Olivet Nazarene University
Digital Commons @ Olivet

Student Scholarship - Computer Science

Computer Science

Spring 3-1-2018

LSAT practicum: an application of human based computation

Seth Rivett
Olivet Nazarene University, sethrivett@gmail.com

Follow this and additional works at: https://digitalcommons.olivet.edu/csis_stsc

Part of the Computer Law Commons, and the Computer Sciences Commons

LSAT PRACTICUM

An Application of Human-Based Computation

Abstract

Human-based computation can be applied to solve problems too hard for a single computer. Crowdsourcing can be applied to ethical modeling by splitting ethical situations among humans. In this senior research project, the crowdsourcing method is applied to produce an ethical model for what web crawlers are allowed to do on websites. By evaluating questions about terms of use on a website, users provide context for the robots. An obstacle to this project is getting the right crowd to participate in the problem. The crowd of potential law students was selected as students typically answer questions to study for a major entrance test into law school. This tool can allow these students to practice legal analysis while letting them build to ethical web knowledge, which is in turn generated into robot-readable code in the form of the Robot Exclusion Protocol. The results were limited by the size of the crowd in this project.

Seth Rivett

sethrivett@gmail.com

Thesis: Web development can be used to create and process ethical models.

Problem Summary

Web Crawling Ethics

Stakeholders

In data mining there are three main stakeholders, the government, the organizations with data and those collecting the data. The members of each entity involved have different perspectives which may conflict in the practice of data mining, and in particular, web crawling. The government has to decide how to judge cases of data mining. There needs to be some standards for judging disputes between the organizations having their data accessed and those searching for and storing the information. The government is not only a regulator of data mining methods, but also a user of data mining methods. The laws in place consider items such as what was done as well as what was intended. These laws do not contain explicit legislation for every computing action; however, some are general enough to be applied to any scenario. Even as the government seeks to regulate the unwarranted or damaging use of internet robots in web crawling, the government may have its own data discovery systems in place. Internal rules can only offer a limited control over government practices in web crawling. This project considers some of the use cases of government data mining, such as the collection of foreign data for intelligence purposes.

Organizations with data to be found may not want their data to be stored by third parties. Some of the unprotected information online may be proprietary. An organization may not want web crawlers to access the system because the increased traffic could increase latency for other clients or, in extreme cases, deny access to the system. Some of these consequences may not be intended, but the data miner may not know the capacity of the system and bring system functionality to a halt before the organization has time to respond to the use of the web crawler. There is a need for a set of standards for fostering peaceable interactions between data miners and the organizations owning the prized data.

The organizations collecting the data can have more profitable ventures if they can access a lot of data at once. Data mining depends on finding the patterns and aggregate results of collections of information that might not be observed in smaller data sets. Data miners can use the information for marketing, security, healthcare and text analysis as well as a number of other applications.

Ethics Language Building

A number of laws govern the internet, but these laws are general. If the laws were more specific, the lawmakers would have trouble keeping up with the technology as they dictated what should be done. Policymakers need to work with inventors and stakeholders. Standards for ethics written by the non-governmental stakeholders specific to the new web technologies can reduce the need for the government to step into the technological sphere for arbitration or adding more restrictive legislation.

The robot exclusion protocol[1] was developed as a way to handle the problem of automated access to networked systems. With this protocol, the system administrators provide a brief document describing who can access the content and which content can be accessed. This allows the system

administrator to informally blacklist any robots web crawling the system. It also can be used to tell a robot which locations of the website should not be visited. To make sure that the robot can easily find the document, it is the job of the administrator to place the document in a file called "robots.txt" at the root of the website.

As web crawling technology develops, there are more items to be considered about what should or should not be crawled. These new items ought to be added to the robot exclusion protocol in order to keep the peace between organizations and data miners as developed data analysis methods are introduced to organizations.

Test Preparation

Studying for tests can become a tedious task. Many graduate tests or entrance exams are long enough to warrant a preparation course just for taking the exam. One such test is the Law School Admission Test (LSAT) provided by the Law School Admission Council. Many people take this test every year, and their placement in a law school is affected by their results. Therefore the preparation for this exam will have an impact on the lives of future law students and potentially alter the beginnings of their career in law. The test involves reading, logic and analysis questions. The analysis of large works is the task of a lawyer. Being able to efficiently read a text and come to a decision about what it implies is a skill that a potential law student ought to practice before taking the LSAT. To do this, one needs a text to read from and a set of items to search for or compare to within the scope of the text. Although general literary analysis may be a useful skill for these students, the ability to quickly sift through legal terminology may prove more helpful.

System Design Overview

Platform as a Service

Microsoft Azure

Politically, the system is a crowdsourcing technique, which could be seen as much as an organizational method as a technical innovation. Practically, the system is a web application built on a virtual platform. Microsoft has developed the platform in use, Microsoft Azure, upon which LSAT Practicum is hosted. Microsoft Azure is a Platform as a Service which allows for the creation and management of small to large-scale applications Azure tools allow developers to create a program which can be executed on multiple systems. This feature is known as cross-platform support. Microsoft Azure also allows for backend development. The platform of Azure stretches then, all the way from the user interface to the information models and their database.

Network Objects

The design uses four distinct network objects in Azure. The first object is the web application. This is what the users will actually see. Since it is a web application, it is linked to a specific URL. For this project, the associated link is lsatpracticum.azurewebsites.net. The second network object is the web service. This is a contract between Microsoft and developers denoting which features within Azure will be allowed to be used by the network objects. The third network object to consider is the SQL database.

This database is attached to a service plan. This allows for monetization by Microsoft as the size of your database is limited by the conditions of the service plan. The fourth network object is the database server. This object virtually hosts the database. This does not significantly contribute to the functionality within the scope of this project, but in an expanded version of the project, this feature would allow for managing multiple databases.

Languages Used

Once Microsoft Azure was established as the platform of choice, the challenge was to get database connectivity working. A Microsoft tutorial was found which used asp.net and C# to connect a Visual Studio project to an SQL server within the Azure system. This tutorial became the determining factor of the languages used in the project. The tutorial used C#, asp.net and cshtml. The design pattern Model-View-Controller (MVC) was also used in this project. The language C# was used to hold the variables and functions of the model portions. As the program was developed, a second model was added to the tutorial. A question was stored in a C# model, and a generated robot protocol was stored in another C# model. When dealing with MVC, each view can be a single webpage. In this website, the webpages were stored as cshtml, which allows for html with additional possibilities for control structures such as a *foreach* loop. This allowed for listing multiple model instances as objects were created or removed.

Frameworks and Tools Used

In this project, the development was done with using Microsoft Visual Studio. This IDE was chosen as a result of following a tutorial for connecting to a back-end database through the Microsoft Azure system. A disadvantage to using Visual Studio was the initial time to install. The installation was also a significant cost in terms of space on the machine. When installing Visual Studio, there are a number of features which can be added to the install but require additional space on a hard drive. The Azure feature was installed as well as a data lake feature (which was never used). Once set up, the IDE was convenient for development. The bottlenecks in developing rested more in network connectivity to Azure than in the Visual Studio editing interface. Even the network time was not always lost, because this gave me a moment to think about whichever problem was at hand or what a possible fix might be.

Another tool used was Adobe Dreamweaver. This tool was used to create the about page for LSAT Practicum. In this development environment, there is a feature for editing a webpage while simultaneously viewing the effects of the edited code on that webpage. Such a layout can improve webpage editing efficiencies. It took some time to learn how to import the finished product from Adobe Dreamweaver into Microsoft Visual Studio and the correct website folders for production. If Adobe Dreamweaver had been used on more than one webpage, the efficiencies of the project might have been streamlined. With the functionality already in place, Adobe Dreamweaver would be a good tool for improving the looks of a website.

System Design Details

The question system is the architecture responsible for managing, storing and serving questions. In *Figure 1*, the question system is depicted to reveal the interplay among the various pieces of the

system. Three of these pieces are the network objects, the web interface, the server and the database. Note that the service plan (not shown) is also a network object in this system. The final piece in this diagram is the end-user. The end-user interacts with the web interface. The web interface, in turn, asks the server and database for the information associated with a question. The web interface also passes the user input to the database server and database to store the results of the end-user's choice for each question. Statistics about each question can be loaded by accessing the specific question model.
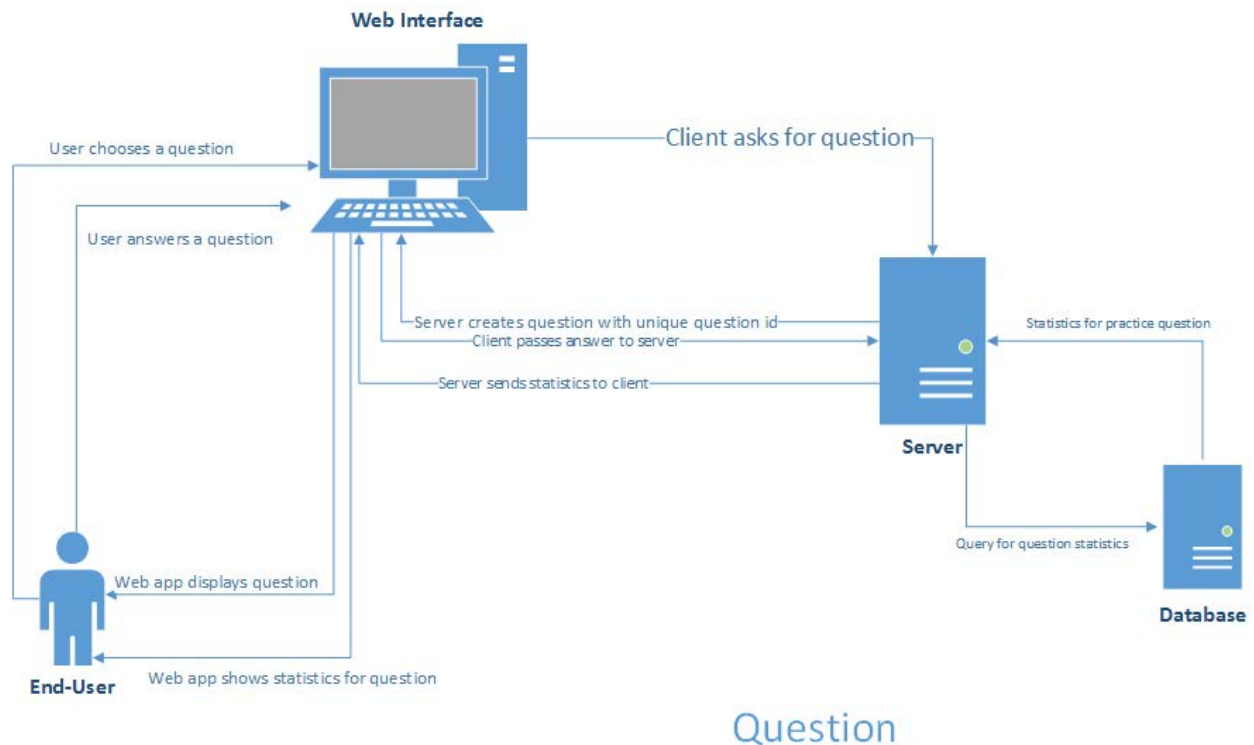
Figure-1 Question System

