National Research Council., 2005, "The Definition and Promise of Network Science", In **Network Science**, The National Academies Press, Washington, DC, pp 26-29.

27. Börner, K., Sanyal, S., and Vespignani, A., 2007, "Network Science", In **Annual Review of Information Science & Technology**. Cronin, B., Information Today, Inc., New Jersey, pp 537-607.

28. Kolb, L., Sehili, Z., and Rahm, E., 2014, "Iterative Computation of Connected Graph Components with Mapreduce", **Datenbank-Spektrum**, Vol. 14, No. 2, pp. 107-117.

29. Zelnik-Manor, L. and Perona, P., 2004, Self-Tuning Spectral Clustering. **Proceedings of the 17th International Conference on Neural Information Processing Systems**. MIT Press, Vancouver, British Columbia, Canada, Vol., pp. 1601–1608.

30. Blondel, V., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E., 2008, "Fast Unfolding of Communities in Large Networks", **Journal of Statistical Mechanics Theory and Experiment**, Vol. 2008, No. 10, p. P10008.

31. Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z., and Wagner, D., 2008, "On Modularity Clustering", **IEEE Transactions on Knowledge and Data Engineering**, Vol. 20, No. 2, pp. 172-188.

32. Diestel, R., 2000, "The Basics", In **Graph Theory**, Springer-Verlag, New York, pp 3-4.

33. Shalabh, **Indicator Variables**, [Online], Available: http://home.iitk.ac.in/~shalab/regression/Chapter8-Regression-IndicatorVariables.pdf [2021, December 1].

34. Shannon, C.E., 1948, "A Mathematical Theory of Communication", **The Bell System Technical Journal**, Vol. 27, No. 3, pp. 379-423.

35. Hagberg, A.A., Schult, D.A., and Swart, P.J.,2008. "Exploring Network Structure, Dynamics, and Function Using Networkx". **Proceedings of the 7th Python in Science conference (SciPy 2008)**, 19-24 August, California, USA, pp., 11-15.

36. Aynaud, T., 2018, **Louvain Community Detection**, [Online], Available: https://github.com/taynaud/python-louvain [November 20].

37. Traag, V.A., Waltman, L., and van Eck, N.J., 2019, "From Louvain to Leiden: Guaranteeing Well-Connected Communities", **Scientific Reports**, Vol. 9, No. 1, p. 5233.

38. Bretto, A., 2013, "Hypergraphs: Basic Concepts", In **Hypergraph Theory: An Introduction**, Springer International Publishing, Switzerland, pp 1-21.

CURRICULUM VITAE

NAME	Mr. Supawit Somsakul
DATE OF BIRTH	24 October 1996

EDUCATIONAL RECORD

HIGH SCHOOL	High School Graduation Saard Phaderm Wittaya School, 2014
BACHELOR’S DEGREE	Bachelor of Engineering (Computer Engineering) King Mongkut's University of Technology Thonburi, 2017
MASTER’S DEGREE	Master of Engineering (Computer Engineering) King Mongkut's University of Technology Thonburi, 2021

SCHOLARSHIP / RESEARCH GRANT	BX Scholarship
PUBLICATION	Somsakul, S. and Prom-on, S., 2020. "On the Network and Topological Analyses of Legal Documents Using Text Mining Approach". **2020 1st International Conference on Big Data Analytics and Practices (IBDAP)**, 25-26 Sept. 2020, Bangkok, pp., 1-6