6-8-2016

Lucem: A Legal Research Tool

Jagjoth Bhullar
Santa Clara University

Nathan Lam
Santa Clara University

Kenneth Pham
Santa Clara University

Adithya Prabhakaran
Santa Clara University

Albert Joseph Santillano
Santa Clara University

Recommended Citation

Bhullar, Jagjoth; Lam, Nathan; Pham, Kenneth; Prabhakaran, Adithya; and Santillano, Albert Joseph, "Lucem: A Legal Research Tool" (2016). *Computer Engineering Senior Theses*. 63.
https://scholarcommons.scu.edu/cseng_senior/63

SANTA CLARA UNIVERSITY

DEPARTMENT OF COMPUTER ENGINEERING

Date: June 3, 2016

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY

Jagjoth Bhullar
Nathan Lam
Kenneth Pham
Adithya Prabhakaran
Albert Joseph Santillano

ENTITLED

Lucem: A Legal Research Tool

BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

BACHELOR OF SCIENCE IN COMPUTER SCIENCE AND ENGINEERING

Thesis Advisor

N. Ling

Department Chair

Lucem: A Legal Research Tool

by

Jagjoth Bhullar
Nathan Lam
Kenneth Pham
Adithya Prabhakaran
Albert Joseph Santillano

Submitted in partial fulfillment of the requirements
for the degree of
Bachelor of Science in Computer Science and Engineering
School of Engineering
Santa Clara University

Santa Clara, California
June 8, 2016

Lucem: A Legal Research Tool

Jagjoth Bhullar
Nathan Lam
Kenneth Pham
Adithya Prabhakaran
Albert Joseph Santillano

Department of Computer Engineering
Santa Clara University

June 8, 2016

ABSTRACT

The legal research industry is currently dominated by two main corporations: WestLaw and LexisNexis. These companies hold a monopoly on the majority of pertinent case law and as a result law firms and law schools are forced to use these systems. However, these systems are incredibly expensive and use antiquated forms of informational retrieval that have suppressed innovation in the field of legal research.

Lucem is a web-based legal research tool that attempts to provide an innovative solution that is easily accessible and intuitive for those wishing to obtain legal information. In creating this product we have leveraged our technical training in order to bring new algorithms and methods of information retrieval to legal research.

In developing this tool, we have attempted to create a system that mirrors the way lawyers approach legal research. In order to do this, we developed visualizations that provide lawyers with an additional research path that serves as a respite to overly textual results. Lucem is an attempt to show what is possible in an industry that desperately needs rejuvenation.

Table of Contents

List of Figures

Chapter 1

Introduction

In this chapter we will define the problem in the industry that Lucem was developed in order to address. After outlining the specific problem, we will provide a discussion of the requirements and constraints of our system. These requirements are divided into functional and non-functional, where functional indicates features of our system and non-functional is related to the heuristics of our system.

1 Problem Statement

The antiquated research system of the legal profession has restricted lawyers to doing inefficient amounts of manual analysis. Traditionally, the research process requires lawyers to locate relevant historic cases and identify precedence, in order to provide a basis to formulate their argument. Lawyers are often required to manually sift through a vast number of documents before successfully finding the applicable cases. Although modern web based systems have alleviated some pain points of the research process, by allowing lawyers to search from a database of previous cases, these systems return an inordinate number of documents that are sometimes only tangentially related to the case at hand. Lawyers are therefore still faced with the daunting task of parsing through a large, undifferentiated list of cases.

Computer Assisted Legal Research (CALR) services, such as Westlaw and Lexisnexis, have supplemented the traditional legal research process by making documents more readily accessible. However, even with the marked increase in efficiency associated with CALR services, these systems still lack the level of automation necessary to eradicate the current problems in the research process, and tend to still inundate lawyers with immense amounts of information. These systems have allowed for minimal improvements in the research process, yet fail to address the crux of the issue.

We have attempted to resolve this problem with a Web-Based solution that leverages document analysis and data visualization to supplement and simplify the current research process for lawyers. Currently, when a lawyer views a specific case they are presented with vast amounts of supplementary information such as the list of case citations. When interacting with this information, there is seemingly no way to gauge the importance of a singular piece of information, which leads to inordinate amounts of manual probing for lawyers. Our system presents this supplementary information in a more usable manner, by consolidating the data and presenting it at a glance. This allows lawyers to extract only the most pertinent information without being overwhelmed with textual data.

Additionally, we use a modified version of Googles PageRank algorithm in order to determine the relevance of a given document based on a query. This provides an improvement on previous systems which typically use less nuanced forms of information retrieval. Our hope is that this will present users with better initial search results. Additionally, we use similarity scores to determine cases that are similar to each other. Thus, allows lawyers to navigate to these cases that we have deemed textually or categorically similar to the one they are currently working on, which provides another path for the lawyer to navigate.

Our solution presents the results of our computations in a detailed, yet concise manner. We utilize visualizations in order to simplify the overall flow of information and mirror the way lawyers approach legal research. While other systems simply return a list of the citations within a case, we present them visually with the purpose of making it more easily discernible as to which citations are the most relevant and precedential. We feel that this tool serves as a supplement to basic legal research methods, and will produce a marked improvement in efficiency.

2 Requirements

This section will discuss the functional and non-functional requirements of our system as well as the design constraints.

The functional requirements describe the specific features that we defined in order to develop a system that addresses the problems inherent with previous systems. Two of the main goals of our system are to provide the user with a relief from overly textual data and provide the user with a clear path that expedites the search process. As a result, we aimed to create visualizations and features such as research sessions that would facilitate a system that behaved in this manner. We defined thirteen functional requirements in total, and they are geared towards creating a searchable legal research tool that expedites the search process.

We defined three non-functional requirements in total. Non-functional describes the heuristics of our system, and the manner in which we want the system to behave. Our non-functional requirements informed the types of features we defined in our functional requirements.

Finally, our system faced two main constraints. Our corpus of documents was limited to those available on CourtListeners public API, and as a web-based system we were required to ensure that the system runs on all browsers.

Functional Requirements

1. Solution will be web-based.

2. The system will allow users to search for legal cases.

3. Users will be able to input case title and be returned a list of relevant cases.

4. The system will provide a visual representation of the returned cases based on their contents.
5. The system will maintain a database of court cases.
6. The system will determine similarity among cases.
7. The system will allow users to bookmark cases.
8. The system will allow users to alternate between multiple cases quickly.
9. The system will maintain research sessions for users.
10. The system will allow users to annotate cases.

Non-functional Requirements

1. The system will be intuitive to use.
2. The results will be returned quickly.
3. The system's interface will please the user.

Design Constraints

1. Cases returned will be limited by those found on CourtListener.
2. The system will run on all major web browsers.

Chapter 2

System Overview

This chapter will provide an overview of our product. We will describe the typical use cases of our system and discuss the goals and conditions related to these use cases. Then, we will discuss the overall flow of the system with our activity diagram, which shows the end-to-end flow of our system. Finally, we will show a conceptual model of our system, which will provide a greater understanding of how our system will look and feel.

1 Use Case Diagram

Figure 2.1 shows our main user (the lawyer) and the different use cases that would compel the user to interact with our system. Below we outline each use case in more detail. Outlining these use cases in this way provided tangible goals and allowed us to easily divide tasks that needed to be completed.

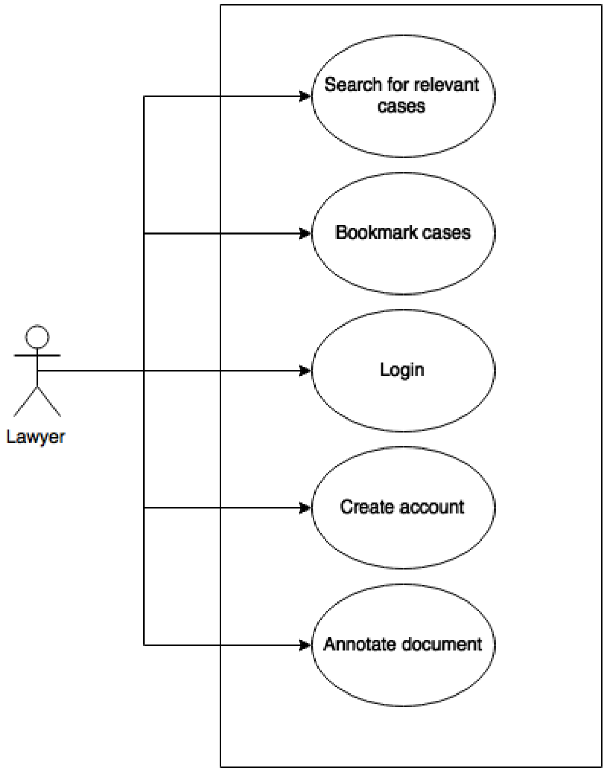


Figure 2.1: Use Case Diagram

Name:	Search for Relevant Cases
Actor:	Lawyer
Goal:	Allow users to find related to their search query.
Precondition:	User has logged in.
Postcondition:	The user is returned a list of relevant cases.
Steps:	1. User inputs his or her search query. 2. User executes search.
Exceptions:	User inputs a string with invalid characters.

Name:	Bookmark Cases
Actor:	Lawyer
Goal:	Allow users to bookmark cases for quick reference later.
Precondition:	The user is looking at a specific case. User is logged in.
Postcondition:	The case is stored in the user's list of bookmarked cases.
Steps:	1. User selects the bookmark option for the case.
Exceptions:	Case already bookmarked by user.

