CS171 Final Project Process Book

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

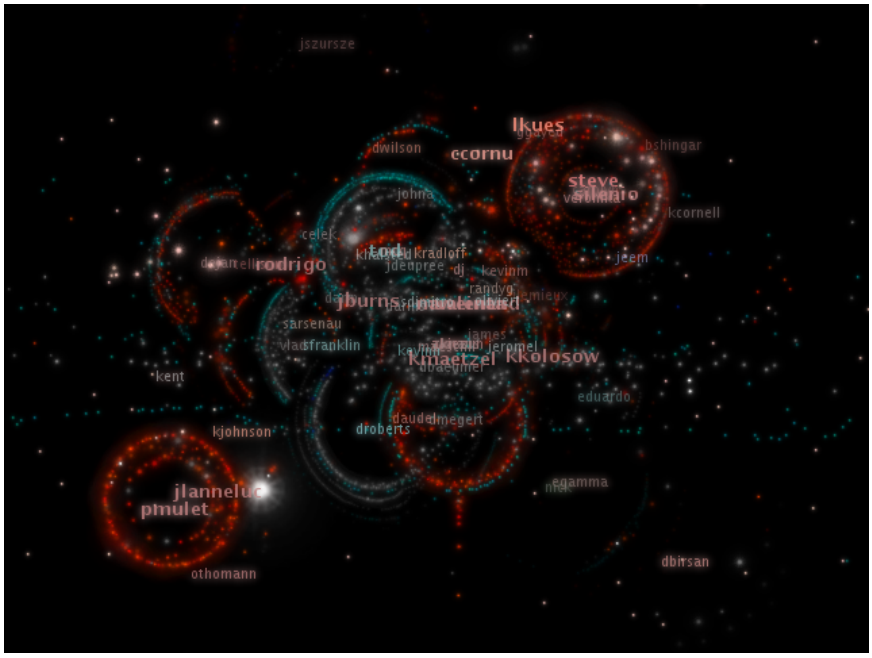
Overview and Motivation

Harvard University stands at a 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. Next, 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 proved to still be too large for D3.
5. Finally, we created a hybrid solution of #3 and #4, creating an index table for the 500 most common "interesting" words, and querying that via AJAX to PHP to SQL and returning a

JSON. This is the final solution that is implemented in our code.

Additionally, after the data is pulled from SQL (after user interaction), we find how frequently words appear on the same day to create our association data set used for the last graph of our visualization.

Exploratory Data Analysis

Initially, we implemented simple line graphs just to get an overall idea of how the frequencies looked over time. We noted that if we looked at time increments that were too small (say, on the weekly/daily level) then our graph would look like a series of spikes with no clear pattern. Therefore, we made the decision to aggregate the data further and implement our visualizations on a scale of years.

We then noted that the line graph showed interesting frequencies over time, but we were also concerned with total counts over our entire time span so we also installed a bar chart implementation showing the integrated line graph (the total counts over our time interval).

We saw that our line graph was typically consistent with intermittent jagged spikes downwards for most years and a final trend up for most words, so we thought it would be interesting to visualize differences between frequencies over consecutive years in a creative manner. Therefore, we implemented the ripple effect visualization to glimpse a quick yet informative picture of how a word has grown or shrunk in frequency over time.

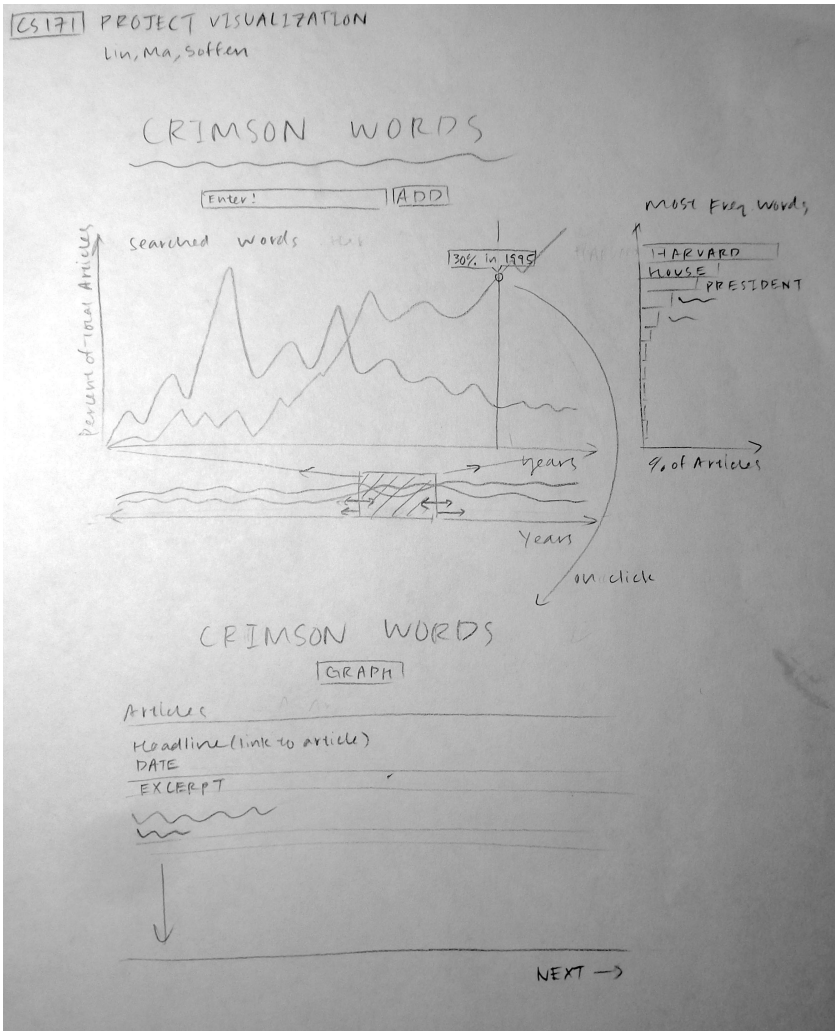
Due to the fact that these were the top words in the Crimson, their frequencies grew with each other as Crimson articles become more frequent or well documented in the Crimson archive. In order to accommodate this high correlation as we implemented our force layout with weighted associations, we made the design decision to not make link lengths linear in order to emphasize differences in the number of associations.

