Legal Terminology Extraction with the Termolator

Nhi Pham
New York University
nhi.pham@nyu.edu

Lachlan Pham
New York University
lp2233@nyu.edu

Adam Meyers
New York University
meyers@cs.nyu.edu

Abstract

Domain-specific terminology is ubiquitous in legal documents. Despite potential utility in populating glossaries and ontologies or as arguments in information extraction and document classification tasks, there has been limited work done for legal terminology extraction. This paper describes some work to remedy this omission. In the described research, we make some modifications to the Termolator, a high-performing, open-source terminology extractor which has been tuned to scientific articles. Our changes are designed to improve the Termolator's results when applied to United States Supreme Court decisions. Unaltered and using the recommended settings, the original Termolator provides a list of terminology with a precision of 23% and 25% for the categories of economic activity (development set) and criminal procedures (test set) respectively. These were the most frequently occurring broad issues in Washington University in St. Louis Database corpus, a database of Supreme Court decisions that have been manually classified by topic. Our contribution includes the introduction of several legal domain-specific filtration steps and changes to the web search relevance score; each incrementally improved precision culminating in a combined precision of 63% and 65%. We also evaluated the baseline version of the Termolator on more specific subcategories and on broad issues with fewer cases. Our results show that a narrowed scope as well as smaller document numbers significantly lower the precision. In both cases, the modifications to the Termolator improve precision.

1 Introduction

Automatic terminology extraction systems identify word sequences which are specific to a domain. The basic approach to terminology extraction involves syntactic processing to identify probable candidates. These are subsequently filtered using statistical techniques. What qualifies as key terminology varies depending on the field and intended purpose, resulting in differences in how terminology extraction is approached. This paper seeks to address the lack of terminology extraction approaches in the legal domain.[1] Specifically, we will tune an existing tool to extract terminology from various categories of US Supreme Court opinions.

Supreme Court opinions are public domain and benefit from comprehensive tagging efforts and, as such, previous natural language processing work on Supreme Court case corpora has included automatic classification using word embeddings and neural networks (Undavia et al., 2018). Supreme Court opinions, as compared with more technical legal documents, like contracts or academic papers, are written to justify decisions to a public audience and are thus written in a less constrained form and with more accessible language. Nevertheless, since decisions are made on recurring legal topics, a regular vocabulary of issue-specific terms is likely to emerge. These are precisely the terms that we target and, in doing so, seek to contribute to the classification problem by providing an explicable identifying feature for categories of legal documents. We use Washington University in St. Louis Database of Supreme Court Decisions, manually annotated with broad and narrow categories for our experiments.

This project uses the Termolator (Meyers et al., 2018), a high-performing open-source terminology extractor. The tool requires a foreground consisting of texts in the target topic and a multi-topic background both of which can be customized to need. The benefit of a curated background as opposed to a general corpus used by many other term extractors is that it intuitively allows for the extraction of the key terminology of a specific subdomain.

[1] We claim this to be the first documented application of terminology extraction to the legal domain. We make no claims about other NLP techniques applied to the legal domain.

Natural Legal Language Processing Workshop 2021, pages 155–162

For this paper, by specifying the foreground as a collection of Supreme Court cases in a given foreground topic and the background as opinions in a set of documents about other, varied topics, the terminology extracted are more likely to be specific to that foreground rather than also including general legal terms present in those cases that aren't necessarily representative of the foreground topic. The Termolator also ranks the terms it finds based on a relevance score determined through web search and well-formedness metrics. The Termolator was designed with a focus on science-oriented texts, so in this paper, we make several adjustments to the tool which improve its baseline precision for at least Supreme Court decisions.

