COMPUTATIONAL ANALYSIS OF LAW

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Master of Science

by

Zhihao Liu, Yuxi Sun

May 2020

ABSTRACT

With the development of technology in a variety of aspects, combining legal analysis with computational research becomes one of the focuses in legal studies. In this year-long specialization project, we study the United States Code with a computational approach, where we build a legal search engine that retrieves relevant sections using multiple signals, extract properties and produce visualization of the citation network, and apply clustering analysis to detect sub-structures of the internal entities.

Keywords: Legal analysis; Legal search; Citation network

ACKNOWLEDGEMENTS

We express our deep and sincere gratitude to our advisor, Professor James Grimmelmann at Cornell Tech and Cornell Law School, for providing invaluable guidance throughout the research. It was a tremendous privilege and honor to work with him, without whom the project will not be complete and successful.

CHAPTER 1

INTRODUCTION

For centuries, law has been an information-intensive profession, where lawyers and information scientists have always pushed against the limits of cutting-edge information retrieval technology. A number of commercial tools have been developed to retrieve and analyze legal information. These tools, designed for legal professionals instead of scientific researchers, however, lack the extensibility to perform in-depth and customized analysis.

The U.S. Code is a consolidation and codification of the general and permanent laws of the United States. It is organized into titles (the number of which varies among different amended versions), each defining a particular substantive domain. Within a title, text is structured as sections, subsections, and clauses. Such structure, however, is loosely defined and not consistent between titles. Normally, sections are considered the finest representation of a cohesive entity of legal information.

We build a legal search engine that allows the user to search sections in the U.S. Code based on multiple signals—keyword occurrence, in-degree, and PageRank. The last two signals are computed using the citation network and measure the internal significance of a certain section. For the citation network, We extract interesting properties by replicating the work of Bommarito and Katz[5], and produce a visualization where each section is represented as a node and each citation as an arc. Provided both static information embedded in the text and dynamic information introduced by the citation network, we apply clustering analysis at different scales—sections, titles, and chapters—to learn the relations between these internal entities of the U.S. Code.

The main contributions of this work are: (1) an efficient and extensible search engine for retrieving information from the U.S. Code; (2) properties and a visualization that provide insights of the citation network; (3) a series of clustering analyses that explore the relations between internal entities of the U.S. Code.

CHAPTER 2
RELATED WORK

There have been approaches for enhancing information retrieval technology in the legal domain. Turtle[9] provided an introduction to text retrieval in the legal domain. Bing used expert systems technology to help users pose requests to standard information retrieval environments[4]. Benjamins et al.[3] explored the legal ontologies and methodologies in legal information retrieval. Rose and Belew developed SCALIR, a legal information retrieval system using combination of symbolic and connectionist artificial intelligence techniques.

Citation networks are also an important subject in legal studies. Bommarito and Katz[5] investigated the properties of the U.S. Code citation network by examining the directed degree distributions. Zhang and Koppaka[10] discussed the use of semantics-based citation networks in a new legal research tool. Regarding network visualization, Batagelj and Mrvar[2] designed a program named Pajek to support abstraction by recursive decomposition of a large network into several smaller networks.

In the field of legal analysis, Ashley[1] described how new legal applications based on artificial intelligence will change the practice of law. Specifically for clustering analysis on legal information, Bommarito et al.[6] developed a distance measure for dynamic citation networks and Lu et al.[8] introduced a

classification-based recursive soft clustering algorithm with built-in topic segmentation.

CHAPTER 3
APPROACH

3.1 Search Engine

The search engine is a web application composed of several computational modules and a user interface. The computational modules developed in Python perform the core search logic and the user interface is built with HTML and JavaScript. We use the October 2019 version of the U.S. Code as our data source, acquired from the website of the Office of the Law Revision Counsel (http://uscode.house.gov). The data are serialized into a set of XML files, each corresponding to a title in the U.S. Code. The content and the XML of an example section are shown in Figures 3.1 and 3.2 respectively. We parse these XML files, excluding appendices, with Python's ElementTree API and then preprocess the data into a single JSON file for fast loading. The JSON structure is shown in Figure 3.3.

