600.466 Information Retrieval & Web Agent

Cross-Lingual Newspaper Analysis

Jin Yong Shin (jshin44)

Ji Won Shin (jshin49)

Abstract

After the South Korean leader Park Geun-Hye became country's first president to be ousted by impeachment, the citizens of Korea received a privilege to hold a new presidential election. The scandal gave the biggest impact on Korea's political history. My partner and I also began to pay close attention to the political and international issues. Beside the intern scandals and political conflicts, Korea has been suffered from international issues including North Korea's consistent provocations, China's economy boycott protests on Korea over THAAD missile, and Korea's political relationship with U.S. after Donald Trump won the election.

After the discussion, we came up with a clever idea to analyze and compare Korean and U.S. medias' viewpoints and standpoints on similar incidents and topics. To accomplish our objectives, we have implemented a newspaper article analyzer that works similar to the vector models used in our HW 2 assignment by employing techniques that we have learned from the course.

The techniques and algorithms that we have implemented in this program are including but not limited to:

1. Automatic Web Agent to retrieve articles from the given domains, which are:
    a. Chicago Tribune
    b. Washington Post Times
    c. Han Post (Korean Newspaper)
2. Query engine to interactively retrieve relevant articles
3. Sentiment Analysis
4. Keyword Analysis
5. Cross-lingual Comparison and Search

All the files are included in our cs466_finalproject_jshin44jshin49.tar.gz file

To run the search engine please type

```
$python3 Analysis.py
```

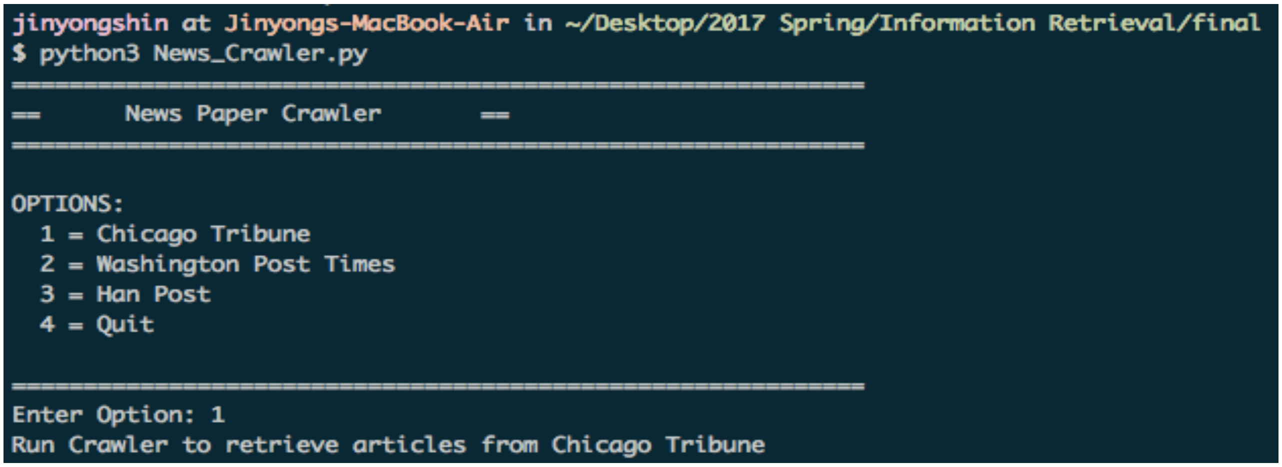
1. **Overall Implementation**

1.1. Newspaper Retrieval Web Agent

We extended web agent from what we learned from HW 4 to retrieve article titles and texts from the newspaper in our domain. The basic web retrieval that we used is BeautifulSoup API from bs4 library and added crawling strategies to visit relevant articles by Breadth First Search like algorithm.