2 Previous Work

2.1 Legal Document Classification

There has been some research on legal document classification using a range of methods such as traditional statistical models, neural networks, and state-of-the-art deep learning classifiers. Previous work includes a study on classifying Supreme Court legal opinions using convolutional neural networks (CNNs) and recurrent neural networks (RNNs) (Undavia et al., 2018). The most successful system in the paper was one that combined word2vec with CNNs. Howe et al. (2019) discusses the performance of different machine learning classifiers for Singapore Supreme Court Judgments. The paper compares state-of-the-art natural language processing methods and statistical models applied to legal documents. The authors found that traditional models outperform neural network classifiers on certain metrics, implying that there is still a need to optimize and improve such tools for legal documents.

These papers serve as a motivation for our work because key terminology can be used as an additional feature to improve the training of legal document classifiers. We are considering this as a topic for future research.

2.2 Terminology extraction

Meyers et al. (2018) describes the Termolator, a high-performing terminology extraction system. The tool uses a chunking procedure which favours chunks containing out-of-vocabulary words, nominalizations, technical adjectives, and other specialized word classes. Subsequently, the tool ranks terms via several metrics including: a distributional score and a relevance score. The distributional score favors terms that are more frequent in the foreground documents than the background documents.[2] The relevance score uses the output of a Yahoo websearch (`www.yahoo.com`) and favors terms that occur more frequently in technical documents (articles, dissertations, etc.) than other documents (online stores, social media, wikipedia, etc.). The Termolator has been applied to scientific articles and patents, achieving upwards of 70% precision in identifying key terminology. We experimented with different domain-specific modifications to the Termolator. The baseline results for comparison come from running the original Termolator. We will be using an identical evaluation method as Meyers et al. (2018), randomly selecting 20 terms from each 20% interval in the ranked output, i.e. 20 from the top 20%, 20 from the 21st to 40th percentile and so forth.

3 EXPERIMENTAL SETUP

3.1 Data set

Our data set consists of 8400 US Supreme Court cases downloaded from the Python Textacy package[3] and labelled using the information available on the Washington University in St. Louis Law, The Supreme Court Database[4]. Each document is classified into 14 broad issues in Table 1, which are further categorized into narrow issues, such as search and seizure (under criminal procedure) or antitrust (under economic activity) [5].

Broad Issue	Number of Documents
01 - Criminal Procedure	1924
02 - Civil Rights	1360
03 - First Amendment	658
04 - Due Process	335
05 - Privacy	110
06 - Attorneys	98
07 - Unions	346
08 - Economic Activity	1667
09 - Judicial Power	1149
10 - Federalism	367
11 - Interstate Relation	58
12 - Federal Taxation	304
13 - Miscellaneous	23
14 - Private Law	1

Table 1: Number of documents in each broad issue

Despite a larger number of existing Supreme Court decisions, the data set was limited by what

[2] Similar to the TF-IDF score used in Information Retrieval.
[3] https://github.com/chartbeat-labs/textacy
[4] http://scdb.wustl.edu/
[5] http://scdb.wustl.edu/documentation.php?var=issue

was made available by the Textacy package which consists of the majority of the decisions issued between 1946 to 2016. We also chose to use the entire data set which is slightly larger, but in a similar order of magnitude to data sets used by the Termolator in previous work in order to maximise the Termolator's effectiveness especially given the likelihood of sparser key terminology when compared with scientific documents.

3.2 Method

A script was produced to generate, as needed, two text documents listing the foreground and background for a specified issue number as required by the Termolator. As an initial baseline, the Termolator was run using the recommended settings. The foreground consisted of opinions categorized under the target issue and the background consisted of the remaining opinions.

Our experiments will focus on the most frequent broad and narrow issues. Among the broad issues, the two most frequent ones for which we evaluated the Termolator are:

- **Issue 01** - Criminal Procedure with 1924 court cases
- **Issue 08** - Economic Activity with 1667 court cases

Among the narrow issues, the two most frequent are subtopics of the aforementioned broad issues, respectively:

- **Issue 10050** - Search and Seizure (other than as pertains to vehicles or Crime Control Act), a subtopic of broad issue 01, with 238 court cases
- **Issue 80010** - Antitrust (except in the context of mergers and union antitrust), a subtopic of broad issue 08, with 216 court cases

We focus on whether our extensions to the Termolator generate better results for legal documents while assessing whether the scope of the issue (broad/narrow) and disparities in the numbers of Court cases affect the precision in legal terminology extraction. To further analyze the results, we perform the experiment on the broad issue 05 - Privacy. This broad issue has 110 Court cases, which is of significantly smaller size compared to the issues 01 and 08, but is within the same range as the most frequent narrow issues 10050 and 80010.