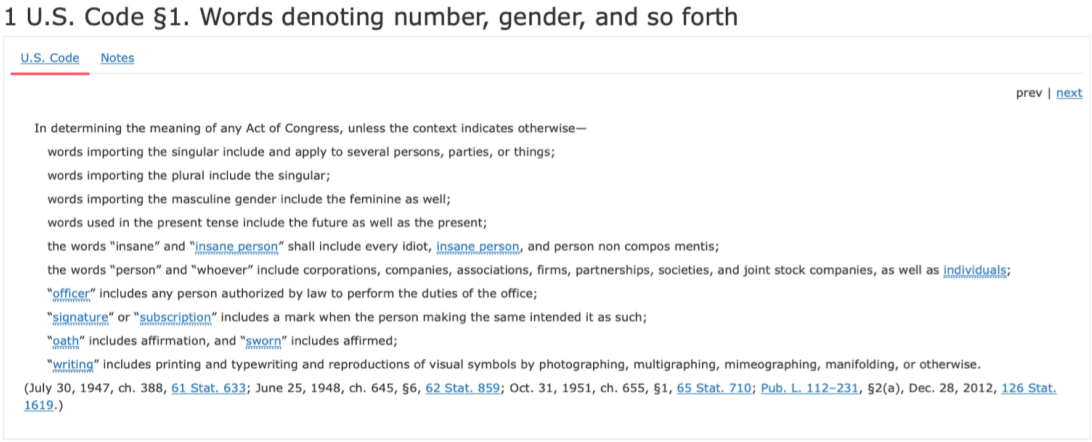

Figure 3.1: Content of a section

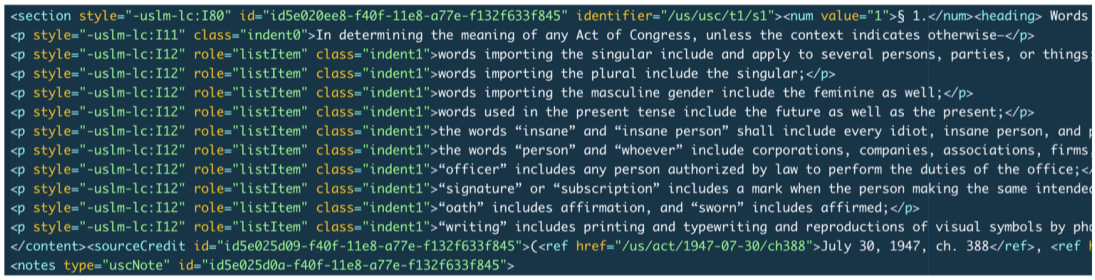

Figure 3.2: XML of a section

```
ROOT: {
    'titles': mapping from TITLE_ID (string) to TITLE (object)
}

TITLE: {
    'id': TITLE_ID (string),
    'sections': mapping from SECTION_ID (string) to SECTION (object)
}

SECTION: {
    'id': SECTION_ID (string),
    'text': TEXT_CONTENT (string),
    // count the number of occurrences of a term
    'terms': mapping from TERM (string) to TERM_COUNT (int),
    // count the number of references to a section
    'refs': mapping from REFERENCE_ID (string) to REFERENCE_COUNT (int)
}
```

Figure 3.3: JSON structure of the data

With the preprocessed data, we construct the citation network of the U.S. Code. The citation network is represented as a weighted directed graph using the NetworkX library. The citation network is both a component for ranking search results and a separate subject for subsequent analyses.

The search engine supports three search modes. Find-by-ID looks up the title number and the section number to find an exact section. Full-text Search iterates through every section and adds it to the results list if it contains the keyword. It takes constant time to check a single section because the preprocessed data store each section as a term counter. Boolean Search first uses the boolean.py library to parse the input query as a Boolean expression composed of keywords and Boolean operators, and then works in a similar way to Full-text Search, except that for each section, instead of checking a single keyword, it checks all the involved keywords and evaluates the Boolean expression accordingly.

Another important process is ranking multiple results returned by Full-text Search or Boolean Search. We take into account three signals for a given section—the number of times a keyword occurs in the section, its in-degree in the citation network, and its PageRank value computed using the citation network. We provide the customizability with which developers can rank the search results by a linear combination of different signals based on weights defined according to their requirements.

On the front end, we have three pages——the search portal, the results page, and the section detail page. The entire user flow of a search request is described as follows. Firstly, the user inputs a query in the search portal, which is passed to the back end through the submission of an HTML form. Secondly, the results are returned by the back-end logic and passed to the results page, where each result is displayed with the title number, the section number, and the content preview. Finally, the user may click on a result on the results page, which triggers the detail page that shows the complete content of the section.

3.2 Properties and Visualization of Citation Network

A citation network is the internal structure of a corpus represented as a weighted directed graph. It is an important concept for corpus with strong internal relations between the documents, of which the U.S. Code is one example. In our case, a vertex represents a section, an arc represents citations from the tail section to the head section, and the weight of an arc is the count of such citations. For instance, an arc from u to v with weight w means that there are w references to Section v within Section u. Note that, since our work is primarily

at the section level, we consider a reference to a sub-entity of a section, e.g. a subsection or a clause, equivalent to a reference to the entire section itself. References to content outside the U.S. Code, e.g. regulations or supreme court cases, are simply ignored.

We extract the following properties from the citation network as similar to the work of Bommarito and Katz[5].

1. Number of vertices: It should match the total number of individual sections in the U.S. Code.
2. Number of arcs: It equals the total number of distinct citations in the U.S. Code.
3. Accumulated weight of all the arcs: It equals the total number of instances of citations within the U.S. Code.
4. Log-log in-/out-degree distributions of all the sections: They provide overall insights to the network.
5. Sections with highest in-/out-degree: Sections with high in-degree are important concepts that are referenced a lot, while sections with high out-degree are complex ideas that involve concepts from a variety of other sections.

We also produce an interactive visualization of the citation network as an additional module of the web application in order that the user can have a quick overview of the citation network structure. In the visualization, a vertex is represented as a circle and an arc as a straight arrow. The user can interact with the network by hovering the cursor over a vertex, at which point a translucent pop-up will appear with the information of the corresponding section.