As for interesting correlations in words, we could see general trends for most words, for example, the consistent association between Harvard and war. However, we would hesitate to name any causation

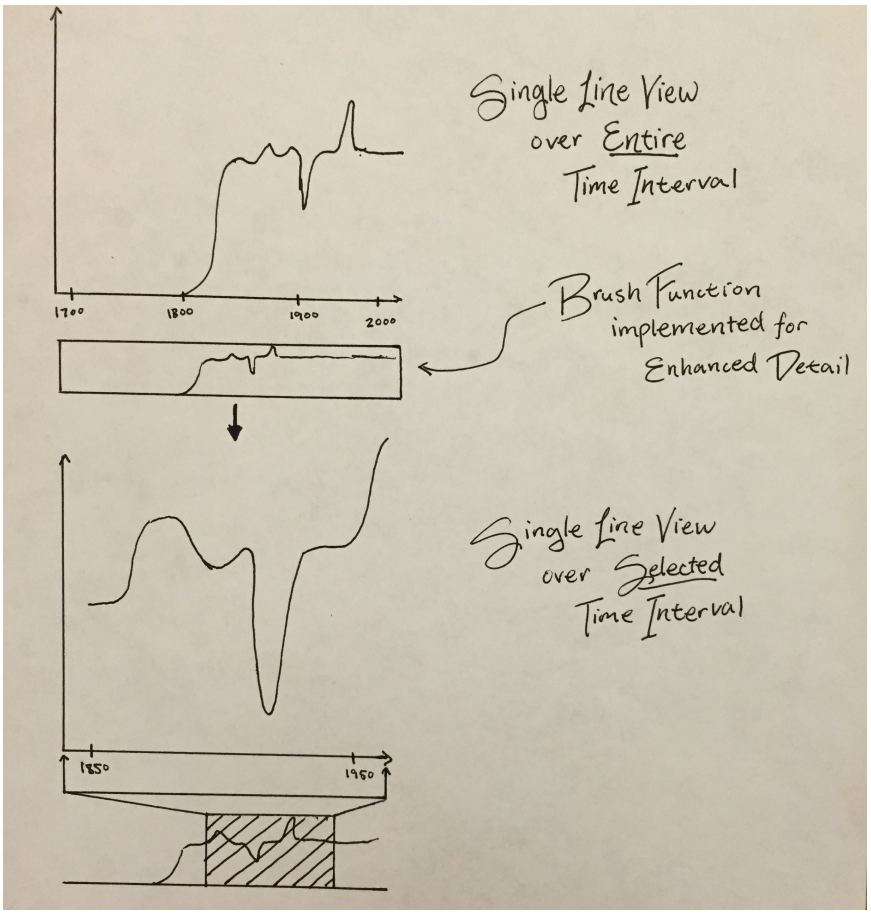
between these words and say that they are nothing more than interesting associations that may be considered.

Design Evolution

Initial Sketches

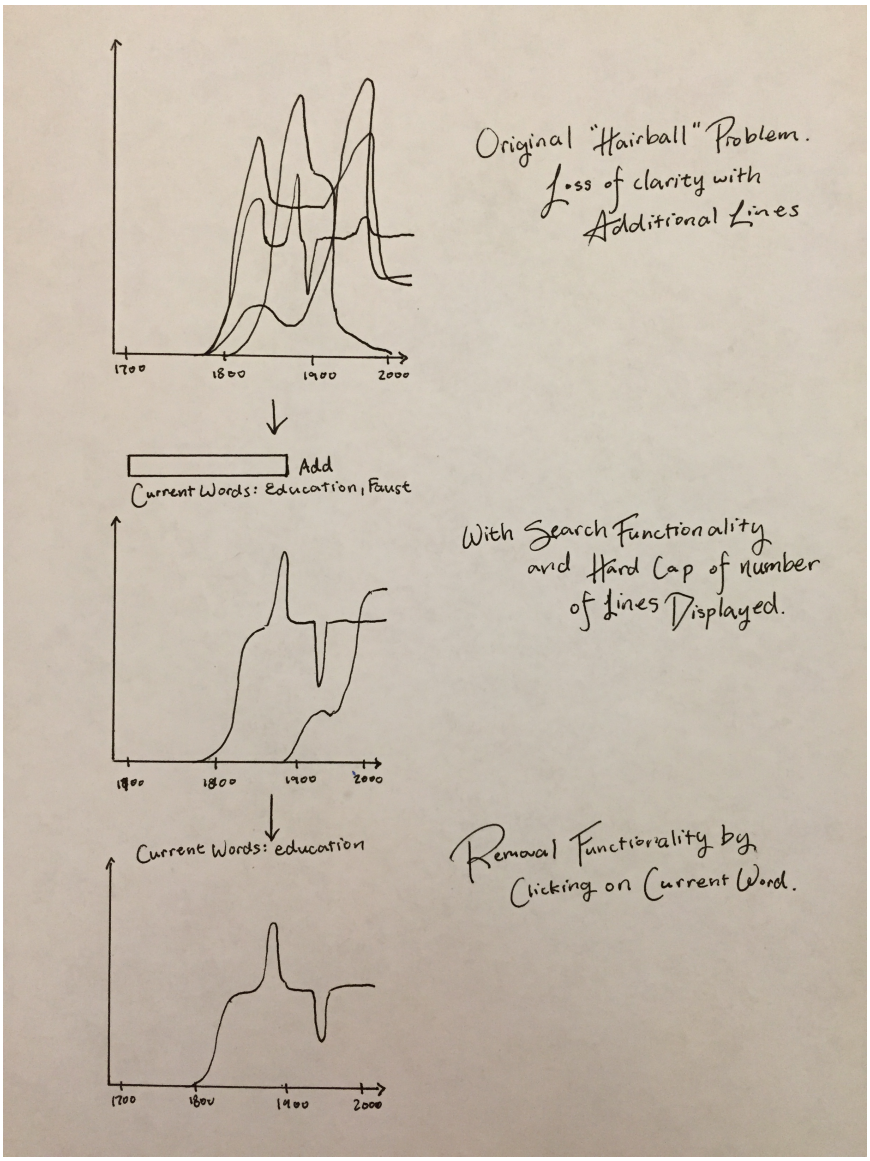


From the start, we hoped to create a visualization that would target an audience of Harvard students who would be inherently related to Harvard's history. We began with a main visualization for clarity and decided upon a line graph to show changes in frequency over time. Due to the large span of time 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.

