Carter Nelson

CIS490 Fall 2018

Dr. Hsu

Subreddit Relationships Project Plan

**Problem Statement:**

On the social media platform Reddit there are many separate sub-communities (subreddits) which draw their own set of individual users. The individual users however can interact with whichever subreddits interest them. Many of subreddits often have political or social stances tied with them due to the like-minded users. For users new to the site who have an interest in a subreddit, they may be off put by those underlying stances that have nothing to do with the subreddit’s original intention. Without having administrative information there is often a difficulty in identifying subreddits with similarities. To do this, one must analyze a dataset of publicly available comments to distinguish which subreddits are closely related.

**Objectives:**

The objective is to create a visual representation for the relationships between subreddits. This will be done by collecting a large set of active user comments. Then convert the comment data into usable data for the D3 visualization. Once converted, use the data and D3 to make the visualization.

**Scope:**

The scope of the visualization will be limited to the restrictions the Reddit admins’ place on the allowed information, and the number of comments to analyze.

**Background:**

Reddit has over 1 million subreddits, each catering to their specific audience. Over time as a subreddit grows, often, the groupthink generated by drawing users of similar mindsets causes for either a subtle or not so subtle underlying political or social stance. These new stances lead to the subreddit losing old subscribers and potential new subscribers, which furthers the groupthink. In the end, a subreddit once created for a specific topic turns into a watered-down version of a similar, yet different topic.

**Methodology:**

Firstly, I will download the Reddit comment data and study the JSON output of the file. After reviewing the layout of the data, I will use PySpark MapReduce to parse the JSON file and insert the needed fields into a Mongodb collection. I will then go through and analyze the collection and generate a new JSON with the values needed for the visualization. After gathering the values, I will use the Reddit API (PRAW) to crawl for the subreddit icons of the collection of subreddits I have. If a subreddit does not have an icon, I will generate one for it using Pillow and lay the subreddit’s name over the created image.

After the data processing is finished, I will start working on the website to display the generated data. First, I will start off by creating a Flask backend to host the website with endpoints to send the generated data to Jinja and then to JavaScript on the frontend. Finally, I will use JavaScript, jQuery, and D3.js to generate nodes for each subreddit and align them in a circle connected to other subreddits they’re related to by lines.

**Data Resources:**

The subreddit comments will be gathered from an opensource repository published by the Reddit user, Stuck\_In\_the\_Matrix.

**List of Algorithms to be Adapted:**

* Doubly linked list
  + Will be used to connect the subreddit nodes on the front end for easier visual manipulation.

**Evaluation Methods:**

Subreddits should be correctly connected to their top related subreddits in an easily viewable format.

**Milestones:**

* Downloading and transferring the data to my remote server.
* Significantly optimizing my PySpark RDDs’ parallelization for the data processing.
  + About 40% of the speedup was due to the parallelization changes and 60% due to code changes.
* Reformatting the Mongodb collection to hold the data being offloaded by PySpark.
* Generating images for subreddits that didn’t have an icon.
* Learning how to use D3.js and SVGs.
* Placing the subreddit nodes in a circle.
* Overlaying the subreddit icon over each of those nodes.
* Bringing SVG elements to the front or back of the document.
* Figuring out how to do all the dynamic events that occur when the mouse is over the node.
* Drawing all the paths between the nodes.

**References:**

u/Stuck\_In\_the\_Matrix. (2015, July 02). I have every publicly available Reddit comment for research. ~ 1.7 billion comments @ 250 GB compressed. Any interest in this? [Blog post]. Retrieved from https://www.reddit.com/r/datasets/comments/ 3bxlg7/i\_have\_every\_publicly\_available\_reddit\_commet/