3.3 Clustering Analysis

For a set of entities, clustering analysis groups them in such a way that entities in the same group are more similar to one another than to those in other groups. The U.S. Code is manually structured into a hierarchy, where smaller entities are grouped into larger entities at each level, e.g. sections are grouped into titles. It is natural to wonder whether such grouping makes sense from a mathematical perspective. In other words, sections in the same title are considered relevant to the same topic by humans, but are they really similar to each other on the mathematical aspect?

We apply a series of hierarchical clusterings on sections and on titles using the following different distance measures:

1. Vectorization distance: It is the euclidean distance between the vector representations of two entities. The vector representation of an entity is a term counter, where each dimension corresponds to a specific term and the value in that dimension is the number of times that term occurs in the text of the entity.
2. Citation-based distance: It considers the number of shared citations between two entities and two entities are closer if they have more citations in common. The distance between entities a and b is calculated as

$$D(a, b) = 1 - \frac{I(a, b)}{U(a, b)}$$

where $I(a, b)$ is the number of shared citations (intersection) between a and b and $U(a, b)$ is the number of total citations (union) in a and b. It is easy to see that this distance measure ranges from 0 to 1, where 1 means no shared citation at all and 0 means two entities having exactly the same citations.

3. Sink-based distance[6]: It is similar to citation-based distance, except for that it is based on shared sinks instead of shared citations. In graph theory, a sink is a vertex whose out-degree is 0. In a citation network, a sink has no citations to other documents and is considered "novel ideas" that depend on nothing else.

The vectorization distance is considered a static measure because it only depends on the text in an entity. The citation-based distance and the sink-based distance are considered dynamic measures because they take into account the dynamic structure of the citation network. The results of hierarchical clustering are illustrated as dendrograms.

We compare the results of different clustering approaches to each other as well as to the original structure of the U.S. Code using the Fowlkes-Mallows index[7]. Given two clusterings X and Y both having k clusters, the FM index is defined as

$$B_k = \sqrt{\frac{TP}{TP + FP} \cdot \frac{TP}{TP + FN}}$$

where

- TP: True positive count, the number of pairs present in the same cluster in both X and Y.
- FP: False positive count, the number of pairs present in the same cluster in X but not in Y.
- FN: False negative count, the number of pairs present in the same cluster in Y but not in X.

If two clusterings are completely the same, i.e. every pair present in the same cluster in X will also be in the same cluster in Y, FP and FN will both be 0 and

the FM index becomes 1. If two clusterings are completely different, i.e. every pair present in the same cluster in X is in different clusters in Y, TP will be 0 and the FM index becomes 0. Essentially, FM index ranges between 0 and 1 and a higher FM index indicates the two clusterings being more similar.

It is clear that the FM index B_k is calculated only for a specific k. In order to compare two hierarchical clusterings, we calculate B_k for every given $k = 2, 3...n - 1$ where n is the number of original samples, and plot B_k against k as a Fowlkes-Mallows curve. The FM curve tells us how similar two hierarchical clusterings are when we cut the dendrograms at a specific point.

The U.S. Code are organized into 54 titles in the highest level of the hierarchy. To evaluate the organization of titles from a mathematical perspective, we apply clustering analysis on chapters (sub-entities of a title) to reorganize them into exactly 54 clusters and compare the result clusters to the original titles using the FM index.

CHAPTER 4

RESULTS

The search portal is illustrated in Figure 4.1, with a query input box, options corresponding to 3 search modes, and a search button. Figure 4.2 shows the search results page for Full-text Search with the query "copyright", each section containing the keyword "copyright" is listed with its title number, section number, and the content preview where the keyword is highlighted in red. If the user clicks on one of the results, he/she will be directed to the result detail page with the complete content of the section, as in Figure 4.3.

