Looking at User Trends Through Their Web Browsing History

Bandr AlSwyan; Ravi Sajjan

Professor Mohammad Owrang Ojaboni
Professor Ericka Menchen-Trevino

American University
4400 Massachusetts Ave NW, Washington, DC 20016

CSC-493-001

April 29, 2019

Table of Contents

Abstract

The purpose of this capstone project is to give academic researchers and marketing firms the information and tools to perform analysis on user web browsing trends with possibility of drawing conclusion about political polarization. The need for this application is greater than ever. As people in the modern era produce more and more data, the demand for a tool to help us understand this data and it affects our lives is needed. Our tool will help lower the barrier to entry of understanding of the data provided. It does this by giving the user easy to understand visualizations. The data that will be integrated into the application will be provided by Professor Ericka Menchen-Trevino and was collected from people with their consent and approval through (Web Historian). The application front end will consist of an HTML, CSS, and JavaScript website, and the data will be represented through R as a backend, a data processing software tool. The purpose of this report is to show the breakdown of what was required to complete the project. There are a number of costs that comes with the features of the project. Some of these costs are due to the large amount of data, it is required to have ample data storage and processing power so that the handling can be smooth. This application is developed with mindset that the user will use it to draw conclusions on things such as human interaction and trends, and explore more information about their dataset.

1. Introduction

The purpose of this capstone project is to give academic researchers and marketing firms the information and tools to perform analysis on user web browsing trends with possibility of drawing conclusion about political polarization. The need for this application is greater than ever. As people in the modern era produce more and more data, the demand for a tool to help us understand this data and it affects our lives is needed. We will apply and support our data with statistical analyses that will reflect better understanding of the data collected.

The data that will be integrated into the application was provided by Professor Ericka Menchen-Trevino and was collected from people with their consent and approval through (Web Historian) a browser extension for Google Chrome that helps visualize the web browsing history. According to the consent that the participant signed and agreed on web historian “By clicking in the "participate" button I certify that I am at least 18 years of age. I have read this consent form and I understand what is being requested of me as a participant in this study. I freely consent to participate in this additional data collection.” we can use the data collected to reach the purpose that was seeked for, and therefore meeting the demand of our project, giving us the permission to use the data.

The application front end will consist of an HTML, CSS, and JavaScript website, and the data will be represented through R as a backend, a data processing software tool. The data was converted into JSON and CSV formats to be used in R for the purpose of our visualizations. The website displays visualizations that include most visited websites by categories and date, Browsing timeline, compare and contrast search and mostly chains of events. When the user clicks on visuals, information regarding metadata will popup. In this popup bubble, the user will be able to see information regarding category/website and visit count, filters will be provided so the user can filter the information displayed as they please. Among other options, there will be a news rating system for the news category that will showcase the polarization effect between two domains.

The objective of this project is to provide academic researchers and marketing firms with a better understanding of the data in an simple and interactive manner. The response to the needs mentioned above is to design a web tool that can be easily accessed and used through an interface that the user can draw conclusions from.

In this report, the following sections will be discussed. First, review of previous projects that are already out there and why this project is different. Then, the design requirements and project details outlined, as well as the costs that follows it. Next, the feasibility of the project which talks about the literatures and justification for the design of the project. After the feasibility discussion, a use case diagram (UML) and Data-flow diagram (DFD) will be shown that outlines the expected user experience of the website. Lastly, the results of the project will be shown with images of the completed application.

2. Review

The Web Historian website provides a similar implementation of one of the features that this project intends to do. But it was soon realized that Web Historian lacks the information needed for users to draw conclusions in an easy manner or supported environments for different browsers since it only runs on Chrome as main base at the moment. If the user wants information about groups of categories or compare they would have to part the data and re submit them in order to distinguish which is which. In our project's implementation, all of this information will be displayed in one place. In our project, users are able to select from a variety of options that will help them view and understand the data in different ways.

3. Design Requirements and Project Details

The design of our project was built upon the need for an interactive tool as mentioned above. Our main objective is to meet the possible support to power any user and benefit from the visit of our application. Moreover, in the terms of the business rules, the academic researchers and marketing firms are considered to be the user and they are represented as our customers. The end user will have the ability to elaborate on the data that comes from people browsing history.

