CS171 Final Project Process Book:

Milestone 1

Kim Soffen
Tianxing (Lisa) Ma
Hsiu-Chi (Jimmy) Lin

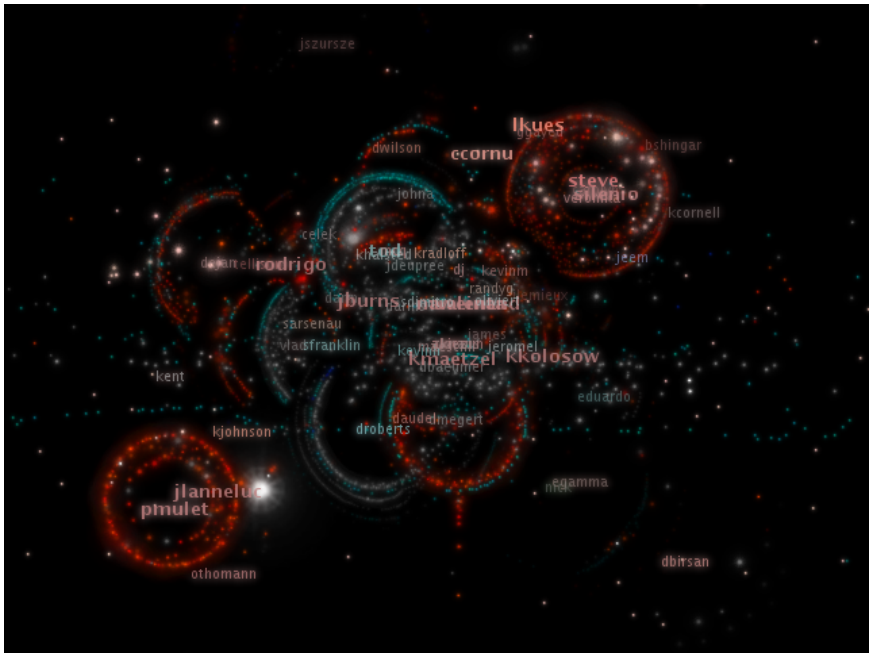
Overview and Motivation

Harvard University stands at an unique position as the oldest university in the United States and as one of the most prominent and influential academic institutions in the world. Likewise, its student publication, *The Crimson*, has special distinction in its age and vast archives of past articles, enough to construct an image of the issues important to the Harvard community and the world at the time of publication. Our team, two of whom are involved with *The Crimson*, were interested in analyzing broad trends in word frequency published by the student newspaper and from there we committed ourselves to creating a visualization that would serve as a tool for that purpose. We hoped to see the rises and falls of words over time as well as to identify words that were most popular over the entirety of Harvard's history.

Related Work

We found several helpful examples on bl.ocks.org that led us to create different features. Some include:
http://bl.ocks.org/mbostock/1667367
http://bl.ocks.org/WillTurman/4631136

In addition, the interactive software visualization in Design Critique 4 somewhat inspired our visualization of the percent changes in frequency of word appearances.



Questions

1. Are there issues that are consistently most important over the entire course of Harvard's history?
2. How are current events reflected in *The Crimson*; can we identify the rise and fall of politicians or pinpoint wars and social movements through word frequencies?
3. Are there words that are unexpectedly associated with each other?

Data

We received a SQL database dump from *The Crimson*; it was a series of 5 tables containing information (content, author, date, headline, URL) on over 500,000 articles that have been published in the paper's history. The database totaled 1.2GB, which was a serious obstacle to putting it in a workable format. The following details the different methods we followed trying to get the data down to a reasonable size.

1. We first converted the SQL dump to a JSON file, and found, not surprisingly, that the JSON was far too large to upload directly in the d3 module; it would crash the browser.
2. We then uploaded the SQL dump file into our own SQL database. Note that this SQL database is organized such that an article's entire text is in one column, so in order to find if a

given word is in that article, the SQL 'IN' operator is used. With this, we attempted to implement a real-time query system, such that a user could 'search' a word, that word would be queried in the SQL database, and then added to the visualization. Technically, this was done by making an AJAX function within the Javascript file that called a PHP script that queried the SQL table, returning the results as a JSON. Unfortunately, given the size of the database, a single query took roughly 30 seconds, which is too long to ask the user to wait.