Webpages

The Q*uestion Interface* (displayed in *Figure-*2) is centered on the question webpage. This webpage contains the URL for the website terms, the context of the question and the answer choices as described below:

Question Interface Fields
- URL: This field is also used for identifying the website to map robot protocol to.
- Passage: This is the background text needed to answer the questions. The LSAT has brief passages included in the questions. The passages on LSAT Practicum are currently longer than LSAT questions (See https://www.lsac.org/). The passage comes from the website being modeled, and is currently taken from a set of terms on the website being modeled.
- Question: This is the one-sentence problem that will guide users to search the passage and choices to select an answer. It may be general to avoid a bias toward an answer.

- Choice A...Choice E: These are the five answer choices to the question given.

At the bottom of the page there are radio buttons. There is also a clear way to submit the answer.

The Question webpage is shown in *Figure-2* below.

LSAT Practicum

Question

Url
https://www.olivet.edu/sites/default/files/pdf/AcceptableUsePolicy.pdf
Passage
12.1. Violations The following list of violations is provided as a list of examples only as is not meant to be exhaustive or complete. 12.1.1. Malicious Activity Violations • Attempting to defeat the network, servers, or administrative computer's security s ystems, mechanisms, or controls. • Possessing, using, distributing, or developing password cracking programs, algorithms , keystroke loggers, or any hacking utility. • Launching or participating in a denial of service attack or any other action with the intent of damaging or disrupting any network, system, or device on or outside of ONUnet. • Disabling, tampering with, or accessing any ONUnet system, network device or cabling. • Any action intended to disrupt normal system services or which adversely affects other u sers' accounts, electronic material, e -mail, or network performance. Ver. 4 .0 Page 10 • Any attempt to probe or scan ONUnet or any network outside of ONUnet. • The creation or intentional distribution of any virus ,worm , or malicious code .
Question
Given the above information about malicious activity, which would be the most accurate statement about ethics for this site?
ChoiceA
Getting a huge group of people to use the website at the same time would be okay.
ChoiceB
Getting a huge group of people to use the website at the same time would be unethical.
ChoiceC
Users are responsible for other people's accounts.
ChoiceD
Distribution of potentially malicious files is okay if done without bad intentions.
ChoiceE
Viewing many webpages at once should not be done.

A ○ B ○ C ○ D ○ E ○

Answer

Back to List

2018 - Olivet Nazarene University - CSIS 492

Figure-2 Question Interface

To see the results of the questions answered, there is a humans' results page, *Results*. This webpage displays the number of votes for each answer. The critical goals of this webpage are as follows:

- Show the top aggregate answer(s), named the "Crowd Answer"
- Display the user's answer
- List the question and answer choices for review
- Give the tally for each other answer to check for close answers
- Provide access to more information about the question

The link to more information currently leads to a statistics page. This page shows the same data as above with the addition of sample size and dominance percentage values as well as the maximum number of votes received for the top answer(s). The statistics webpage is in *Figure-3* below.

Figure-3 Statistics Webpage

6. Robot Results Page
   c. Website List
   d. Extensions of Robot Exclusion Protocol

There is also a webpage for displaying the data which can be read by robots. The Robot Exclusion Protocol was used as the starting point for this output. The features of this webpage, *Robots*, are explained below:

- Each website is listed on the webpage.
- Multiple questions for the same website merge into one distinct robot protocol for that website.
- Duplicate protocols are generated for questions associated with the same website.
- The generated protocols can be edited, copied and pasted from the text area on the webpage.

This project uses some of existing proposed extensions to the Robot Exclusion Protocol as well as some new additions defined in this web application. Below are the items to map to each website for use in developing a "robots.txt".

Variables to Include in Web Ethics Model

Law/Ethics Principle	Variable in Model	Mapping to Robots.txt
Security: Principle of Least Privilege	boolean PortScanningAllowed	**Disable Port Scanning**
Computer Fraud and Abuse Act	boolean HasCrawlDelay	used to write next line (or not)
Computer Fraud and Abuse Act	int CrawlDelayTime (Seconds)	CrawlDelay: CrawlDelayTime
Computer Fraud and Abuse Act	int HitsAllowed	**Simultaneous: HitsAllowed**
Digital Millennium Copyright Act	boolean AllowsCommercialIndexing	used to write next line (or not)
Digital Millennium Copyright Act	boolean IndexBySitemap	**Sell: directory**
End-User License Agreements	String: SitemapURL	Sitemap: <SitemapURL>
End-User License Agreements	boolean MustReviewTerms	**Update Terms: TermsURL**
End-User License Agreements	String: TermsURL	**Terms: TermsURL**
Digital Millennium Copyright Act	String: Copyright Directory Content	Copyright Content: Directory
Digital Millennium Copyright Act	String: Copyright Structure Content	Copyright Structure: Directory
Electronic Communications Privacy Act of 1986	String: AbandonedDirectoryURL	**Abandoned:** abandoned URL
Patriot Act	String: CommunicationMetadataDirectoryURL	**Communication Metadata: metadata directory URL**
FISA Amendments Act	String: ForeignDataDirectory	**Foreign: directory**
Communications Decency Act (Section 230)	Boolean: BotResponsibleForSpeech	Used to determine if next field should be written
Communications Decency Act (Section 230)	String: InteractiveDirectoryURL	**Interactive Directory: interactiveURL**

*Bolded listings are new LSAT Practicum proposals for extensions to the Robot Exclusion Protocol.

**The previously proposed extensions used are listed on Wikipedia at https://en.wikipedia.org/wiki/Robots_exclusion_standard#Nonstandard_extensions.[3]

With the current design, the table above can be used to design questions about a website which map to actions generating particular portions of the robot code. There is a webpage *Create* which allows a question to be formed by a user. Each answer can be assigned to an action and a value which will be used to write the robot protocol for that website. For example, in asking whether a robot ought to write data to a website, a question can be written which maps the answer "Robots cannot make up information submitted to the website." to the action "BotResponsibleForSpeech" with a value of "Yes". Then, when test-takers answer the question with the above answer as the *CrowdAnswer*, accessing the robot protocol for the website will show generate the field "Interactive Directory:" in the protocol. The output webpage for the robot protocols in the webpage *Robots* is shown below:

LSAT Practicum

Generated Robots.txt
Robot Exclusions by Website

Robots.txt

Url https://www.olivet.edu/sites/default/files/pdf/AcceptableUsePolicy.pdf

RobotsTxt | User-agent: *
Disallow:

Disable Port Scanning
Simultaneous: 1
Disable Commercial Indexing
Sitemap:
Update Terms:
Copyright Content Directory:
Copyright Structure Directory:
Abandoned:
Communication Metadata:
Foreign:

Robots.txt

Url https://savethestorks.com/terms-conditions/

RobotsTxt | User-agent: *
Disallow:

Robots.txt

Url https://www.wycliffe.org/terms-and-conditions

RobotsTxt | User-agent: *
Disallow:

Robots.txt

Url https://www.youtube.com/static?gl=GB&template=terms

RobotsTxt | User-agent: *
Disallow:

Proposed Extensions to the Robot Exclusion Protocol

Law/Ethics Principle | Variable to model idea | Mapping to Robots.txt | Piece of Information to Connect to Website

Security: Principle of Least Privilege | boolean PortScanningAllowed | Disable Port Scanning | This indicates whether or not the robot can browse to hidden links.

Computer Fraud and Abuse Act | boolean HasCrawlDelay | used to write next line (or not) | This tells the robot whether or not it must pause while searching to keep the network running quickly.

Computer Fraud and Abuse Act | int CrawlDelayTime (Seconds) | CrawlDelay: CrawlDelayTime | This is the timed pause in-between searching pages.

Computer Fraud and Abuse Act | int HitsAllowed | Simultaneous: HitsAllowed | The number here denotes how many times a robot may open a page at the same time.

Digital Millennium Copyright Act | boolean AllowsCommercialIndexing | used to write next line (or not) | This indicates whether indexed information should not be sold.

Digital Millennium Copyright Act | boolean IndexBySitemap | Sell: directory | This blacklists the dark web for the web crawler.

End-User License Agreements | String: SitemapURL | Sitemap: | This gives the web crawler a whitelist for indexing the website.