Our newspaper agent web crawler will traverse the first input page and look for 'href's for the next iteration. To avoid self-referencing and non-local web pages, we implemented similar techniques from HW4. Additionally, we have added parse function to avoid non-article pages such as pages with lists of other article pages and non-relevant news articles. To find relevant articles, we added helper function to calculate relevance scores by parsing full context by using regex and key words within http addresses. Therefore, all the relevant articles are pushed to the top of the list for iteration and the web crawler will recursively traverse high relevant web pages to extract titles and body text.

All of the retrieved data are stored in resources directory and you can run our crawler to collect data by running **$python3 News_Crawler.py** and interact with the program to choose your choice of domain.



```
jinyongshin at Jinyongs-MacBook-Air in ~/Desktop/2017 Spring/Information Retrieval/final
$ python3 News_Crawler.py
============================================================
==      News Paper Crawler       ==
============================================================

OPTIONS:
  1 = Chicago Tribune
  2 = Washington Post Times
  3 = Han Post
  4 = Quit

============================================================
Enter Option: 1
Run Crawler to retrieve articles from Chicago Tribune
```

1.2. Keyword Retrieval by Normalized Frequency Score

One critical part of our implementation is collecting top 20 keywords from the article titles and body-text. We used the scoring scheme based on words frequency model and then normalize the frequency by dividing the length of entire article body-text length. Then we compute frequency score by multiplying weight to find the top 20 key words from article titles and body-text.

1.3. Sentiment Analysis

To view how U.S. and Korean media analyze similar topic respectively, we also implemented sentiment analysis algorithm. The basic algorithm that we used in our implementation is iterating the title and body-text to give positive (+1) and negative (-1) value to the words by identifying from our pre-determined positive & negative word lists. Additionally, we also empowered the sentiment in particular words by searching incrementing and decrementing words in previous or post (distance by 1) position. Finally, we also reversed the sentiment score if each word has inverter in previous or post position.

In the Cross Lingual Information Retrieval analysis, collection of these sentiment results were used to calculate the sentiment score for the query.

1.4. Term Clustering Methods / Query Expansion

As mentioned in the lecture, stemming is the most basic and simplest term clustering methods. We developed Python programs "stemmer_en.py" and "stemmer_kr.py" We used koNLPy, a Python package for natural language processing of the Korean language, library to tokenize Korean words. We used NLTK, a Python natural language toolkit package, SnowballStemmer class to stem English words. As our main data source was news articles, words were represented in various forms and we needed to cluster these various forms into one group depending on their meaning.

We also utilized query expansion to increase recall. By using a preexisting thesaurus, we fleshed out sparse queries with related words. We used PyDictionary package to implement query expansion.

