# DEVELOPING A SEARCH ENGINE LINKED TO AN ENHANCED TEXT MINING FOR AN OPTIMIZED SEARCH OF LEGAL DOCUMENTS

**Ariz Abbas Naqvi**

*Department of Liberal Arts, Aligarh Muslim University, Aligarh, India*

## ABSTRACT

*Accessing the right information quickly and accurately can make all the difference in the law, which is a vast and complicated body of knowledge. Providing clients with the best possible legal advice and representation necessitates having access to information. Legal professionals and everyday people alike must conduct in-depth investigation into their case. In order to accomplish this, they must read extremely lengthy judgments and attempt to extract useful information from them. The currently available search engine provides judgments for this purpose, but it is extremely difficult to locate a specific judgment from this list. Therefore, we have created and proposed a search engine that will make it simpler to locate a specific verdict.*

## INTRODUCTION

The law is a dynamic system that is constantly evolving to meet the demands of society. Legal professionals must therefore have access to current information to stay ahead of the curve. For instance, lawyers must be familiar with the specifics of any new legislation in order to provide their clients with the best possible legal counsel. Lawyers may be giving their clients outdated advice that could have a negative impact on their cases if they do not have access to the most recent information.

In addition, having access to information is essential for lawyers to be able to examine precedent and previous cases to determine the most effective approach to take with their clients' cases. Legal professionals need access to information in order to identify and address legal issues that their clients may not be aware of. Last but not least, legal professionals place a high value on having easy access to information given the complexity of court cases and other legal proceedings. Without this access, the legal process could take too long and be difficult to manage. Legal professionals can get the right information quickly and effectively if they have access to it, which ultimately saves time and money in the long run.

## METHODOLOGY

A tool that lets users search for legal documents and cases by keyword, phrase, or topic is known as a legal search engine. A legal search engine aims to make legal information more searchable and accessible to researchers, professionals, and the general public. The following are the steps that must be taken when creating a search engine:

1) Collecting Data: Gather a sizable collection of legal documents like statutes, court rulings, and legal briefs.

2) Pre-treatment: Pre-processing the text data is the next step. Tokenizing the text, removing punctuation and stop words, and stemming or lemmatizing are all examples of this. To ensure that the text is in a format that can be used for analysis, these steps must be taken.

3) Vectorization of Text: Create numerical vectors from the text so that it can be represented numerically.

4) Calculation of similarity: The legal documents' similarity can be calculated using four different algorithms: Cosine Similarity, Levenshtein Distance, Latent Dirichlet Allocation (LDA), and Vector Space Model (VSM). Based on the characteristics of the problem and the nature of the data, choose the algorithm with the best results after comparing them.

5) Recovery: To find legal documents that match a user's query, use the index.

6) The Finding: The user should receive the judgments that ranked highest as search results.

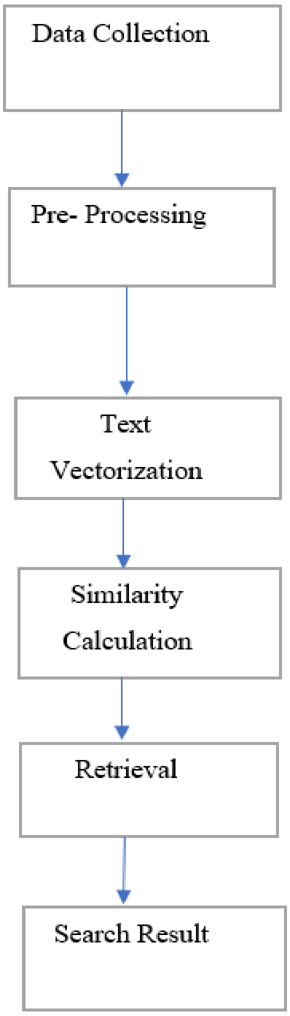


Fig 1: Flowchart

## THE PROPOSED ARCHITECTURE CONSISTS OF THE FOLLOWING STAGES

A. Document Extraction The indiankanoon.org website is used to extract legal documents. The extracted judgments are available as a pdf file.

B. Conversion After the judgment data is extracted, it is converted into text.

C. Pre-processing is a data mining technique for converting raw data into a format that is both efficient and useful. To clean the data, the following steps were taken:

1) Imitation: A lot of text is broken up into smaller parts called tokens through this method. The sentences are broken up into words here.