End-User License Agreements | boolean MustReviewTerms | Update Terms: TermsURL | The bot must check for changes to the website terms.

End-User License Agreements | String: TermsURL | Terms: TermsURL | This denotes the link to the website terms.

Digital Millennium Copyright Act | String: Copyright Directory Content | Copyright Content: Directory | To avoid breaking intellectual property rights for any web party a copyright folder is created.

Digital Millennium Copyright Act | String: Copyright Structure Content | Copyright Structure: Directory | This is used to avoid storing an original web structure.

Electronic Communications Privacy Act of 1986 | String: AbandonedDirectoryURL | Abandoned: abandoned URL | This could simplify the legality of web crawling for government usage.

Patriot Act | String: CommunicationMetadataDirectoryURL | Communication Metadata: metadata directory URL | This specifies the directory for compliance with the Patriot Act.

FISA Amendments Act | String: ForeignDataDirectory | Foreign: directory | This allows for unwarranted data collection of foreign intelligence.

Communications Decency Act (Section 230) | Boolean: BotResponsibleForSpeech | Used to determine if next field should be written | This monitors web crawlers performing webpage input.

Communications Decency Act (Section 230) | String: InteractiveDirectoryURL | Interactive Directory: interactiveURL | All items in subdirectories contain user input which the robot will be responsible for.

Back to List

To create the questions which will later be mapped to the *Robots* webpage, the webpage *Create* (modified from the tutorial *Create* webpage) is used.[2] Drop-down variables for mapping ethical actions are included to speed up entry of the questions. The webpage *Create* is shown in the following figure:

LSAT Practicum

Description	
Created Date	02/17/2018
Url	
Passage	
Question	
ChoiceA	
ActionA	MustReviewTerms ▼
ValueA	Yes
ChoiceB	
ActionB	MustReviewTerms ▼
ValueB	
ChoiceC	
ActionC	AllowsCommercialIndexing ▼
ValueC	
ChoiceD	
ActionD	IndexBySitemap ▼
ValueD	
ChoiceE	
ActionE	BotResponsibleForSpeech ▼
ValueE	

Create

Back to List

A simple modified home page from the Microsoft tutorial links each of the webpages together as shown below.[2]

LSAT Practicum

Questions

Create New Question

About LSAT Practicum

Robot Exclusions

Description

Olivet Web Ethics

Try Question | Details | Results |

Save the Storks: Story Submission

Try Question | Details | Results |

Wycliffe: Crawler Requirements

Try Question | Details | Results |

YouTube content

Try Question | Details | Results |

2018 - Olivet Nazarene University - CSIS 492

Analysis of Results

An email explaining LSAT Practicum and including a link to the proof of concept website was published on January 8, 2018 to Dr. Bareiss, Dr. Vail and the students within the computer science department at Olivet Nazarene University. On January 17, 2018, the results were investigated.

The human results for each of the four questions are listed below:

LSAT Practicum

Statistics

Aggregate Information for Question

CrowdAnswer

B

MostRecentAnswer

A

Description	Olivet Web Ethics
Created Date	2018-01-09
Url	https://www.olivet.edu/sites/default/files/pdf/AcceptableUsePolicy.pdf
Passage	12.1. Violations The following list of violations is provided as a list of examples only as is not meant to be exhaustive or complete. 12.1.1. Malicious Activity Violations • Attempting to defeat the network, servers, or administrative computer's security s ystems, mechanisms, or controls. • Possessing, using, distributing, or developing password cracking programs, algorithms , keystroke loggers, or any hacking utility. • Launching or participating in a denial of service attack or any other action with the intent of damaging or disrupting any network, system, or device on or outside of ONUnet. • Disabling, tampering with, or accessing any ONUnet system, network device or cabling. • Any action intended to disrupt normal system services or which adversely affects other u sers' accounts, electronic material, e -mail, or network performance. Ver. 4 .0 Page 10 • Any attempt to probe or scan ONUnet or any network outside of ONUnet. • The creation or intentional distribution of any virus ,worm , or malicious code .
Question	Given the above information about malicious activity, which would be the most accurate statement about ethics for this site?
ChoiceA	Getting a huge group of people to use the website at the same time would be okay.
ChoiceB	Getting a huge group of people to use the website at the same time would be unethical.
ChoiceC	Users are responsible for other people's accounts.
ChoiceD	Distribution of potentially malicious files is okay if done without bad intentions.
ChoiceE	Viewing many webpages at once should not be done.
CountA	1
CountB	4
CountC	0
CountD	0
CountE	0
Sample Size	5
MaxAnswer	4
DominancePercentageForCrowdAnswer	80
DominancePercentageForMostRecentAnswer	20
DominanceA	20
DominanceB	80
DominanceC	0
DominanceD	0
DominanceE	0

Edit | Back to Results | Back to List

2018 - Olivet Nazarene University - CSIS 492

Figure-4 "Olivet Web Ethics" Results – January 17, 2018

LSAT Practicum

Statistics
Aggregate Information for Question

CrowdAnswer
B C

MostRecentAnswer
B

Description	Save the Storks: Story Submission
Created Date	2018-01-09
Url	https://savethestorks.com/terms-conditions/
Passage	" BEFORE SUBMITTING YOUR STORY YOU NEED TO AGREE TO THESE TERMS The story you're sharing has to be accurate and honest, as far as you know. The story cannot mention people that don't want to be mentioned. We are in the business of empowering and loving people and want to make sure no one will be unknowingly mentioned and embarrassed by what you're sharing. By sharing your story you're giving us permission to share your name on our blog as well as social media platforms, unless otherwise stated. We understand that not everyone gets paid to correct grammar, but some of our staff do and they might adjust some commas and quotation marks on your content so you have to be okay with that. If you do choose to upload photos, you're giving us permission to share those photos on our sites along with printed materials and other promotional usage. Any photo you release to us needs to be owned by yourself or you have permission to share it, again, we're just trying not to blindside people! If any of these terms seem weird, please email us at mystory@savethestorks.com and we'll talk you through it! BEFORE SUBMITTING YOUR STORY YOU NEED TO AGREE TO THESE TERMS The story you're sharing has to be accurate and honest, as far as you know. The story cannot mention people that don't want to be mentioned. We are in the business of empowering and loving people and want to make sure no one will be unknowingly mentioned and embarrassed by what you're sharing. By sharing your story you're giving us permission to share your name on our blog as well as social media platforms, unless otherwise stated. We understand that not everyone gets paid to correct grammar, but some of our staff do and they might adjust some commas and quotation marks on your content so you have to be okay with that. If you do choose to upload photos, you're giving us permission to share those photos on our sites along with printed materials and other promotional usage. Any photo you release to us needs to be owned by yourself or you have permission to share it, again, we're just trying not to blindside people! If any of these terms seem weird, please email us at mystory@savethestorks.com and we'll talk you through it! "
Question	The passage above provides an informal description of what should not be done on the website. What should be an ethical extension of these terms when considering web robots?
ChoiceA	Robots can submit any information they want.
ChoiceB	Robots must submit accurate information.
ChoiceC	Once submitted, indexing or storing other people's stories is okay.
ChoiceD	Robots should not store any stories they read.
ChoiceE	Terms must be reviewed regularly for updates by sending an email to "mystory@savethestorks.com".
CountA	0
CountB	1
CountC	1
CountD	0
CountE	0
Sample Size	2
MaxAnswer	1
DominancePercentageForCrowdAnswer	50
DominancePercentageForMostRecentAnswer	50
DominanceA	0
DominanceB	50
DominanceC	50
DominanceD	0
DominanceE	0

Edit | Back to Results | Back to List

Figure-5 "Save the Storks: Story Submission" Results – January 17, 2018

LSAT Practicum

Statistics

Aggregate Information for Question

CrowdAnswer

MostRecentAnswer

Description	Wycliffe: Crawler Requirements
Created Date	2018-01-09
Url	https://www.wycliffe.org/terms-and-conditions
Passage	At this time, www.wycliffe.org does not include interactive areas. Should this site come to include such areas, the following will apply:You are welcome to post, transmit, or submit messages and other materials (which include uploading files, inputting data or any other materials, or engaging in any form of communication in connection with this Site) (collectively "Messages") to bulletin boards, chat rooms, or other public areas within, or in connection with, this Site (collectively "Forums"). However, Wycliffe accepts no responsibility whatsoever in connection with or arising from such Messages. In addition, children under the age of 13 are not authorized to post Messages or participate in Forums on this Site. Wycliffe does not endorse and has no control over the content of Messages submitted by others to Forums. Messages submitted to Forums are not necessarily reviewed by Wycliffe prior to posting, and do not necessarily reflect the opinions or policies of Wycliffe. Wycliffe makes no warranties, express or implied, as to the content of the Messages in the Forums or the accuracy and reliability of any Messages and other materials in the Forums. Nonetheless, Wycliffe reserves the right to prevent you from submitting Materials to Forums and to edit, restrict, or remove such Messages for any reason at any time.
Question	This clause would require a web crawler to:
ChoiceA	Store information about communication on the website.
ChoiceB	Monitor what the robot says on the website.
ChoiceC	Check for a change in website terms.
ChoiceD	Keep a record of old data collected from the website.
ChoiceE	Index foreign data separately to be used for U.S. counter-intelligence.
CountA	0
CountB	0
CountC	0
CountD	0
CountE	0
Sample Size	0
MaxAnswer	0
DominancePercentageForCrowdAnswer	0
DominancePercentageForMostRecentAnswer	0
DominanceA	0
DominanceB	0
DominanceC	0
DominanceD	0
DominanceE	0

