New York Times Article Abstract Analysis using Hadoop and NLTK

By: Lucas Nunno

Part 1: Data Acquisition

Data acquisition

- Used the python **requests** module.
  - Used the offset parameter to load new pages of abstracts and slept 1/8th of a second between each request to abide by the NYT API terms of use.
- Loaded JSON response into python dictionary and then exported as a single large JSON file containing all the articles and all metadata. (**~40,000**)
- In a separate script, I export this JSON data to a CSV file with the docIDs, URLs, and abstracts.
  - This is also where I check for **duplicates**. I have a set of URLs that the exporter has seen, if this URL is in this set the program prints a warning and does not export it.

Part 2: Preprocessing and tf-idf

Preprocessing

- Used the python natural language toolkit (NLTK) module for most of the preprocessing tasks. The algorithm is as follows:
  1. Convert text to lowercase.
  2. Remove punctuation and numbers.
     - Simple regex substitution:

     ```
     remove_pattern = re.compile(r'[,.;&#!\-\'"\(\)\|0-9]')
     ```
  3. Remove stopwords.
     - See: `nltk.corpus.stopwords`
  4. Stem all the remaining words.
     1. See: http://www.nltk.org/api/nltk.stem.html#module-nltk.stem.porter
  5. Output the cleaned abstract.

tf-idf

Part 3: Clustering and Visualization

Clustering

Visualization