In the following subsections, we describe the adjustments in detail. We choose broad issue 08 for our initial experiments as our development set, for purposes of evaluation. Issue 08 has the advantage of being the second highest labelled issue, allowing for a relatively robust foreground from which to extract key terminology. In addition, the topic of economic activity was slightly easier to differentiate key terminology on the basis of relation to the topic i.e. anything relating to a large-scale organization's or a government's finances.

3.3 Baseline and Evaluation

In line with the evaluation procedure of Meyers et al. (2018), a sample of 100 terms were taken from each output, 20 terms were randomly selected from each fifth of the 5000 terms (in rank order) and evaluated as correct or incorrect. A term is considered correct if it:

- has an obvious relationship to the issue area; and
- is not a term that a layperson would understand in a legal context nor could it be frequently used in other court opinions, except in reference to cases within the issue area

We did not consider names of legislation and citations to previous cases as correct terms. We removed these automatically, using a regular expression-based procedure.[6] A selection of output terminology is shown accompanied by its evaluation and justification in Table 2. Some common attributes of incorrect terms are if it:

- includes an incoherent sequence of letters or if it is an acronym (wwsp, cftb)
- is the name of a case or legal act
- is a geographical name (Merrimack River), an organisation name (Trinko) or a person's name (John Brunsman)
- contains any digit (t-2-1)
- is not a noun group (quasi-suspect)

[6]Named entities are sometimes included in terminology detection output and sometimes not. For example, Termeval 2020 (Rigouts Terryn et al., 2020) calculated terminology detection scores both ways. Citations to laws or court decisions have a similar status. Citations, like terminology are characteristic of particular subfields. However, they arguably have a separate status from terms.

Rank	Term	Evaluation	Justification
0	derivative work	✓	related to copyright laws
327	ppg	✗	acronym
1018	engelgau	✗	case name
1070	class certification	✗	not specific to economic activity
1629	atomic energy	✗	general, widely-understood term
1665	licensor	✓	label for economic actor
3070	job freeze	✓	economic phenomenon
3735	quasi-suspect	✗	adjective, general term (note hyphen)
3896	California tax	✗	reference to a specific set of laws
4792	howey test	✓	assesses nature of transactions

Table 2: Selection of key terminology output, evaluation and justification for baseline run on legal issue 08, economic activity.

The evaluation criteria may seem subjective. Some output terms were not obviously associable with the analyzed issue; for example, a layperson could view "flow control" and be able to conjure a logical non-economic definition; however, the term has a specific economic denotation, which given the right context would be considered domain specific. We erred toward labelling these ambiguous terms as valid. An improved evaluation criteria, perhaps created in consultation with a legal expert, could assist in the delineation of key terminology thereby improving inter-annotator agreement.

For the remainder of this section, we describe domain-specific modifications to the Termolator system which improve the Termolator's performance compared to the baseline system.

3.4 Parameter Adjustments

We adjusted the suggested maximum number of terms considered by the Termolator from 30000 to 6000, and number of terms kept from 5000 to 1000. While the suggested parameters were tested successfully on science-oriented topics, we are working with a data set with far fewer key terms. This intuition was confirmed when the baseline output for issue 08 produced a far higher rank-weighted precision ($\sim 35\%$) than overall precision ($\sim 23\%$) indicating that higher-ranked terms were much more frequently identified as correct.