The data was outputted as JSON and CSV formats. Once all of the data was cleaned, there was a total of 2.5M URLs. In order to make parsing and data manipulation easier, all of these files had to be combined into one set. This was done using R using a left join command on the "domain" field. The following command was used:

```
data <- left_join(browsing.data, domain.cat)
```

Once the CSV files were merged, it was time to explore the data to see what was being dealt with. It was discovered that the data had issue. The issues encountered were the following

- Domain name consistency
- Timestamp format
- Categorization
- Personal data

- Paywalls
- Unicode/ASCII conversion issues
- HTML consistency

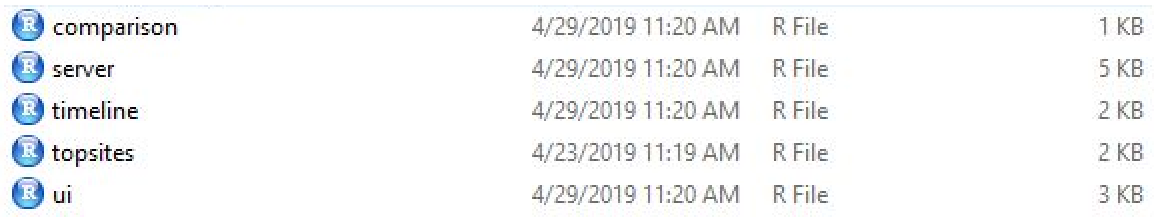
These issues were discovered during the parsing of websites provided in the dataset and when attempting to visualize the data. A number of Python scripts were used to fix each of the above obstacles. The script that reduced the dataset by 20% ran into these same issues. It was also discovered that some websites did not have any categorization. Because of this, a new scheme had to be made to be able to identify the type of these URLs. Also contained within the data were a number of unnecessary columns that only made the data more difficult to work with due to the size increase.

For the statistics portion of the project, various statistics were discussed to meet the desired information that is going to be shown to the end user.

The statistics determined are below:

- Top N Most Visited Domains
- Domains Sorted by Category
- Browsing Timeline
- Site Comparison
- Political Polarization

For the website, number of files were designed. As shown below:



comparison	4/29/2019 11:20 AM	R File	1 KB
server	4/29/2019 11:20 AM	R File	5 KB
timeline	4/29/2019 11:20 AM	R File	2 KB
topsites	4/23/2019 11:19 AM	R File	2 KB
ui	4/29/2019 11:20 AM	R File	3 KB

In order to show R in the main webpage, the workbooks had to be published to the Shiny server, which then allows the website to display the work that has been done in the R Studio desktop application in the web environment.

3.1 Costs

There are a number of costs that comes with the requirements of the project. One of the main costs is the performance of the application. Due to the size of some of the data, it is recommended to the user to have strong hardware, specifically RAM that can handle the application. Another cost of the application is privacy, upon a worry that if the data was not clean properly, there may be a chance of personal information being leaked. This application is not intended to help or assist any inhumane act or whatever kind if it reflect harm. It should be noted that the data is static not dynamic, meaning that it requires manual updating.

After doing the project in R, we came to realize a better option would have been Python. While R is a great language, we overestimated our abilities with R and that caused us to not be able to implement certain features that we would have liked to. We felt that our Python skills would have been more than enough to process the data in the fashion we wished.

3.2 Data

Large amount of data can limit the performance of processing which in addition can have a side effect include but no limited to: Longer response time, crashing of browser(use Firefox, duh!), and most notably loss of track of info from the current session.

3.3 Security

One of the possible limitation of our application is security. The web socket can be sniffed and tracked from the traffic that might be sent between the user data and the network request. While the website is useful for letting people view the data publically, it is recommended to run our application locally so that no data can be sniffed.

