The CMU 2008 Political Blog Corpus

Jacob Eisenstein and Eric Xing

March 26, 2010
CMU-ML-10-101

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

This report describes a collection of political blogs on the subject of American politics in the year 2008. The collection was obtained by crawling blog archives in November and December 2009. It is available at http://sailing.cs.cmu.edu/socialmedia/blog2008.html.
Keywords: datasets, social media
The blogs in the corpus are shown in Table 1. They were selected by the following criteria: the Technorati rankings of blog “authority,” ideological balance, coverage for the full year 2008, and ease of access to blog archives.

In the general election for U.S. President in 2008, the following blogs supported Barack Obama: Digby, ThinkProgress, and Talking Points Memo. John McCain was supported by American Thinker, Hot Air, and Michelle Malkin. In general, the blogs that supported Obama in the election tend to advocate for similar policies and candidates as the Democratic party; and the blogs that supported McCain tend to advocate Republican policies and candidates. Digby, Hot Air and Michelle Malkin are single-author blogs; the others have multiple authors.

2 Data

The corpus includes all blog posts with more than 200 words (rough word counts were obtained by splitting the text on all whitespace tokens). For each post, there are several files; the shared part of the filename indicates the date of the post, under the format: TTYYDDD00_NN: TT is a 2 or 3 character abbreviation for the blog title (in parentheses in Table 1); YY is a two digit year marker; DDD is a three digit indicator of the date of the year (from 000 to 365); NN numbers the posts on that day. Note that numbers are skipped when posts are filtered due to having too few words.

There are three suffixes for each post: .text, .xml, and .info. The first suffix is for files that contain the blog text. The text was scraped using blog-specific XPath and regular expressions, designed to exclude boilerplate text, advertisements, and comments. Blockquotes are included.

The .xml file contains a subtree of the original document that is guaranteed to include all of the text. Finally, a file with the suffix .info contains a set of URLs, one per line: the first is the

---

1 http://technorati.com
3 Acknowledgments

Thanks to Dan Wheeler and Tae Yano for helpful discussions about gathering this data.