2) Removal of a stop word: The raw data is devoid of words like "the," "in," and "an."

Third, Lemmatization: the process of condensing various word forms into a single form.

D. Text Vectorization Turn the text into numerical vectors that can be used to represent it numerically.

E. Calculating the similarity Between the legal documents, the four algorithms (Cosine Similarity, Levenshtein Distance, Latent Dirichlet Allocation (LDA), and Vector Space Model (VSM)) are used to calculate the similarity.

F. Retrieval The legal documents that match the user's query are retrieved in this step using the created index.

G. Search result A web application that lets users search for legal documents is built using the flask framework.

The following is the proposed architecture:

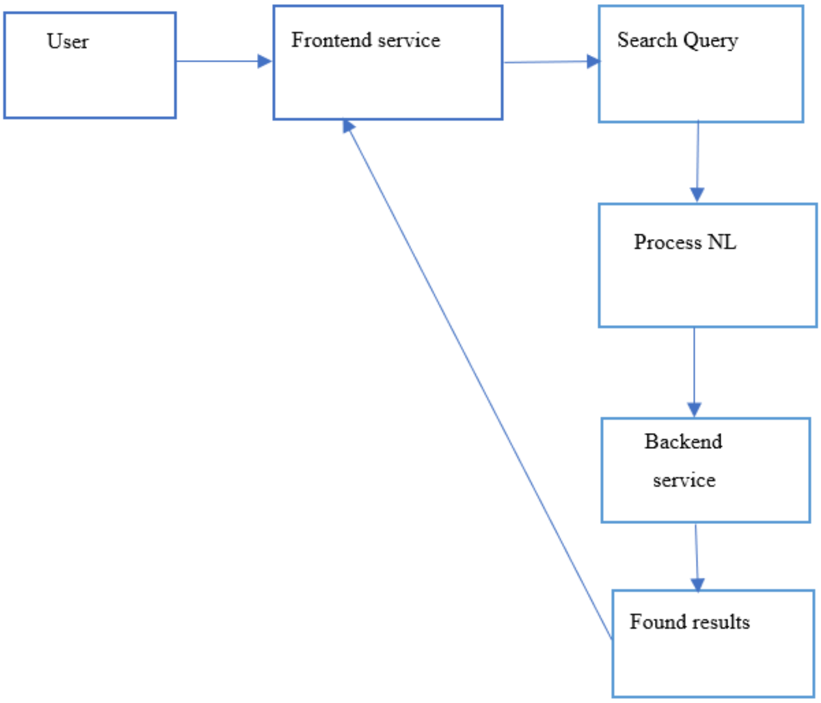
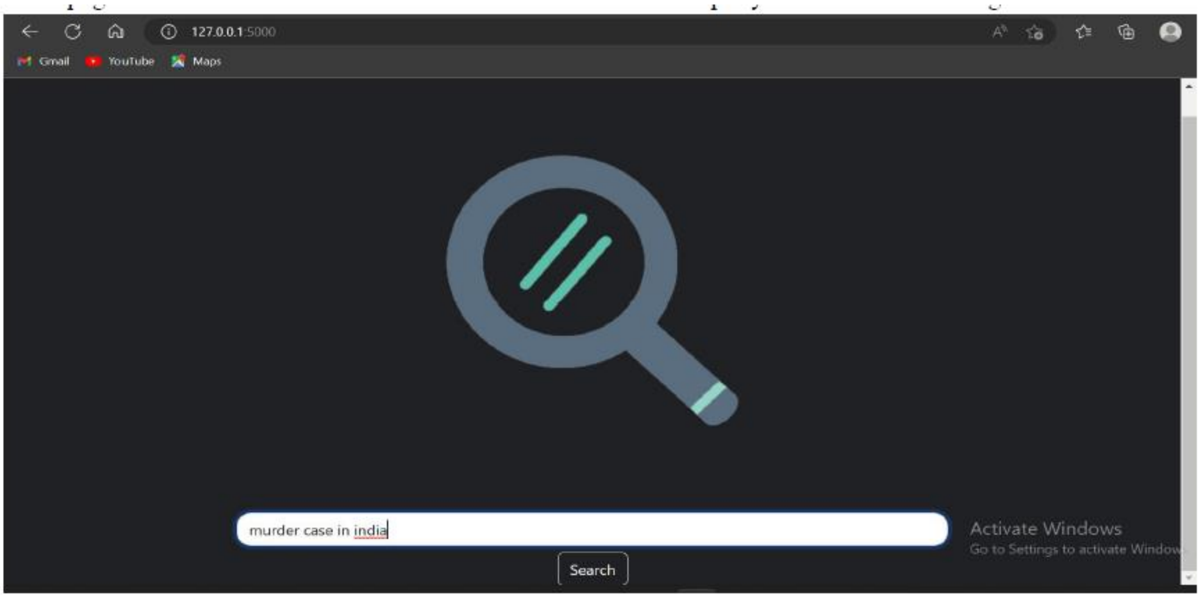