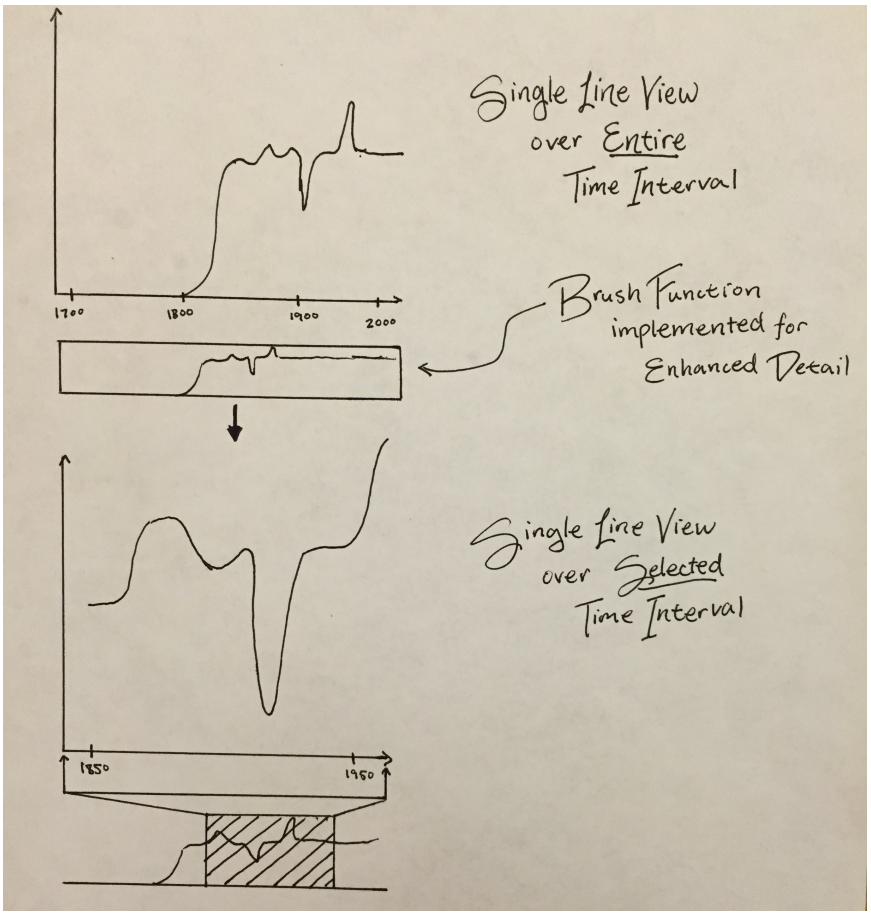
3. Given that the previous queries, which relied on "IN", were performing too slowly, we thought that if we indexed the SQL table by word (essentially creating a new table where the word is a non-unique key and the article's unique identifier is the value), it would be more efficient. This way, when the user live-queried a word, the SQL query (still using the AJAX to PHP framework from attempt #2) would be looking for key equality, rather than "IN", which would certainly be faster. However, we were unable to ever create this indexed table. Using the most efficient algorithm we could create, we estimated it would still take roughly 5 days (assuming the algorithm runs 24 hours per day) for the indexing algorithm to run, which would not give us enough time to actually build the visualization before the first milestone. (Note we chose not to parallelize the algorithm, as it has almost no computation outside of SQL insertions, so the locking/unlocking of the SQL table, necessary to prevent errors with parallelized programs, would have spent more time than parallelizing saved). Therefore, we had to abandon this method.
4. Finally, we decided we had to limit the scope of the problem, and only examine the most commonly used "interesting" (ie: manually excluded words like "the") words. To do this, we wrote a script that found the most common 500 words and created a JSON in the format of [{"word": "yourwordhere", "count":[{"web-publish-date":"date1here"},{"web-publish-date":"date2here"},...]},...], which is uploaded into d3 normally. This is the final format of our data, which does not involve any live queries to our SQL database.

Exploratory Data Analysis

At this current point in time we are only working with a partial subset of our dataset, so we have not yet drawn any conclusions.