From Milestone One

Our data set now changes dynamically according to the user's interaction with the line graph. When the user adds a word, word is added to the currently used data set; similarly, when the user deletes the word, the word is deleted from the data set. Since we use the same data set for all of our visualizations, we needed to decide upon a number of words that would work well for all of them.

Due to this change in the way we managed data, we increased the limit of line graphs to six, since six line graphs still look reasonably distinct and six nodes for our word association visualization would allow for a reasonable number of comparisons.

Frequency of Words over Time:
The overall structure of this visualization (see Figure 1) portrayed the information we wanted clearly, so we decided to simply build upon it. However, the initial empty graph was not appealing visually and did not properly guide the user. Therefore, we decided to only show the "Enter Word" input box when the user reaches the visualization part of the page. After the user enters a word, all the visualization components become visible.

Figure 1. Before

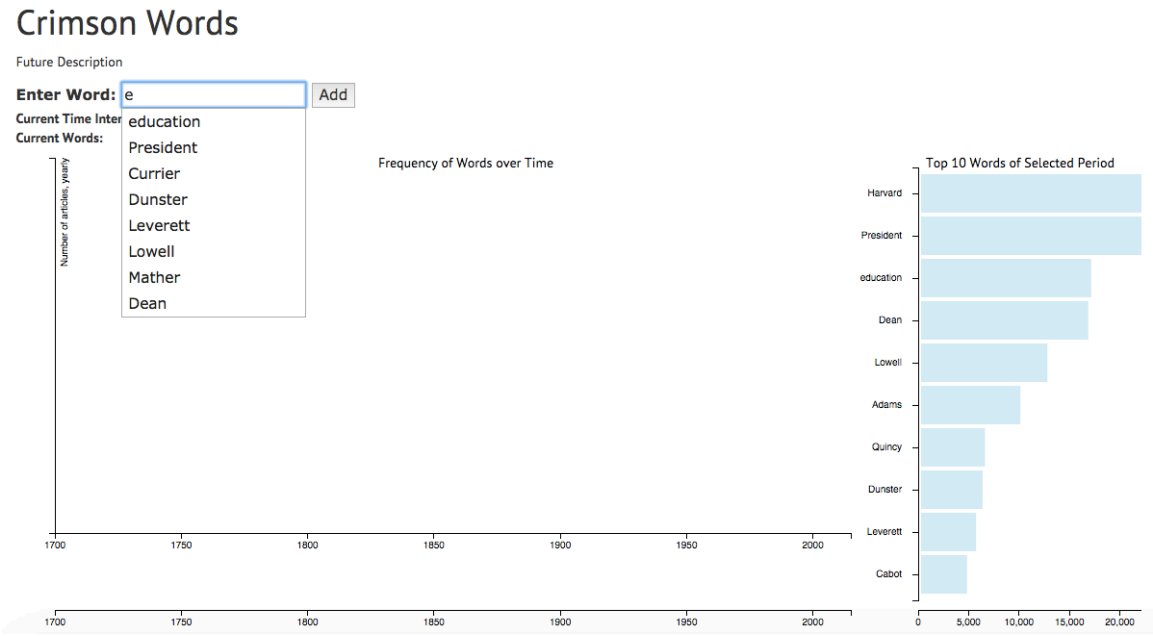
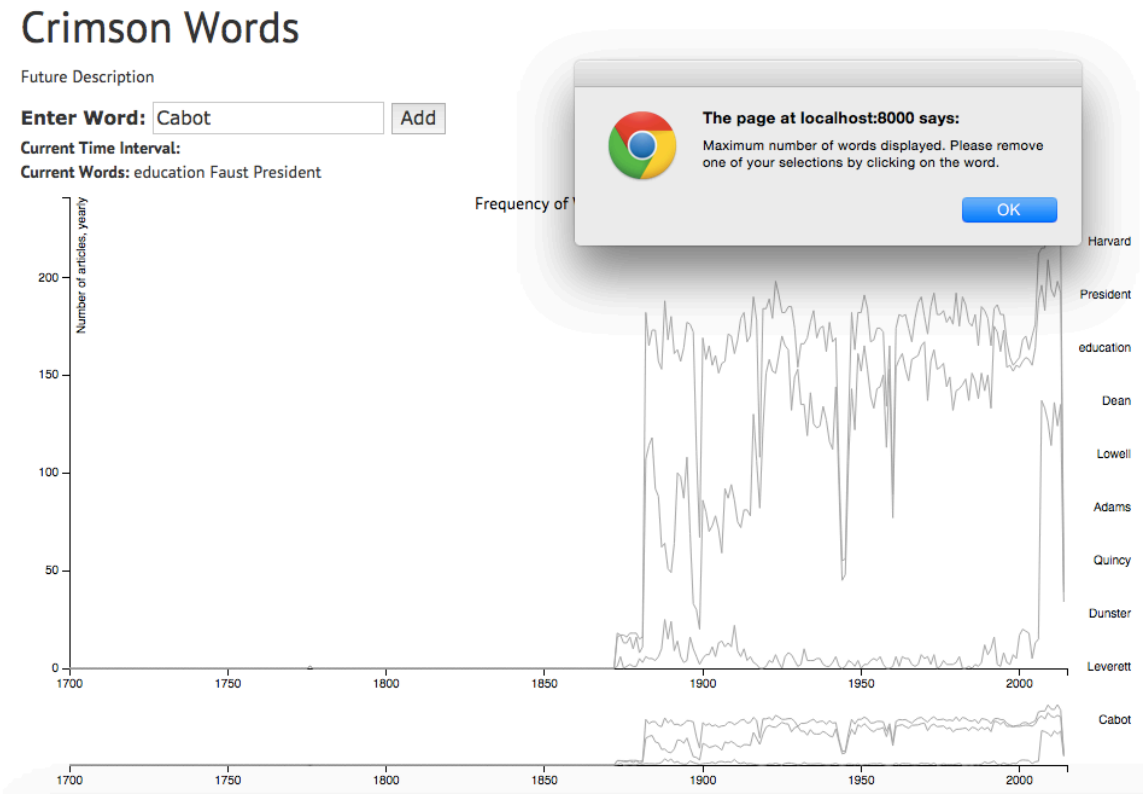
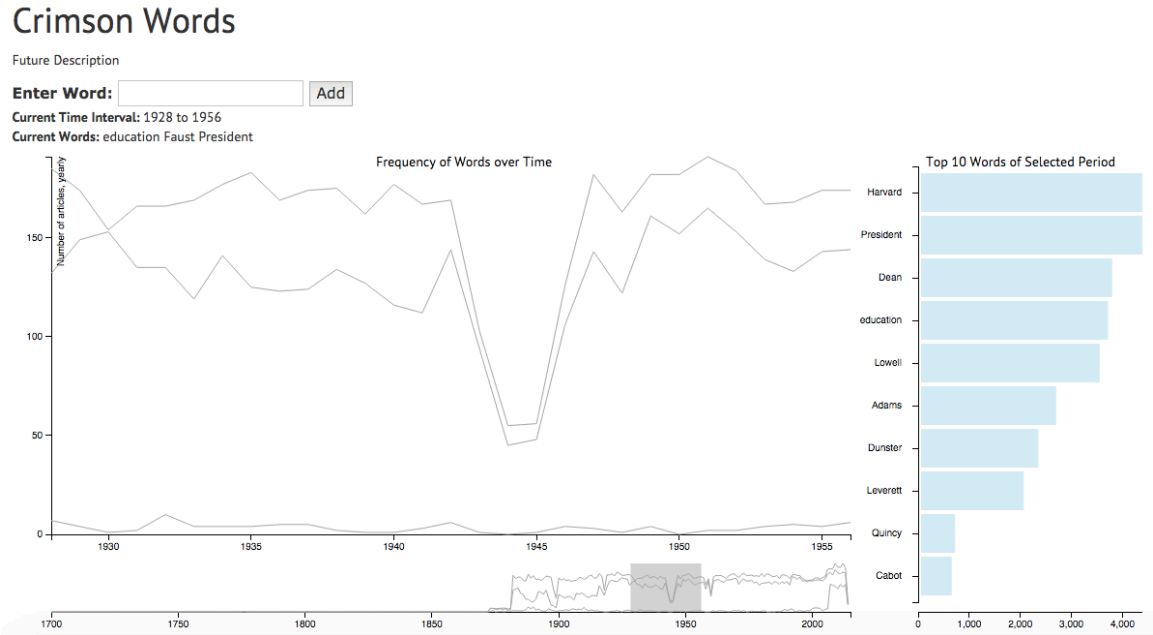


