Baleen is a tool for ingesting formal natural language data from the discourse of professional and amateur writers: e.g. bloggers and news outlets. Rather than performing web scraping, Baleen focuses on data ingestion through the use of RSS feeds. It performs as much raw data collection as it can, saving data into a Mongo document store.

Natural Language Graph Analysis: Data Modeling

1.5 M documents, 7,500 jobs, 524 GB (uncompressed)
Baleen Keyphrase Graph
Number of nodes: 2,682,624
Number of edges: 46,958,599
Average degree: 35.0095

Baleen's objective is simple: given an OPML file of RSS feeds, download all the posts from those feeds and save them to MongoDB storage. While this task seems like it could be easily completed with a single function, once you start integrating the parts of the program, things get more complex.