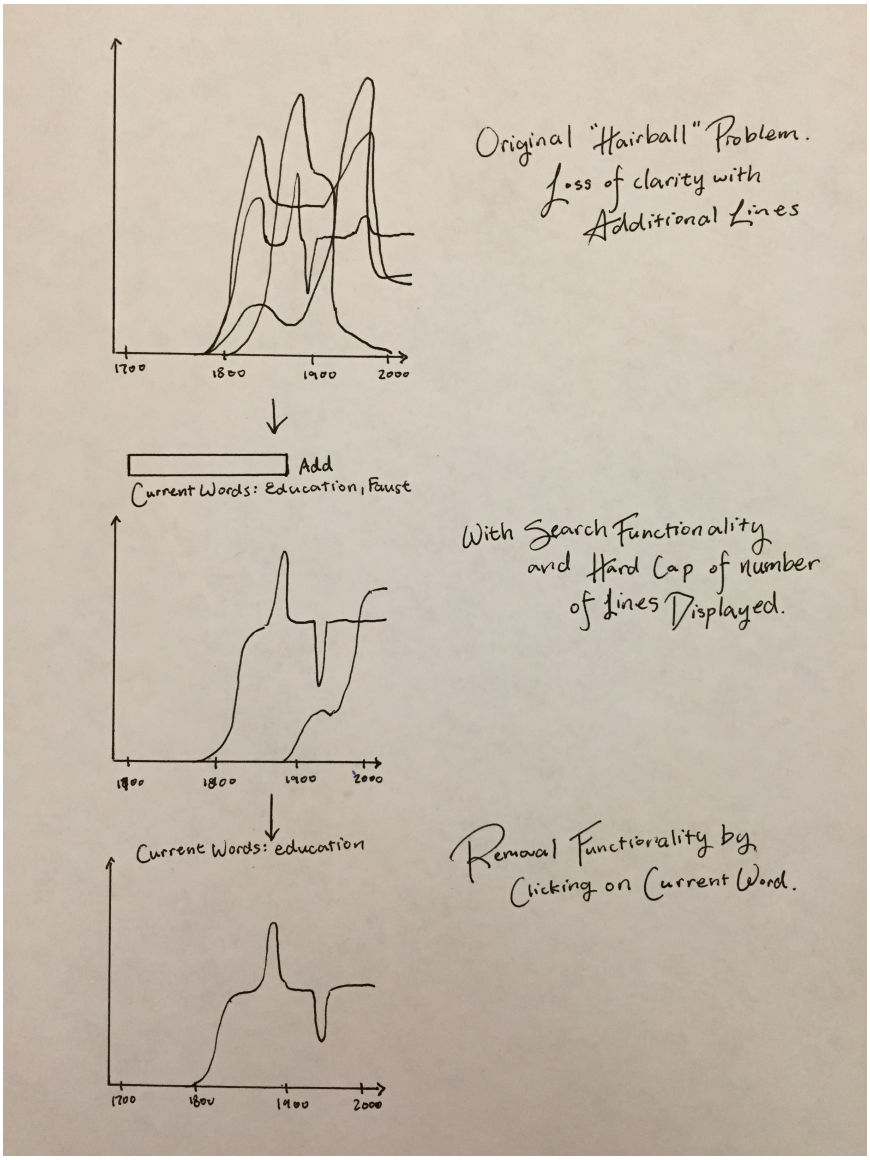
Design Evolution

From the start, we hoped to create a visualization that would target an audience of Harvard students who would be inherently related to Harvard's history. We began with a main visualization for clarity and decided upon a line graph to show changes in frequency over time. Due to the large span of time covered by Harvard's history, we also decided to implement a time brushing functionality in order to give detail on the data over smaller time spans.



With our basic setup determined, we were also interested in how to best show the specific words that were of interest to our audience, so we implemented a search query functionality. Unfortunately, we were forced to limit our possible word selections with an autocomplete search bar because scraping the archives resulted in a data file over 1.2GB in size, searching over which would have slowed down the visualization to unreasonable levels.

To prevent our visualization from becoming too messy with the addition of multiple search queries, we placed a hard cap on the number of words that could be displayed at a time and implemented functionality to remove displayed lines as well.



