William Tyler Nebel - Alex Meyers - Mateo Acebedo
Cornell University - INFO 3300 : Data Driven Web Design
March 3, 2014
Project 1 : Reddit Live Unigram Trends

Data

The data used for our visualization is sourced live from reddit. Behind the scenes, the website is executing API calls to reddit.com/new.json using a timed AJAX call every 2 seconds, as long as reddit hasn't throttled our requests(as it occasionally does) this will retrieve 250 of the most recent posts to reddit.. For each post, we extract the titles and determine unigram frequency for all post titles. Once the total distribution is known, posts are separated into time windows. The first time window is plotted relative to the total word frequencies from all times collected, while each subsequent time window is compared against the time window that came before. By comparing words in time slices, this allows one to see how the word frequency trends from time window to time window. In the scope of this project, this tells us current most popular words on reddit, and how they vary. We also have a larger dataset that we used for testing purposes; it contains 1000 posts, generally viewed this data with 5 minute time windows. Clicking the buttons above the graph allows one to switch between viewing the stored data, and new live data.

Mapping

The visualization maps title unigrams to a time, relative frequency, and absolute frequency. The x-axis corresponds to time window, showing all unigrams at each time, and how their frequency changed. The y-axis corresponds to unigram frequency relative to time; so the unigram frequency plotted at any given x, is the unigram frequency at that x is actually how the unigram frequency changed from the previous x. The "points" plotted are the unigrams themselves, their color is related to being above/below the x-axis, and their transparency is varied by how far above/below the x-axis they are. The font size for each "point" is relative to that unigram's absolute frequency.

Visualization

The visualization helps to paint a picture of what is currently trending on Reddit surrounding the time of the visualization load. Reddit is a unique environment because it is a hub for all types of internet content. What is also interesting about Reddit, is one can find trends in post subject depending on when one views the site. For example, surrounding the USA vs. Canada hockey game in the Olympics, one could find many posts discussing the event. Not all Reddit trends revolve around current events. Sometimes visitors to the site find the most random topics trending, such as posts about bananas. With all of this in mind, it occurred to our group that there is not a great visualization tool for depicting the trending words [topics] on Reddit, so we set out to use what we have learned about JavaScript and D3.js to collect the most recent Reddit posts, and plot the most popular words in given time segments. This strategy is deeper than simply plotting the number of times a word appeared in the most recent Reddit posts because it allows for the viewer to see how words are trending; whether they are increasing in popularity over the collected time period, or decreasing. The x-axis displays a specific time

segment being accounted for, the y-axis tells the rate of change of the word, and the word size shows the relative number of times the word occurred relative to the other words trending on Reddit in a given collection period. All of this information gives an auto-updating (on page refresh) representation of what is occurring on Reddit at the present moment, which is fascinating for any Reddit connoisseur. What's surprising about the live feed is what it tells us about is being posted to reddit minute to minute; it's generally easy to see how reddit fads move over the course of a day as one checks it, but minute by minute is more difficult. While our visualization gives insight into reddit over a short period of time, there's definitely potential in a more robust database that could store and keep track of reddit posts over a longer period of time, and provide the information for a visualization like this.