Patterns in text corpora

CS294-10 Visualization
Nicholas Kong
nkong@eecs.berkeley.edu

Discovering significantly co-occurring words in the NYT headlines

Problem

We sought to use visualization to address the problem of finding patterns in large text corpora. Specifically, we addressed discovering significantly co-occurring words in the NTY headlines.

We also wished to integrate this visualization into our collaborative visual analysis system as we believe large-scale text analysis could prove a compelling task for group analysis.

Motivation

The internet brought with it access to large volumes of massive text corpora and the concomitant difficulty of making sense of it all. Visualization, and particularly interactive visualization, provides a way to consolidate this information. Examples include aiding search and finding trends in time in news.

However, documents are highly multidimensional and language is inherently ambiguous, making this problem a difficult one.

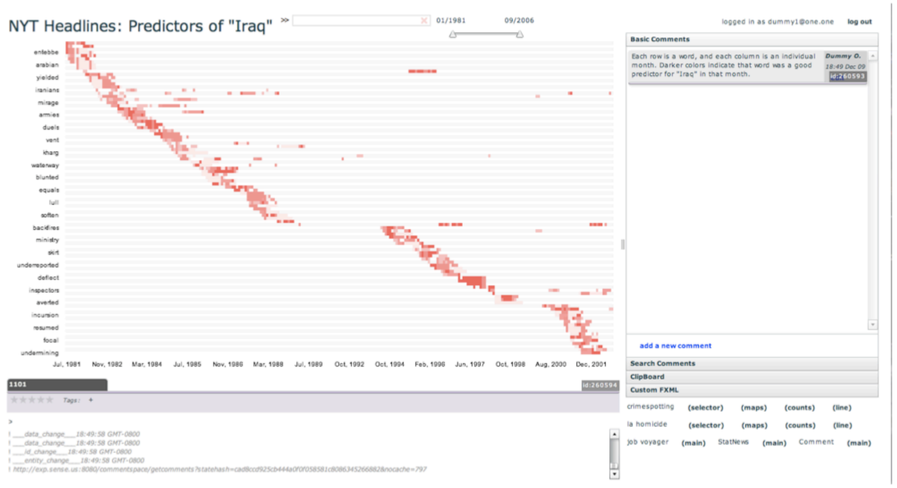
Approach

We worked closely with Professor Laurent El Ghaoui of UC Berkeley EECS. His **StatNews** project provided the statistical data from the text analysis which we subsequently visualized.

We used an early result of StatNews, that of *"predictors"* of the word "Iraq" in New York Times headlines from 1981 to 2006. Using sparse linear regression, StatNews recovers words that predict Iraq in the same headline. We visualize this using an interactive matrix display. This reveals trends over time ("sticky" words) as well as information about the typical lifetime of a story. The visualization is also integrated with our collaborative analysis tool, allowing comments on different states.

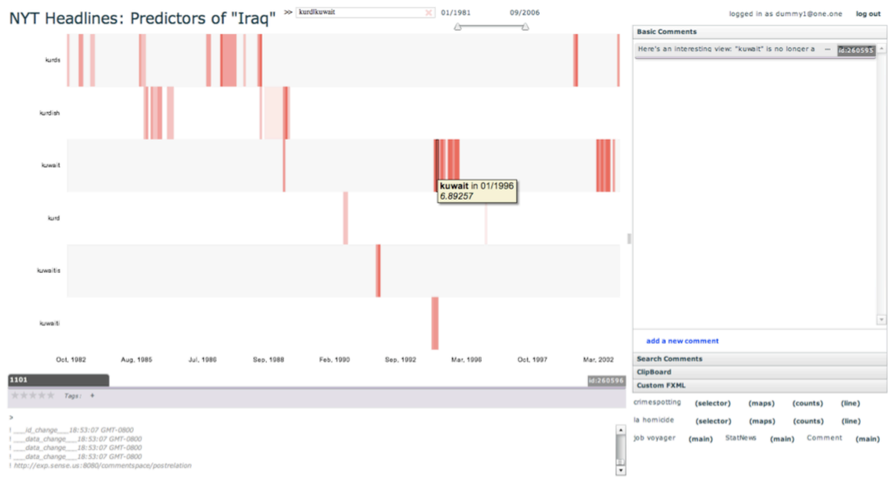
Implementation

- Flex 3 and the flare toolkit
- The visualization is specified via XML and each state has a unique XML representation.
- Only displays top 100 most "significant" words at any view, where "significant" is defined by the total number of nonzero values a word has over the whole dataset.



Initial view

Comment pane on the right is tied to the current state of the visualization, as defined by the search term and the date range.



Drill-down view

The search is for "kurds" and "kuwait". Rollover allows the user to recover the regression coefficient. Note that during the Gulf War (1991), "kuwait" is no longer a strong predictor, presumably because many words co-occur with Iraq in that time period.

Future Work

- Allow for more direct input of date ranges
- Integrate with the forthcoming StatNews API, thus allowing for custom word queries. Currently the turnaround time is below interactive rates.
- Scalability. The system already struggles with around 7000 points, but we should be able to handle much more.
- Allow for comparison of news sources. This tool could be used to reveal bias.