3.4 Privacy

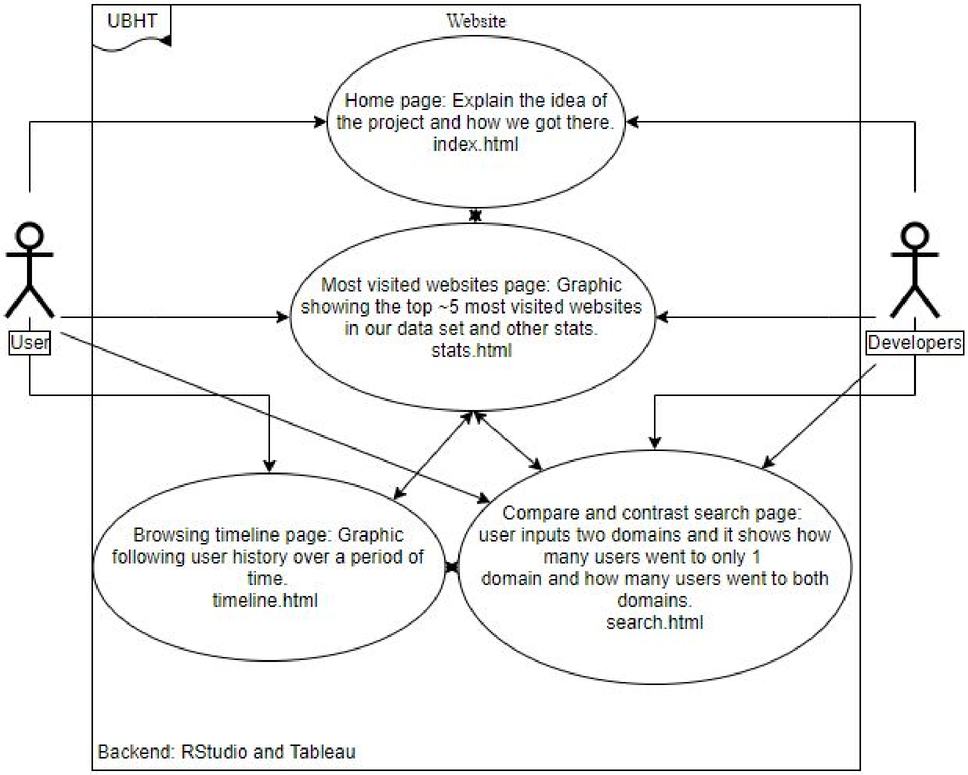
The concern for privacy is raised due to the fact that the data and information being retrieved comes from real humans. If the data were to be leaked or not properly anonymized it would be possible to identify users and see some of their personal information. We strongly advise to review the data manually and personally before uploading any data and to make sure that it serve the purpose it was created with minimum reference to any personal info.

3.5 Feasibility Discussion

At the start of this project, it was very difficult for the team to identify an idea regarding big data visualization. We were informed of a research being conducted by Professor Ericka Menchen-Trevino of the Communications department and we quickly became involved with the project. The challenges were mostly based on the idea that there was no unique topic to cover, since most of the big data that is on the internet had already been visualized by other people. Various brainstorming sessions have been done to decide how the line of our capstone would be a good fit for Professor Ericka Menchen-Trevino project. It was ultimately decided that pursuing the web history trends was the way to move forward on our Capstone. The Web Historian website was encountered when searching the topic. When the website was viewed, it was clear that Web Historian shares some similarities to our project, but its purpose was more geared towards analyzing a single user’s history rather than a collection of users. In addition, it was observed that there was a possibility of generating more information from the data offered which had clear potential. Different resources were checked to evaluate the performance of our model example “("CORDIS | European Commission", 2019)”. As it was mentioned above, some of the ethical questions that were encountered was the use of data; will people take advantage of it? Will ad firms leverage this to their favour into dark path? The answer is not clear yet as more testing need to be done. The data that was created or collected can be used by third parties (outsiders), and therefore granting us the permission of use the data.