Name:	Login
Actor:	Lawyer
Goal:	Allow users to log into their accounts to access/add bookmarks.
Precondition:	User is on the login page.
Postcondition:	The user is logged in.
Steps:	1. User inputs his or her username. 2. User inputs his or her password. 3. User clicks the login button.
Exceptions:	User inputs invalid credentials.

Name:	Create Account
Actor:	Lawyer
Goal:	Allow users to create accounts to use the system.
Precondition:	User is on the account creation page.
Postcondition:	An account with the provided information is created.
Steps:	1. User provides account information. 2. User clicks the create account button.
Exceptions:	User provides invalid information or a user name already in use.

Name:	Annotate Document
Actor:	Lawyer
Goal:	Allow users to highlight text with a document and take notes on it.
Precondition:	User is viewing a document.
Postcondition:	User's annotations for a given document are saved.
Steps:	1. Select annotate option from menu. 2. Select text to highlight. 3. Add note associated with highlighted text.
Exceptions:	No note input for highlighted text.

2 Activity Diagram

The activity diagram can be found in figure 2.2. Here we show the typical sequence of actions users take when using our system.

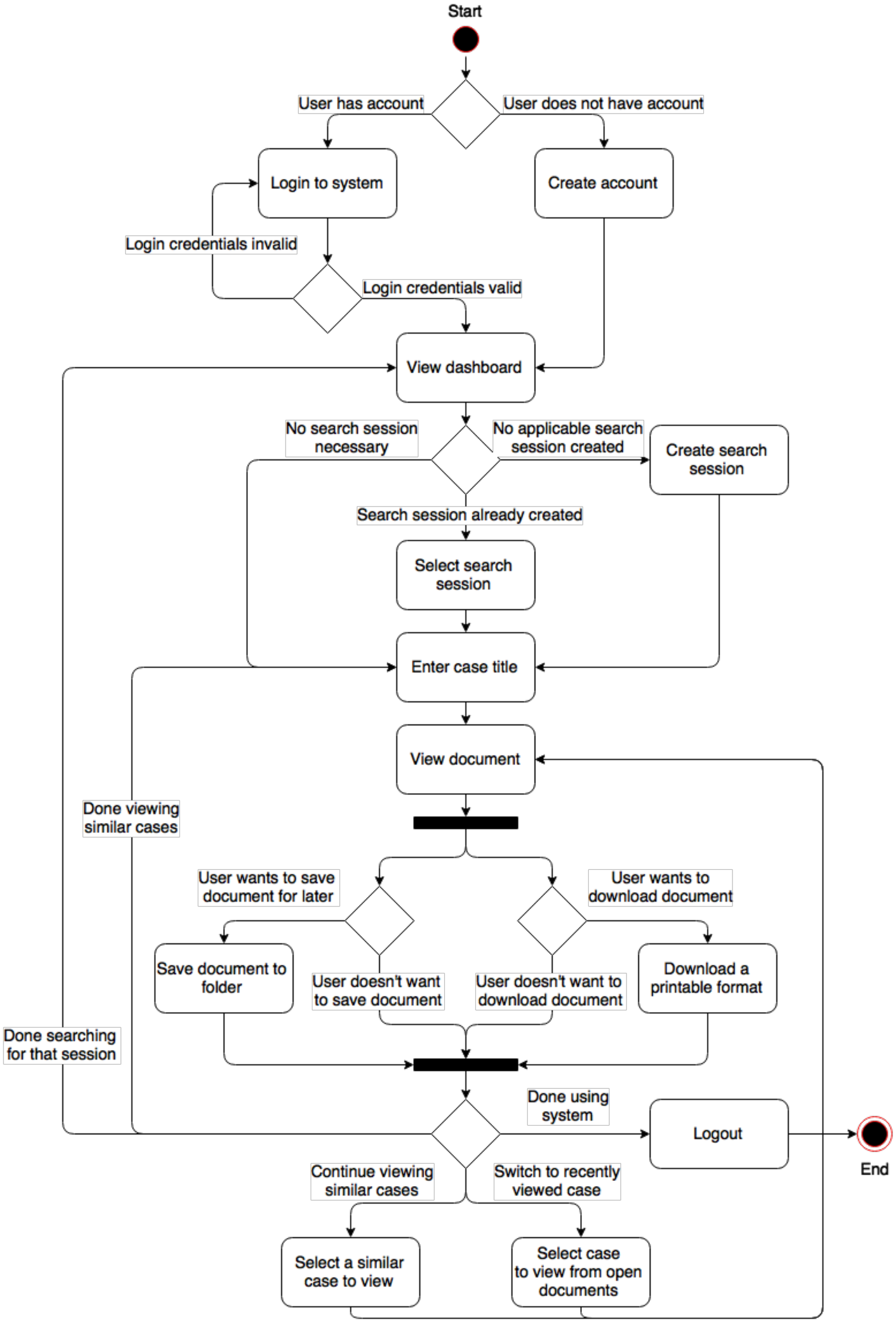

Start
User has account
User does not have account
Login to system
Create account
Login credentials invalid
Login credentials valid
View dashboard
No search session necessary
No applicable search session created
Create search session
Search session already created
Select search session
Enter case title
View document
Done viewing similar cases
User wants to save document for later
User wants to download document
Save document to folder
User doesn't want to save document
User doesn't want to download document
Download a printable format
Done searching for that session
Done using system
Logout
End
Continue viewing similar cases
Switch to recently viewed case
Select a similar case to view
Select case to view from open documents

Figure 2.2: Activity Diagram

3 Conceptual Model

This section provides visual screenshots of every page that is part of our system. With this section, we aim to provide more clarity as to the specific nature of our product.

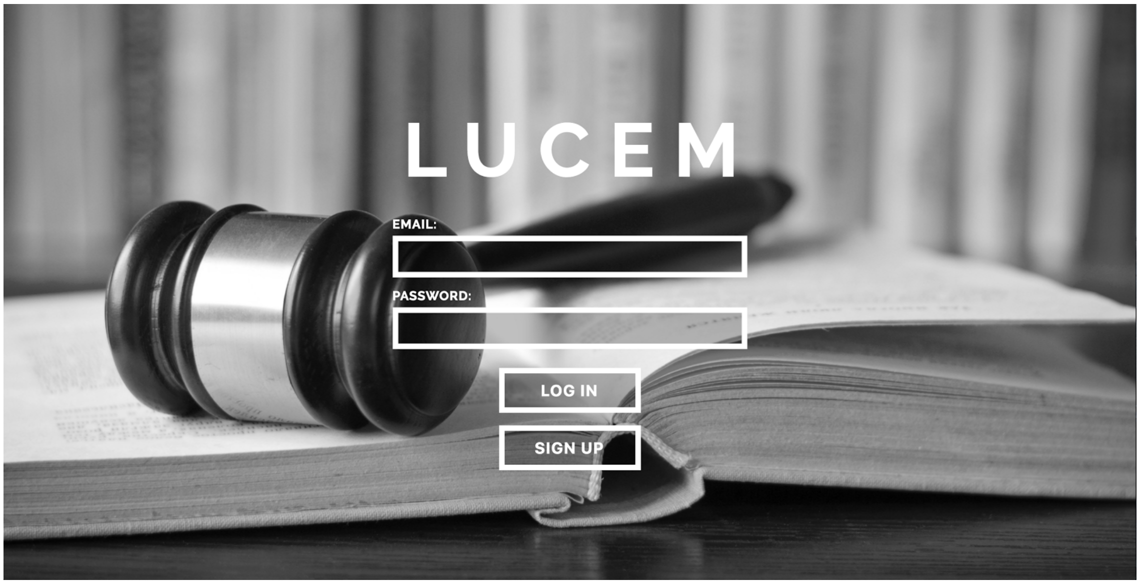


Figure 2.3: Login screen

Access to our system requires authentication to provide the user with customized recommendations while also preventing malicious activity. Consequently, the first page users encounter when interacting with our system is the login page, as shown in Figure 2.3.

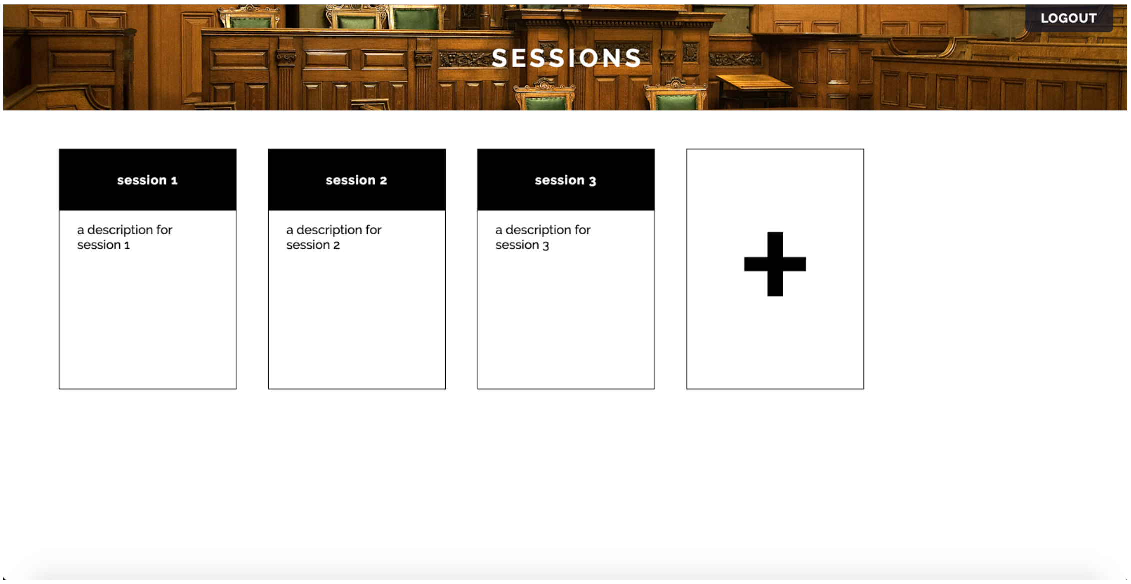

LOGOUT
SESSIONS
session 1
a description for session 1
session 2
a description for session 2
session 3
a description for session 3
+

Figure 2.4: Sessions list

Before being able to search, the user must select a session, as shown in Figure 2.4 Sessions give users a way to organize their research. A user might be doing research on multiple different cases or assignments at once, so these research assignments would be broken up into different sessions. Once a session has been selected, the user can start searching. Any documents that the user views or saves while searching will be be stored in the session.