3.5 Case and Legislation Name Filter

Case names and legislation names are essentially named entities. They often contain abbreviations and numbers and are patterned in a way that makes them difficult to process by both standard noun group chunkers, as well as the Termolator's noun group detection component. We used a regular-expression based system to identify these entities and to remove them from the text. This prevents these entities or parts thereof from being considered as terminology candidates by the noun chunker and subsequent stages of the Termolator. As noted in section 3.3, we define terminology to exclude named entities and citations, as these do not serve the same function as the legal terminology we are considering.

There were two separate sets of regular expressions used: one set of patterns to recognize citations to legislation and one set of patterns for citations to previous court cases (case names). Legislation patterns combine keywords and abbreviations (code, act, const. laws, etc.), indicators of subsections (art, article, amendment, etc.), enumerating expressions (Arabic numbers, Roman numbers and consecutive letters) and the section symbol (§) with keywords that are typically part of US legislation references (statute, code, section, etc.). For citatons to court decisions, the two most common patterns are "standard" and "X v Y". Standard case citations are a combination of standard abbreviations for court reporters (cal., reptr., supp., u.s., etc.), additional abbreviations (e.g. for U.S states), punctuation and numbers. For example, "410 U.S. 113 (1973)" is a standard citation (U.S. refers to a court reporter and the numbers refer to years, pages and volumes). The "X v Y" variety of citation is essentially 2 names joined by a **v** indicating "versus". Our regular expressions were quite long and incorporated various lists of key items (e.g. a list of court reporters). The construction of these expressions was guided by information found in the "the Blue Book"(The Harvard Law Review and The University of Pennsylvania Law Review and The Yale Law Journal, 2010), as well as trial and error using a small set of court opinions (e.g. *Roe v Wade*)[7].

We observed that this change had a noticeable effect on the baseline system. The baseline test output included 28 legislation and case names, such as: Nike, Teleflex, Phillipsburg-Easton. These names were likely identified by the baseline system because they are fairly infrequent and are likely to be more highly concentrated in cases related to the legal field to which the legislation or cases pertain: legislation and cases related to economic activity are going to be cited and referred to in cases about economic activity due to the need for similar precedence-related interpretations of laws. This filter is part of the latest version of Termolator.[8]

[7]The actual regular expressions and the system for using them will be part of our Github release of this system.

[8]https://github.com/AdamMeyers/The_Termolator

3.6 Digit and Hyphen Filter

Our error analysis of the baseline system suggested that terms containing digits and hyphenated words were incorrect. In contrast, hyphenated terms were given additional weight in the Termolator ranking system owing to the fact that in scientific articles such terms often combine two scientific words composing a singular more specific word (e.g. "X-ray" or "wavelength-variable"). The Supreme Court documents were more likely to use hyphens to combine two common, non-jargon words and the combined word was often an adjective ("Texas-Mexican") or a name ("pre-Jones act"). Additionally, any identified key terminology containing a digit was incoherent out of context ("t-2-1") and likely refers to a section or clause of a cited law or case.

3.7 Customization of Websearch Filter for Legal Documents

As part of its relevance score ranking, the original Termolator incorporates the results of a Yahoo search. Based on the websearch, the Termolator calculates a relevance score between 0 and 1. The relevance score combines the number of hits (on a logarithmic scale) with the percentage of the top 10 hits that are "technical", where it is assumed that a technical hit contains a keyword like "thesis", "article" or "dissertation". A high websearch score indicates that a candidate term is found in many scholarly sources on the Internet. This works well for scientific articles, but not so well for legal documents since it is much easier to find and differentiate scientific texts than legal documents using websearches. Furthermore, websearches often find many documents that would be of little relevance. For example, websearching category-specific terms, like the "foreclosure sale" in the economic activity domain, would often return commercial texts related to that category rather than the legal documents we would need to validate terms.