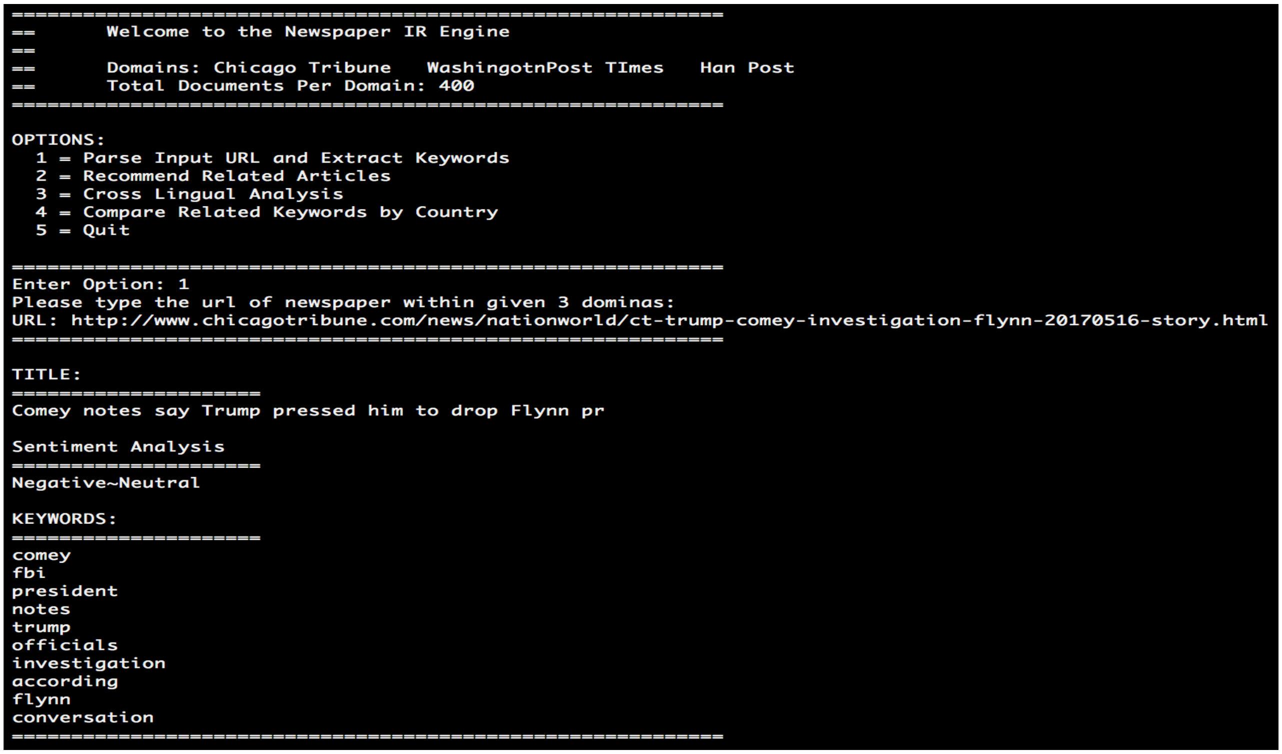
1.5. Cross Lingual Information Retrieval

We took the approach of CLIR where we translate all the data into English beforehand and perform retrieval in the English language. We developed a Python program "kor_to_eng.py" to translate from Korean to English. In doing so, we used a Korean to English dictionary contained in the file "Korean2.tsv.stemmed." We stemmed each English word in the dictionary using the same stemmer mentioned above to maintain consistency. We used the regex matching function to translate the stemmed HanPost file into English. With the translated HanPost file, we performed Cross Lingual Information Retrieval and various sentiment and associated word comparisons.