Like Google Search, the search page, shown in Figure 2.5, is minimalistic devoid of an overabundance of options and animations to make the search function efficient and intuitive. Users start by entering a search query which may be a general keyword or a particular case, and our system returns a list of cases that are deemed related to the search query.

The immediate results of a search are displayed as a list of links to documents that are textually related to the search query, as shown in Figure 2.6. Each result includes attributes related to the case such as the date it was argued, the legal issue in the case, the type of respondent, the type of petitioner, and the Chief Justice at that time. The legal issue of a cases can range from things like states rights to search and seizure. The terms petitioner and respondent are synonymous with the terms plaintiff and defendant; the petitioner refers

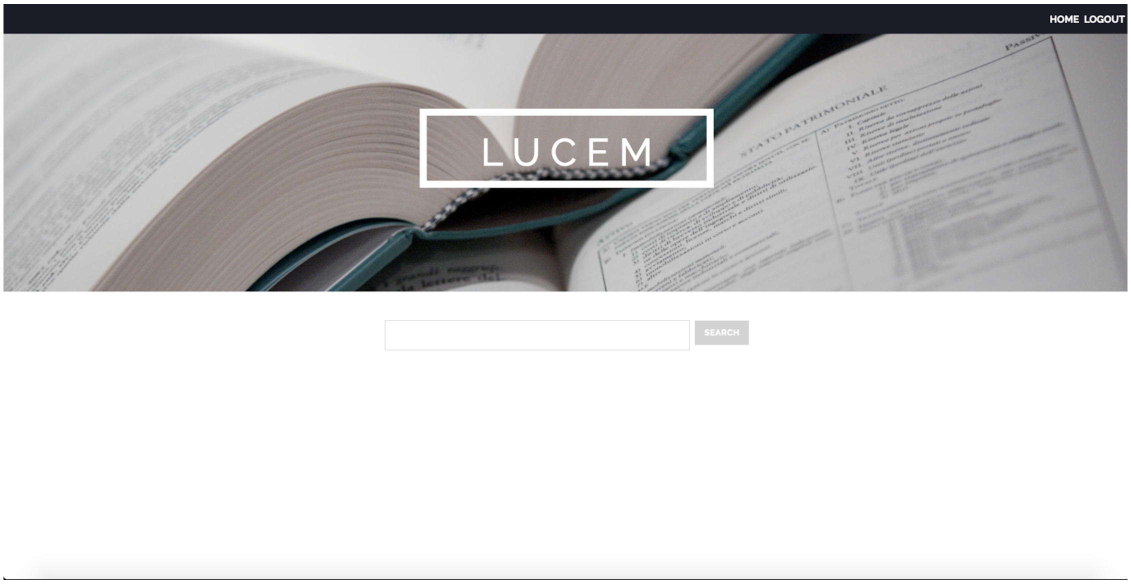
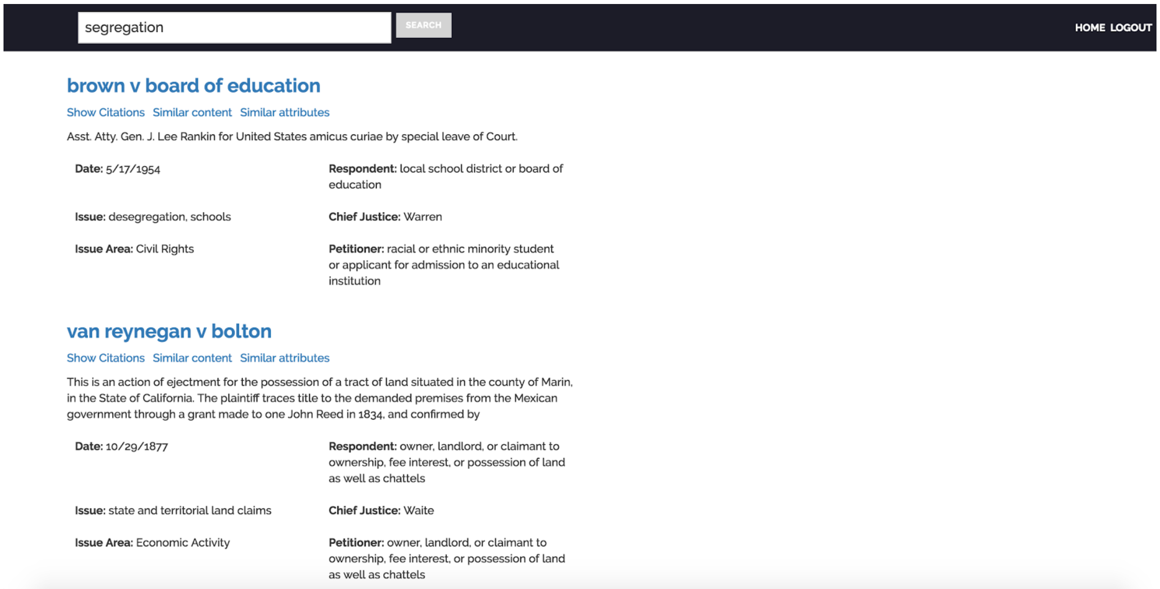


Figure 2.5: Search screen



brown v board of education

Show Citations Similar content Similar attributes

Asst. Atty. Gen. J. Lee Rankin for United States amicus curiae by special leave of Court.

Date: 5/17/1954

Respondent: local school district or board of education

Issue: desegregation, schools

Chief Justice: Warren

Issue Area: Civil Rights

Petitioner: racial or ethnic minority student or applicant for admission to an educational institution

van reynegan v bolton

Show Citations Similar content Similar attributes

This is an action of ejectment for the possession of a tract of land situated in the county of Marin, in the State of California. The plaintiff traces title to the demanded premises from the Mexican government through a grant made to one John Reed in 1834, and confirmed by

Date: 10/29/1877

Respondent: owner, landlord, or claimant to ownership, fee interest, or possession of land as well as chattels

Issue: state and territorial land claims

Chief Justice: Waite

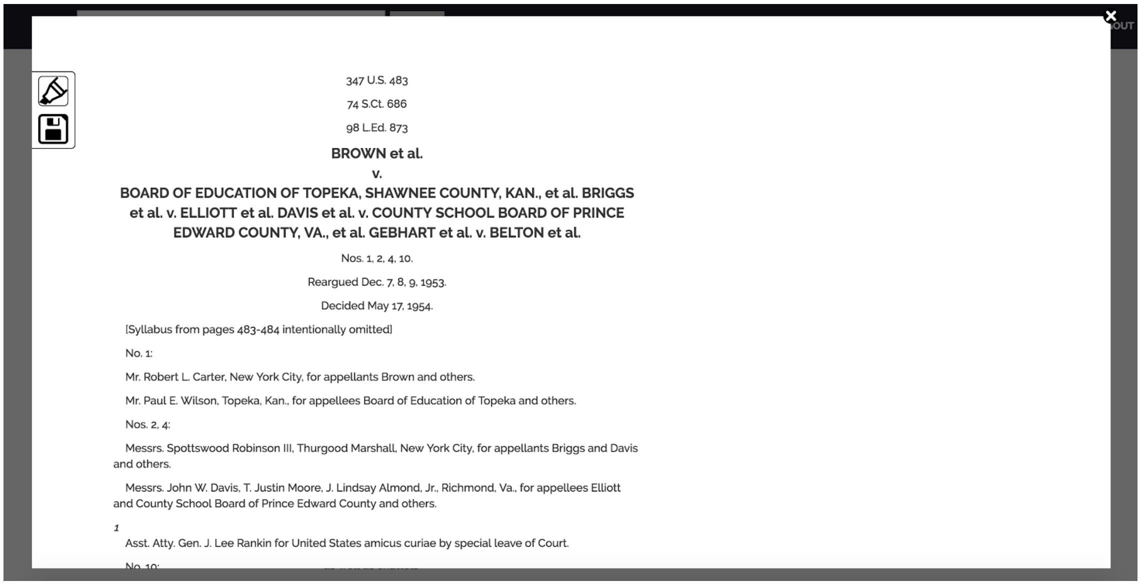
Issue Area: Economic Activity

Petitioner: owner, landlord, or claimant to ownership, fee interest, or possession of land as well as chattels

Figure 2.6: Initial search results

to the party bringing the case to the court, while the respondent refers to the party defending itself in court. Some common values for these fields include individuals, states, and corporations. Clicking on a search result opens a modal box showing the documents text.