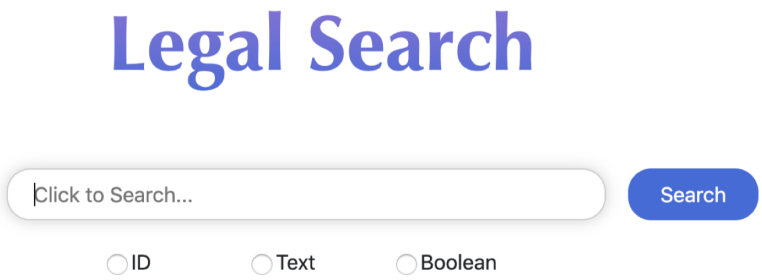

Figure 4.1: Search portal

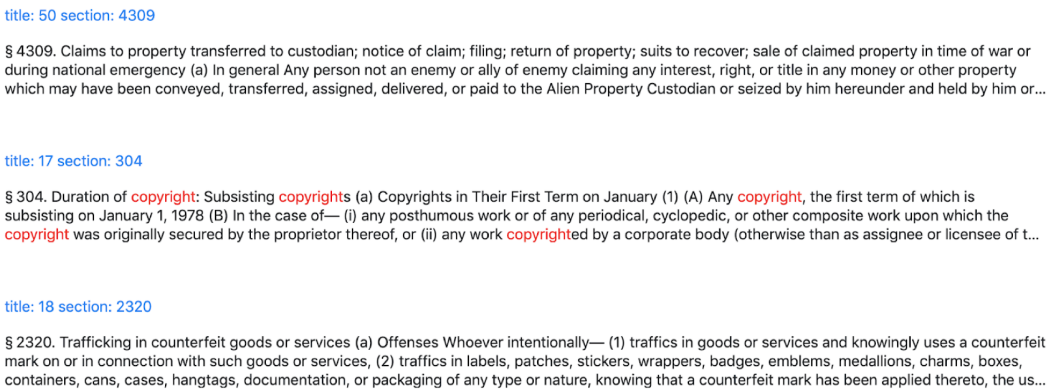

Figure 4.2: Search results page

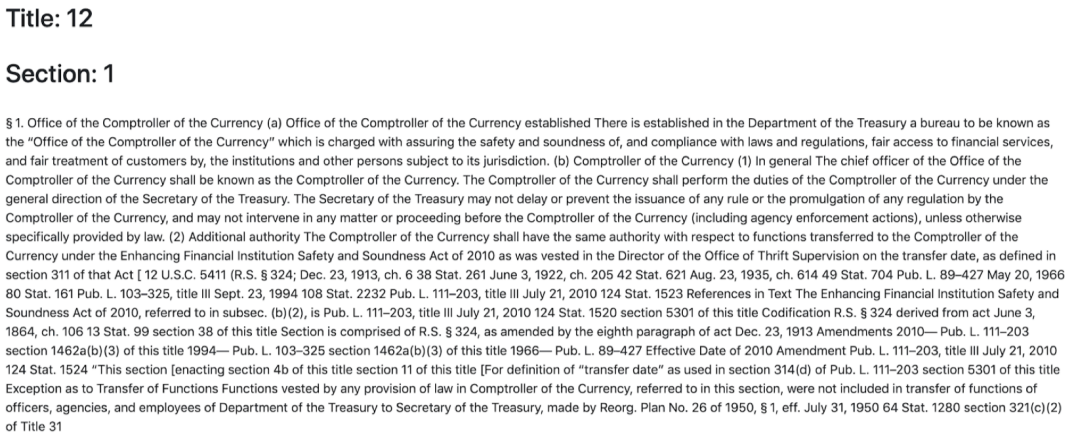

Figure 4.3: Result detail page

The basic properties of the citation network is shown in Table 4.1. The log-log in-/out-degree distributions of all sections are illustrated in Figure 4.4. The top 5 sections highest in-/out-degree are listed in Table 4.2.

Table 4.1: Basic properties of citation network

Vertices	64,134
Arcs	150,688
Accumulated weight	235,100

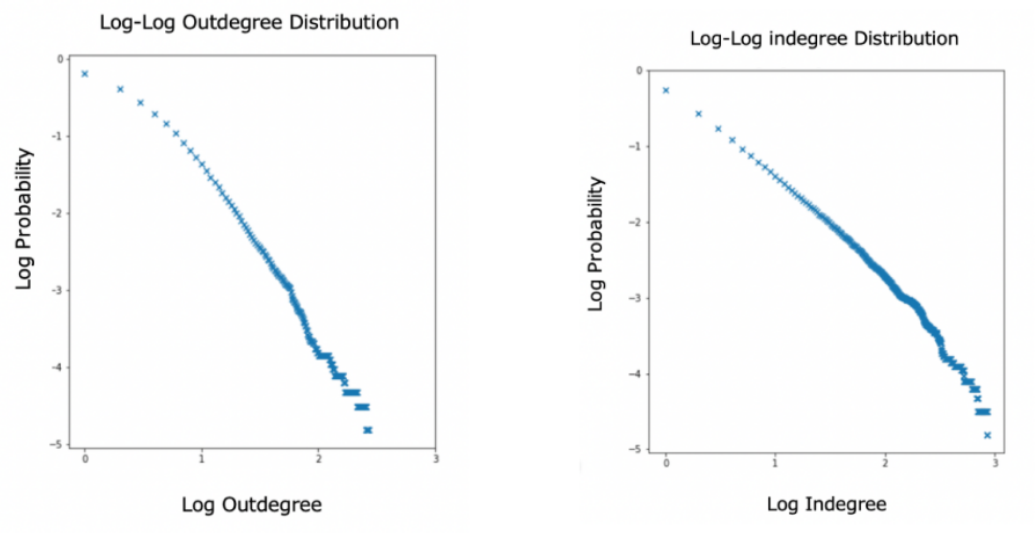

Figure 4.4: Log-log in-/out-degree distributions

Table 4.2: Top 5 sections with highest in-/out-degree

Title	Section	In-degree	Out-degree
6	542	856	20
10	3001	851	16
7	8701	700	36
26	1	688	169
10	101	616	51

Title	Section	In-degree	Out-degree
31	1113	525	269
42	201	222	257
3	301	395	214
26	1	688	169
42	1396	98	163

The interactive visualization of the citation network is illustrated in Figure 4.5.