Edit | Back to Results | Back to List

2018 - Olivet Nazarene University - CSIS 492

Figure-6 "Wycliffe Crawler Requirements" Results – January 17, 2018

LSAT Practicum

CrowdAnswer
A B C D E

MostRecentAnswer

Description	YouTube content
Created Date	2018-01-09
Url	https://www.youtube.com/static?gl=GB&template=terms
Passage	"7. Content 7.1 As a YouTube account holder you may submit Content. You understand that whether or not Content is published, YouTube does not guarantee any confidentiality with respect to Content. 7.2 You retain all of your ownership rights in your Content, but you are required to grant limited licence rights to YouTube and other users of the Service. These are described in paragraph 8 of these Terms (Rights you licence). 7.3 You understand and agree that you are solely responsible for your own Content and the consequences of posting or publishing it. YouTube does not endorse any Content or any opinion, recommendation, or advice expressed therein, and YouTube expressly disclaims any and all liability in connection with Content. 7.4 You represent and warrant that you have (and will continue to have during your use of the Service) all necessary licenses, rights, consents, and permissions which are required to enable YouTube to use your Content for the purposes of the provision of the Service by YouTube, and otherwise to use your Content in the manner contemplated by the Service and these Terms. 7.5 You agree that your conduct on the site will comply with (and you agree that the content of all of your Content shall comply with) the YouTube Community Guidelines, found at https://www.youtube.co.uk/t/community_guidelines, as updated from time to time. 7.6 You agree that you will not post or upload any Content which contains material which it is unlawful for you to possess in the country in which you are resident, or which it would be unlawful for YouTube to use or possess in connection with the provision of the Service. 7.7 You agree that Content you submit to the Service will not contain any third party copyright material, or material that is subject to other third party proprietary rights (including rights of privacy or rights of publicity), unless you have a formal licence or permission from the rightful owner, or are otherwise legally entitled, to post the material in question and to grant YouTube the licence referred to in paragraph 8.1 below. 7.8 On becoming aware of any potential violation of these Terms, YouTube reserves the right (but shall have no obligation) to decide whether Content complies with the content requirements set out in these Terms and may remove such Content and/or terminate a User's access for uploading Content which is in violation of these Terms at any time, without prior notice and at its sole discretion. 7.9 You further understand and acknowledge that in using the Service, you may be exposed to Content that is factually inaccurate, offensive, indecent, or otherwise objectionable to you. You agree to waive, and hereby do waive, any legal or equitable rights or remedies you have or may have against YouTube with respect to any such Content."
Question	Which of the following best describes Youtube's terms regarding internet content?
ChoiceA	YouTube is held responsible for anything you post.
ChoiceB	Any user is liable for unlawful content posted.
ChoiceC	Any user may download and store content from youtube.
ChoiceD	Content at YouTube downloaded from other countries must be stored separately.
ChoiceE	YouTube content may only be accessed through the use of the provided search menu or following the links given by the sitemap.
CountA	0
CountB	0
CountC	0
CountD	0
CountE	0
SampleSize	0
MaxAnswer	0
DominancePercentageForCrowdAnswer	0
DominancePercentageForMostRecentAnswer	0
DominanceA	0
DominanceB	0
DominanceC	0
DominanceD	0
DominanceE	0

Edit | Back to Results | Back to List

Figure-7 "YouTube content" Results – January 17, 2018

The first question which was about Olivet Web Ethics produced the largest sample size: 5. Based on the reading of the web terms section, the crowd answer was that "Getting a huge group of people to use the website at the same time would be unethical." The dominance of this answer was 80%. One person answered "Getting a huge group of people to use the website at the same time would be okay." Two types of information were provided in this passage which could help users to give an ethical analysis. One is the consideration of intent of the user. Another is the actual action chosen by the user. The most common answer was the one that limited Olivet web users the most specifically and disregarded intentions.

The next question was about the organization Save the Storks. Upon review, the passage text for this question was accidentally listed twice. This may have discouraged people from answering the question. The sample size for the question was two. The two users chose different answers. One favored limiting the robots' activities. The other permitted other people's stories to be stored as data.

The questions of Wycliffe and YouTube ethics remained unanswered during this test.

The robot exclusion protocol generated for each website is displayed in *Figure-8* below:
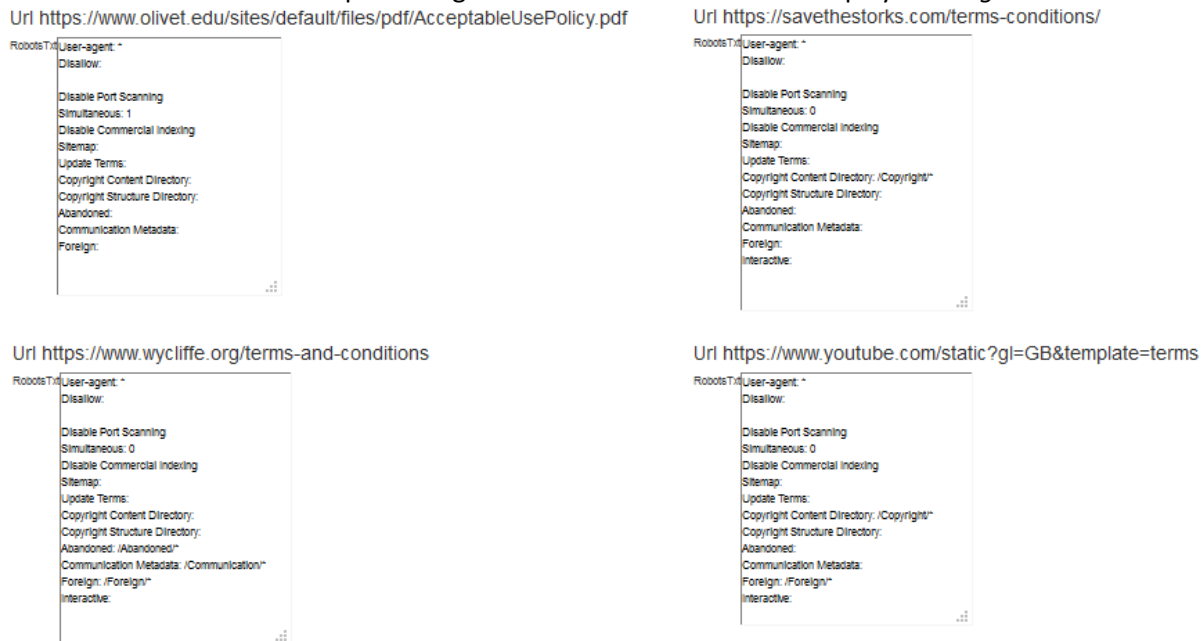

Url https://www.olivet.edu/sites/default/files/pdf/AcceptableUsePolicy.pdf

```
RobotsTxt User-agent: *
          Disallow:

          Disable Port Scanning
          Simultaneous: 1
          Disable Commercial Indexing
          Sitemap:
          Update Terms:
          Copyright Content Directory:
          Copyright Structure Directory:
          Abandoned:
          Communication Metadata:
          Foreign:
```

Url https://savethestorks.com/terms-conditions/

```
RobotsTxt User-agent: *
          Disallow:

          Disable Port Scanning
          Simultaneous: 0
          Disable Commercial Indexing
          Sitemap:
          Update Terms:
          Copyright Content Directory: /Copyright/*
          Copyright Structure Directory:
          Abandoned:
          Communication Metadata:
          Foreign:
          Interactive:
```

Url https://www.wycliffe.org/terms-and-conditions

```
RobotsTxt User-agent: *
          Disallow:

          Disable Port Scanning
          Simultaneous: 0
          Disable Commercial Indexing
          Sitemap:
          Update Terms:
          Copyright Content Directory:
          Copyright Structure Directory:
          Abandoned: /Abandoned/*
          Communication Metadata: /Communication/*
          Foreign: /Foreign/*
          Interactive:
```

Url https://www.youtube.com/static?gl=GB&template=terms

```
RobotsTxt User-agent: *
          Disallow:

          Disable Port Scanning
          Simultaneous: 0
          Disable Commercial Indexing
          Sitemap:
          Update Terms:
          Copyright Content Directory: /Copyright/*
          Copyright Structure Directory:
          Abandoned:
          Communication Metadata:
          Foreign: /Foreign/*
          Interactive:
```

Figure-8 Robot Exclusion Protocols Generated – January 17, 2018

Note that the protocol for Olivet includes the clause "Simultaneous: 1". This means that the ethical decision was made to only allow a robot to scan one webpage at a time on the Olivet website (olivet.edu). Unlike the other protocols, the Olivet protocol did not include an interactive directory clause (Note that none of these clauses had a specific interactive URL directory).

For the Save the Storks robot protocol, notice that the copyright content directory got set to "/Copyright/*". Also an interactive directory clause was added, although the URL was not specified. There was a tie for the highest answer to this question, so both corresponding protocol actions (set copyright directory and include interactive directory) took place.

For the remaining two websites, no questions were answered. This produced a five-way tie among the answer choices, which led to five protocol actions being executed for each website. Since the actions were defined differently for each website, the resulting protocols were slightly different. For instance, the Wycliffe protocol includes specific URLs for its Abandoned and Communication Metadata clauses but the YouTube protocol does not.

It appears then, that a high volume of data as well as adequately written questions are necessary for more relevant results in crowdsourcing.

To increase the data in the results, an Olivet Nazarene University class, Career Seminar, tested the application by attempting the questions. The results from this data addition are pictured and investigated below.