Drilling into a search result, a user is served the document page, which allows the user


347 U.S. 483

74 S.Ct. 686

98 L.Ed. 873

BROWN et al.
v.
BOARD OF EDUCATION OF TOPEKA, SHAWNEE COUNTY, KAN., et al. BRIGGS et al. v. ELLIOTT et al. DAVIS et al. v. COUNTY SCHOOL BOARD OF PRINCE EDWARD COUNTY, VA., et al. GEBHART et al. v. BELTON et al.

Nos. 1, 2, 4, 10.

Reargued Dec. 7, 8, 9, 1953.

Decided May 17, 1954.

[Syllabus from pages 483-484 intentionally omitted]

No. 1:

Mr. Robert L. Carter, New York City, for appellants Brown and others.

Mr. Paul E. Wilson, Topeka, Kan., for appellees Board of Education of Topeka and others.

Nos. 2, 4:

Messrs. Spottswood Robinson III, Thurgood Marshall, New York City, for appellants Briggs and Davis and others.

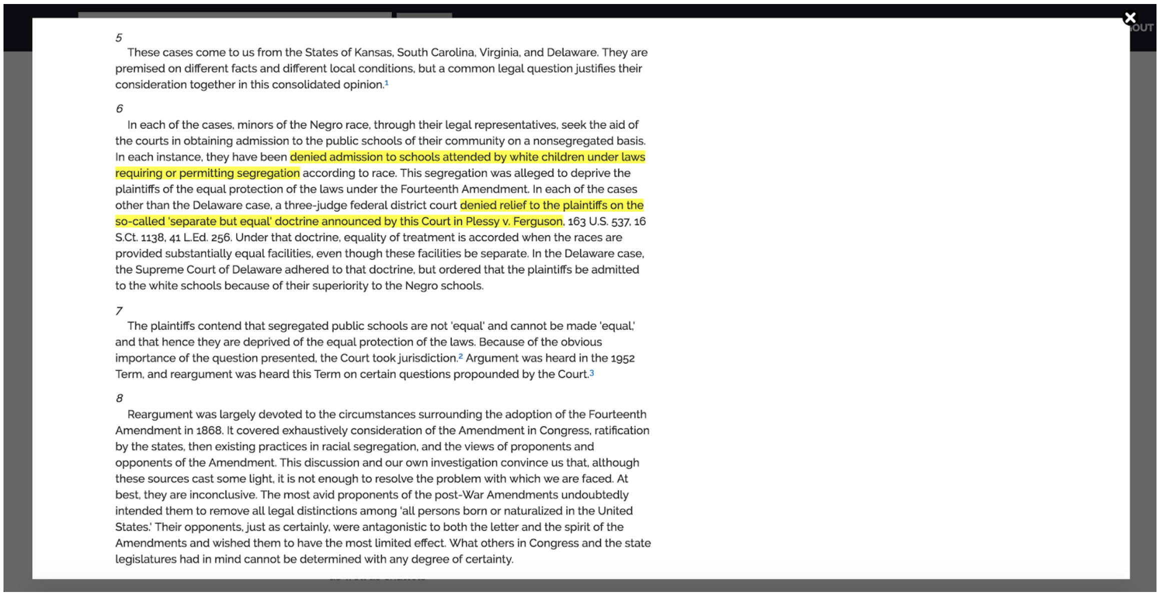
Messrs. John W. Davis, T. Justin Moore, J. Lindsay Almond, Jr., Richmond, Va., for appellees Elliott and County School Board of Prince Edward County and others.

1
Asst. Atty. Gen. J. Lee Rankin for United States amicus curiae by special leave of Court.

No. 10:

Figure 2.7: Viewing the case

to view the case document in its entirety, as shown in Figure 2.7. The document viewer also allows users to highlight the documents text, as shown in Figure 2.8. The highlights are persistent and are associated with a particular document in a session


5
These cases come to us from the States of Kansas, South Carolina, Virginia, and Delaware. They are premised on different facts and different local conditions, but a common legal question justifies their consideration together in this consolidated opinion.[1]

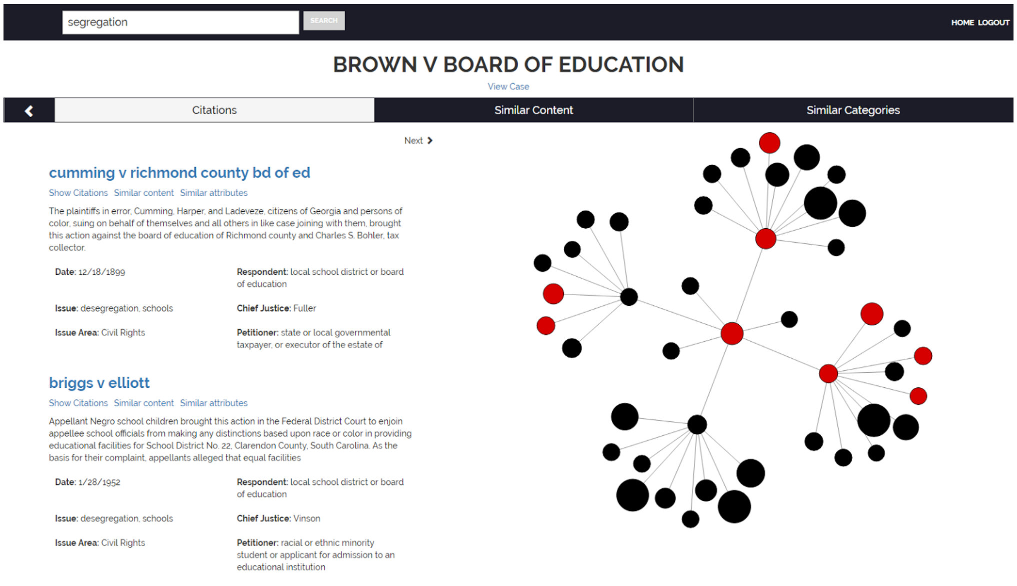
6
In each of the cases, minors of the Negro race, through their legal representatives, seek the aid of the courts in obtaining admission to the public schools of their community on a nonsegregated basis. In each instance, they have been denied admission to schools attended by white children under laws requiring or permitting segregation according to race. This segregation was alleged to deprive the plaintiffs of the equal protection of the laws under the Fourteenth Amendment. In each of the cases other than the Delaware case, a three-judge federal district court denied relief to the plaintiffs on the so-called 'separate but equal' doctrine announced by this Court in Plessy v. Ferguson, 163 U.S. 537, 16 S.Ct. 1138, 41 L.Ed. 256. Under that doctrine, equality of treatment is accorded when the races are provided substantially equal facilities, even though these facilities be separate. In the Delaware case, the Supreme Court of Delaware adhered to that doctrine, but ordered that the plaintiffs be admitted to the white schools because of their superiority to the Negro schools.

7
The plaintiffs contend that segregated public schools are not 'equal' and cannot be made 'equal,' and that hence they are deprived of the equal protection of the laws. Because of the obvious importance of the question presented, the Court took jurisdiction.[2] Argument was heard in the 1952 Term, and reargument was heard this Term on certain questions propounded by the Court.[3]

8
Reargument was largely devoted to the circumstances surrounding the adoption of the Fourteenth Amendment in 1868. It covered exhaustively consideration of the Amendment in Congress, ratification by the states, then existing practices in racial segregation, and the views of proponents and opponents of the Amendment. This discussion and our own investigation convince us that, although these sources cast some light, it is not enough to resolve the problem with which we are faced. At best, they are inconclusive. The most avid proponents of the post-War Amendments undoubtedly intended them to remove all legal distinctions among 'all persons born or naturalized in the United States.' Their opponents, just as certainly, were antagonistic to both the letter and the spirit of the Amendments and wished them to have the most limited effect. What others in Congress and the state legislatures had in mind cannot be determined with any degree of certainty.

Figure 2.8: Highlighting text

In addition to viewing a particular case, the user can also look at citations for that case,



BROWN V BOARD OF EDUCATION

View Case

cumming v richmond county bd of ed

Show Citations Similar content Similar attributes

The plaintiffs in error, Cumming, Harper, and Ladeveze, citizens of Georgia and persons of color, suing on behalf of themselves and all others in like case joining with them, brought this action against the board of education of Richmond county and Charles S. Bohler, tax collector.