2. Search Engine

2.1. Interactive Keyword Query

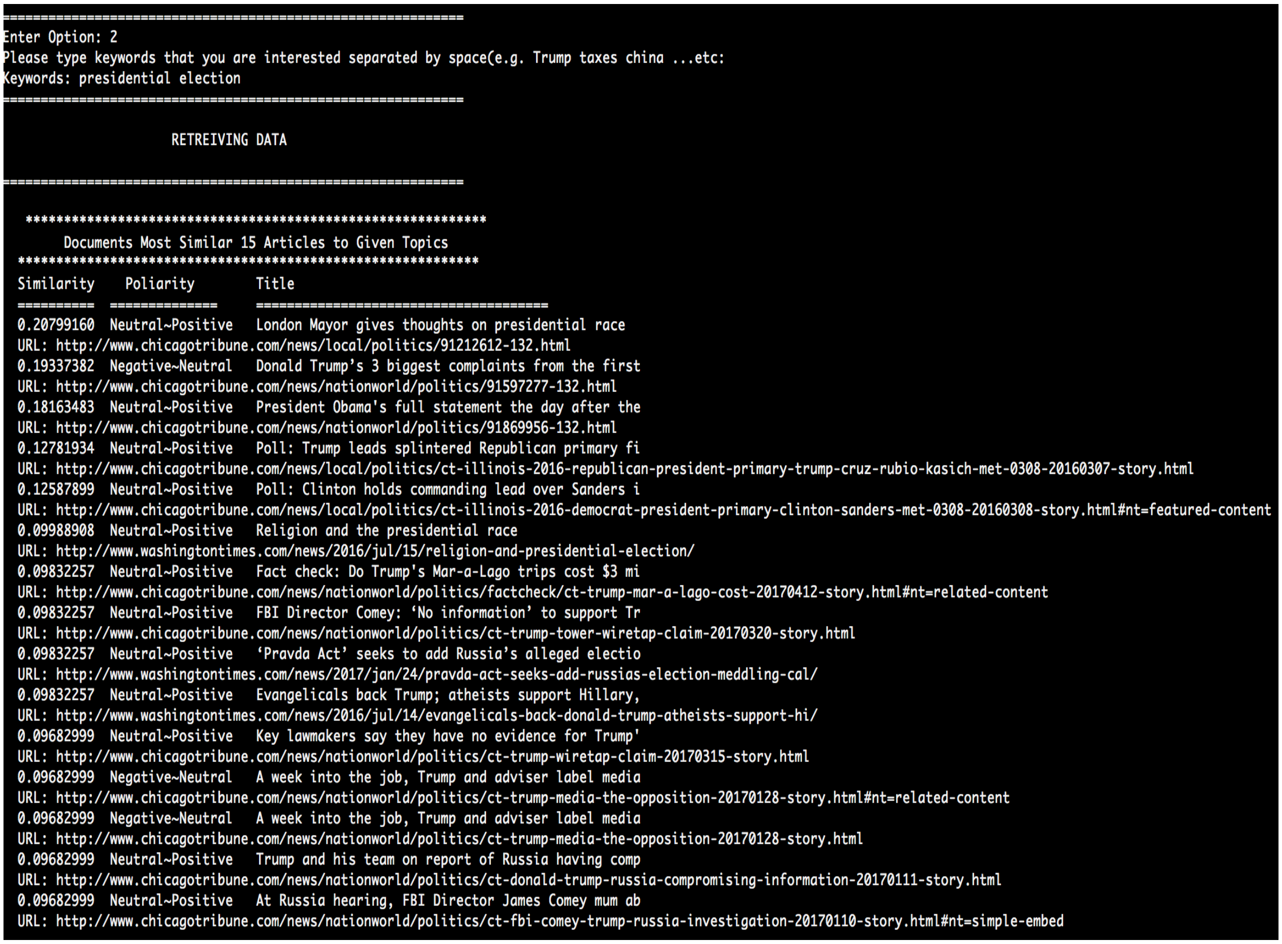
The first option of our search engine is interactive keyword query with given URL. If particular article is not in our database and user wants to see the analysis of the recent article chosen by user, we provide option to interact with our search engine to compute the result ad-hoc basis. This is just similar work that has done in articles in our database. We wanted to extend the usability of our analysis tool in real time and ad-hoc basis.



```
============================================================
==      Welcome to the Newspaper IR Engine
==
==      Domains: Chicago Tribune   WashingotnPost TImes   Han Post
==      Total Documents Per Domain: 400
============================================================

OPTIONS:
  1 = Parse Input URL and Extract Keywords
  2 = Recommend Related Articles
  3 = Cross Lingual Analysis
  4 = Compare Related Keywords by Country
  5 = Quit

============================================================
Enter Option: 1
Please type the url of newspaper within given 3 dominas:
URL: http://www.chicagotribune.com/news/nationworld/ct-trump-comey-investigation-flynn-20170516-story.html
============================================================

TITLE:
=====================
Comey notes say Trump pressed him to drop Flynn pr

Sentiment Analysis
=====================
Negative~Neutral

KEYWORDS:
=====================
comey
fbi
president
notes
trump
officials
investigation
according
flynn
conversation
============================================================
```

2.2. Article Recommendation System

In second option, user can now interact with the search engine to find the most relevant articles by providing query in set of keywords. Similar to the techniques in our HW2 to find the relevant document in each query, we implemented interactive version of recommendation system given user's query to find relevant articles in our database. As explained, we employed PyDictionary API to find thesaurus-like synonyms to do query expansion and assign smaller weights in the synonym words. Then we used similar scheme as HW 2 to compute cosine similarity and retrieve top 15 relevant articles from our database.