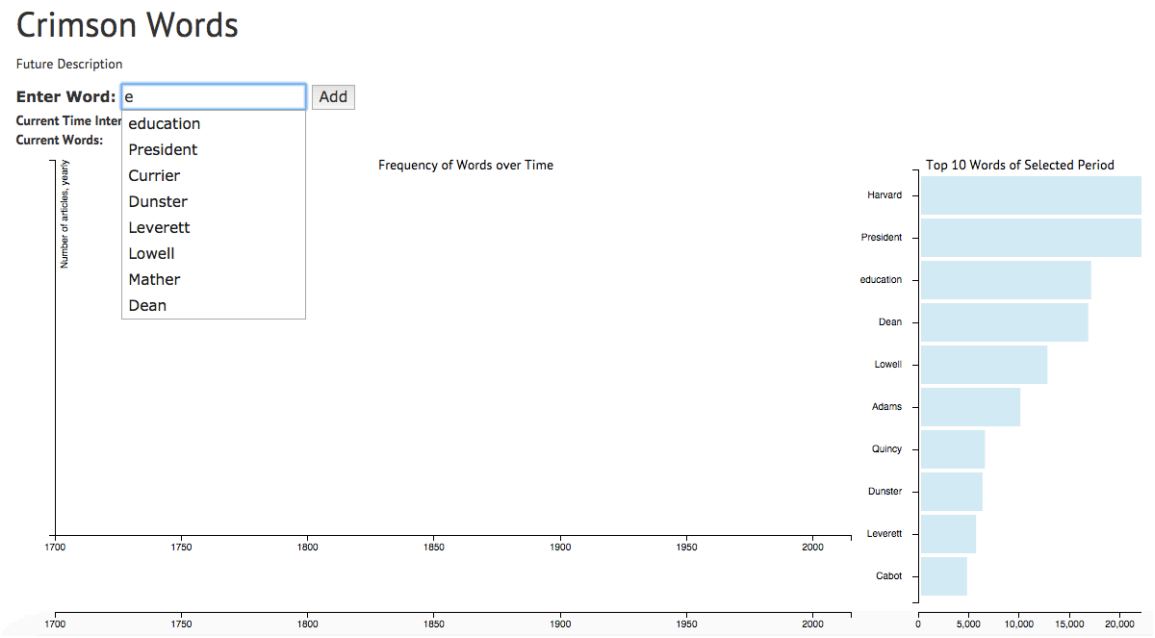
We added in an overall top frequency visualization to the right in response to design critiques that overall trends should be displayed to give users an idea of what they should specifically be looking for.

Implementation

Frequency of Words over Time:

Overall
In this visualization, the user can enter a word in the search box to see its frequency of appearances in articles (as a line graph) over the years. When the user starts typing into the box, a drop-down menu of words containing those letters appears.[1]

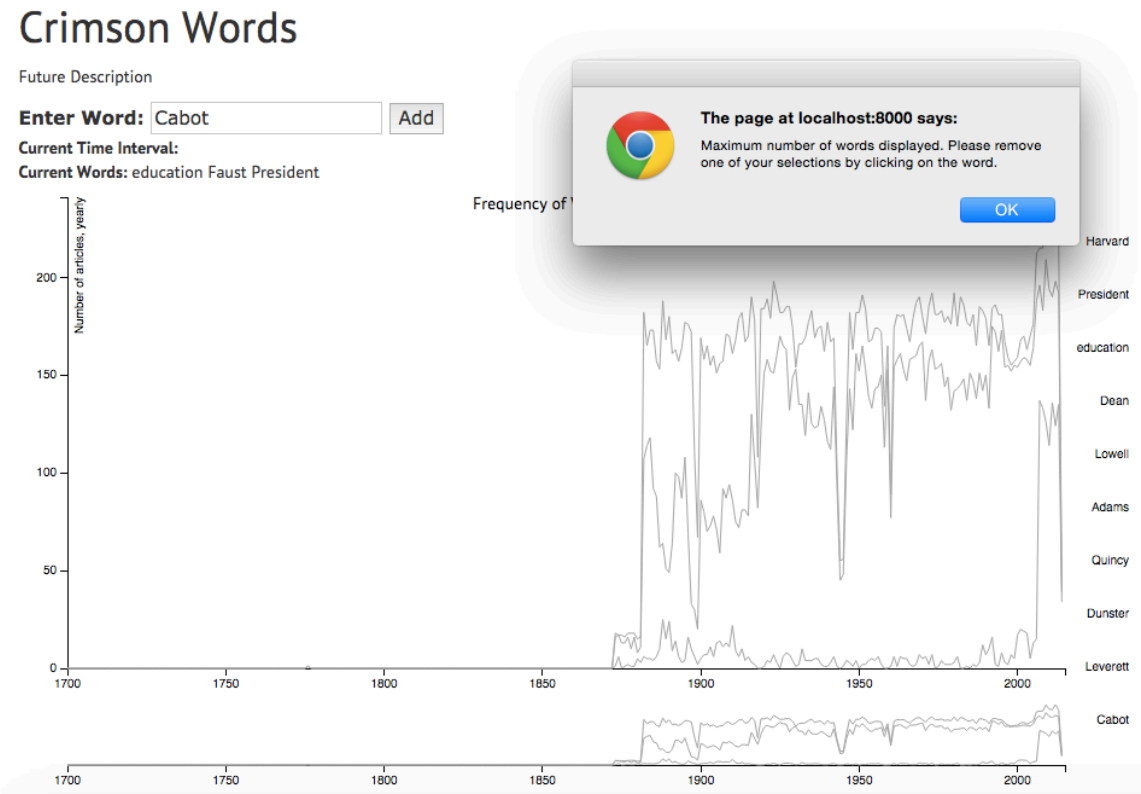
Figure 1.