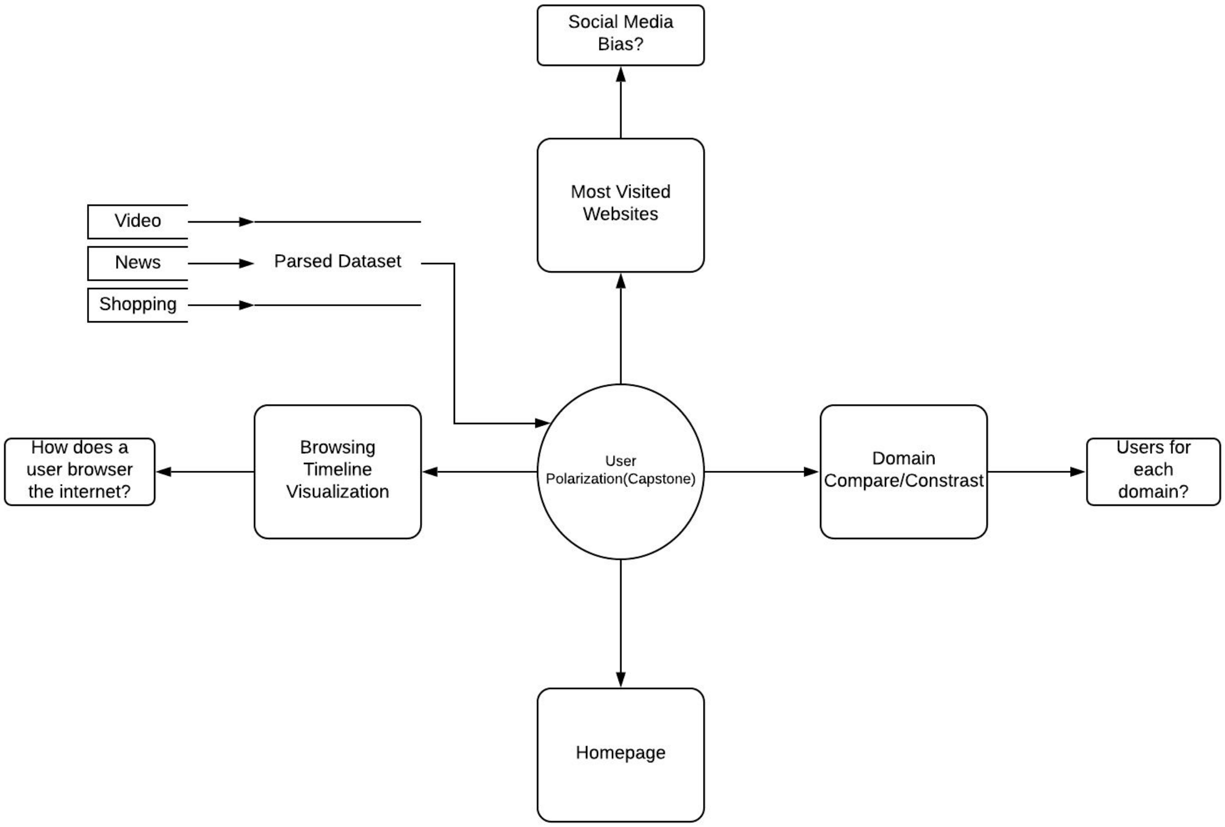
4. Final Implementation

In addition to what was mentioned above, the expected Use case diagram (UML) for the end user using the website is shown below.


UBHT
Website
Home page: Explain the idea of the project and how we got there. index.html
Most visited websites page: Graphic showing the top ~5 most visited websites in our data set and other stats. stats.html
Browsing timeline page: Graphic following user history over a period of time. timeline.html
Compare and contrast search page: user inputs two domains and it shows how many users went to only 1 domain and how many users went to both domains. search.html
User
Developers
Backend: RStudio and Tableau