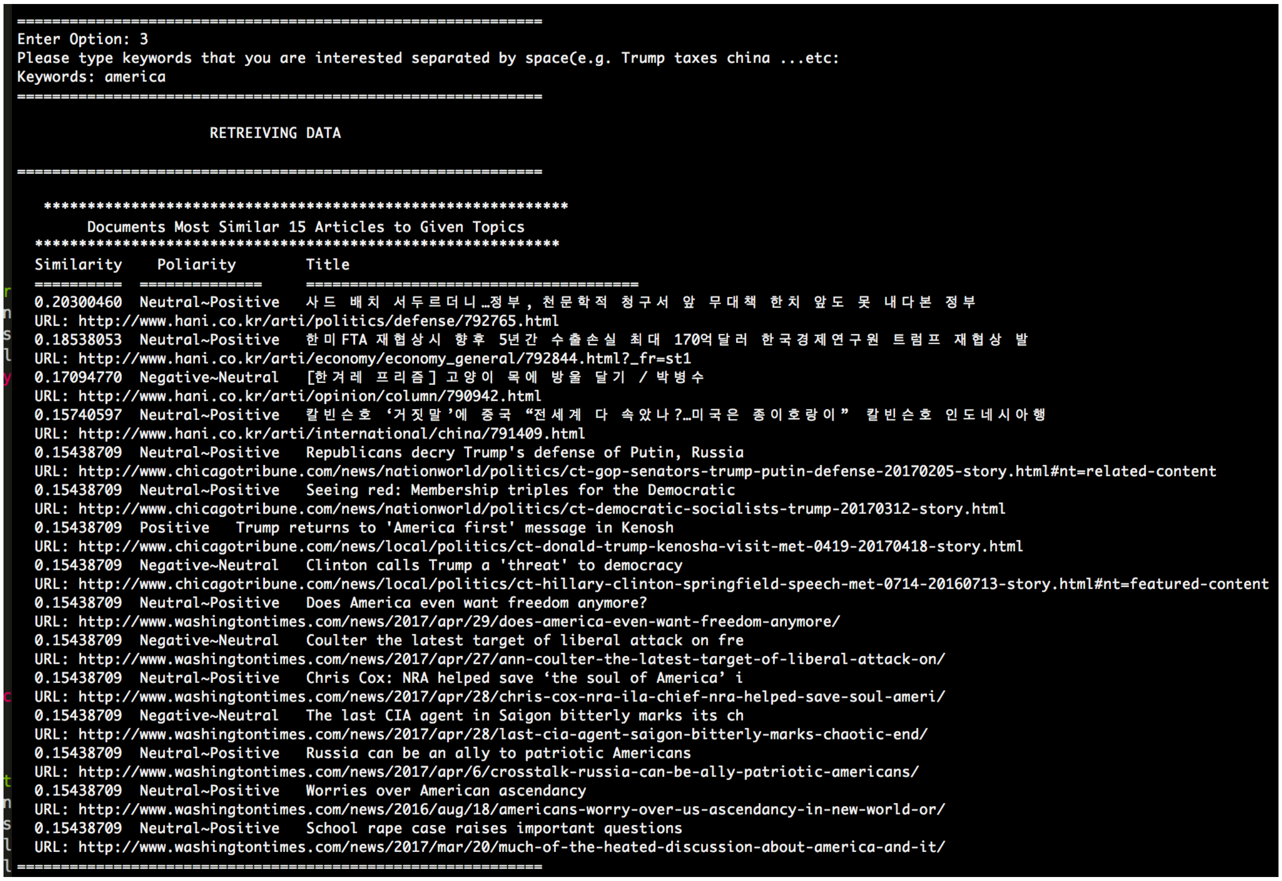


```
============================================================
Enter Option: 2
Please type keywords that you are interested separated by space(e.g. Trump taxes china ...etc:
Keywords: presidential election
============================================================

                      RETREIVING DATA

============================================================

   ************************************************************
        Documents Most Similar 15 Articles to Given Topics
  ************************************************************
  Similarity    Poliarity       Title
  ==========  ==============    ======================================
  0.20799160  Neutral~Positive   London Mayor gives thoughts on presidential race
  URL: http://www.chicagotribune.com/news/local/politics/91212612-132.html
  0.19337382  Negative~Neutral   Donald Trump’s 3 biggest complaints from the first
  URL: http://www.chicagotribune.com/news/nationworld/politics/91597277-132.html
  0.18163483  Neutral~Positive   President Obama's full statement the day after the
  URL: http://www.chicagotribune.com/news/nationworld/politics/91869956-132.html
  0.12781934  Neutral~Positive   Poll: Trump leads splintered Republican primary fi
  URL: http://www.chicagotribune.com/news/local/politics/ct-illinois-2016-republican-president-primary-trump-cruz-rubio-kasich-met-0308-20160307-story.html
  0.12587899  Neutral~Positive   Poll: Clinton holds commanding lead over Sanders i
  URL: http://www.chicagotribune.com/news/local/politics/ct-illinois-2016-democrat-president-primary-clinton-sanders-met-0308-20160308-story.html#nt=featured-content
  0.09988908  Neutral~Positive   Religion and the presidential race
  URL: http://www.washingtontimes.com/news/2016/jul/15/religion-and-presidential-election/
  0.09832257  Neutral~Positive   Fact check: Do Trump's Mar-a-Lago trips cost $3 mi
  URL: http://www.chicagotribune.com/news/nationworld/politics/factcheck/ct-trump-mar-a-lago-cost-20170412-story.html#nt=related-content
