Project Proposal:

Crimson Words Archive

Jimmy Lin, (Lisa) Tianxing Ma, Kim Soffen

Background and Motivation.

We are all interested in journalism and current events, especially in their ability to give inferences into what was important and urgent to a population in their contemporary period. We thought it would be interesting to create a visualization for The Harvard Crimson, to be able to examine the trends in the university's history via the words that appear in the newspaper. All of us agree that Harvard is particularly rich in history and frequently involved in global current events, and we strongly believe that this visualization provides a gateway to tap into the dynamics of the Harvard community.

Project Objectives.

We are interested in investigating word frequencies in the publications of The Harvard Crimson. We hope to present the data in a way that may reveal trends in the current events and issues important to the Harvard community in terms of both time and magnitude. Ultimately, our project will provide the benefits of giving insight into the social and political culture of Harvard at a given point in time and providing a summarized snapshot of Harvard thought. Additionally, we believe that there is a strong interest amongst Harvard students to see these trends and we hope to generate student interest in both data visualization and the Crimson publication through our work.

Data.

The data (i.e. all Crimson articles that have been put online) will be scraped from the Crimson archives here.

Data Processing.

Data processing will involve indexing all of the articles in the Crimson archives. Given the text of all articles published by the Crimson (the raw data), we will create a JSON data structure. The key will be a phrase, and the value will be a list of article objects that contain that phrase. An article object will contain the article's date of publication (so we can aggregate words over timespans), URL, title, and excerpt (all so we can give basic information about the article after interaction).

Visualization.

Our visualization will have three components. The main component will be a line graph, with time on the x-axis, % of articles the word appears in on the y axis, and every word is a line on the graph. The graph begins with no lines (or we may pre-generate one TBD) on it, and the user can enter search terms to add them as lines on the graph. On this graph, when you click on one of the lines, we will display a list of titles, URLs, and excerpts from each of the articles that contain that word and were published at the time you clicked. Our second component is this same graph in a compressed version below; here, brushing is allowed so the user can zoom in on a certain time frame in the main graph above. Our third component is a horizontal bar chart to the right, which will have the most common words during the brushed time period (besides the industry standard stop words) listed with a bar chart of their frequency.

See attached sketch.

Must-Have Features.

- Main graph of word frequency over time, with each word displayed as a line
- A 'search' feature so the user can add words to the graph
- A compressed version of the main graph that allows for brushing to select a time period for the main graph.

Optional Features.

- Graph of most frequent words in a selected time period
- Click option on the graph, where the user is redirected to the articles containing the searched word at that particular time selected
- Tooltip that displays word, percent, and time period

Project Schedule.

Process book will be completed throughout this period.

Week of April 5th

- Scrape data

Week of April 12th (and April 17th milestone)

- Data processing complete

- Working line graph visualization and compressed version with brushing visualization without full user interaction functionality.

Week of April 19th (and TF check-in)

- Working 'search' feature to add words to the graph
- Tooltip to display words

Week of April 26th

- Graph of most frequent words
- If remaining time is sufficient, implement 'click' option to show article information.
- Test user interactivity functions
- Gitpage website for project setup
- Complete screencast

Final Project Deadline: May 5th

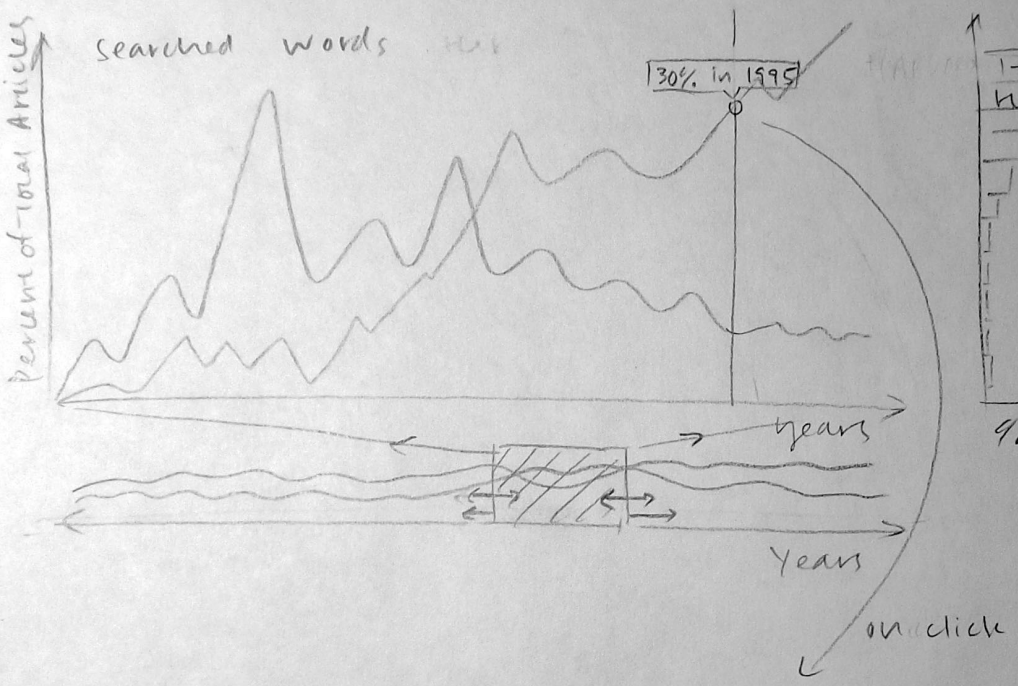
- All components complete

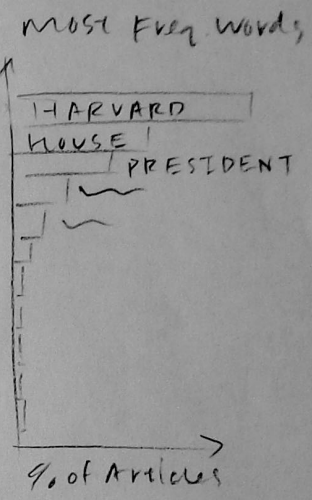
CS171 PROJECT VISUALIZATION

Lin, Ma, Soffen

CRIMSON WORDS

Enter! | ADD





CRIMSON WORDS

GRAPH

Articles

Headline (link to article)

DATE

EXCERPT

NEXT →