Date: 12/18/1899

Respondent: local school district or board of education

Issue: desegregation, schools

Chief Justice: Fuller

Issue Area: Civil Rights

Petitioner: state or local governmental taxpayer, or executor of the estate of

briggs v elliott

Show Citations Similar content Similar attributes

Appellant Negro school children brought this action in the Federal District Court to enjoin appellee school officials from making any distinctions based upon race or color in providing educational facilities for School District No. 22, Clarendon County, South Carolina. As the basis for their complaint, appellants alleged that equal facilities

Date: 1/28/1952

Respondent: local school district or board of education

Issue: desegregation, schools

Chief Justice: Vinson

Issue Area: Civil Rights

Petitioner: racial or ethnic minority student or applicant for admission to an educational institution

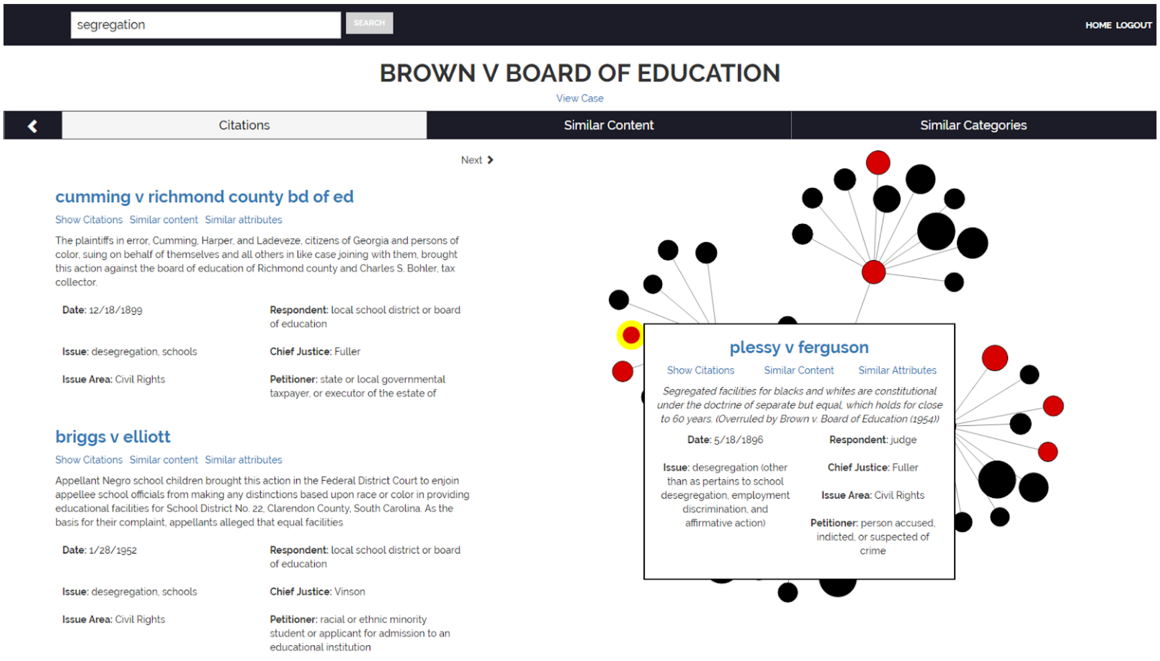
Figure 2.9: Viewing citations for a case

cases that have similar content, and cases that have similar attributes. These three options represent different ways for users to discover documents rather than inputting a query.

When looking at citations for a case, users are presented with the cases in a list format as well as in a visual format, as shown in Figure 2.9. The central node in the visualization represents the current case being looked at. Nodes that are linked to the central node represent the cases that the current case cites. The other nodes linked to these nodes represent the cases that those cases cite. There are two levels of citations being displayed in the visualization, the cases that are directly cited by the current cases as well as the cases that those cases cite.

Red nodes in our citation visualization indicate landmark cases. These are cases that changed precedent or are universally considered to be paramount decisions in law. These serve to guide the user to important cases in the system. The box of information appears whenever the user hovers over one of the nodes. The same information found in the list is also present here. For landmark cases, the holding information is also displayed, which describes the kind of precedent that was set.

The user can also navigate to the similar content page, as shown in Figure 2.11, which



cumming v richmond county bd of ed

Show Citations Similar content Similar attributes

The plaintiffs in error, Cumming, Harper, and Ladeveze, citizens of Georgia and persons of color, suing on behalf of themselves and all others in like case joining with them, brought this action against the board of education of Richmond county and Charles S. Bohler, tax collector.

Date: 12/18/1899 **Respondent:** local school district or board of education

Issue: desegregation, schools **Chief Justice:** Fuller

Issue Area: Civil Rights **Petitioner:** state or local governmental taxpayer, or executor of the estate of

briggs v elliott

Show Citations Similar content Similar attributes

Appellant Negro school children brought this action in the Federal District Court to enjoin appellee school officials from making any distinctions based upon race or color in providing educational facilities for School District No. 22, Clarendon County, South Carolina. As the basis for their complaint, appellants alleged that equal facilities

Date: 1/28/1952 **Respondent:** local school district or board of education

Issue: desegregation, schools **Chief Justice:** Vinson

Issue Area: Civil Rights **Petitioner:** racial or ethnic minority student or applicant for admission to an educational institution

plessy v ferguson

Show Citations Similar Content Similar Attributes

Segregated facilities for blacks and whites are constitutional under the doctrine of separate but equal, which holds for close to 60 years. (Overruled by Brown v. Board of Education (1954))

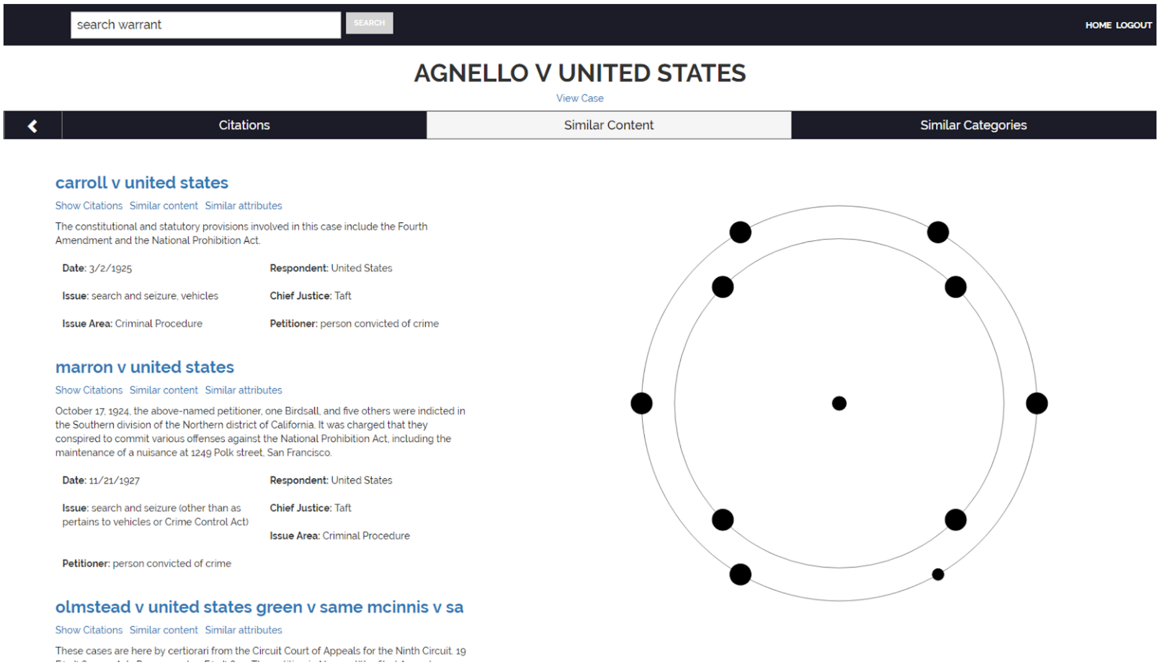
Date: 5/18/1896 **Respondent:** judge

Issue: desegregation (other than as pertains to school desegregation, employment discrimination, and affirmative action) **Chief Justice:** Fuller

Issue Area: Civil Rights

Petitioner: person accused, indicted, or suspected of crime

Figure 2.10: Viewing landmark cases



carroll v united states

Show Citations Similar content Similar attributes

The constitutional and statutory provisions involved in this case include the Fourth Amendment and the National Prohibition Act.

Date: 3/2/1925 **Respondent:** United States

Issue: search and seizure, vehicles **Chief Justice:** Taft

Issue Area: Criminal Procedure **Petitioner:** person convicted of crime

marron v united states

Show Citations Similar content Similar attributes

October 17, 1924, the above-named petitioner, one Birdsall, and five others were indicted in the Southern division of the Northern district of California. It was charged that they conspired to commit various offenses against the National Prohibition Act, including the maintenance of a nuisance at 1249 Polk street, San Francisco.

Date: 11/21/1927 **Respondent:** United States

Issue: search and seizure (other than as pertains to vehicles or Crime Control Act) **Chief Justice:** Taft

Issue Area: Criminal Procedure

Petitioner: person convicted of crime

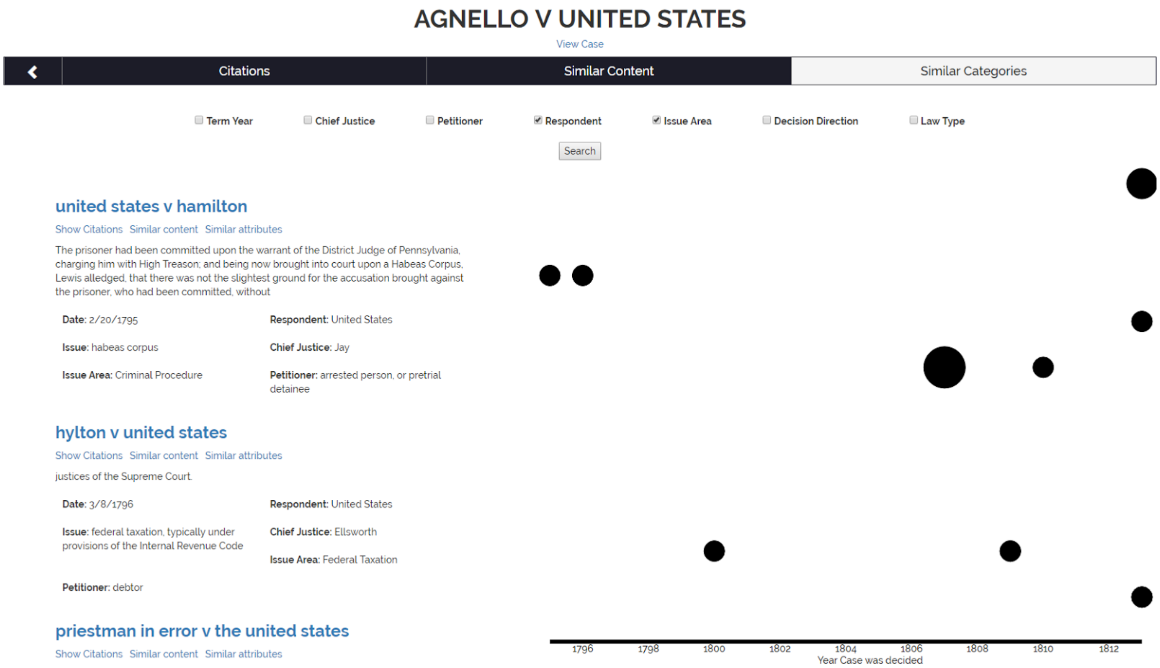
olmstead v united states green v same mcinnis v sa