Therefore, we replaced the Termolator's websearch ranking system with our own program in which we use Harvard's CaseLaw Access Project API [9] as the replacement "search engine". The API is composed of 6,725,065 official, book-published United States state and federal case law ranging from 1658 to 2018. The case law is labelled with pertinent metadata and allows searching by case name, case text and court.

$$\textbf{score} = \frac{log_{10}(all\_count)}{log_{10}(ALL)} \times \frac{log_{10}(SC\_count)}{log_{10}(all\_count)} = \frac{log_{10}(SC\_count)}{log_{10}(ALL)} \simeq \frac{log_{10}(SC\_count)}{7}$$

where **score** $\in [0, 1]$, *SC_count* is the number of Supreme Court cases that the term appears in, *all_count* is the number of all cases, including all state courts, federal courts, and territorial courts, in which the term appears, and *ALL* is simply the size of the API - 6,725,065.

This scoring is parallel to the quantity and quality metrics used in the original Termolator's relevance score. The first part of the product, $\frac{log_{10}(all\_count)}{log_{10}(ALL)}$, essentially reflects the number of times the term appears in the Harvard's CaseLaw data set thereby measuring a term's relevance across the entire database - a 'quantity' metric. And the second part of the product, $\frac{log_{10}(SC\_count)}{log_{10}(all\_count)}$, identifies what fraction of returned court cases are Supreme Court cases thereby measuring a term's relevance among the returned results - a 'quality' metric. This score formula simplifies to approximately $\frac{log_{10}(SC\_count)}{7}$. This formula also ensures the output to be in the range [0,1], guaranteeing compatibility with the Termolator. Given that the input (candidate terminology) and output (score in range $[0, 1]$) are identical in both websearch programs, the Yahoo script was easily replaced by our proposed script.

However, it is important to note that the web relevance scoring function in the Termolator uses the top 10 search results from Yahoo for each term and calculates the proportion of these top 10 which are scientifically relevant. Unlike Yahoo, the Harvard's CaseLaw Access Project API search lacks a ranking algorithm. Without some sorting, the first ten results of a search from the CaseLaw Access Project API are not of greater relevance as compared to the remaining results; additionally, all search results are of the same document type - U.S. cases - as opposed to Yahoo search which considers all kinds of websites and document types. So, in our investigation, we considered a crude measure of quality to be the proportion of results which were Supreme Court documents. However, if there were very few Supreme Court cases relative to the total number of cases containing the search term, the relevance score would be low even though it may simply relate to an issue that is more com-

[9] https://case.law/

monly dealt with in lower federal courts. Instead, an analogous analysis of the top search results from a web search engine but for legal documents could overcome these limitations by taking advantage of the automatic ranking of the search engine, better capturing the intuition that something is key terminology if a high proportion of the top search results link to appropriate sites or documents.

4 RESULTS

Modification	Precision Score
0 - Baseline	23%
1 - Parameter Adjustments	35%
2 - Case/legislation filter	44%
3 - Digits/hyphen filter	31%
4 - Combination of 1, 2, and 3	50%
5 - 4 + Legal Search Customization	63%

Table 3: Precision scores of each modification for broad issue 08 - Economic Activity

We first interpret the result for broad issue 08, our development set. All parameter tuning and calculations of contributions are based on issue 08. We validate these results by testing on additional issues, and demonstrating similar results, at least for other broad issues like broad issue 01.

In the baseline, a precision of 23% was found. After decreasing output size, the various steps of filtration incrementally increased precision. Just excluding terms with both digits or hyphens increased the precision to 31%, with the majority of the incorrect terms being case and legislation names (37 of the 69 labelled incorrect). We found that the removal of hyphenated terms, sometimes at the cost of incorrectly omitting a small number of valid key terminology, was superior to merely decreasing the weight applied to the hyphenated terms. Just excluding references to case names and legislation increased precision to 44%. Combined, these changes improved precision to 50%. This significant increase in precision indicates that that the Termolator is already a good tool at identifying key terminology and that many of the correct terminology is found but the quality of the output is merely diluted by the presence of easily filtered terms. Upon the replacement of the Yahoo web-search relevance score with Harvard's CaseLaw Access Project's API search, alongside the above filtration, precision improved to 63%. A sample evaluation can be seen in table 3. The detailed annotation for all considered issues can be found at: Legal Termolator Annotation Results. The annotators were 2 of the authors (both undergratuate NYU students). Inter-annotator agreement scores varied between 90-95%, largely due to disagreement about whether a layperson would know the meaning of an identified key terminology. [10]