Statistics

Aggregate Information for Question

CrowdAnswer
B

MostRecentAnswer
E

Description	Olivet Web Ethics
Created Date	2018-01-09
Url	https://www.olivet.edu/sites/default/files/pdf/AcceptableUsePolicy.pdf
Passage	12.1. Violations The following list of violations is provided as a list of examples only as is not meant to be exhaustive or complete. 12.1.1. Malicious Activity Violations • Attempting to defeat the network, servers, or administrative computer's security s ystems, mechanisms, or controls. • Possessing, using, distributing, or developing password cracking programs, algorithms , keystroke loggers, or any hacking utility. • Launching or participating in a denial of service attack or any other action with the intent of damaging or disrupting any network, system, or device on or outside of ONUnet. • Disabling, tampering with, or accessing any ONUnet system, network device or cabling. • Any action intended to disrupt normal system services or which adversely affects other u sers' accounts, electronic material, e -mail, or network performance. Ver. 4 .0 Page 10 • Any attempt to probe or scan ONUnet or any network outside of ONUnet. • The creation or intentional distribution of any virus ,worm , or malicious code .
Question	Given the above information about malicious activity, which would be the most accurate statement about ethics for this site?
ChoiceA	Getting a huge group of people to use the website at the same time would be okay.
ChoiceB	Getting a huge group of people to use the website at the same time would be unethical.
ChoiceC	Users are responsible for other people's accounts.
ChoiceD	Distribution of potentially malicious files is okay if done without bad intentions.
ChoiceE	Viewing many webpages at once should not be done.
CountA	1
CountB	13
CountC	7
CountD	1
CountE	11
Sample Size	33
MaxAnswer	13
DominancePercentageForCrowdAnswer	39
DominancePercentageForMostRecentAnswer	33
DominanceA	3
DominanceB	39
DominanceC	21
DominanceD	3
DominanceE	33

Edit | Back to Results | Back to List

MostRecentAnswer
B

Description	Save the Storks: Story Submission
Created Date	2018-01-09
Url	https://savethestorks.com/terms-conditions/
Passage	" BEFORE SUBMITTING YOUR STORY YOU NEED TO AGREE TO THESE TERMS The story you're sharing has to be accurate and honest, as far as you know. The story cannot mention people that don't want to be mentioned. We are in the business of empowering and loving people and want to make sure no one will be unknowingly mentioned and embarrassed by what you're sharing. By sharing your story you're giving us permission to share your name on our blog as well as social media platforms, unless otherwise stated. We understand that not everyone gets paid to correct grammar, but some of our staff do and they might adjust some commas and quotation marks on your content so you have to be okay with that. If you do choose to upload photos, you're giving us permission to share those photos on our sites along with printed materials and other promotional usage. Any photo you release to us needs to be owned by yourself or you have permission to share it, again, we're just trying not to blindside people! If any of these terms seem weird, please email us at mystory@savethestorks.com and we'll talk you through it!"
Question	The passage above provides an informal description of what should not be done on the website. What should be an ethical extension of these terms when considering web robots?
ChoiceA	Robots can submit any information they want.
ChoiceB	Robots must submit accurate information.
ChoiceC	Once submitted, indexing or storing other people's stories is okay.
ChoiceD	Robots should not store any stories they read.
ChoiceE	Terms must be reviewed regularly for updates by sending an email to "mystory@savethestorks.com".
CountA	0
CountB	5
CountC	1
CountD	0
CountE	0
SampleSize	6
MaxAnswer	5
DominancePercentageForCrowdAnswer	83
DominancePercentageForMostRecentAnswer	83
DominanceA	0
DominanceB	83
DominanceC	16
DominanceD	0
DominanceE	0

Edit | Back to Results | Back to List

LSAT Practicum

Statistics
Aggregate Information for Question

CrowdAnswer
A

MostRecentAnswer
A

Description	Wycliffe: Crawler Requirements
Created Date	2018-01-09
Url	https://www.wycliffe.org/terms-and-conditions
Passage	At this time, www.wycliffe.org does not include interactive areas. Should this site come to include such areas, the following will apply:You are welcome to post, transmit, or submit messages and other materials (which include uploading files, inputting data or any other materials, or engaging in any form of communication in connection with this Site) (collectively "Messages") to bulletin boards, chat rooms, or other public areas within, or in connection with, this Site (collectively "Forums"). However, Wycliffe accepts no responsibility whatsoever in connection with or arising from such Messages. In addition, children under the age of 13 are not authorized to post Messages or participate in Forums on this Site. Wycliffe does not endorse and has no control over the content of Messages submitted by others to Forums. Messages submitted to Forums are not necessarily reviewed by Wycliffe prior to posting, and do not necessarily reflect the opinions or policies of Wycliffe. Wycliffe makes no warranties, express or implied, as to the content of the Messages in the Forums or the accuracy and reliability of any Messages and other materials in the Forums. Nonetheless, Wycliffe reserves the right to prevent you from submitting Materials to Forums and to edit, restrict, or remove such Messages for any reason at any time.
Question	This clause would require a web crawler to:
ChoiceA	Store information about communication on the website.
ChoiceB	Monitor what the robot says on the website.
ChoiceC	Check for a change in website terms.
ChoiceD	Keep a record of old data collected from the website.
ChoiceE	Index foreign data separately to be used for U.S. counter-intelligence.
CountA	2
CountB	1
CountC	0
CountD	0
CountE	0
Sample Size	3
MaxAnswer	2
DominancePercentageForCrowdAnswer	66
DominancePercentageForMostRecentAnswer	66
DominanceA	66
DominanceB	33
DominanceC	0
DominanceD	0
DominanceE	0

Edit | Back to Results | Back to List

LSAT Practicum

Results
Question Analysis

CrowdAnswer
B

MostRecentAnswer
B

Description	YouTube content
Created Date	2018-01-09
Url	https://www.youtube.com/static?gl=GB&template=terms
Passage	"7. Content 7.1 As a YouTube account holder you may submit Content. You understand that whether or not Content is published, YouTube does not guarantee any confidentiality with respect to Content. 7.2 You retain all of your ownership rights in your Content, but you are required to grant limited licence rights to YouTube and other users of the Service. These are described in paragraph 8 of these Terms (Rights you licence). 7.3 You understand and agree that you are solely responsible for your own Content and the consequences of posting or publishing it. YouTube does not endorse any Content or any opinion, recommendation, or advice expressed therein, and YouTube expressly disclaims any and all liability in connection with Content. 7.4 You represent and warrant that you have (and will continue to have during your use of the Service) all necessary licenses, rights, consents, and permissions which are required to enable YouTube to use your Content for the purposes of the provision of the Service by YouTube, and otherwise to use your Content in the manner contemplated by the Service and these Terms. 7.5 You agree that your conduct on the site will comply with (and you agree that the content of all of your Content shall comply with) the YouTube Community Guidelines, found at https://www.youtube.co.uk/t/community_guidelines, as updated from time to time. 7.6 You agree that you will not post or upload any Content which contains material which it is unlawful for you to possess in the country in which you are resident, or which it would be unlawful for YouTube to use or possess in connection with the provision of the Service. 7.7 You agree that Content you submit to the Service will not contain any third party copyright material, or material that is subject to other third party proprietary rights (including rights of privacy or rights of publicity), unless you have a formal licence or permission from the rightful owner, or are otherwise legally entitled, to post the material in question and to grant YouTube the licence referred to in paragraph 8.1 below. 7.8 On becoming aware of any potential violation of these Terms, YouTube reserves the right (but shall have no obligation) to decide whether Content complies with the content requirements set out in these Terms and may remove such Content and/or terminate a User's access for uploading Content which is in violation of these Terms at any time, without prior notice and at its sole discretion. 7.9 You further understand and acknowledge that in using the Service, you may be exposed to Content that is factually inaccurate, offensive, indecent, or otherwise objectionable to you. You agree to waive, and hereby do waive, any legal or equitable rights or remedies you have or may have against YouTube with respect to any such Content."
Question	Which of the following best describes Youtube's terms regarding internet content?
ChoiceA	YouTube is held responsible for anything you post.
ChoiceB	Any user is liable for unlawful content posted.
ChoiceC	Any user may download and store content from youtube.
ChoiceD	Content at YouTube downloaded from other countries must be stored separately.
ChoiceE	YouTube content may only be accessed through the use of the provided search menu or following the links given by the sitemap.
CountA	0
CountB	9
CountC	0
CountD	0
CountE	1

Statistics | Back to List

One of the most exciting parts of this data is that there are now at least some responses to all of the questions. This means that there should be no robot texts based only on defaults.

Below are the results of the robots.txt files generated:

Robots.txt

Url https://www.olivet.edu/sites/default/files/pdf/AcceptableUsePolicy.pdf

RobotsTxt
```
User-agent: *
Disallow:

Disable Port Scanning
Simultaneous: 1
Disable Commercial Indexing
Sitemap:
Update Terms:
Copyright Content Directory:
Copyright Structure Directory:
Abandoned:
Communication Metadata:
Foreign:
```

Robots.txt

Url https://savethestorks.com/terms-conditions/

RobotsTxt
```
User-agent: *
Disallow:

Disable Port Scanning
Simultaneous: 0
Disable Commercial Indexing
Sitemap:
Update Terms:
Copyright Content Directory:
Copyright Structure Directory:
Abandoned:
Communication Metadata:
Foreign:
Interactive:
```

Robots.txt

Url https://www.wycliffe.org/terms-and-conditions

RobotsTxt | User-agent: *
Disallow:

Disable Port Scanning
Simultaneous: 0
Disable Commercial Indexing
Sitemap:
Update Terms:
Copyright Content Directory:
Copyright Structure Directory:
Abandoned:
Communication Metadata:
/Communication/*
Foreign:

Robots.txt

Url https://www.youtube.com/static?gl=GB&template=terms

RobotsTxt | User-agent: *
Disallow:

Disable Port Scanning
Simultaneous: 0
Disable Commercial Indexing
Sitemap:
Update Terms:
Copyright Content Directory:
Copyright Structure Directory:
Abandoned:
Communication Metadata:
Foreign:
Interactive:

Here are the combined results of the robot-generated text after the career seminar class tried the questions:

Url https://www.olivet.edu/sites/default/files/pdf/AcceptableUsePolicy.pdf

RobotsTxt User-agent: *
Disallow:

Disable Port Scanning
Simultaneous: 1
Disable Commercial Indexing
Sitemap:
Update Terms:
Copyright Content Directory:
Copyright Structure Directory:
Abandoned:
Communication Metadata:
Foreign:

Url https://savethestorks.com/terms-conditions/

RobotsTxt User-agent: *
Disallow:

Disable Port Scanning
Simultaneous: 0
Disable Commercial Indexing
Sitemap:
Update Terms:
Copyright Content Directory:
Copyright Structure Directory:
Abandoned:
Communication Metadata:
Foreign:
Interactive:

Url https://www.wycliffe.org/terms-and-conditions

RobotsTxt User-agent: *
Disallow:

Disable Port Scanning
Simultaneous: 0
Disable Commercial Indexing
Sitemap:
Update Terms:
Copyright Content Directory:
Copyright Structure Directory:
Abandoned:
Communication Metadata:
/Communication/*
Foreign:

Url https://www.youtube.com/static?gl=GB&template=terms

RobotsTxt User-agent: *
Disallow:

Disable Port Scanning
Simultaneous: 0
Disable Commercial Indexing
Sitemap:
Update Terms:
Copyright Content Directory:
Copyright Structure Directory:
Abandoned:
Communication Metadata:
Foreign:
Interactive:

Before/After Comparisons for added data from class addition:

The copyright directory was turned off for the YouTube question. The abandoned directory was turned off for the Wycliffe question, as well as the foreign folder. So defaults were

changed in the last two categories. Some data was lost during the collection of this input. It was discovered that an error in concurrency was the likely cause of the loss of some simultaneous data input. Although there were data errors in testing the application with the Career Seminar class at Olivet, the experiment allowed for students to have a framework for discussing ethical questions as mentioned by the class's teacher, Dr. Larry Vail.

Management Report

The project's plans were developed beginning in the spring of 2017. The area of human-based computation was considered for a field of research. Developing a human-based computation game was considered, but other applications within the field were considered. Eventually I chose the problem of web ethics. With this problem domain, there was still the problem noted by Dr. Bareiss of where to get the crowd. Eventually, I decided on potential law students preparing for the LSAT. The plans to research were further refined throughout the fall of 2017 throughout the class CSIS 492. As the semester progressed, we shared reports of where each of our individual research projects were are. It was noted by other students that there already existed a robot protocol for internet web ethics. This observation of my classmates influenced the project. Rather than avoiding web ethics, the project involved this protocol in the project. Below is a timeline of the project.

Project Timeline

September 12, 2017. Gain Knowledge and refine topic.

September 15, 2017. Determine how the project is going to be completed.

September 29, 2017. Design the software thoroughly.

October 20, 2017. Implement the software.

October 27, 2017. Deploy the software.

November 27, 2017. Soft link distribution deadline.

December 4, 2017. Hard link distribution deadline.

December 11, 2017. Analysis and poster due.

December 5, 2017. Present at C.S. open house.

March 1, 2018. Submit report.

April 17, 2018 Senior Capstone Report Presentation

Thirteen hours in September went into the project. By the end of the month, I had chosen and set up Azure as the platform to use. To get to this point, I had begun a few branches of work for considering what tools might be used in development. To set up the database was the next big step.

Halfway through October, the basic tutorial was working with my Azure account. From this point, edits could be made to try to shape the tutorial to the intended application. A little over 9 hours were spent on the project in October.

A little over 25 hours were spent on the project in November of 2017. During this stage of development, the question page was modified. In this phase small changes or building of webpage features were added to get a question to correctly store results into the database with an appropriate entry method. Some ambitions for the project had to be scaled down. It could take a couple of hours or more to add a single feature or fix a bug. Development was slow because of the unfamiliarity of the tools and languages used. The statistics page was also developed during this timeframe.

Beginning at the end of November, the translation of the question data into a workable robot system was developed. This required a considerable dive away from developing back into the problem domain to understand how laws and robot protocol should work together. My original goal for testing this project was to use a legal case as an example for the project and see if the application would come to the same results as determined in court given similar ethical data. This would have helped to test the accuracy of the program. Throughout December the ethical language was defined and built into code which could generate robot protocol based on aggregate user responses. Twenty-four hours were spent during this phase as well as an additional hour at the Computer Science Open House to present the project. The project was run on a lab computer at the event, and a poster accompanied it which both aided in the explanations of the project.

In January 2018, a little over 18 hours were spent preparing the project for to deploy to computer scientists at Olivet Nazarene University as well as investigating the results and beginning the report.

In February 2018 class data was investigated and the report was written. The work reached over 100 hours total for the project in this month.

Lessons Learned

It was learned how to set up a system. The initial time to learn a system requires a long time at first. This was the case with using the Azure system. It took almost eleven hours to set up Azure and figure out how to get a basic tutorial to work on it. The initial time to learn a system is made even harder by the fact that you often pursue a few branches while learning. For example, I spent some of the time trying to get a separate repository system to push files to Azure. But after exploring that option without success, I eventually switched over to a different tool, Visual Studio. Having a software product created by the same company as Azure made for a working solution to publish source code.

The time to publish was reduced as time went on. At first, I would remove and add again most of the system components used in Azure. As I learned more about how to use the system, I found that it was possible to publish with minimal or no changes to the Azure components. When the database was changed during development it would still be removed and rebuilt, but the rest of the system would be left untouched during development.

Incremental development was critical to solving the many small bugs that arose throughout the software's construction. Sometimes it is faster to go back to an earlier version of a working system to fix an error. The main trick to finding a bug is isolating the code which caused it. The more sequential code edits, the harder it is to isolate and solve a bug. By developing in tiny pieces you can eliminate the amount of code logic and definitions of library methods you have to scan through to find an error.

Budgeting time was a challenge throughout this project. It seems that I have the tendency to expect more from myself than I can do. The lesson is to budget time realistically. The mistakes I would make in budgeting time were to underestimate how long a task would take, poorly defining or missing a task to be done. Even the small tasks need to be accounted for, because the details of these tasks take up time. Another big mistake I would make was to incorrectly estimate the amount of time I could spend on the project each week. Seldom does anyone have the ability to work on exactly one project at a time. Other academic responsibilities were often prioritized over the research, and I poorly followed through on weekly research time goals. In future work, it would be worth spending more time considering more effects on time estimates such as holidays, breaks, academic projects and tests. The known things which can be planned for should be planned for. Then some time for flexibility should be added in to the longer-term goals. A low baseline of time, such as two hours a week, for all but one or two scheduled but flexible break weeks would be an example of an achievable and realistic goal for a long-term research project with other activities. Then the tasks could be organized into each week using a general estimate of academic workload per week.

For future students, it should be noted that the development and maintenance should be focused on if the project is going to be developed in the future.

Future Development and Maintenance

The modeling system created in this project is a proof of concept. It serves to show what can be done. A Microsoft tutorial was used in the building of this application.[2] The tutorial was iteratively modified until it fit the functionality of the design requirements. In the original tutorial, to-do lists were created. The variables were modified to represent questions instead of to-do lists. Webpages were added to the program as well as a new model to hold the robot protocol generated for each website. As such, this application would most likely need to have its source code rewritten for future, especially non-academic, use.

Concurrency is an issue within networked applications. Solving concurrency is critical for a web application involving a voting system. The accuracy must be high, and a locking system is not practical for a crowdsourcing solution. The scalability of the system should be considered as the program is developed.

A student Azure account was used in the production of this application. This licensing allows for quick student development in a virtual environment. However, this subscription will end after the student graduates, presenting a problem to the system maintenance. Microsoft Azure is designed as a tiered Platform as a Service in which you pay for different features as you develop virtual systems. To maintain the website, the platform subscription would need to be changed from the student version to a regular account for commercial endeavors. However, this problem could be avoided by keeping the

program strictly academic. The project could be handed down to a younger student with an Azure account. This would allow for future development of the project. The concept of a system which serves to mutually benefit student test-takers and web ethicists is the principle that should be maintained from this project to its expansion through redevelopment whether for commercial or educational purposes.

Each new system administrator should see that the following objectives are met:

- Students administering or developing the project should learn new technology while working.
- Interface questions need to be continuously designed or shaped to benefit students.
- Output to a modern, robot-readable protocol should be included.
- A public interface with a project description, data report and contributor list should be maintained to explain the project's purpose and document its development.
- The development of the project should stimulate ethical language building for peacemaking.

Since the LSAT Practicum is designed to be used by students, LSAT Practicum would likely best be kept entirely within the academic community. This possibility would allow for student and professor maintenance. Such an approach may avoid complicating the development of the application. There are tradeoffs for keeping such an application purely academic. There are a number of stakeholders who might benefit from a more extensive version of the application. These include data mining organizations and test preparation companies. However, government, academia and non-profit could still benefit from an academic version of the website.

Additional features could be added to LSAT Practicum that would improve its usability. One of the most obvious features to add is crowdsourced explanations for answers chosen. That is, the crowd should not only answer questions on the website and check answers, but they should be enabled to discuss why an answer is a certain way to learn more about what makes an answer right or wrong. A descriptive forum explaining each answer ought to be linked to each question.

To increase user motivation, question passages should be trimmed. Also, putting two or three questions on one webpage with the same passage should be considered. Such a change may be better for the desktop user. The mobile user may prefer to answer one question at a time.
One advantage to providing multiple questions on the same passage is less text for the user to read overall. A number of improvements need to be made to keep the crowd interested in the program. As usage of the website increases, users are likely to want a way to keep track of how many questions they have gotten right. More question attempts means more data. Experiments could be done to see how to encourage the crowd to answer more questions at one time or for a longer period (multiple website visits).

