PORTLAND, OREGON
PYCON 2017

ON THE HOUR DATA INGESTION FROM THE WEB TO A MONGO DATABASE

By Will Voorhees, Benjamin Bengfort, Rebecca Bilbro, and Tony Ojeda

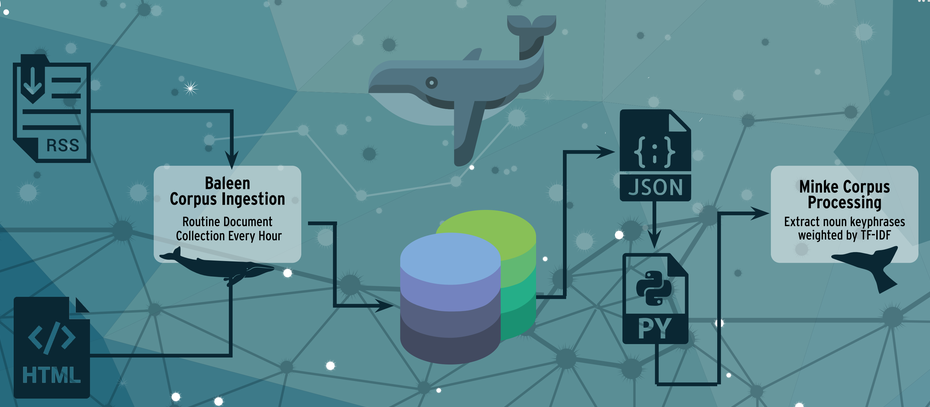
DistrictDataLabs
pycon.districtdatalabs.com

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.

```
import os
import requests
import feedparser

feed = "http://feeds.washingtonpost.com/rss/national"

for entry in feedparser.parse(feed)['entries']:
    r = requests.get(entry['link'])
    path = entry['title'].lower().replace(" ","-") + ".html"

    with open(path, 'wb') as f:
        f.write(r.content)
```

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

Configuration: how do you connect to the database, handle logging, where paths are, or manage passwords?

Ingest: how do you manage feed synchronization and post wrangling, schedule when it occurs, or how to deal with errors?

Utilities

Logging: how do you keep track of what's happening in your system to handle errors?

Feed Sync: how do you handle syncing feeds and finding duplicates?

Post Wrangler: how do you parse different types of feeds/entries into a standard form?

<OPML />

OPML Reader: how do you get the input feeds and parse it, or add new feeds?

Post and Feed: how do you represent things in the database, index and fetch them?

Corpus

Exporter: how do you create a corpus with which to do analytics on?

Feed Sync: how do you handle syncing feeds and finding dupli-cates?

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.