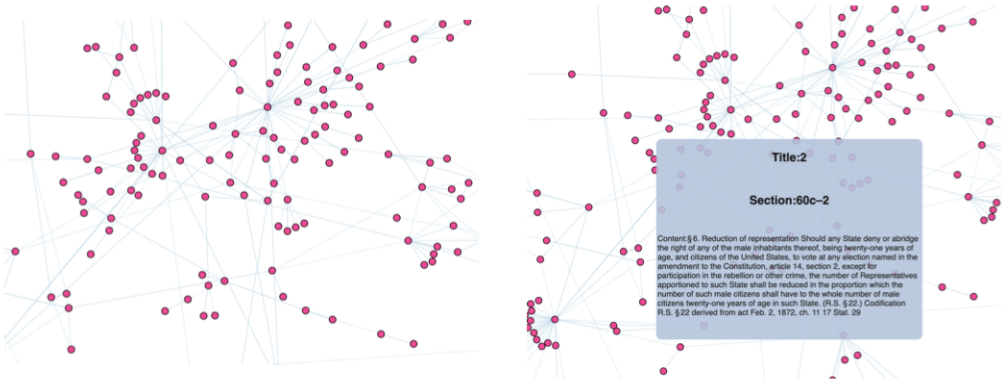

Figure 4.5: Citation network visualization

Considering the total number of sections in the U.S. Code is huge, applying clustering analysis on all of them is not necessary and it would be difficult to present the results. We apply clustering analysis on sections of Title 17 using different distance measures and illustrate the dendrograms resepectively in the following figures. The FM curves between them are plotted in Figure 4.9.