Show Citations Similar content Similar attributes

These cases are here by certiorari from the Circuit Court of Appeals for the Ninth Circuit. 19

Figure 2.11: Viewing cases with similar content

will provide the user with a list of cases that our algorithms have deemed to be similar in textual content to their current case. In this visualization the central node represents the



united states v hamilton

Show Citations Similar content Similar attributes

The prisoner had been committed upon the warrant of the District Judge of Pennsylvania, charging him with High Treason; and being now brought into court upon a Habeas Corpus, Lewis alledged, that there was not the slightest ground for the accusation brought against the prisoner, who had been committed, without

Date: 2/20/1795 **Respondent:** United States

Issue: habeas corpus **Chief Justice:** Jay

Issue Area: Criminal Procedure **Petitioner:** arrested person, or pretrial detainee

hylton v united states

Show Citations Similar content Similar attributes

justices of the Supreme Court.

Date: 3/8/1796 **Respondent:** United States

Issue: federal taxation, typically under provisions of the Internal Revenue Code **Chief Justice:** Ellsworth

Issue Area: Federal Taxation

Petitioner: debtor

priestman in error v the united states

Show Citations Similar content Similar attributes

Figure 2.12: Viewing cases with similar attributes

current case while surrounding nodes represent similar cases. The closer to the center a case is, the more similar to the current case it is. This kind of visualization allows users to quickly determine which cases are most similar to the current case as well as make comparisons between cases.

Finally, users can navigate to the similar attributes page, as shown in Figure 2.12, which will give users the ability to find cases that have one or more attributes in common with the case they are currently working on. These attributes refer to the same information found in the list of results. For example, a user would be able to find all the cases with the same legal issue as the current case. This allows the user another layer of navigation and has the potential to narrow the scope of their search. The visualization accompanying this page displays cases as a scatter plot with the year the case was argued along the x-axis. There is no data being represented along the y-axis.

Chapter 3

Design and Implementation

With the use cases and conceptual model described in the previous section as an overview of the project itself, this section turns to more specific details regarding the underlying implementation behind Lucem. These implementation details include both technology specifics as well as a discussion regarding the rationale behind the choices made for the user interface. As a web-based project, many of the basic technology decisions were implicitly made for us (HTML/CSS/Javascript are the de facto languages of the web), but actual choices needed to be made in the use of libraries such as d3.js, or in the selection of the Play Framework for our backend. These choices were generally made based on which resources we found to have the least learning curve, most robust development community, and high degrees of usage in the industry at large.

1 Technology Used and Design Rationale

Java: *General purpose, object-oriented programming language which runs on a variety of systems.*

The two members of our team who are primarily responsible for the backend of the system are very familiar with Java, and felt the most comfortable with the language. Given prior experience writing web applications in Java, it presented the least learning curve, and with a relatively short development time, intensive resources could not be allocated to learning an entirely new programming language.

Play Framework: *A popular, well-supported Java web framework.*

Compared to other Java backend frameworks such as J2EE, Play is easy to set up, free, and easy to learn. Play provides all the resources needed to set up a REST API architecture quickly, and allowed us to focus on the actual logic of the code. The Play Framework also rapidly sped up development time with its on-the-fly compilation feature, which meant that with every code change on the server, the code auto-compiled and instantly became live on the website. This allowed us to rapidly prototype new features for testing, rather than waiting for long build times to check if the code written worked.

HTML: *Markup language used to display content on web pages*

CSS: *Styling language used to modify the appearance of web pages*

Javascript: *Interpreted programming language that provides interactivity to web pages*

HTML, CSS, and Javascript are the de facto languages of the web, and essential in building any web based application. While frameworks such as Twitter Bootstrap exist to provide pre-built templates in CSS, we opted to use the vanilla versions of these technologies, building our system from scratch. This gave us the maximum control over the look and feel of the website.

Python: *A popular interpreted and high-level programming language.*

The first major technical challenge in building the Lucem platform was the sanitization and extraction of data from the legal documents themselves. Thus, we needed a way to write scripts to run on our massive collections of documents. Despite using Java for our backend, it did not work well as a scripting language, as it was far too verbose. We settled on Python because of its simple learning curve, which allowed us to quickly write scripts without knowing every intricacy of the language. Python is used in our system for scripts that do everything from sanitizing data for Solr, calculating the importance of documents, and extracting citation data from the source text.

Solr: *An open-source enterprise search platform that uses Java. It is built on Apache Lucene which is an information retrieval library.*
Solr provides indexing and other features that will allow us to easily implement search functionality. The two popular choices in the industry for open-source search platforms are essentially just Apache Solr and Elasticsearch. Both seemed to suit our needs, but with support for the JSON format of our documents built into Solr with little configuration, it was the better tool for the project out of the box.

MongoDB: *Open source NoSQL database that replaces table-based databases with a document-based approach.*
A NoSQL database which allows us to store user data. Given the emphasis on using JSON throughout the system, using Mongo was a natural extension.

CourtListener: *Free legal database that allows for bulk downloads of legal cases through a REST API.*
CourtListener is the primary open-source provider of legal documents in the world, and it provided a REST API that allowed us to download the documents in bulk.

Amazon Elastic Cloud Compute: *A web service that provides cloud compute capability in the form of a remote Linux instance.*

From a user experience and interface perspective, the system is generally designed to facilitate quick interactions. That is, lawyers should be able to find the document they need in the least amount of time possible. Any design choices made from a UI perspective are intended to allow lawyers to clearly see the logical path their research had taken in a particular session, and allow lawyers to jump from document to document without losing a train of thought. Minimalism or the absence of elements besides the core necessities, complements this goal as the user is not inundated with options or distracted by animations or flashy colors. An additional consequence is that the system projects the professionalism expected of a lawyers tool.

Our system requires users to create accounts, which enables the system to leverage their document viewing history for one of its hallmark features: sessions. Unlike Google searches, in which the results of a single search may be used in determining the results of future searches, legal document searches vary by the case that the lawyer is researching. The notion of sessions separate search contexts by the case for which it is being researched. This prevents heuristics for a previously researched case from influencing those of the current case. Furthermore, it can be used as a placeholder to mark the users current place in research for later visitation.

The search results page displays the results accompanied by a case map in order to show relationships between cases, specifically by citations. As civil law is heavily based on precedence the outcome of similar historic cases influences the current cases verdict visualizing the linkages between cases would be beneficial for discovering patterns between rulings.

The document page is typically reached by clicking on a search result or by opening a saved session. On the document page, the user is presented with the document in its entirety such that the user may deem it relevant or irrelevant to his or her research. The user may also follow a clickable citation to view the document from which the current document draws authority. Once the user has consumed the document, he or she may choose to view another document with a similar fact pattern to the current document, which our system suggests under the recommended tab. If the user is continuing his or her research from a previous point in time, he or she may view the document history to see the logical path of documents taken to arrive at the current document.

2 Architectural Design

We have chosen to use a client-server architecture for our system. Users will interact with a web browser which will send requests to a REST API on our server utilizing the Java Play Framework. The server will render HTML and serve them to the client. The server will also interact with a MongoDB instance to handle persistent storage and a Solr instance to handle searching through data.

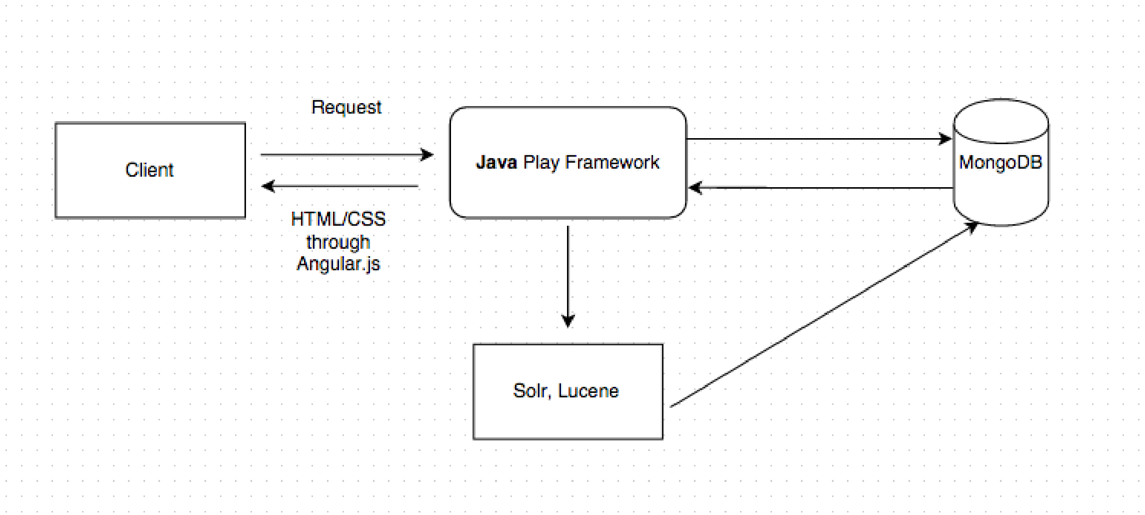


Figure 3.1: Architectural Design

Chapter 4

Testing

1 Test Results

In the design of our system, we aimed to create a tool that is intuitive, easy to use, and accurate. As engineers, our knowledge of the field of law is certainly limited, compared to those working in the field. In order to test our system, we conducted tests with law students from Santa Clara University, as well as research attorneys from the Santa Clara Public Defenders Office. Both groups are used to conducting legal research with the current tools on the market, which was important for us to get feedback on the research process and how well the system. We used two phases in our testing: open-ended and task-based testing. These were done to learn more about two primary aspects of our system: intuition and functionality.