Figure 2. Before



The functionality of the brushing stayed the same (see below).

Figure 3. Before



Details (of the “focus” graph):
In addition to showing the frequency of the last mouse-overed line graph, we have added the year and the word that is associated with the line.

Figure 4. Before

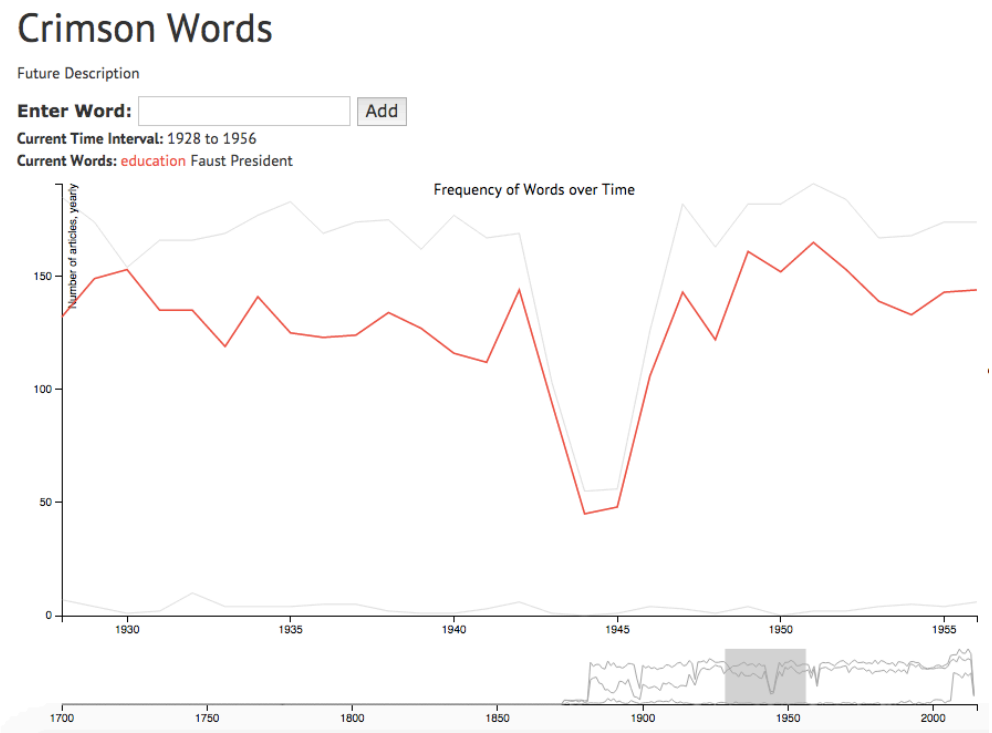
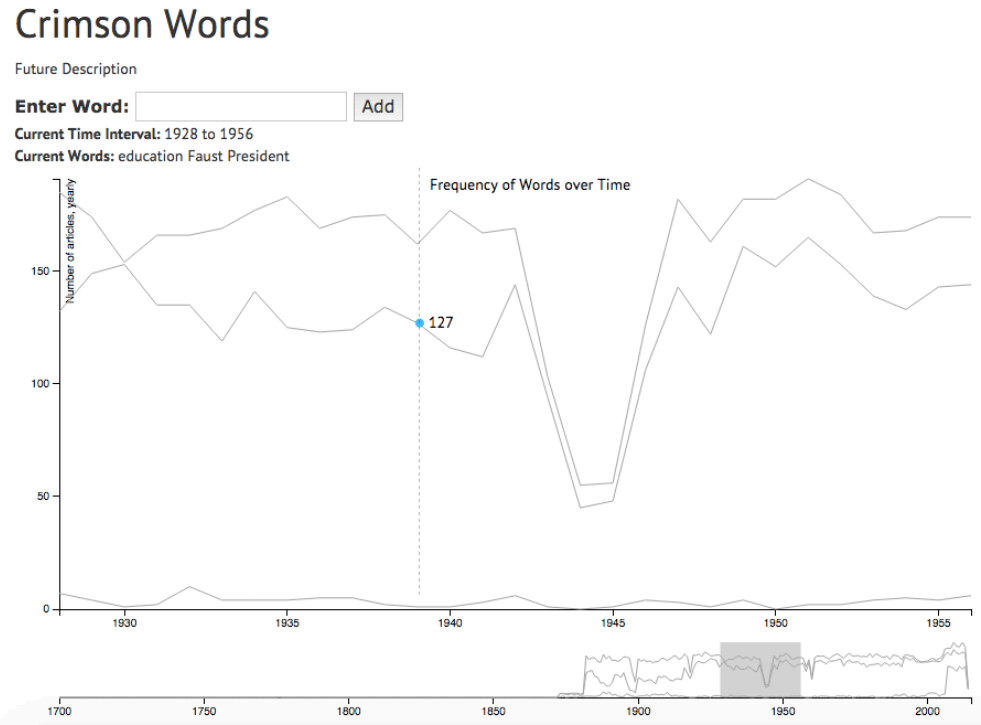