Conclusions

The growth of technology requires a continued discussion of ethical applications. Human-based computation was used in this project to interpret the terms of use regulating web crawler ethics. The method of user input was a web application with multiple choice questions. The target audience for the crowdsourcing portion of the application was law students. However, computer scientists were involved instead for the testing of this application.

Appendices

References

[1] **http://www.robotstxt.org/robotstxt.html**

[2] https://docs.microsoft.com/en-us/azure/app-service/app-service-web-tutorial-dotnet-sqldatabase

[3] https://en.wikipedia.org/wiki/Robots_exclusion_standard#Nonstandard_extensions

Annotated Bibliography

Choosing the Research Problem

http://www.users.csbsju.edu/~lziegler/CS338/NP-Complete%20Problems.html

The above link provides a list of NP-Complete problems and gives examples of applications of these problems. Earlier in the research process, I was considering breaking such problems apart. As I continued thinking about which problem would be picked, I realized I would need to solve a problem harder than NP-complete. But time-reductions still matter in the computational process. As we develop a human-based computation system we need to recognize which parts of a problem can be quickly be solved by computers and which cannot.

https://en.wikipedia.org/wiki/Wicked_problem

Wicked problems were considered as a type of problem to investigate.

https://en.wikipedia.org/wiki/Promise_problem

Another problem class explored was the promise problem.

https://en.wikipedia.org/wiki/AI-complete

An overview of AI-complete problems can be found above. These are the types of problems, I ultimately decided to solve.

http://www.academia.edu/1419272/AI-Complete_AI-Hard_or_AI-Easy_Classification_of_Problems_in_Artificial

The above link also refers to AI-complete problems, categorizing problems as generally accepted within the field of computing. There are also some examples of and philosophical considerations for applications of solutions to AI-complete problems.

Human-Based Computation (Crowdsourcing)

https://en.wikipedia.org/wiki/The_Wisdom_of_Crowds

Here are some general ideas about crowds when compared to experts in solving problems.

https://hbr.org/2013/04/using-the-crowd-as-an-innovation-partner

This 2013 article from Harvard Business Review provides some useful insights into effective uses of crowds to accomplish work. The author categorizes crowds into different types and considers what motivates each crowd. Applying these principles is necessary to make the most of a crowdsourcing endeavor.

Michelucci, P. (2013; 2014 ;). *Handbook of human computation* (1; 2013; Ed.). New York: Springer. Doi: 10.1007/978-1-4614-8806-4

This is a great resource for exploring human-based computation and its various applications.

Selecting the Crowd

https://www.census.gov/popclock/

Here is a way to estimate how many people are in the world. This provides an upper limit for the size of the crowd chosen. The workforce potential to collectively work on the same problem is limited by the global population.

https://www.bls.gov/ooh/legal/lawyers.htm

The above link contains statistical information about the law profession from the Bureau of Labor Statistics within the United States Department of Labor. Within the United States, this document can be used to consider if the law crowd is useful for solving problem types by noting the number in the profession. Later, such statistics could be used for targeted marketing.

https://www.lsac.org/lsacresources/data/lsats-administered

These data were referenced to display the number of people who take the Law School Admission Test every year. The data show that there is a high number of annual test-takers. This information indicates that the potential law student crowd is relatively large.

https://books.google.com/books?id=0TQlDwAAQBAJ&pg=PA391&lpg=PA391&dq=student+motivations+for+using+apps&source=bl&ots=QKd43I0vnr&sig=gkbEaExPMYfEfgS03sx1ebSEIag&hl=en&sa=X&ved=0ahUKEwiRpryBxqjWAhUl5oMKHQu-BYwQ6AEITjAG#v=onepage&q=student%20motivations%20for%20using%20apps&f=false

Here is some information as to why a student would use an application.

LSAT Test Preparation

https://www.kaptest.com/lsat?&mkwid=sG2kLDPpY_dc&pcrid=90825190956&pmt=b&pkw=+lsat%20+preparation&gclid=CjwKCAjw9O3NBRB3EiwAK6wPT6IZpQ37kC32Xf-Wk6_isPqtXBzqvYosyviPu9R-K3IbyHidhd4MQxoCb40QAvD_BwE

Kaplan provides a number of test preparation tools and services. Study tools or services and their marketed prices are found at the link above. This source is used to consider what test preparation tools are already on the market, which is necessary in developing and branding new test tools. This website was also used to help develop an estimate of how much time is needed to study for the LSAT. This information can be used to design the volume and accessibility of content which should be included in the program. This information was incorporated into the feasibility study.

https://lawschooli.com/how-much-time-do-you-spend-on-each-lsat-logical-reasoning-question/

This information can be used to consider the question: "How long will a student take to answer questions on the LSAT?" This affects how a realistic preparation for the LSAT should be made as well as the volume of content necessary.

https://www.usnews.com/education/blogs/law-admissions-lowdown/2015/02/16/set-a-4-month-lsat-study-plan-for-3-types-of-test-takers

Here we can find the necessary duration of a study program for the LSAT. Note that there may exist a bias among the commercial test preparation experts toward longer study times due to marketing efforts to sell products to students.

http://www.opencolleges.edu.au/informed/features/30-tricks-for-capturing-students-attention/

Attention span decreases over time for students, and this article notes how this occurs. Considering student attention span can help in developing a product that stays interesting to test-takers. Studying attention spans helps us to understand how many questions we can expect a user to answer in a single test-preparation session.

https://play.google.com/store

See the Google Play Store for information about a number of LSAT Preparation tools. Noting the number of downloads on the LSAT specific applications confirmed that a significant crowd size could be reached for the potential law student crowd.

https://www.lsac.org/jd/lsat/prep/analytical-reasoning

https://www.lsac.org/jd/lsat/prep/logical-reasoning

The two links above mention some of the question types found on the LSAT and what to expect as well as examples for each type.

Legal Information

http://www.cfaainstitute.org/wp-content/uploads/2015/04/Qvc-Inc.-v.-Resultly-Llc26.pdf

As a document showing part of the legal action between QVC and Resulty, there is information here describing the arguments and discussion of intent as well as application of the Computer Fraud and Abuse Act. A list of relevant cases is included in this document under the heading "Table of Authorities".

https://www.digitaltrends.com/web/laws-every-internet-user-should-know/

This website article provided a context for what laws are associated with web ethics. It was heavily used in developing my proposed extensions to the Robot Exclusion Protocol.

See the following within the U.S. law code for more information: Computer Fraud and Abuse Act, Digital Millennium Copyright Act, Electronic Communications Privacy Act of 1986, Patriot Act, FISA Amendments Act, Communications Decency Act (Section 230).

Interface Design

Shneiderman, B., Plaisant, C., Cohen, M., Jacobs, S., Elmqvist, N. (2017). Designing the User Interface (6th Edition): Strategies for Effective Human-computer Interaction. Hoboken, NJ: Pearson Education, Inc.

The effectiveness of design was considered according to standards of usability defined in this book. It was determined that time to learn was the critical usability function to be included in the design for this project. The principles of this book were kept in mind as the interface was developed.

Ethics Modeling

https://en.wikipedia.org/wiki/Web_crawler#Politeness_policy

A politeness policy is the term used to describe the logic inhibiting a web crawler from acting indiscriminately toward websites and web data while browsing.

http://blog.mischel.com/2011/12/20/writing-a-web-crawler-politeness/

Here is an article which tells some general things a person implementing a politeness policy would need to know.

https://en.wikipedia.org/wiki/Robots_exclusion_standard

The above link can be used to develop an understanding of the robot exclusion protocol.

http://www.robotstxt.org/orig.html

Above lists some information about the original robots.txt standard.

http://www.robotstxt.org/robotstxt.html

Here is a more official representation of the robots.txt model.

https://en.wikipedia.org/wiki/Robots_exclusion_standard#Nonstandard_extensions

The current extensions to the robot exclusion protocol were consulted in developing a tool which could output to an accepted protocol.

https://en.wikipedia.org/wiki/Media_type#Common_examples

This explains the types of files found in the web. We need to know what formats are available to know which types of content we ought to block from web crawlers. The scope of this project only allowed for the filtering of content by directory. For example, if any content could be written on by a web user, it would be put in a certain directory and the web crawler would be told which directory followed that rule. The same logic could be used to separate media content by type.

https://www.w3schools.com/tags/tag_meta.asp

Filtering through the use of *meta* tags was considered. Not incorporating this fine-tuned filtering allowed for quicker development of the proof of concept.

https://en.wikipedia.org/wiki/Glob_%28programming%29

The term glob was explored to understand the types of users or delimiters allowed in the robot exclusion protocol.

https://stackoverflow.com/questions/23702202/what-are-the-differences-between-glob-style-pattern-and-regular-expression

Comparing the use of glob programming to regular expressions could help with understanding the exact functionality in the robot exclusion protocol.

Olivet Documentation

http://twiki.cs.olivet.edu/twiki/bin/view/DepartmentInformation/ProjectRequrements

Requirements for developing projects can be found at the above link within the Olivet twiki.

Azure Development

https://docs.microsoft.com/en-us/azure/app-service/app-service-web-tutorial-dotnet-sqldatabase

This is the main tutorial from which the website was developed. I modified this website by repeatedly adding variables and features.

https://social.msdn.microsoft.com/Forums/azure/en-US/2c45de40-91f5-4af7-a640-0971f0d25ca1/aspnet-core-web-app-the-edition-free-does-not-support-the-database-max-size-1073741824?forum=ssdsgetstarted

This link provided a solution to the initial building problem. There was a glitch in attaching a SQL database to an Azure account with student licensing. This fix allowed Visual Studio and Microsoft Azure to allow for publishing to an existing SQL database.