The visualization permits a maximum of three line graphs to avoid clutter. When the user attempts to add a fourth word, a message pops up to inform the user about the maximum.

[1] Note for future changes: Do not allow users to add words that are not included in our list. Currently, misspelled or excluded words show up on the page but no line graphs are added.

Figure 2.

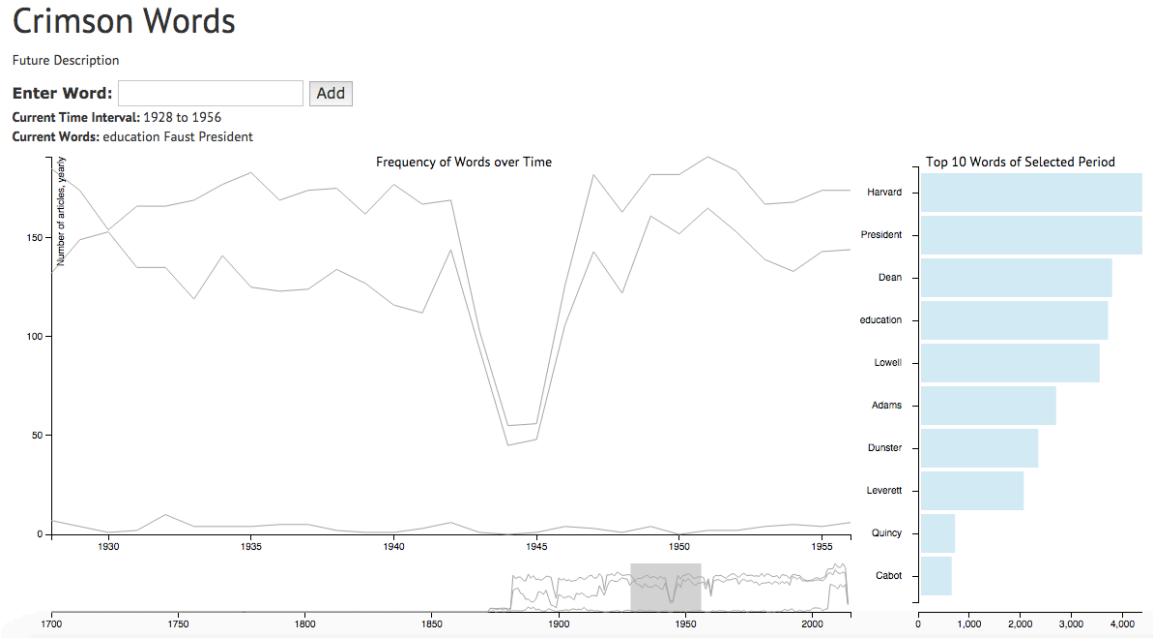


Clicking on the word deletes the line graph.

Brushing

The smaller graph on the bottom serves as the context of the larger graph above, which presents the focused area. The user can select an area on the bottom of the graph and the data for the selected time interval will be displayed in the graph above.