Figure 5. Before



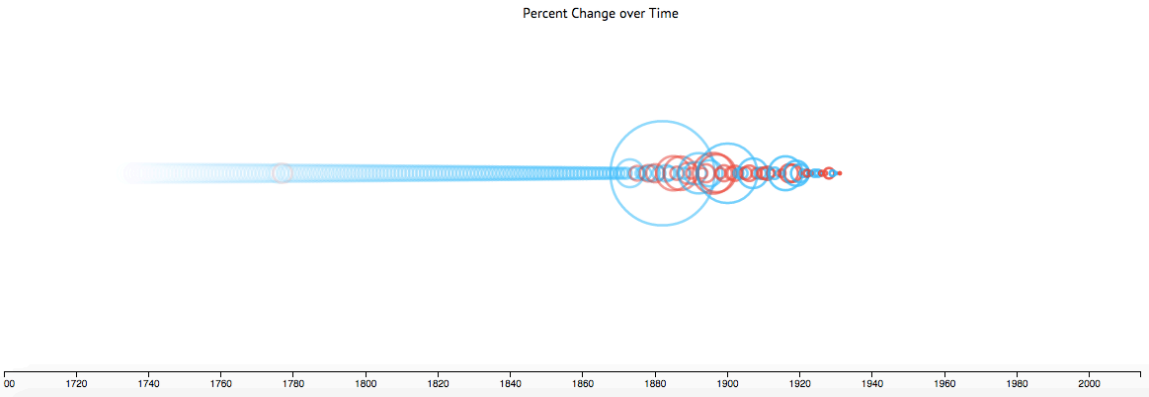
Top Words → Overall Frequency

This visualization no longer shows the top ten words in the selected time period and the number of their appearances in articles over that period as bars. Instead, it simply orders the words displayed on the line graph from the one with the highest count to the one with the lowest. (See Figures 1 and 3.)

Percent Change over Time:

While the size of the circle still shows the magnitude of change, we changed the colors because of our red background. (Blue is now for negative change and white is for positive. If no change occurred over a year, nothing is displayed.) The user can now select which word to examine by clicking on the line graph of that word.

Figure 6. Before



Beyond Milestone One

Word Associations

We added a force layout to display the relationship between words. We chose to focus on the association between a selected word and the other words in the data set. The size of a node connected to the center node represent the number of days in the selected year that both words (represented by the two nodes) appeared in an article on the same day. If the link distance is weighted by association, the nodes that represent words that are more "associated" (or appear more frequently together) with the selected word move closer to the center node.

The size is scaled across the years; therefore, when the count increases from one year to another, the node also increases. The size is not relative according to the year to allow users to more easily see changes across the years. Similarly, the link distance is scaled across the years for similar reasons.

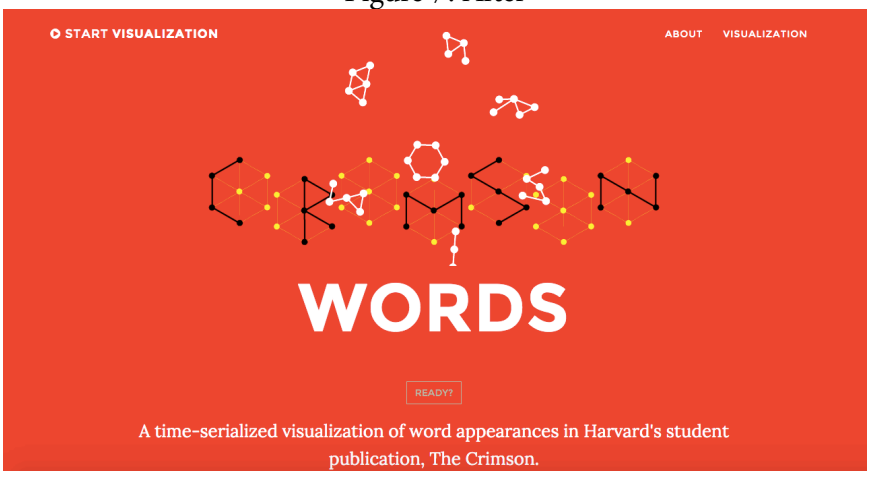
Because of our large source data file, we could only select a few years to find association data in order to keep processing speed reasonable. Therefore, we choose the years 2000, 2005 and 2010 to allow users to find patterns in word association in recent years.

