Looking at User Trends Through Their Web Browsing History

Bandr AlSwyan; Ravi Sajjan

Professor Ericka Menchen-Trevino

American University

4400 Massachusetts Ave NW, Washington, DC 20016

CSC-493-001

January 31, 2019

Setting/Context

Our project will run on a browser based application that the user will be able to interact, search with and reflect information they are looking for.

Capstone Problem and Rationale

The need for this kind of technology is greater than ever. Looking at the United States, we can see that the political battle between the two major parties is fierce. Each side in the political war has become more and more polarized causing turmoil across the world. The United States is not alone in this political battle, countries around the globe are experiencing similar issues. With our project, we will apply and support our data with statistical analysis that will help reflect our findings to the end user. Our user will consist of academic researchers and those who are looking to understand the general trends of people, and how data is spread amongst the user(s) which will help enhance tons of aspect of and not limited to marketing teams, political teams , etc.

The sample data that we will be using for our project has been collected by Professor Menchen-Trevino and contains the web browsing history of a few hundred users. The full dataset, collected across globe by other members of the research group, contains ~9,000 unique users spanned across 3 different countries(USA, Netherlands, Poland) and each user has ~90 days' worth of web history. To maintain user privacy we will only be able to show general conclusions, nothing specific to a user. Some attribute of the data are: URL, Domain, Timestamp, Title, etc.

The application will consist of an HTML, CSS, and JavaScript website and the data will be parsed and handled by Python, C#, R. and Tableau to help ensure fast and reliable data storage,

we will be relying on a relational database using SQL. The website server will run on Digital Ocean, a platform that is designed for website hosting. The website will provide the end-user with statistics based on the dataset. For example, the user will be able to sort (filter) the data to see which domains are most frequently visited amongst the users, another feature will be that the user will enter in two domains and they will be able to see how many users visited a single domain or both domains, this will allow for them to compare and contrast user trends. For example, they could put in “CNN” and “FOX” to see if users are going to both websites, or just one to get their political views.

Objectives

The purpose of this project is to provide the end-user with a better understanding of web trends as mentioned above in the rationale.

Deliverables

- Website
    - Home page
        - Explaining the idea of the project and how we got there
    - Most visited websites
        - Graphic showing the top ~5 most visited websites in our dataset
    - Browsing timeline
        - Graphic following user history over a period of time
            - EX: On Facebook -> Bloomberg -> Facebook -> Twitter -> CNN
    - Compare and contrast search

- User inputs two domains and it shows how many users went to only 1 domain and how many users went to both domains

Timeline:

1. Finding and getting to know the data

    a. Exploring the data to know what we will work with and to check if it's legal/free to use

    b. Figuring out the format of the data and how it will be collected

2. Collecting data

    a. Figuring out how to collect and input the data into our application

    b. Group the data into workable format

3. Processing data

    a. Converting and organizing data so that it meets the requirement of the tools that we will work with

    b. Outputting new information driven from the collected data through statistical transformation

4. Representing data

    a. Create an interactive template of the website that will show the data

    b. Display the outputted statistical data

5. Enhancement and optimization of the application

    a. Making the website functional, ready for user interaction and creating the necessary direction for the user

    b. Debugging

6. Presentation

Resources:

Hardware	Computers, Digital Ocean Server
Software	Browsers, RStudio, Tableau, HTML/CSS, Javascript, Python, C#, MySQL, Unix server
Data formats	Json, CSV, SQL, R