Issue	Category	Baseline Precision	Adjustment Precision
Issue 01 - Criminal Procedure	Broad	25%	65%
Issue 08 - Economic Activity (development set)	Broad	23%	63%
Issue 05 - Privacy	Broad	27%	40%
Issue 10050 - Search and Seizure (Other)	Narrow	19%	30%
Issue 80010 - Antitrust	Narrow	14%	28%

Table 4: Precision Score Summary

In general, our adjusted Termolator is able to produce better results for both broad and narrow categories (see Table 4). Although we tuned our system to the development set (issue 08), the results seem to be the same when tested on the held-out issue 01 data, an issue with similar broadness. We observe the approximately 2.5 times higher precision of 65% compared to the baseline 25% for issue 01. For the narrow issues, the difference between the precision scores of the adjusted Termolator and the baseline for issue 10050 and 80010 is not as high as that of the broad issues, but still marks a 50–100% relative improvement (30% and 28% with adjustments, 19% and 14% baseline). With fewer cases in broad issue 05 in comparison to narrow issues 10050 and 80010, our adjusted Termolator is still able to generate a higher precision of 40% for broad issue 05. These results validate our hypothesis that both the number of cases in an issue and the nature of the issue (broad/narrow) contribute to the change in precision. The fewer cases available to produce a foreground mean there is a smaller range of total key terms and the frequency of any potential term is diminished. The smaller scope of a narrow issue means that fewer outputs would be labelled key terminology as compared to the more encompassing scope of its parent broad issue.

To further investigate the latter hypothesis, we inspected how frequently an output for a narrow issue was labelled incorrect even if it would otherwise have been labelled correct had we been evaluating it for its parent broad issue. For example, the term "exchange commission rate" is not considered key terminology for narrow issue 80010 since there

[10] One of the reviewers was concerned about the expertise of the annotators. By posting these annotations online, we are being transparent.

is no clear association with the topic of antitrust but would have been labelled correct for the corresponding broad issue 08, economic activity. This is another indicator that the broadness of the issue plays a role in the precision score of terminology extraction.

In comparison to the results in Meyers et al. (2018) for scientific documents, the precisions we obtained for broad legal issues with a sufficient number of court cases are somewhat comparable. When run using a 5000 document background and a 500 document foreground on refrigeration patents, the Termolator achieved a 70% precision and when run on a 500 document foreground of semi-conductor patents, the Termolator achieved a 79% precision. The precision increased further when the Termolator was run on larger foregrounds, which is also in line with our adjusted Termolator's results (see Table 4). Beyond the Termolator being developed with a focus on scientific documents and key terminology, the evaluation criteria used in Meyers et al. (2018) could be considered more lenient than our own. The evaluation procedure included terms special to any particular field or subfield, not necessarily the field of the document being annotated. In our investigation, a more analogous evaluation would involve the inclusion of terminology that related not only to the issue area but legal documents in general which would increase the number of terminology identifiable as correct. Additionally, Supreme Court opinions are written with a general audience in mind since they exist to explain the justices' decisions to the public. As such, jargon phrases in need of defining and worthy of classification as key terminology are relatively rarer. This contrasts significantly with the patent and academic article document types which the Termolator was originally designed to analyze and which are far denser with scientific jargon since they are directed at experts in the relevant fields. A more comparable set of legal documents might be contracts whose intended audience is also more specialized or educated. Ultimately, the adjustments to the Termolator certainly improved its precision for the legal field, although it has yet to achieve the same precision as when it was applied to patents by the Termolator's creators. However, this may be attributable to either differences in evaluation methods or irreconcilable differences in data sets.