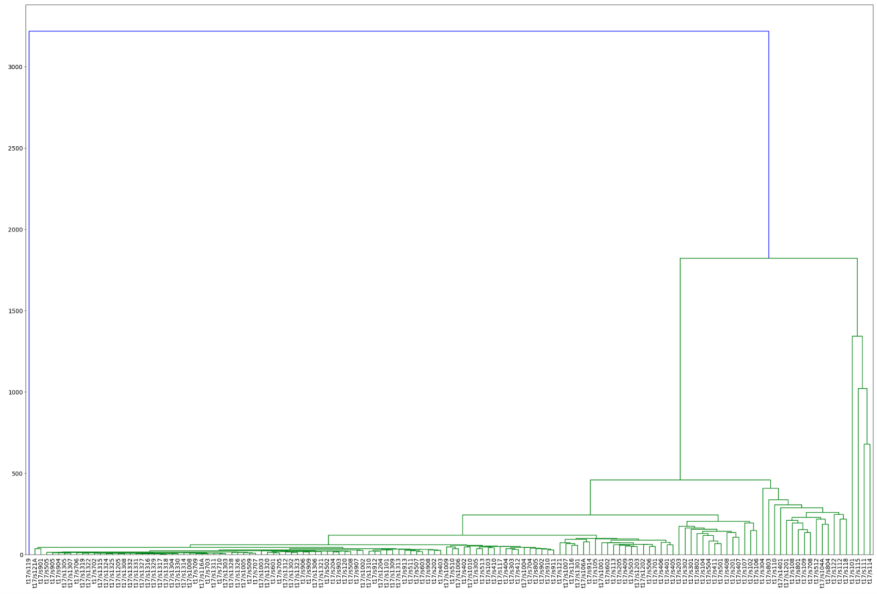

Figure 4.6: Section clustering using vectorization distance

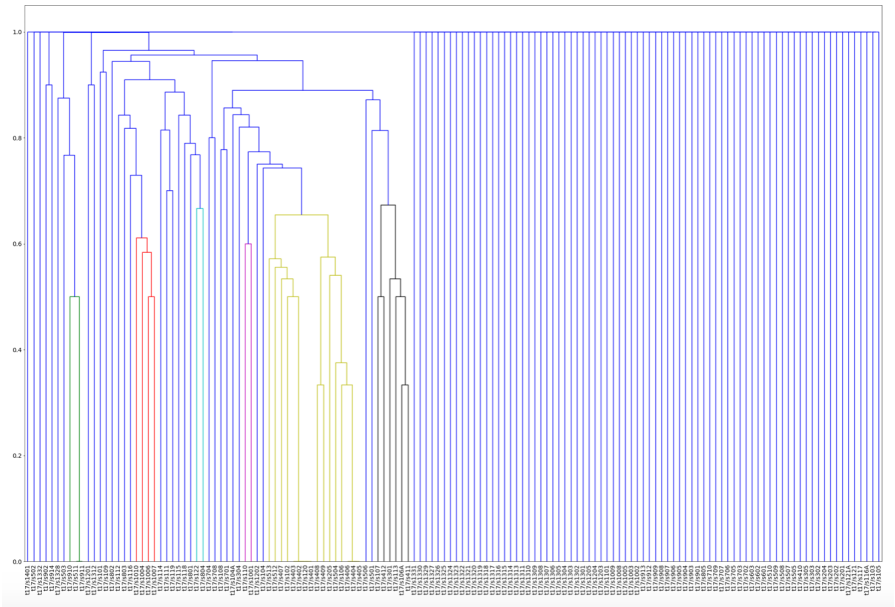

Figure 4.7: Section clustering using citation-based distance

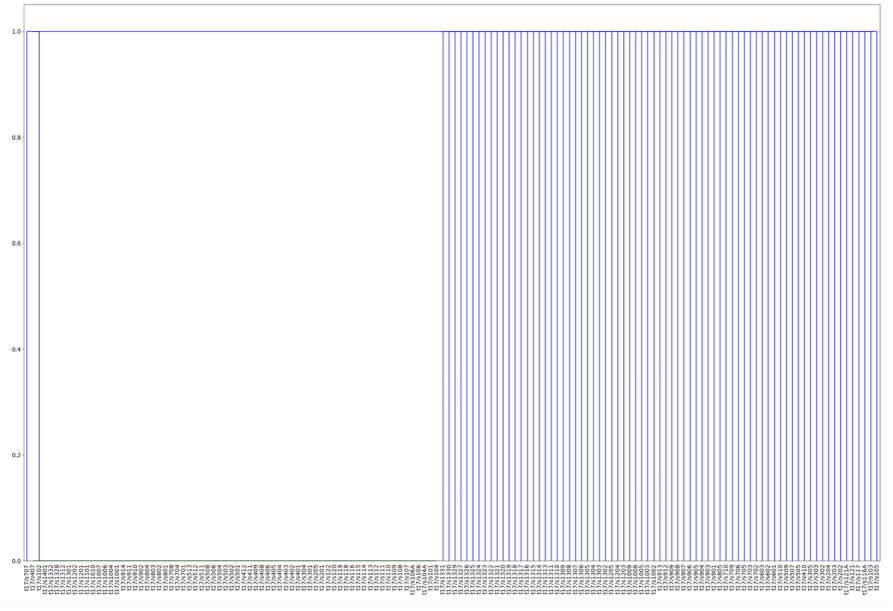

Figure 4.8: Section clustering using sink-based distance

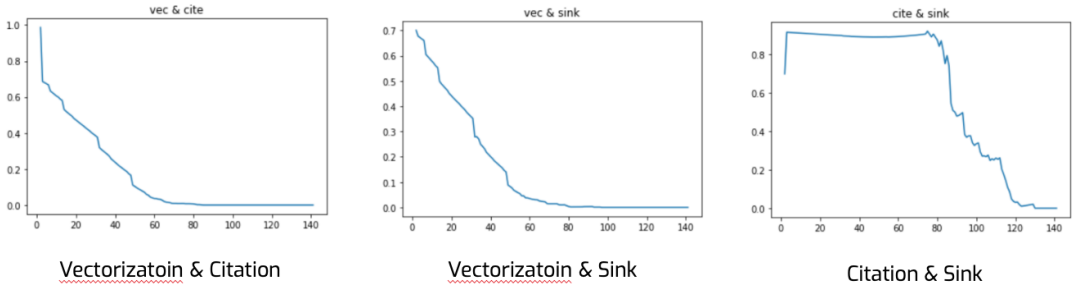

Figure 4.9: FM curves between section clusterings

The dendrograms for applying clustering analysis on titles using different distance measures are illustrated resepectively in the following figures. The FM curves between them are plotted in Figure 4.13.