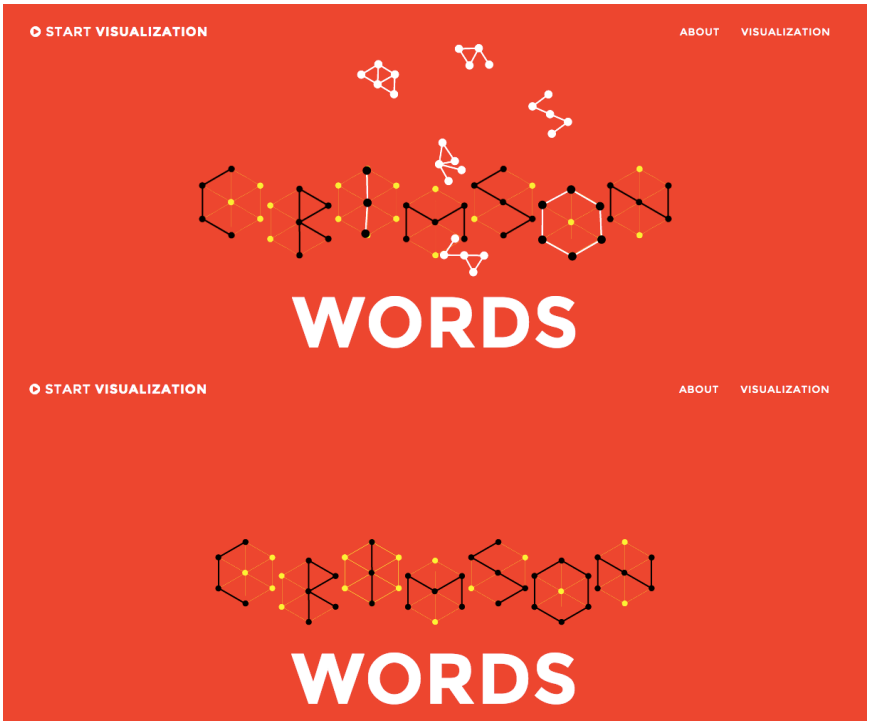
Implementation

Start

Our page first takes users to an interactive feature involving our title. The floating elements (created using d3 force layout) represent the letters in the word "crimson." The style of these elements and the letters of the background image are similar to the way we present word association in our data visualization. The user can play around with these elements. Instructions appear by hovering over the "Ready?" button. Ultimately, we wanted users to complete the word "crimson" in the background, which is missing the letters "i" and "o," by dragging the floating "i" and "o" to their right places. Once the user fixes all nine nodes, by clicking the button below the title, the background image reloads to display a completed "crimson."

Figure 7. After

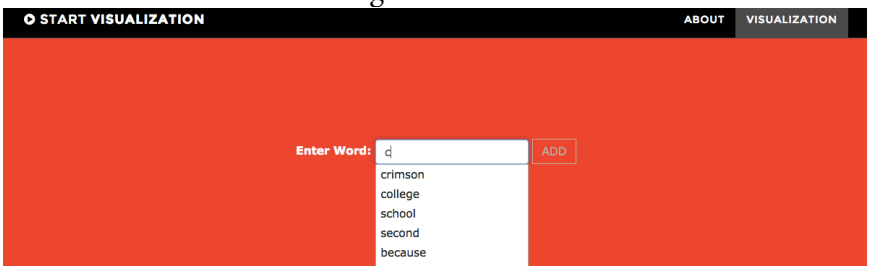




Frequency of Words over Time:

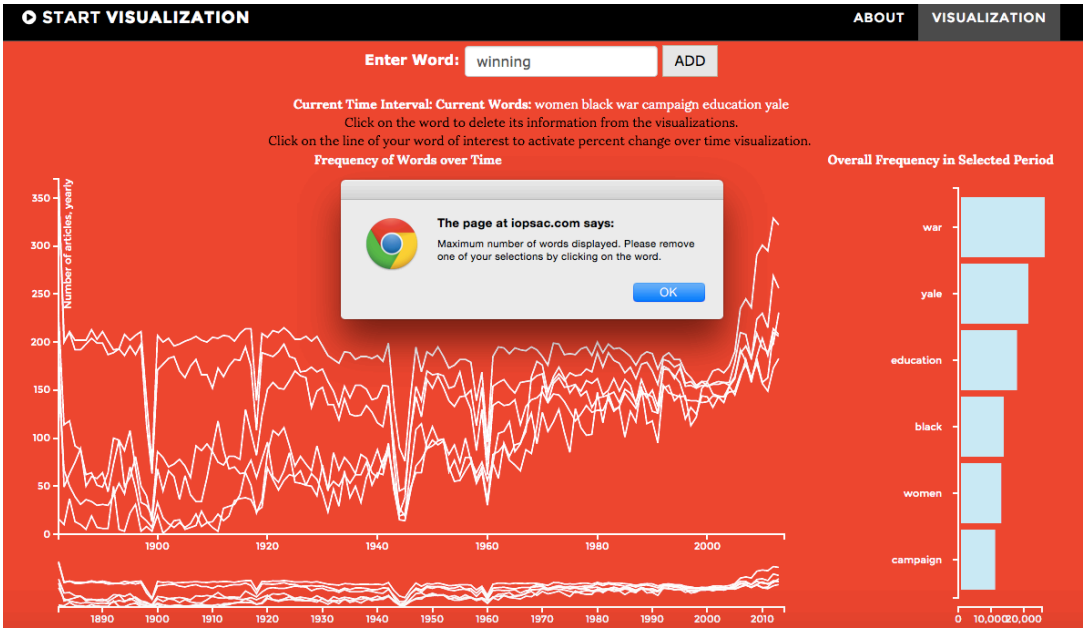
Overall:
In this visualization, the user can enter a word in the search box to see its frequency of appearances in articles (as a line graph) over the years. When the user starts typing into the box, a drop-down menu of words containing those letters appears (see Figure 8).

Figure 8. After



The visualization permits a maximum of six line graphs to avoid clutter. When the user attempts to add a seventh word, a message pops up to inform the user about the maximum (see Figure 9). If the user tries to enter a word not in our list of 500 or a word that has already been added, the action is denied.