Development and Assorted References

http://www.c-sharpcorner.com/uploadfile/b19d5a/how-to-create-radio-button-in-Asp-Net-mvc3-razor-application/

This was used in developing radio buttons in Model View Controller. It is simple to be able to model a field for a variable without explicitly specifying the dimensions or sizing or style of every control on a web form.

https://msdn.microsoft.com/en-us/library/ms228360(v=vs.90).aspx#Compound%20data%20types

The main language I have worked with at Olivet so far is Java. This article gave insight into how Java is similar to C#, the language the tutorial I was editing used.

https://www.codeproject.com/Articles/5912/Easy-to-use-Hit-Counter

This article was used as I considered how to track a count of answers to the questions.

http://www.c-sharpcorner.com/UploadFile/cd7c2e/creating-click-counter-using-knockout-in-mvc4/

This article made me think of putting code in a different class. MVC was new to me at the time.

https://stackoverflow.com/questions/19213342/switch-statement-inside-razor-cshtml

To understand switch statements look at the above link. Knowing which functionality in Java would be transferrable over to the languages used in the project was important to development.

https://stackoverflow.com/questions/11913011/how-to-prevent-session-value-reset-of-global-asax-file-in-asp-net

http://www.csharpcorner.com/uploadfile/abhikumarvatsa/jquery%2Dajax%2Dget%2Dand%2Dpost%2Dcalls%2Dto%2Dcontrollers%2Dmethod%2Din%2Dmvc/

https://stackoverflow.com/questions/37531271/asp-net-mvc-directly-assign-value-to-model-inside-razor-view

https://stackoverflow.com/questions/16814119/how-do-i-conditionally-show-a-field-in-asp-net-mvc-razor

The above links were less helpful in investigating some of the problems at hand.

https://stackoverflow.com/questions/7142961/mvc3-dropdownlistfor-a-simple-example

https://msdn.microsoft.com/en-us/library/system.web.mvc.selectlist.selectlist(v=vs.118).aspx

The documentation at the above link gave me the information I needed to pick the single parameter for creating the select list.

https://docs.microsoft.com/en-us/aspnet/web-pages/overview/ui-layouts-and-themes/9-working-with-images

http://www.bing.com/search?q=what+to+include+in+professional+email+signature&qs=n&form=QBRE&sp=1&pq=-what+to+include+in+professional+email+signature&sc=0-47&sk=&cvid=0776480782914746A3872EAC0FF1BF42

https://www.wisestamp.com/goodies/email-signature-for/college-students/

http://www.html.am/html-codes/links/email-link.cfm

http://go.microsoft.com/fwlink/?LinkID=322105

https://docs.microsoft.com/en-us/dotnet/standard/design-guidelines/general-naming-conventions

This was used to understand how to develop consistent variable names within the .asp framework.

https://en.wikipedia.org/wiki/Hungarian_notation#Examples

This is a description of Hungarian notation, which was avoided in this project.

http://www.tutorialsteacher.com/csharp/csharp-keywords

This is a list of identifiers reserved in the language of C#.

https://www.codeproject.com/questions/250989/what-is-int-in-csharp

https://docs.microsoft.com/en-us/dotnet/csharp/programming-guide/nullable-types/index

The above two references show that "int?" refers to an integer that can be equal to null.

www.quackit.com/html/tags/html_dt_tag.cfm

This was used as a reference for reading html.

http://www.tutorialsteacher.com/mvc/htmlhelper-radiobutton-radiobuttonfor

https://stackoverflow.com/questions/27863174/how-to-explicitly-set-value-for-a-model-in-mvc-partial-view

https://stackoverflow.com/questions/7347989/implementing-a-switch-statement-in-a-cshtml-page

https://msdn.microsoft.com/en-us/library/ttw7t8t6.aspx

This was used in understanding *foreach* in C#.

https://docs.microsoft.com/en-us/dotnet/csharp/language-reference/operators/conditional-operator

In Adobe Dreamweaver, I used a template to create an about page.

https://moz.com/blog/9-simple-tips-for-making-an-about-us-page-that-works-for-your-brand

Here was a page useful in considering branding for the about page.

https://stackoverflow.com/questions/19929990/redirect-to-action-by-parameter-mvc

https://msdn.microsoft.com/en-us/library/d00bd51t(v=vs.110).aspx

https://stackoverflow.com/questions/15958225/redirecttoaction-with-error-message

https://www.jud.ct.gov/legalterms.htm#I

This was used to understand legal terms useful for marketing to the law crowd.

https://nlp.stanford.edu/IR-book/pdf/20crawl.pdf

https://en.wikipedia.org/wiki/Sitemaps

https://softwareengineering.stackexchange.com/questions/208114/mvc-architecture-how-many-controllers-do-i-need

This was used while evaluating whether I ought to use more than one controller in MVC.

https://stackoverflow.com/questions/4279353/c-sharp-casting-from-string-to-int-or-int32-possible

portal.azure.com

This was the platform used for development. Microsoft Dream Spark was used for the Azure Account, and the subscription was provided through Olivet Nazarene University.

https://www.olivet.edu/sites/default/files/pdf/AcceptableUsePolicy.pdf

Olivet has a set of website terms which system users have agreed to. It is good to test a tool on the organization with which the project is affiliated for relevance to those testing as well as to first improve or acknowledge the organization contributing to the project.

https://savethestorks.com/terms-conditions/

Save the Storks is a pro-life pregnancy resource center giving choices to women through sonogram-equipped and professionally staffed mobile "Stork buses". The story submission allowed online provide an example of how robot-generated text could present an ethical dilemma.

mystory@savethestorks.com

This is where to email to get more information about the terms for Save the Storks story submissions.

https://www.wycliffe.org/terms-and-conditions

Wycliffe is a Bible translation organization. It represents a non-profit sector website which may be data mined.

https://www.youtube.com/static?gl=GB&template=terms

YouTube terms are used to represent the website terms a business would have.

http://csharppad.com/

http://zetcode.com/lang/csharp/io/

https://stackoverflow.com/questions/3853700/c-sharp-switch-case-string-starting-with

https://www.google.com/webmasters/verification/home?hl=en&theme=wmt&continue=https://www.google.com/webmasters/tools/dashboard?hl%3Den%26sig%3DALjLGbObxffroJU1pjkUmSgElrJJxabXYg&pli=1

Google has made a robots.txt checker. It is at the above link.

http://www.javawithus.com/tutorial/using-ellipsis-to-accept-variable-number-of-arguments

https://stackoverflow.com/questions/9528276/does-c-sharp-support-a-variable-number-of-arguments-and-how

Java Design Pattern Essentials, Tony Bevis, Ability FIRST, 2nd edition, 2012.

This book was consulted in understanding and explaining MVC. The flyweight pattern described in the book was considered but not used in the project.

Game Studio was used to begin a proof of concept.

http://www.qr-code-generator.com/a1/?PID=1704&msclkid=85e3b7d997e410822959ed8a7f516aad&utm_source=bing&utm_medium=cpc&utm_campaign=USA%20%7C%20TOP%20KW&utm_term=qr%20code&utm_content=QR%20Code%20%7C%20Exact

https://docs.microsoft.com/en-us/azure/mysql/connect-java

https://docs.microsoft.com/en-us/java/azure/java-quickstart-maven-webapps

http://microsoftazurewebsitescheatsheet.info/

http://markheath.net/post/azure-website-deploy-onedrive

https://help.github.com/articles/adding-an-existing-project-to-github-using-the-command-line/

https://docs.microsoft.com/en-us/azure/app-service/app-service-deploy-local-git

http://www.howtosolutions.net/2013/05/explore-and-understand-visual-studio-ide-areas-for-beginners/

https://en.wikipedia.org/wiki/.NET_Framework

https://www.myenglishteacher.eu/blog/legal-terms/

https://www.lsac.org/

https://www.merriam-webster.com/dictionary/adjourn

https://stackoverflow.com/questions/19929990/redirect-to-action-by-parameter-mvc

https://stackoverflow.com/questions/271743/whats-the-difference-between-b-and-strong-i-and-em

https://docs.microsoft.com/en-us/aspnet/web-pages/overview/data/7-displaying-data-in-a-chart

stephenwalther.com/archive/2009/03/03/chapter-6-understanding-html-helpers

https://stackoverflow.com/questions/2030539/how-does-asp-net-mvc-link-views-and-controllers

Future Work

https://infoscience.epfl.ch/record/204725/files/crowdexpert.pdf

This article presents a way to confirm the answers of the crowds by weighting the answers of experts into the crowdsourced answer. This process could be used in the context of the LSAT Practicum by having the input of law professionals, government officials or experts in the field of data analytics be verified as experts and by weighting the answers of these experts. New motivations would have to be explored for encouraging the experts to contribute to the solution.

https://docs.microsoft.com/en-us/aspnet/core/tutorials/first-mvc-app/start-mvc?tabs=aspnetcore2x

This link would be helpful in making an application from scratch.

https://stackoverflow.com/questions/19707885/c-sharp-copy-to-clipboard

A click-to-copy feature could be added as in the above example for the robot output page.

https://www.phpbb.com/downloads/

https://www.phpbb.com/community/docs/INSTALL.html#quickinstall

The above two links were investigated as a possibility for the future feature of adding a forum to allow a discussion of the answers.

As I was developing, Visual Studio recommended using Bootstrap Snippet Pack and Glyphfriend 2017.

https://stackoverflow.com/questions/1557507/what-is-a-good-statistical-math-package-for-net

I thought about using a statistical package in the project. Some such packages are found in the above link. This could be done in future work and may help with data analysis as the project expands.