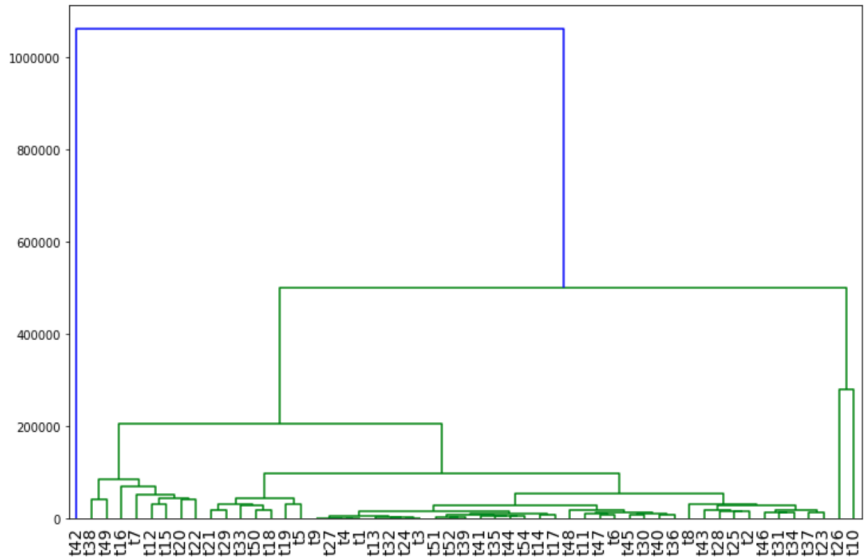

Figure 4.10: Title clustering using vectorization distance

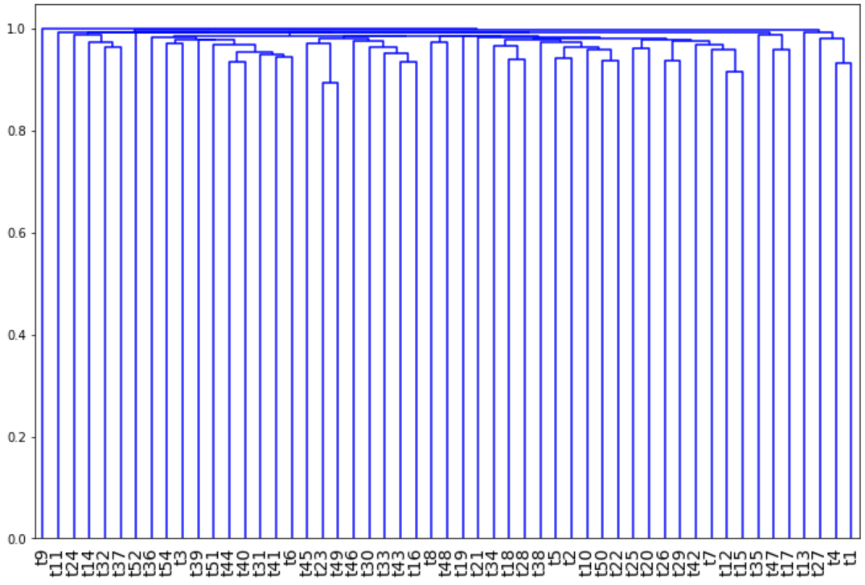

Figure 4.11: Title clustering using citation-based distance

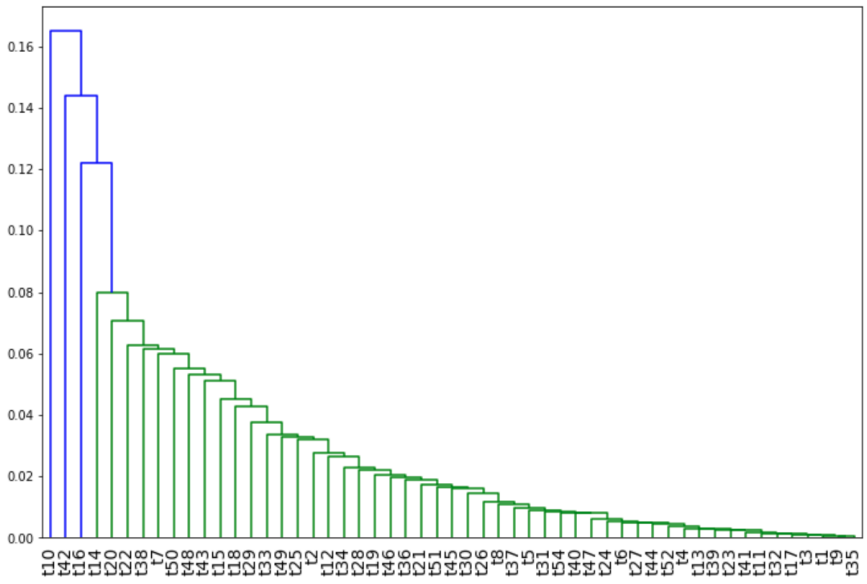

Figure 4.12: Title clustering using sink-based distance

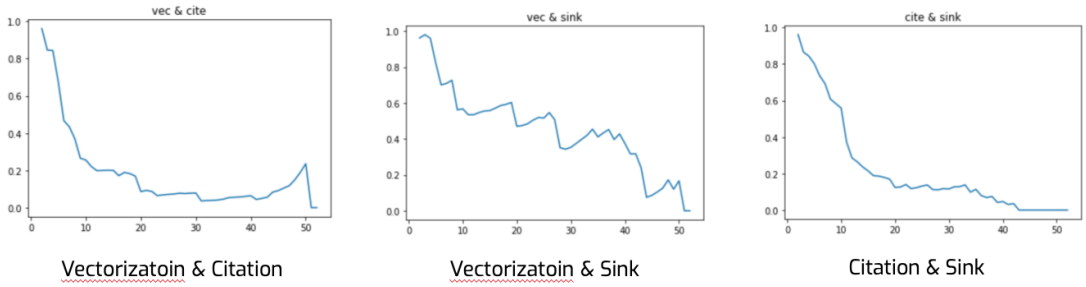

Figure 4.13: FM curves between title clusterings

The FM indices between reorganizations of chapters and the original titles are listed in Table 4.3.

Table 4.3: FM indices between reorganizations and original titles

Original Titles & Vectorization Clusters	0.192
Original Titles & Citation Clusters	0.186
Vectorization Clusters & Citation Clusters	0.844

CHAPTER 5
DISCUSSION

The search engine functions as expected. With the preprocessed data structure, the search engine is sufficiently performant and takes minimal amount of time to complete a query. The results returned are valid with regards to the query and the ranking is justified using legal expertise, i.e. sections are ordered by their significance in the U.S. Code.

The properties extracted from the citation network are very similar to the replicated work, with slight discrepancy that can be accounted for the variation between different amendments. The sections with highest in-/out-degree are justified by legal expertise, i.e. sections with highest in-degree are significant and sections with highest out-degree are of complex topics.

The visualization of the citation network is clear and the interaction functions smoothly.

The clustering results vary significantly between different approaches. For clusterings on sections, we interpret them as follows.

1. Section clustering using vectorization distance: In Figure 4.6, a lot of sections are very close to one another and go into one big cluster, while others are grouped into several small clusters. This indicates that most sections have similar text content.
2. Section clustering using citation-based distance: In Figure 4.7, a fair number of sections are grouped into different clusters with similar sizes, while the other are totally unrelated with one another (the distance being 1 means no shared citations between them). This indicates that, while some sections are related to one another, a visible number of sections are very isolated from the others.
3. Section clustering using sink-based distance: In Figure 4.8, a fair number of sections are grouped into one huge cluster, while the other are totally unrelated with one another. This result conforms with the previous one. Since sinks are more abstract than citations, the small clusters are consolidated into a huge one. The isolated sections remain the same.