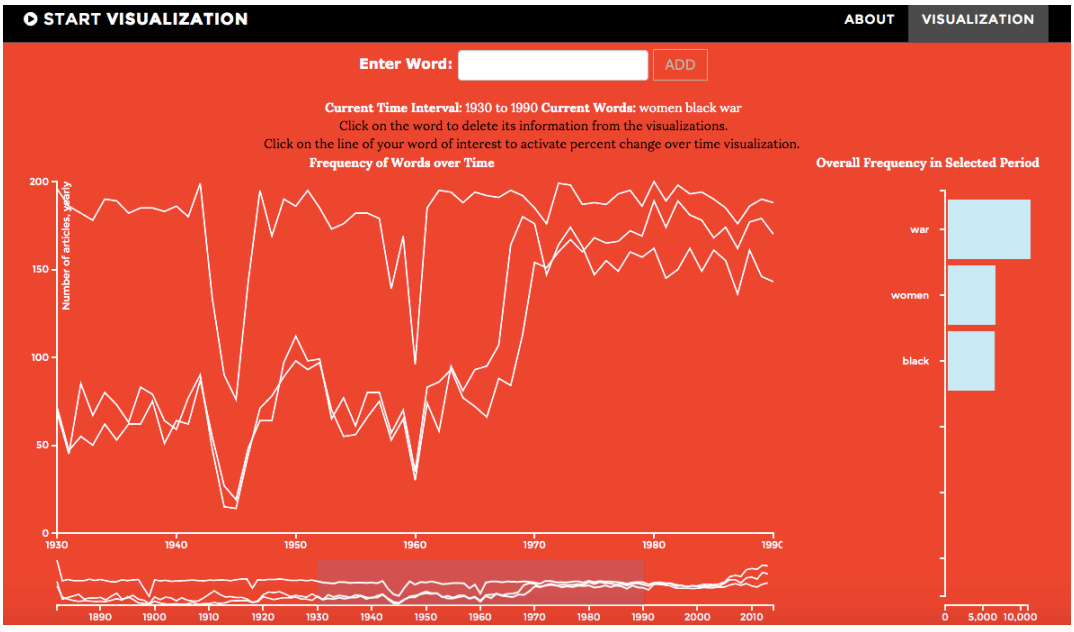
Figure 9. After



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 (see Figure 10).

Figure 10. After



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 (see Figure 11). Mousing over a line graph highlights that graph and its word; mousing over the word highlights the word and its graph (see Figure 12)

Figure 11. After

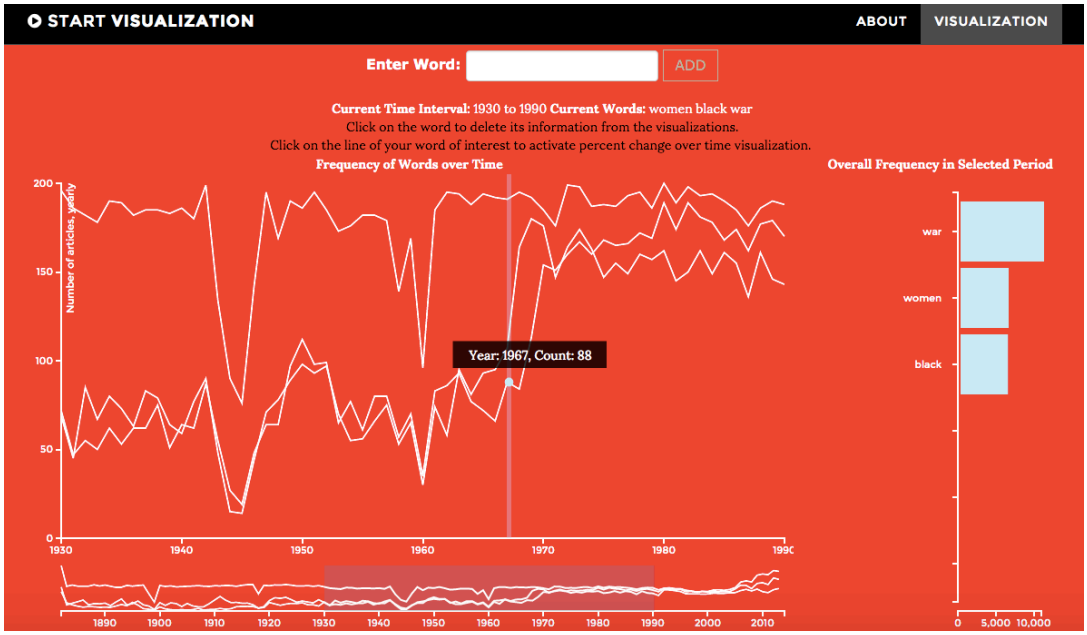
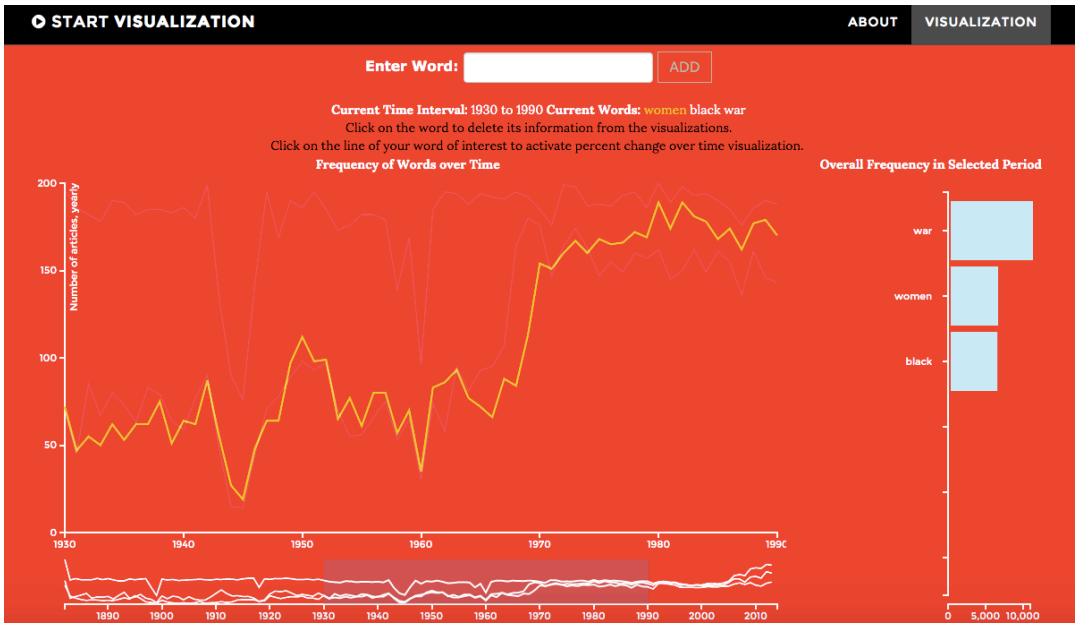


