**Project URL**: <http://cims.nyu.edu/~kfg217/WSE/>

**Objective**: A search engine that characterizes a journalist based on his/her published articles, displaying key topics the journalist writes about.

**Description**: A search engine featured around (primarily print) journalists. A user can submit a journalist's name to the search engine, which will seek out that journalist. The engine collects information from pre-determined sites at query time, presenting options if multiple journalists are found under that name. The articles under that author are then collected, analyzed and the user received an output page with the journalist's covered key topics and the engine's analysis of the journalist's opinions.

**Architecture**:

1. Client-facing webpage to submit journalist query

2. Search of the journalist at query time from pre-determined sites

3. Collection of queried journalist's published articles

4. Analysis of key topics covered by the queried journalist based on their published articles

5. Present results

**External software**:

- AlchemyAPI (A NLP analysis software)

**Web Resources**

- Directories of contributing writers from major US publicications and media outlets:

- NYT, Los Angeles Times

- External aggregators of journalists:

- Muckrack.com

**System features**:

- Hosting on web services provided by Courant

- Client-side: HTML, CSS

- Server-side: PHP

**Example Results**:

Anderson Cooper (via MuckRack): Failure

Mr. Cooper’s keyword extraction resulted almost exclusively in names. When examining the articles obtained via the news aggregator used for crawling, many of the documents used for keyword extract were transcripts or links to short, low quality articles.

Gary Levin (via MuckRack): Failure

Mr. Levin’s articles primarily concern entertainment media across various mediums, though he also occasionally covers other topics such as healthcare technology. Mr. Levin’s wide range of article topics results in sparse keyword counts that do not properly encapsulate his writing in keyword terms. This issue could be somewhat improved by increasing the sample size of Mr. Levin’s articles.

Dan Mangan (via MuckRack): Mixed

Mr. Mangan has wrote about healthcare related topics such as Obamacare and healthcare services and costs. The resulting analysis for his articles is mixed; while Mr. Mangan’s tone is generally positive about Obamacare, correctly characterized by positive sentiment about the keyword Affordable Care Act, it also assigned negative sentiment to keywords like “open enrollment” and “special enrollment season”. The keyword extractor also picked up some generic words such as “people” and “percent”.

Ben Welsh (via LATimes): Success

Mr. Welsh has written exclusively about the Los Angeles Fire Department (LAFD) since January. The keywords extracted from the document reflect this topic. Mr. Welsh’s characterization of the LAFD has been very critical and this sentiment is correctly reflected in the keyword analysis. The merging of certain terms such as “Los Angeles” and “Los Angeles County” could help reveal other keywords of interest.

Kate Zernike (via New York Times): Success

Ms. Zernike has wrote articles about New Jersey and Gov. Chris Christie for the past 2 months. The keywords obtained from her articles correctly reflect the topics of her writing. Furthermore, the system worked well in categorizing Ms. Zernike’s writing as distinctly negative towards the New Jersey governor and other entities involved. While the system does a good job on picking out the key subjects and on sentiment analysis, it could benefit from merging keywords such as “Mr. Christie” and “Gov. Chris Christie” and possibly “Christie administration” as well.

**Discussion**

**Issues**

A number of issues were uncovered while undertaking this project. These included both technical challenges and issues with resource availability and tools used. One of the primary issues in querying and crawling is determining which articles belong to which author and how it is best to locate those articles. This issue was overcome by utilizing directories from publications when available and limiting selection of a journalist based only from their listing in a single directory. Finding articles based on a directory allows a guarantee that all articles used in the system analysis are representative of the same author and avoids potential issues when two authors share the same name.

Keyword analysis presented a few challenges. One unexpected issue was from the tool itself, which was limited to 1000 transactions per day. Each article requires a transaction for keyword analysis and another for sentiment analysis. Analyzing hundreds of a journalist’s published works was thus not feasible. This was dealt with by keeping a low sample size of 10 articles per journalist query at maximum, however the quality of topic ranking based on keywords suffered from the small sample size.

Additionally, many like terms were emitted as unique keywords, so that two keywords representing the same concept were not joined. Deciding the strictness of the keyword extractor and choices such as if n-gram keywords should be reduced to uni-grams were interesting and important factors in keyword extraction quality that were not initially considered.

Originally the project hoped to move forward from analyzing a journalist not only based on keywords but also on a larger political spectrum based on issues that those keywords represented. Due to the limited sample size per author and lack of a general classifier, however, this effort proved too ambitious.

**Possible Improvements**

Though issuing the crawl and the time of query was initially appealing, it seems like it would be best to attempt a more comprehensive crawl, either by a more involved web crawl or by collecting articles from certain endpoints over a long period of time and storing this into a database. Additionally, a means of resolving the same articles hosted on different sites would remove the bias of having keywords mentioned in those articles over-weighed.

The quality of the analysis was greatly hampered by the small sample size of articles for each queried journalist. Though the AlchemyAPI NLP tool by IBM provided good keyword results, it may have been more beneficial to have a weaker keyword extractor which operated over a much larger sample size. In the previously mentioned example of Dan Mangan, increasing his sample size to 25 articles made more representative healthcare terms, such as “Obamacare” take a higher precedence.

Reducing n-gram keywords to uni-grams was briefly experimented with some success in obtaining more representative keyword topics. This gain, however, was at the cost of properly represented sentiment analysis. Experimenting with techniques to reduce n-gram keywords while maintaining a rough estimate of sentiment analysis on those keywords would result in improvements to the system.

In terms of achieving the additional goals originally set out, it might be best if the search engine was structured based on a pre-compiled set of journalists instead of an open query. This might allow us an easier method of training a classifier for higher order concepts based on extracted keywords.

Ultimately this project encountered a number of non-trivial natural language processing hurdles. A more robust crawler that was able to have some resolution of the articles to the correct author could obtain a much more representative sample of a journalist’s writings, particularly now that online journalists often publish on multiple platforms. Ranking of topics could greatly benefit from a larger sample of articles and other potential techniques to resolve similar keywords. Finally, adding a means of performing semantic analysis on keywords could allow for synthesis of more general topics and sentiment towards those topics, greatly improving the usefulness of this tool in analyzing a journalist’s agenda. Improving the system at any of these points could increase the utility of the system.