The comparison between section clusterings shows that citation-based and sink-based clusterings are highly similar when the number of clusters is small (roughly 2 80). This conforms with our intuition because both two are based on citations with only difference in the level of abstraction. When the number of clusters is small, the level of abstraction is sufficiently high that such difference becomes insignificant. Vectorization clustering shows no visible similarity to the other two. The FM curves drop naturally as the number of clusters increases.

Similarly, we interpret the clustering results on titles.

1. Title clustering using vectorization distance: In Figure 4.10, a fair number

of titles are very close to one another and go into one big cluster, while others are grouped into several small clusters. The indication is similar to that for sections.

2. Title clustering using citation-based distance: In Figure 4.11, titles are reasonably distant to one another and grouped into clusters of similar sizes. This indicates no sub-structure identified among the titles.

3. Title clustering using sink-based distance: In Figure 4.12, each title gradually goes into one single cluster. This result, despite the phenomenon being strange and interesting, indicates the same as the previous one, i.e. no sub-structure is identified.

The comparison between title clusterings shows no significant similarity between any of them. The FM curve between vetorization clustering and sink-based clustering drops slightly slower than the other two, but the difference is not sufficient to draw any conclusions.

In Table 4.3, neither vetorization clustering nor citation clustering is similar to the original title organization. However, the two clusterings are highly similar to each other (with an FM index of 0.844). We conclude that the original title organization does not well reflect the mathematical relevance of the content, while clustering analysis does, either using static text-based information or dynamic citation-based information.

5.1 Limitations

For the search engine, we did not implement reverse indexing as most search engines should do because the U.S. Code is a small data set and with the pre-processed data it is sufficiently efficient. This could be an issue if we were to extend the search engine to other legal data sets, e.g. state statutes or supreme court cases.

As to the properties extracted from the citation network, although our results are very close to the replicated work and the minimal discrepancy can be accounted for by the difference between versions. However, as the authors did not mention what version they used for the study, we could not precisely verify whether our work will produce the exact same results if the same version is used. Regarding the visualization, there is still potential improvement in the implementation to provide a smoother experience. Right now, the visualization is not sufficiently efficient to display the entire network at the same time.

Regarding clustering analysis, we have only practiced some approaches selected based on legal expertise. There are a huge number of clustering algorithms out there and also other distance measures. Our work mainly demonstrates a process of applying this traditional statistical method on a set of legal information, but does not exhaust all possible practices.

5.2 Future work

To improve the performance and extensibility of the search engine, reverse indexing is definitely a great option as the next start. As to verifying the prop-

erties extracted from the citation network, it is possible to repeat the process on different versions and compare the results with authoritative sources. For the performance of the visualization, tiling at different scales may be leveraged. Furthermore, it is still a good idea to try other settings or algorithms for clustering analysis to provide deeper insights to the internal relations of the U.S. Code.

CHAPTER 6
CONCLUSION

By leveraging information retrieval technology on legal information, we have developed a efficient and extensible search engine for the U.S. Code, designed for both legal professionals and scientific researchers. We have constructed the U.S. Code citation network, from which we have extracted important properties and produced an interactive visualization, providing insights to the internal relations of the U.S. Code entities. Through clustering analysis on the U.S. Code entities, we have demonstrated a general process of analyzing the structure of legal information with a computational approach.

BIBLIOGRAPHY

[1] Kevin D Ashley. *Artificial intelligence and legal analytics: new tools for law practice in the digital age*. Cambridge University Press, 2017.

[2] Vladimir Batagelj and Andrej Mrvar. Pajek—analysis and visualization of large networks. In *Graph drawing software*, pages 77–103. Springer, 2004.

[3] V Richard Benjamins, Pompeu Casanovas, Joost Breuker, and Aldo Gangemi. *Law and the semantic web: legal ontologies, methodologies, legal information retrieval, and applications*, volume 3369. Springer, 2005.

[4] Jon Bing. Designing text retrieval systems for conceptual searching. In *Proceedings of the 1st international conference on Artificial intelligence and law*, pages 43–51. ACM, 1987.

[5] Michael James Bommarito and Daniel Martin Katz. Properties of the united states code citation network. *SSRN 1502927*, 2009.

[6] Michael J Bommarito II, Daniel Martin Katz, Jonathan L Zelner, and James H Fowler. Distance measures for dynamic citation networks. *Physica A: Statistical Mechanics and its Applications*, 389(19):4201–4208, 2010.

[7] Edward B Fowlkes and Colin L Mallows. A method for comparing two hierarchical clusterings. *Journal of the American statistical association*, 78(383):553–569, 1983.

[8] Qiang Lu, Jack G Conrad, Khalid Al-Kofahi, and William Keenan. Legal document clustering with built-in topic segmentation. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 383–392. ACM, 2011.

[9] Howard Turtle. Text retrieval in the legal world. *Artificial Intelligence and Law*, 3(1-2):5–54, 1995.

[10] Paul Zhang and Lavanya Koppaka. Semantics-based legal citation network. In *Proceedings of the 11th international conference on Artificial intelligence and law*, pages 123–130. ACM, 2007.