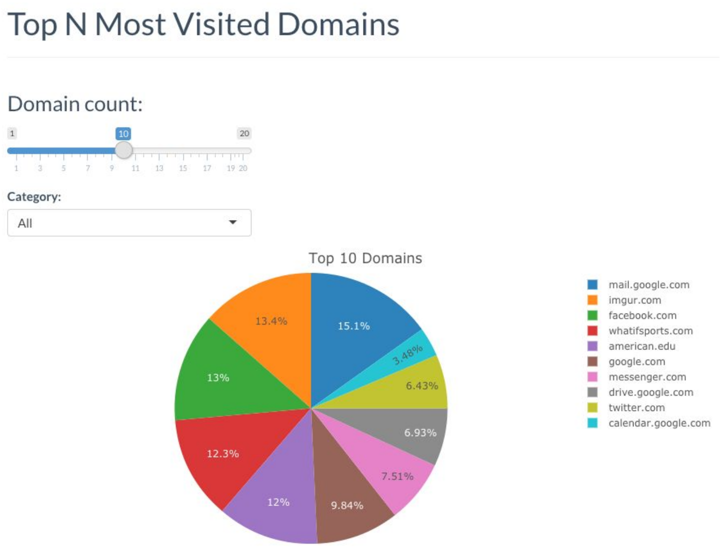
The user is expected to be able to browse all the given four pages seen in the diagram. They are expected to interact with the index page which includes the dashboard stats regarding the data. They can also explore the other features. In addition, the developers of the project have read and write access to these pages and it content.

Also here is the Data-flow diagram (DFD):



5. Results (Images of final work)

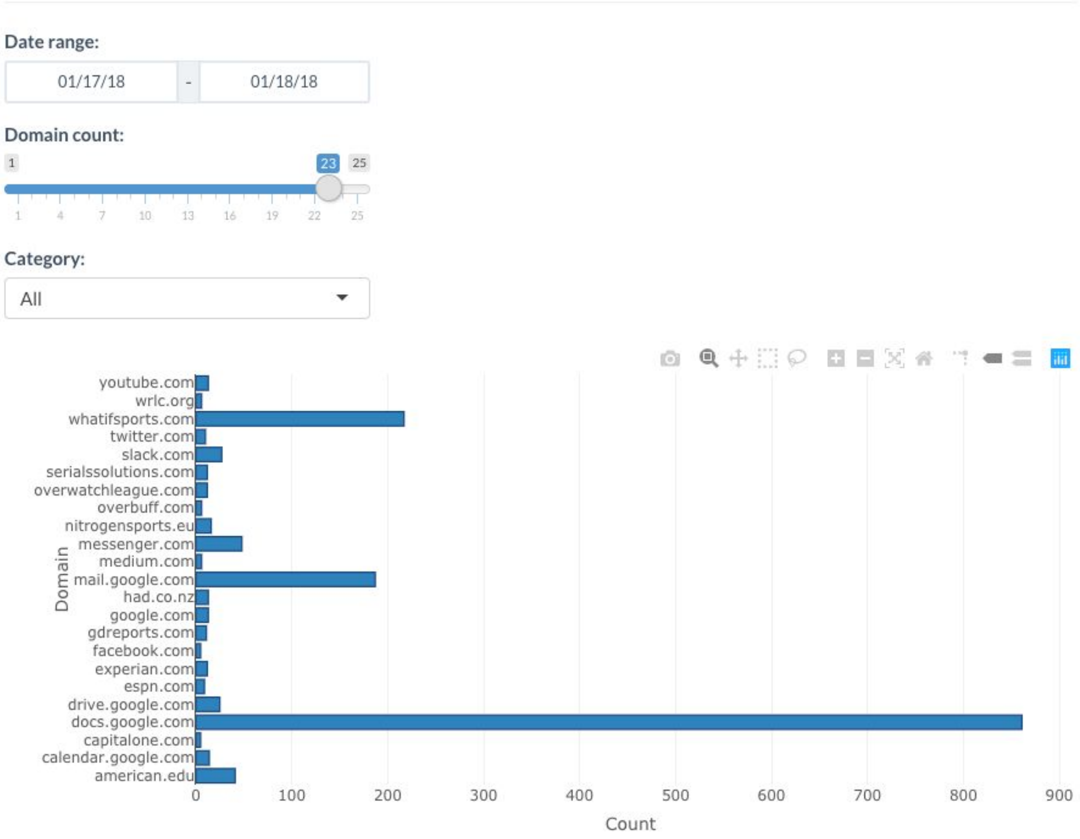
The screenshots in this section are to show the outputs of the project. The same screenshots are provided in the “OUTPUTS” folder in the zip file.