Figure 3.



Details (of the "focus" graph)
A vertical bar tool helps display the exact frequency of a word in a particular year based on the location of the mouse. The frequency of the last mouse-overed line graph is shown; the circle attached to the vertical bar follows the last mouse-overed graph. Mousing over a line graph highlights that graph and its word; mousing over the word highlights the word and its graph.[2]

[2] Note for future changes: Eliminate errors from the tool when no line graphs.

Figure 4.

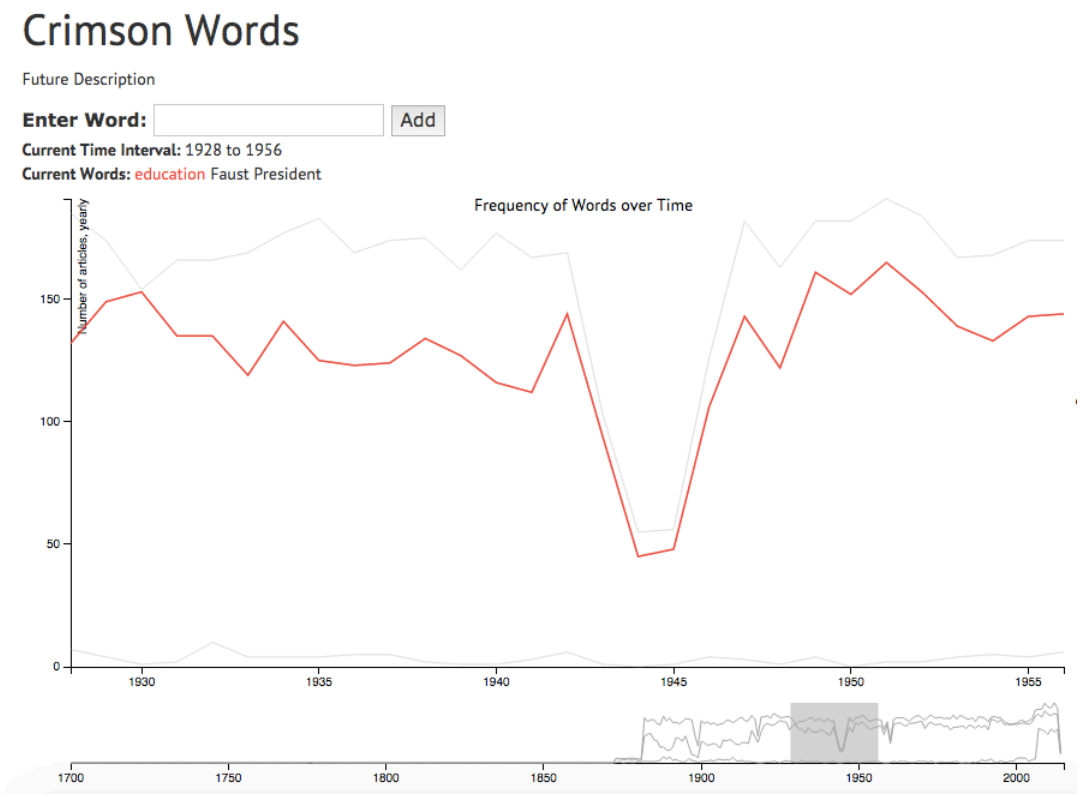
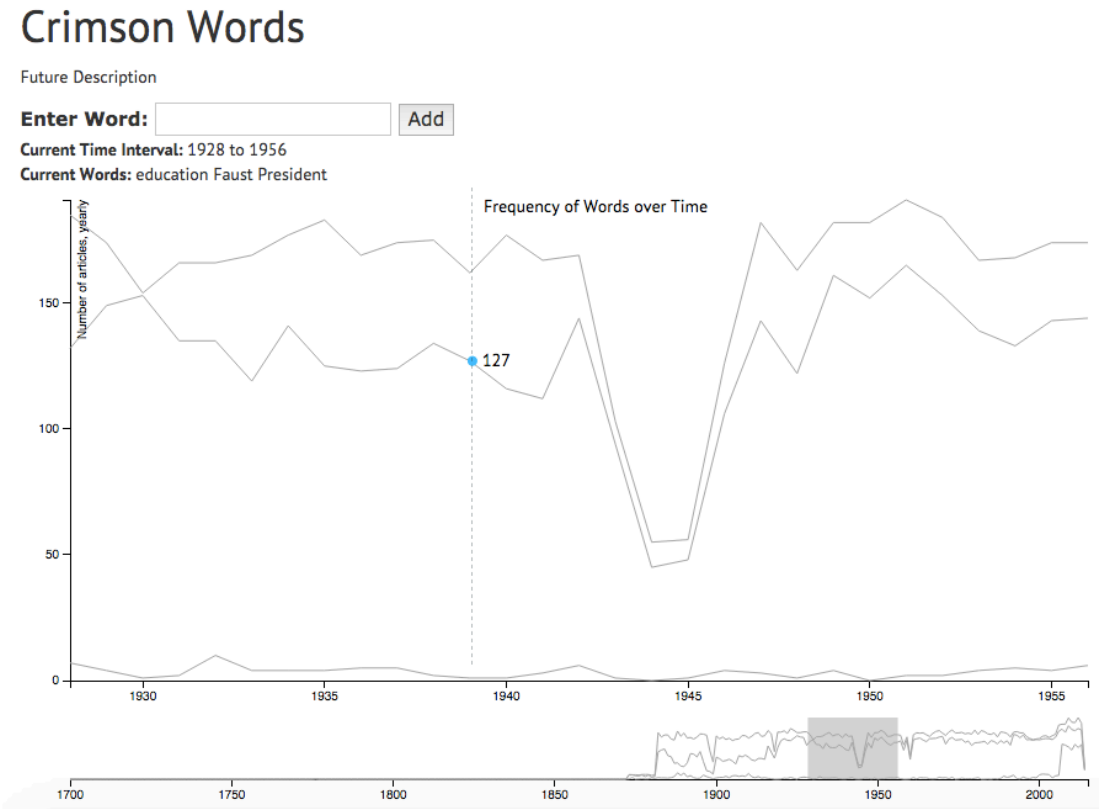