Fig. 2 System Architecture

## RESULTS

The following search engine was created with a search bar for users to use to find information. As depicted in the figures below, the search bar directs you to the result page, which contains documents related to the entered search query.


127.0.0.1:5000

murder case in india

The figures above demonstrate that when a user types in a search query, the resulting search returns a list of documents that are similar to that query. The legal decisions that were found include the case-specific legal information that was found in the search results. These decisions can be used to learn more about the law and study a particular case in depth.

## CONCLUSION

A search engine is being developed in this location. Users of the developed search engine have access to Indian legal judgments that are unique to the country's legal system. It is ideal for legal professionals, scholars, students, and the general public who require legal information to use because it provides access to a wide range of judgments. Legal professionals and users alike will find this helpful in keeping up with legal information. Additionally, assisting legal professionals in responding to client inquiries promptly.

## REFERENCES

[1] Lahitani, Alfirna & Permanasari, Adhistya & Setiawan, Noor Akhmad. (2016). Cosine similarity to determine similarity measure: Study case in online essay assessment. 1-6. 10.1109/CITSM.2016.7577578.

[2] Maake Benard Magara; Sunday O. Ojo; Tranos Zuva, (2018), A comparative analysis of text similarity measures and algorithms in research paper recommender system" DOI: 10.1109/ICTAS.2018.8368766

[3] Veena G, Jalaja G, Levenshtein, (2015), Distance based Information Retrieval, *International Journal of Scientific & Engineering Research*, Volume 6, Issue 5, May- 2015 113 ISSN 2229-5518

[4] Vasile Rus, Nobal Niraula, Rajendra Banjade (2013), Similarity Measures based on Latent Dirichlet Allocation, *International Conference on Intelligent Text Processing and Computational Lingusitics*, The University of Memphis USA

[5] Berger B, Waterman MS, Yu YW. (2020) Levenshtein Distance, Sequence Comparison and Biological Database Search. *IEEE Trans Inf Theory*. 2021 Jun;67(6):3287-3294. doi: 10.1109/tit.2020.2996543. Epub 2020 May 21. PMID: 34257466; PMCID: PMC8274556.

[6] Mohammed Firdhous, (2010), Automating Legal Research through Data Mining, *(IJACSA) International Journal of Advanced Computer Science and Applications*, Vol. 1, No. 6, December 2010

[7] Shahmin Sharafat, Zara Nasar and Syed Waqar Jaffry, (2019), Data mining for smart legal systems, *Computers & Electrical Engineering*, Volume 78, September 2019, Pages 328-342 https://doi.org/10.1016/j.compeleceng.2019.07.017

[8] V. Vaissnave and P. Deepalakshmi, (2019), An Artificial Intelligence based Analysis in Legal domain, *International Journal of Innovative Technology and Exploring Engineering (IJITEE)* ISSN: 2278-3075, Volume-9 Issue-2S2, Retrieval Number: B11131292S219/2019 doi: 10.35940/ijitee.B1113.1292S219

[9] Tiedan Zhu, Kan Li, (2012), The Similarity Measure Based on LDA for Automatic Summarization, *Procedia Engineering*, Vol 29, pg 2944-2940; doi.org/10.1016/j.proeng.2012.01.419

[10] Yu Zhang; Mengdong Chen; Lianzhong Liu, (2015), A review on text mining, *6th IEEE International Conference on Software Engineering and Service Science (ICSESS)* DOI:10.1109/ICSESS.2015.7339149