Top N Most Visited Domains

Domain count: 10

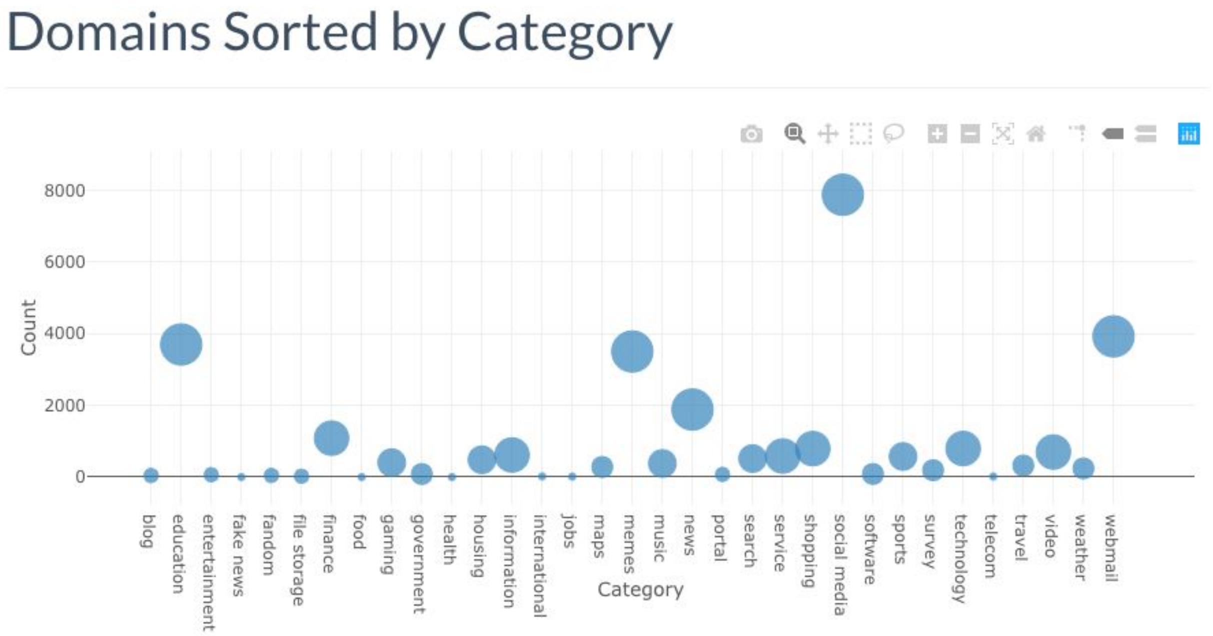
Category: All

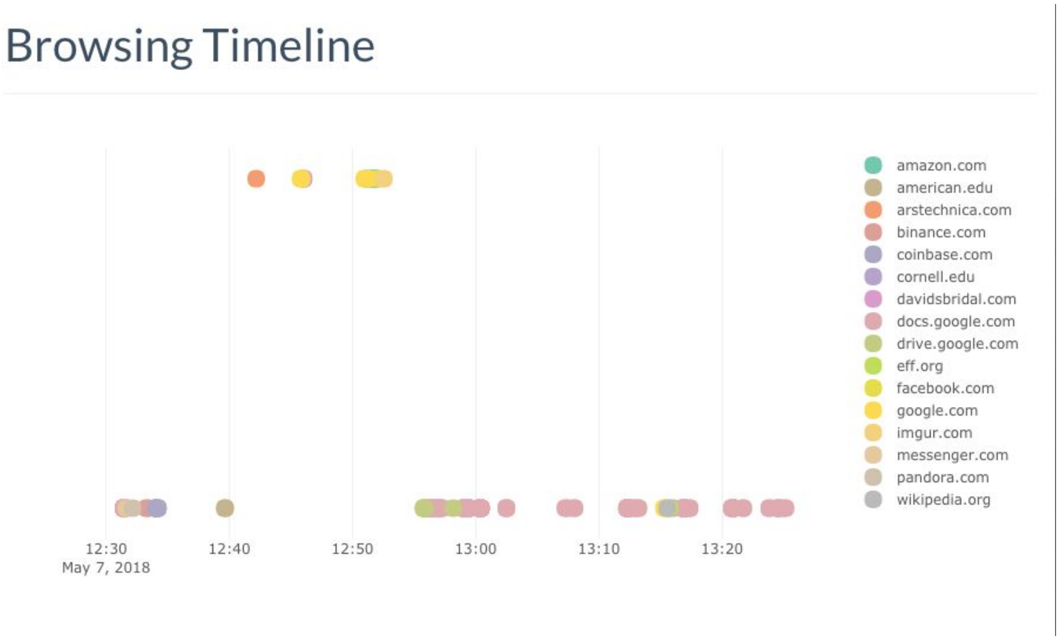


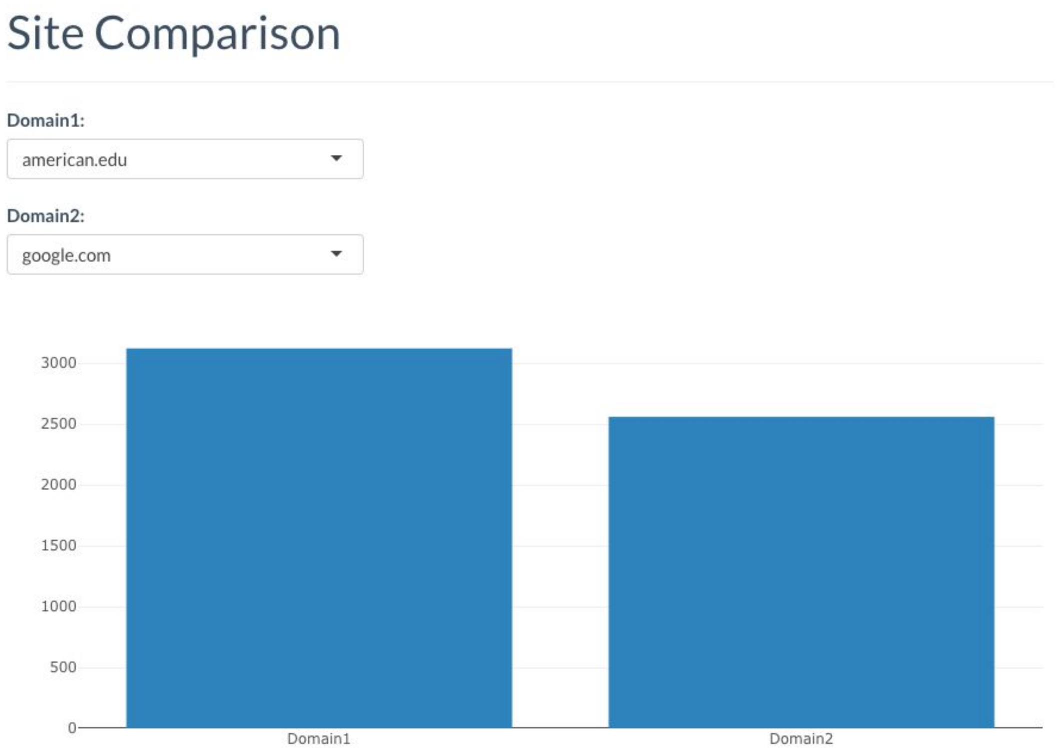
Date range: 01/17/18 - 01/18/18

Domain count: 23

Category: All


Domains Sorted by Category
Count
8000
6000
4000
2000
0
blog
education
entertainment
fake news
fandom
file storage
finance
food
gaming
government
health
housing
information
international
jobs
maps
memes
music
news
portal
search
service
shopping
social media
software
sports
survey
technology
telecom
travel
video
weather
webmail
Category