Figure 5.



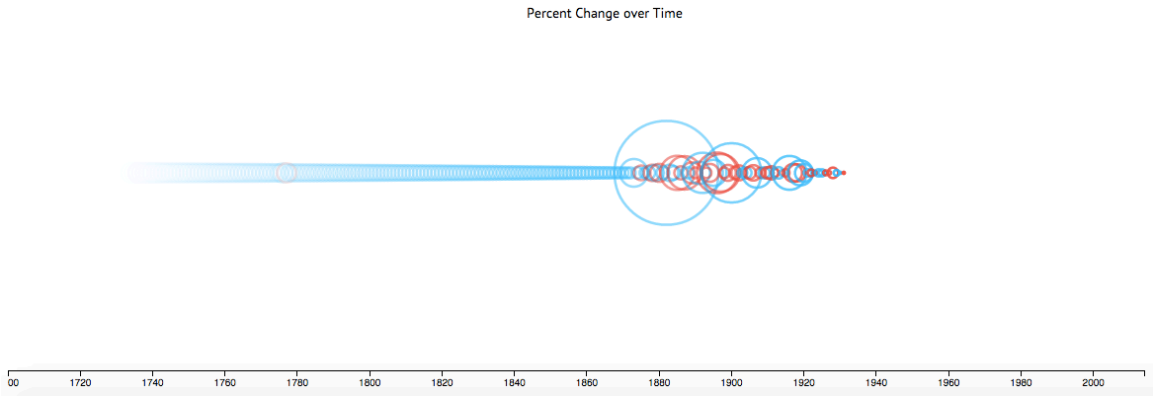
Top 10 Words:

This visualization shows the top ten words in the selected time period and the number of their appearances in articles over that period as bars. The words and their corresponding bars are sorted from the highest count to the lowest. (See Figures 1 and 3.)

Percent Change over Time:

This visualization displays the percent change of the frequency of appearances in articles from one year to the next. The size of the circles shows the magnitude of change while the color shows the direction (red is for negative change and blue is for positive). Click anywhere on the area to start the animation.[3]

Figure 6.



Evaluation

We are still in the process of creating our visualizations!

[3] Note for future changes: Allow users to choose which word they would like to see for this visualization. Currently, the data for the word "education" is used.