In the open-ended testing, we asked users to simply use our system without any instruction or guidance from us. We did this in order to analyze how users interact with our system. We wanted to be able to see whether different elements of each webpage was placed in such a way that users could figure things out without any help. Our procedure started with a basic introduction to the system. We explained to the users why we were building the system and its intended use. We then gave them ten minutes to do whatever they wanted to explore the system. We kept this process as simple so that the users' experience would feel as natural as possible as they use the system for their first time. Here, we just collected qualitative data, primarily examining how well the users could learn the system independently.

After the open-ended testing we had the task-based testing. We gave the users a walk-through of our system, showcasing all the things it could do and how to make use of these functionalities so that they could then do our task-based testing. We would go through every feature in the system, such as the citation view, similar cases, and sessions, and make sure that the users knew why and how they should use them. In task-based testing, we presented users with a basic research assignment, and asked them to use our system to do certain tasks. Some of the tasks that we gave them were:

- Find cases with the same petitioner as the case brown v board of education
- Find a landmark case after querying police search

This was done so that we could get further insight on the functionality of the system, by learning more about their thought processes as they did research. This gave us more insight into what the user thinks of and what they need as they do their research.

We obtained the majority of our feedback verbally simultaneously with the testing. This worked best for us because when it comes to usability and functionality it can be difficult to prepare questions since the users can have anything to say. We found that most of the feedback from the attorneys pertained to practicality and availability of information in cases while the Law students were more concerned with the visualizations and usability of the system. For example, we got the idea to include landmark cases from an attorney. These types of improvements were things that we did not know exist (e.g. landmark cases) or did not think of. In terms of knowing that things were working fairly well as they were, one of the law students informed us that the citation maps mirrored the way [he] thinks. Because of the testing, we were able to make changes to the system to make it more useful for our intended users.

Chapter 5

Project Management

1 Risk Analysis

In the context of software engineering, risks are unintentional events that result in negative consequences. The goal of risk analysis is to identify and quantify risks and plan solutions for the event they are encountered. When we initially designed Lucem, we tried to ascertain the risks associated with developing a large scale system. However, in the process of building our system, the risks we encountered were related more specifically to our problems domain.

One of the risks we encountered while developing Lucem was getting access to legal documents. LexisNexis and Westlaw hold copyrights on a majority of legal documents. This is permissible because both companies publish legal documents and they devote a large amount of resources towards maintaining their collection of documents. While each company has a public API, the APIs do not give access to the legal documents.

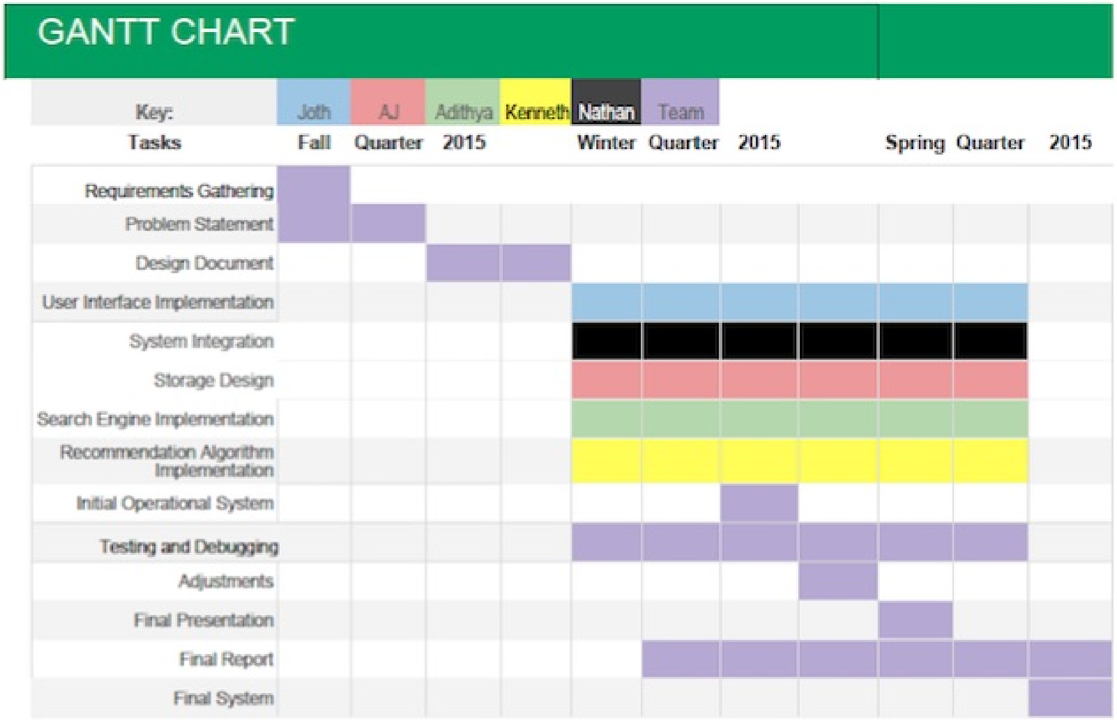
Another risk we encountered while developing Lucem was that of anomalies in the way the documents are structured. While CourtListener returns the documents and their associated metadata with a particular format, the contents of the documents themselves do not adhere to any structure.

On the following page, figure 5.1 shows our risk table in its entirety. This table describes the overall impact we anticipated with a given risk, and the mitigation strategies we developed in order to address these problems.

Risk	Consequences	Probability (0-1)	Severity (0-10)	Impact (prob x sev)	Mitigation Strategy 1	Mitigation Strategy 2
Propagating bug in the code	Would delay timeline and set back project timeline	.9	5	4.5	Conduct peer review code sections	Ensure code is edited from a central location
Failure to meet requirements	Building a system that does not align with the users desires	.3	9	2.7	Thoroughly document and analyze requirements before moving onto implementation	Maintain consistent contact with client to ensure understanding of requirement
Missed deadline	Would set project back and greatly reduce chances of success	.2	8	1.6	Set team deadlines days in advance of actual deadline, in order to resolve potential difficulties	Meet with advisers and clients early in the process of every deliverable, so goals are readily understandable
Failure to gather enough user feedback	Would lead to a system that lacks thorough validation	.4	3	1.2	Provide incentives for users to participate in testing	Allow for users to provide feedback remotely
Loss of Data	Forces aspects of project to be redone	.1	10	1	Store data across multiple locations and platforms	Make copies of data to prevent impact of any loss of information

Figure 5.1: Risk Analysis Table

2 Development Timeline



GANTT CHART

Key:	🟦 Joth	🟥 AJ	🟩 Adithya	🟨 Kenneth	⬛ Nathan	🟪 Team					
Tasks	Fall	Quarter	2015		Winter	Quarter	2015		Spring	Quarter	2015
Requirements Gathering	🟪										
Problem Statement	🟪	🟪									
Design Document			🟪	🟪							
User Interface Implementation					🟦	🟦	🟦	🟦	🟦	🟦	
System Integration					⬛	⬛	⬛	⬛	⬛	⬛	
Storage Design					🟥	🟥	🟥	🟥	🟥	🟥	
Search Engine Implementation					🟩	🟩	🟩	🟩	🟩	🟩	
Recommendation Algorithm Implementation					🟨	🟨	🟨	🟨	🟨	🟨	
Initial Operational System							🟪				
Testing and Debugging					🟪	🟪	🟪	🟪	🟪	🟪	
Adjustments								🟪			
Final Presentation									🟪		
Final Report						🟪	🟪	🟪	🟪	🟪	🟪
Final System											🟪

Figure 5.2: Development Timeline

Chapter 6

Societal Issues

1 Ethical

One of the oldest institutions in society is the concept of codified ethics and the use of legal codes to serve as the foundation for guidance in society. Thus, having unrestricted access to these laws that define ones society serves as a fundamental human right in a democratic world.

Because of the duopoly that WestLaw and LexisNexis have on all pertinent legal information, this transparent access is typically restricted behind an unaffordable paywall. As a result, the majority of citizens lack an understanding of the very laws that are supposed to rule their behavior.

Thinking of it more abstractly, these are two private organizations attempting to consolidate the means of legal production and information, and as a result they are suppressing the publics access to a fundamental right of access. Therefore, this is not merely a legal problem, this is an ethical societal problem that Lucem attempts to address.

2 Social

The social dynamics of Lucem are inherently linked to the ethical concerns that have driven our project. While the ethical concern is that access to the law is being suppressed, the social issue is that access to law has become a commodity that only the elite of society can truly access. As the quality of your legal defense is directly proportional to the amount of money you can afford to spend on such a defense, often those from impoverished communities receive inadequate legal defense.