Figure 12. After



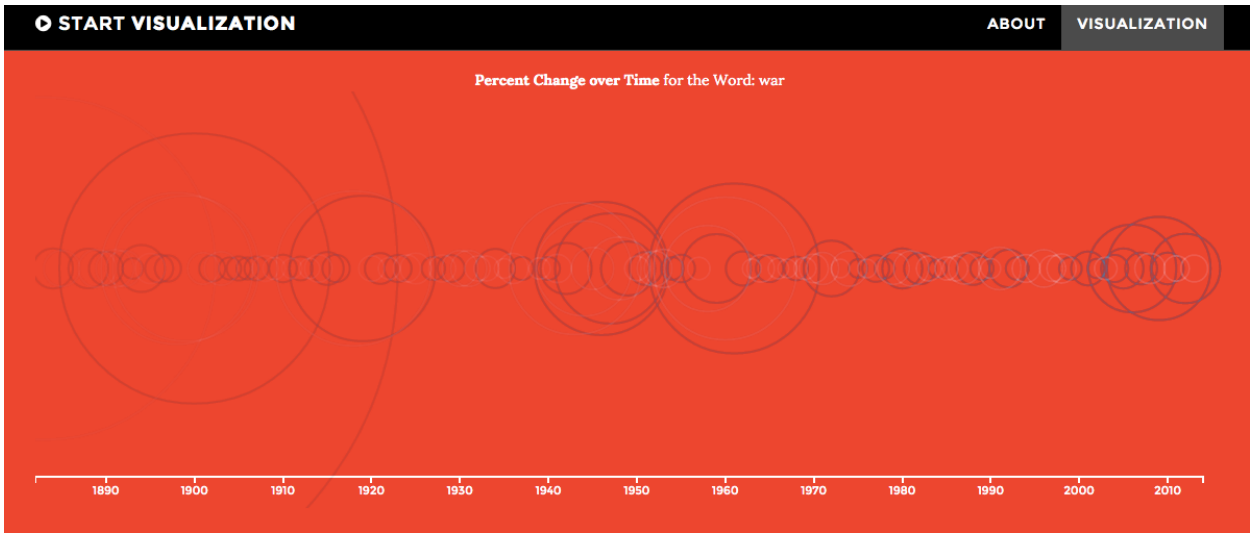
Overall Frequency:

This visualization graphs the area under the line graphs for each word. The words and their corresponding bars are sorted from the highest count to the lowest. (See figures above.)

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 (blue is for negative change and white is for positive). Click anywhere on the area to start the animation (see Figure 13).

Figure 13. After



Word Associations

The default force layout presents all of the words shown in the line graph as nodes (see Figure 14). Clicking on a node links the node to all other nodes. All the nodes, except the center node, change size to represent the strength of association between itself and the center, the greater the size, the stronger the association. If "Weighted by Association" is clicked, the link distances change depending on the strength of the relationship, the closer, the stronger the association. Hovering over the nodes give users the word and number of "connections." The slider permits the user to see changes across the years.

Figure 14.

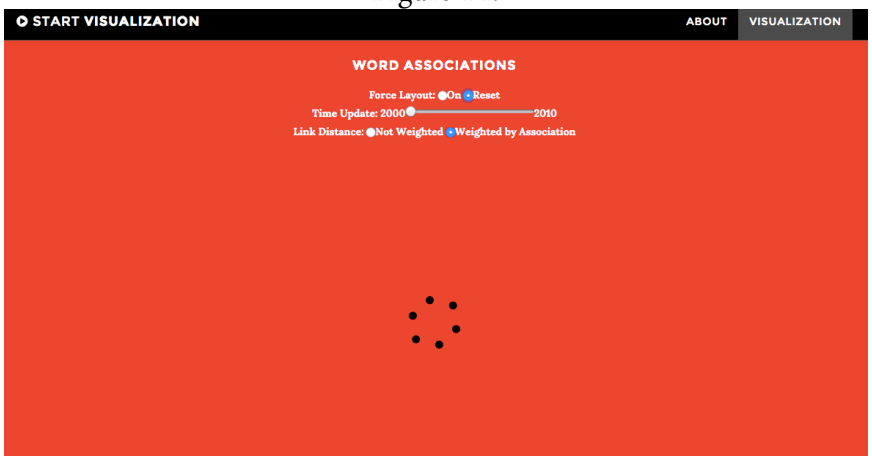


Figure 15.

The reset button allows the user to return to the original force layout.

Evaluation

Our visualization has achieved what we envisioned at the beginning of the project and serves as an interesting tool to discover and explore patterns of words as a reflection of beliefs and events over time.

Improvements
We noted that though the data was represented accurately, it might have been nice to be able to do some more custom queries into the Crimson Archives for words that may have significant historical trends. We were having a lot of trouble scraping from the archives due to frequent server crashes, and we had to rely on a live query method because the data file was too large for D3 to load otherwise.

Additionally, more data points could visually improve the force layout. Because we could only obtain association data after we solved our issue of the large size of the data set, we did not have as much time to develop the word association visualization as we would have liked. For any future developments, we can increase the complexity of the force layout to increase links between all nodes.

In general, given more time, we might have been able to work around the technical issues of selecting, managing, and deleting a larger set of data. For instance, the current method of obtaining the data sets could be changed to allow for more flexibility in showing different aspects of the data. Instead of deleting from the currently in-use data set, we could keep building on top of it (as the user keeps adding words) and pull subsets from as deemed necessary for each visualization.

Smaller notes:
Our live query is constrained by a drop down menu to let the user know what words are available. Unfortunately, we were not able to figure out how to truncate this list when a short query, say the letter "a", was inputted, leading to a large list of options that is annoyingly long. Since the list shortens as a more specific query is inputted, we consider this a minor inconvenience.

The tooltip of the line graphs sometimes move out of range of the SVG and we can no longer see the information. We were unsure how to solve this problem.

Observations

To answer the questions we raised at the beginning of this project, it is safe to say that we can get a grasp of some issues that were important to Harvard over time.

Of note, we saw that the words "black" and "women" rose substantially in frequency during the late '60s and the early '70s. We found this to be historically interesting as those were times of the Civil Rights Movement. From our word association visualization, we observe that “black” and “white” is most closely associated “men” in the current century, perhaps illustrating the disproportionate representation of sex in the news today.

Just as interesting, we note that mentions of "radcliffe" drop substantially after its merge with Harvard in 1999.

Perhaps more tragically, if we make the assumption that the word "harvard" is the baseline (mentioned in most if not all articles) and models the total number of articles, then it is a sad conclusion that “war” closely approximates the "harvard" line, making it one of the most frequent and constant terms in the Crimson literature.

From these issues, we noted that our visualization, given some time and attention, can reflect interesting historical trends in one of the oldest student publications in the United States.