In addition, it is difficult to measure recall score (or related measures like MAP scores). It is not practical to annotate an entire foreground given that most broad issues currently consist of more than 300 documents and there will be future work on expanding the data set. To understand if we could instead simulate a good measure for recall score, we manually annotated 3 moderate-length documents in the foreground of issue 08 - development set. We observe that more than half of correct terms are only specific to the topic of the document or appear with extremely low frequency in the foreground, and thus not representative enough to be considered "good" terms by the Termolator. In fact, such terms are often ignored since the Termolator models the term importance using some statistical scores, i.e. terms that occur more frequently in the topic-related documents are more important than the off-topic ones. A given case on a highly-specific topic may consist of key terms found nowhere else in the foreground and receive a very low recall score, which would not be a representative sample.

5 Future Work

To build upon the proposed adjustments we made to the Termolator, we aim to extend our data set to a larger size. The current data set contains 8400 Supreme Court cases with a significant variance in the number issue areas. Having a more balanced number of cases in each broad issue could improve our results. A larger set in each category could mean that rarer key terminology is successfully identified, each potentially appearing with greater frequency in the foreground. Similarly, a larger background set could improve results by providing more evidence that given phrases are not foreground-related terms. The difficulty in this lies in the lack of labelled cases; the data set used was manually labelled and likely incredibly labour-intensive demanding a reading thousands of court cases. In tandem, the tasks can be mutually reinforcing, with a higher number of labelled cases, assuming high enough accuracy, allowing for larger data sets for key terminology extraction and, similarly, higher quality terminology can be used for greater accuracy in categorization.

We are simultaneously looking into ways of automatically extending the manual classifications to more data. Towards this end, we are currently working on extensions to the work published in (Undavia et al., 2018). It is possible, that we could obtain better results if we used more data, even if

some or all of it were automatically classified. Note that of the approximately 64,000 Supreme Court cases, about half are unusable (really short) and 8400 are annotated. Thus there are about usable 22,000 remaining. However, it is, possible, that other opinions, e.g., circuit court opinions could be used as well.

6 Conclusion

Our adjustments to the Termolator have been shown to improve the precision of the output for Supreme Court cases. These adjustments involved, firstly, the filtering of the most frequently-identified errors followed by a more substantive change to the web-search contribution to the relevance score rankings. We further find that greater precision is achieved when the modified Termolator is applied to broader-scope categories and issues with a greater number of cases. There is significant opportunity to improve and extend upon the investigation of this project to address the relative lack of research in the field of legal terminology extraction.

The modifications to the Termolator described in this paper will be released to the public via Github.

References

Jerrold Soh Tsin Howe, Lim How Khang, and Ian Ernst Chai. 2019. Legal area classification: a comparative study of text classifiers on singapore supreme court judgments. *arXiv preprint arXiv:1904.06470*.

Adam L Meyers, Yifan He, Zachary Glass, John Ortega, Shasha Liao, Angus Grieve-Smith, Ralph Grishman, and Olga Babko-Malaya. 2018. The termolator: terminology recognition based on chunking, statistical and search-based scores. *Frontiers in Research Metrics and Analytics*, 3:19.

Ayla Rigouts Terryn, Veronique Hoste, Patrick Drouin, and Els Lefever. 2020. TermEval 2020: Shared task on automatic term extraction using the annotated corpora for term extraction research (ACTER) dataset. In *Proceedings of the 6th International Workshop on Computational Terminology*, pages 85–94, Marseille, France. European Language Resources Association.

The Harvard Law Review and The University of Pennsylvania Law Review and The Yale Law Journal, editor. 2010. *The bluebook : a uniform system of citation (19th edition.)*. Harvard Law Review Association.

Samir Undavia, Adam Meyers, and John E Ortega. 2018. A comparative study of classifying legal documents with neural networks. In *2018 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 515–522. IEEE.