The quality of your legal defense and your ability to have a fair trial should not be a function of your wealth or influence. Lucem is an attempt to level the playing field by bringing transparent access to the field of law and removing the necessity for exorbitant defense fees.

3 Economic

The idea of billable hours has become the downfall of the legal industry and the quality of defense being provided. Often times those in industry are more interested in maximizing their profits than truly providing the best and most expedient advice for their clients.

One of Lucems core requirements is to expedite the legal research process, and aid lawyers in their goal of legal discovery. As a result of our system, we hope to reduce the overall cost of legal aid, and thus level the economic concerns that are tangentially linked to legal defense.

4 Political, Health and Safety, Manufacturablility

Our project is inherently linked to the political dynamics of the legal system and the influence it carries in regards to society. WestLaw and LexisNexis are two private organizations attempting to consolidate the means of legal production and as a result are suppressing the ability of the public access to a fundamental right. This is an inherently political problem as its implications reach far beyond the spectrum of law.

Lucem is a web-based application that is hosted and maintained by our team of engineers. As a result, there is very little societal concern of manufacturability, as this is taken out of the hands of the user. Additionally, as our project is linked to the legal defense and research process, it is not related to health and safety issues in society.

5 Sustainability and Environmental Impact

As our project is a web based solution, we utilized an EC2 instance from Amazon Web Services. As a result, our solution will continue to be sustainable as long we continue to pay for our instance. However, as our corpus of documents continues to increase, we may face the challenge of reaching the maximum storage capacity.

Our environmental impact is simply a function of the practices Amazon already uses to maintain these Web Services. Amazon has asserted that their goal is to have a system of 100% renewable energy. As of April 2015, 25% of the power consumed in their global infrastructure comes from renewable energy sources and they have a goal to reach 40% by the end of 2016. These virtual web servers serve as a stark improvement over physical servers that generate energy waste. According to Amazon, virtual cloud services only need 16% of the power as compared to an on-premises infrastructure. This represents an 84% reduction in the amount of total power required. Thus, we feel that Lucem is sustainable, manageable, and efficient.

6 Usability

The primary goal of Lucem was to expedite the legal research process and bring transparency to the field of law. Having these goals in mind, we prioritized the usability of our system. We aimed to develop a system that both mirrored the way lawyers already approach case law and that was approachable.

In order to confirm that our system was as usable as we aimed to design, we performed user testing with lawyers and law students. These testers helped validate the heuristics of our system and reported a system that was user friendly and intuitive to use. However, we will conduct more comprehensive testing and attempt to extract numerical data that supports this belief.

7 Lifelong Learning

The concept of lifelong learning serves as the foundation of the field of computer engineering where products continue to evolve to keep pace with ever changing technological landscape. With Lucem, our team is continuously attempting to learn the most effective and efficient methods of data collection. The main problem with the legal field is that it lacks uniformity, and often times pertinent data is scattered across multiple platforms.

In developing this tool, collecting and uniforming the disparate amounts of data became a tedious task. So, one aspect of our lifelong learning is to continue our engagement with the field of legal research. We will attempt to develop, or uncover, new methods of research in order to simplify this process. In addition, we will continue to keep pace with modern technologies in order to further streamline our overall product.

Chapter 7

Conclusion

1 Challenges and Lessons Learned

In developing Lucem we faced an array of different challenges, both technical and non-technical, and as a result we have come away with a greater understanding of the overall process of taking a product from initial inception to final deliverable. From a non-technical perspective, we faced three main challenges: defining effective test plans, managing the scope of our project, and managing ourselves with regards to time and deadlines.

After performing informal testing, we found that the tests provided superior results when we laid out a plan for the user and described what we hoped to gain out of our tests. So, when we moved on to our formal testing with the Santa Clara Public Defenders Office and the Santa Clara University school of law, we created a thorough test plan which outlined the specifics of our test plan. We found that providing the user with more formal instruction helped them provide more specific answers and also allowed us to extract pertinent information that improved our ability to update our system.

The other non-technical challenges we faced were both in regards to management. When our project began, we had many ideas as to how to improve the legal research industry, but we found that at times we talked ourselves into a project that would exceed our bandwidth as a team. The key lesson we learned here was that it is important to build your core functionality first, and then see what extra features are possible to add. When you approach a project with the opposite mentality, as we had, it becomes more difficult to decipher which features are truly important and which are superfluous. The challenge of managing ourselves is in many ways related to this issue of scope. At times we struggled to fully utilize our manpower effectively, because we misunderstood the difficulty (or ease) of certain tasks. The lesson we learned here, is that we should formally define all of our remaining work and discuss how much time each requirement would take to develop. This would alleviate the problem of mismanagement and would allow us to more effectively tackle the complexities of our project.

From a technical perspective we faced two main challenges of data collection and documentation. The challenge of data collection was related to the problem of getting access to legal documents. Lexisnexis and Westlaw hold copyrights on a majority of legal documents. They are able to do so because they are publishers of case law and they devote a majority of their financial and human resources into maintaining their collection of documents. Unfortunately, they have not made these documents available as part of their public API and we, a group of 5 with limited funding, could not hope to build a collection of documents as vast as theirs. Instead, we focused on building the best search system and data visualizations we could and relied on Courtlistener to collect legal documents. Courtlistener is not yet as comprehensive as LexisNexis or Westlaw but they are making swift progress and hope to make all legal documents publicly accessible in the next few years. Integrating each new batch of documents is as simple as re-calculating the inverse document frequencies and performing our relevance calculations. This was one of the main challenges of our system going into the project, however we are confident that as legal documents become more accessible, the magnitude of this challenge will begin to dissipate.

Another challenge we faced was related to the legal documents themselves and the anomalies in the way these documents are structured. While CourtListener returns the documents and their associated metadata with a particular format, the contents of the documents themselves do not adhere to any structure whatsoever. In other words, we could not write code based on the assumption a particular piece of information is at the same place in every document, rather we developed scripts to extract information from the documents based on the patterns we were able to find and we supplemented this with categorical data from the Supreme Court database. This lack of uniformity in legal documents created an extra layer of complexity to our project, however this challenge forced us to familiarize ourselves with the content of the legal documents themselves. As a result, we became more equipped to develop a tool that leverages the content that is present in these documents.

2 Future Plans

Although we have fulfilled the requirements we had initially laid out for our project, we are not yet done developing and improving Lucem. We have identified three main ways that we can improve our system in the future: implementation of a user feedback feature, expanding our document corpus, and extending our application to mobile platforms.

We feel that a feature that allows users to give us feedback on our search results would enhance the system performance of Lucem for all users. Ideally, users would notify us of cases that were inherently valuable to building their specific case and using these notifications we would promote the importance of these cases in subsequent searches. This would be system wide and provide all users of our system recommendations for relevant cases based on previous searches.

Considering that one of the drawbacks of Lucem is the lack of documents, one of our main points of improvement will focus on expanding our corpus. Currently, we are only utilizing the public API of CourtListener, and furthermore we are currently limited to the sixty four thousand documents in Supreme Court History. Our first move will be to expand our corpus from only the Supreme Court to also include state and appellate courts. We will continue to research techniques to improve our database of documents in hopes to create a more effective system.

One of the ways in which we will enhance the efficiency of our system is through more user testing. Specifically, we will aim to create a user comparison between our system and that of WestLaw and LexisNexis, in order to see where our system exceeds these systems and where we can possibly improve. This testing will allow us to pinpoint our efficiency, effectiveness, and satisfaction in the legal research industry.

Finally, although our system is a web application currently, we plan to expand our system to mobile platforms. We have found from lawyers that they often work on their tablets or personal phones, and this upgrade would be extremely beneficial for our users.

In addition to expanding the systems functionality, we also plan on performing more stringent testing on our system. While we have already performed user testing, we have yet to conduct A/B testing and elicit quantitative data from the testers. To A/B test our system, we plan to compare the efficiency of our system and the benchmark, LexisNexis and WestLaw, by having users perform equivalent tasks on each and recording the amount of time required. We will rely on the users legal intuition to inform us whether the search results we provide are as accurate as the current systems. Finally, we will also determine means to test and quantify the degree of user satisfaction with our system compared to existing systems.

3 Summary

At Santa Clara University, our engineering program is rooted in the belief that the beauty of engineering is in its ability to affect change. In developing Lucem we were driven by this same belief. We believe that the duopoly that WestLaw and LexisNexis have on pertinent legal information serves as a fundamental encroachment on our civil liberties and our right to have transparent access to the laws that define our society.

The importance of a tool that brings transparent access to the field of law can be seen from one of our test groups: the Santa Clara Public Defenders Office. The mission statement of the Santa Clara PD is to provide legal counsel to those that would not otherwise be able to afford it. By concealing functionality and documentation behind an unaffordable paywall, WestLaw and LexisNexis are imposing a heavy burden on these public institutions.

Thus, funds that would otherwise be used to better society are instead being used to pay for access to documentation and functionality that would otherwise be free. When thinking of this more abstractly, two private organizations are attempting to consolidate the legal means of production and as a result, a public organization whose main goal is to aid society is being suppressed from fulfilling their duty. Now, this is not merely a legal problem, this is a societal problem.

There is a movement called the FreeLaw initiative, which consists of universities and startups across across the nation who are attempting to open source all legal information. We are proud to provide this free legal research tool, and contribute to this movement.

What Lucem is at its core is an attempt to bring transparency and access back to the field of law, and we hope that we have been able to shine a light on what is possible in an industry that desperately needs a rejuvenation.