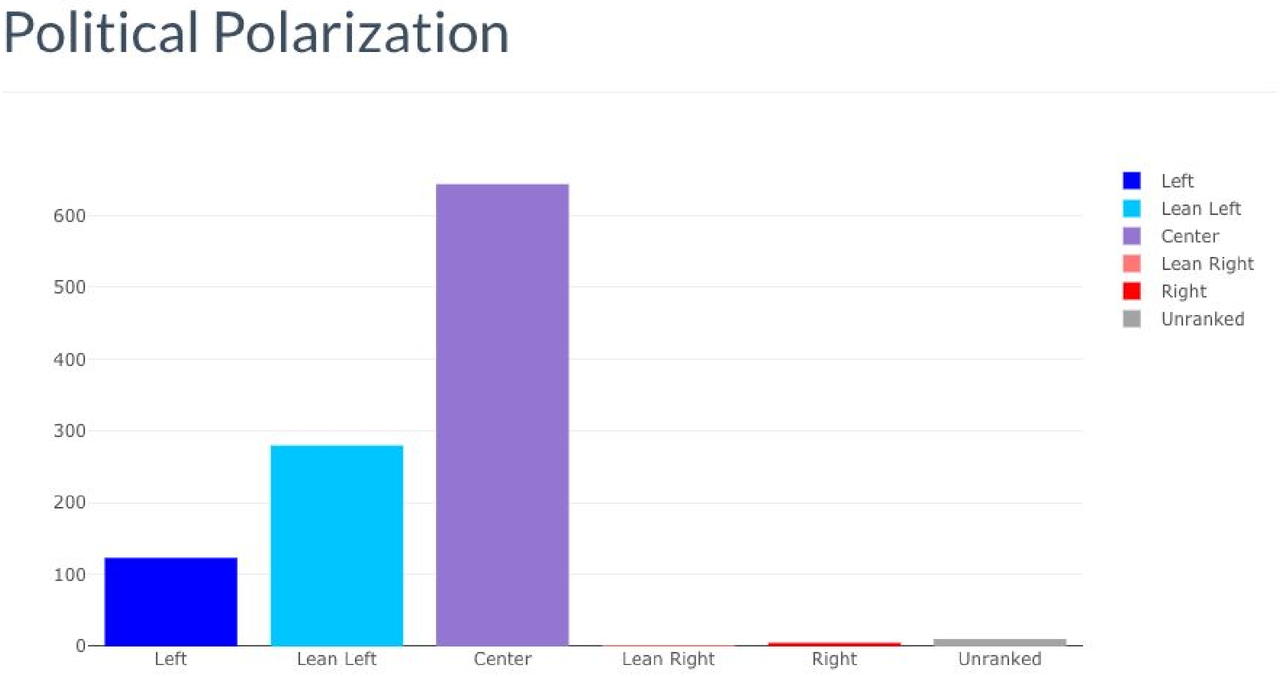

Browsing Timeline
amazon.com
american.edu
arstechnica.com
binance.com
coinbase.com
cornell.edu
davidsbridal.com
docs.google.com
drive.google.com
eff.org
facebook.com
google.com
imgur.com
messenger.com
pandora.com
wikipedia.org
12:30
May 7, 2018
12:40
12:50
13:00
13:10
13:20



Site Comparison

Domain1: american.edu

Domain2: google.com



Political Polarization

6. Conclusions

Throughout the project our main goal was to solve the problem which was the lack of coherent understanding that the Web Historian system had, and provide more information (understanding) about the data that surround the browsing history and web trends . As a team we saw potential and possibility to do more; the data was enhanced, statistic had been done and more over a visualization that carry suitable information about each record with more guided direction to explore. The project application(not the data), is available as open source software under the GPL license. Its functional and usable and live to public ("Capstone", 2019).

Through the course of this project, a number of lessons have been obtained that will be considered for future projects. An important lesson learned was, “premature optimization is the root of all evil” meaning, do not try to write perfect code the first time, just get the project working first and then worry about optimizing of the code later. Putting this project together was a fun experience that the whole team enjoyed.

7. References

Capstone. (2019). Retrieved from https://rsajjan.shinyapps.io/capstone/

CORDIS | European Commission. (2019). Retrieved from https://cordis.europa.eu/project/rcn/214597/factsheet/en

Ericka Menchen-Trevino – Research methods, digital media, political communication. (2019). Retrieved from http://www.ericka.cc/

Web Historian. (2019). Retrieved from http://www.webhistorian.org/