  0.09832257  Neutral~Positive   FBI Director Comey: ‘No information’ to support Tr
  URL: http://www.chicagotribune.com/news/nationworld/politics/ct-trump-tower-wiretap-claim-20170320-story.html
  0.09832257  Neutral~Positive   ‘Pravda Act’ seeks to add Russia’s alleged electio
  URL: http://www.washingtontimes.com/news/2017/jan/24/pravda-act-seeks-add-russias-election-meddling-cal/
  0.09832257  Neutral~Positive   Evangelicals back Trump; atheists support Hillary,
  URL: http://www.washingtontimes.com/news/2016/jul/14/evangelicals-back-donald-trump-atheists-support-hi/
  0.09682999  Neutral~Positive   Key lawmakers say they have no evidence for Trump'
  URL: http://www.chicagotribune.com/news/nationworld/politics/ct-trump-wiretap-claim-20170315-story.html
  0.09682999  Negative~Neutral   A week into the job, Trump and adviser label media
  URL: http://www.chicagotribune.com/news/nationworld/politics/ct-trump-media-the-opposition-20170128-story.html#nt=related-content
  0.09682999  Negative~Neutral   A week into the job, Trump and adviser label media
  URL: http://www.chicagotribune.com/news/nationworld/politics/ct-trump-media-the-opposition-20170128-story.html
  0.09682999  Neutral~Positive   Trump and his team on report of Russia having comp
  URL: http://www.chicagotribune.com/news/nationworld/politics/ct-donald-trump-russia-compromising-information-20170111-story.html
  0.09682999  Neutral~Positive   At Russia hearing, FBI Director James Comey mum ab
  URL: http://www.chicagotribune.com/news/nationworld/politics/ct-fbi-comey-trump-russia-investigation-20170110-story.html#nt=simple-embed
```

2.3. Cross Lingual Article Retrieval

This is an add-on to the implementation introduced in 2.2. This implement makes use of the cross lingual information retrieval of a given query. In addition to the American newspaper articles, we included the translated articles from HanPost, a popular news media company based in Seoul. We could see what kind of news articles media companies based in different countries write regarding the same topic.

The example below shoes the retrieval results for the query “america.”



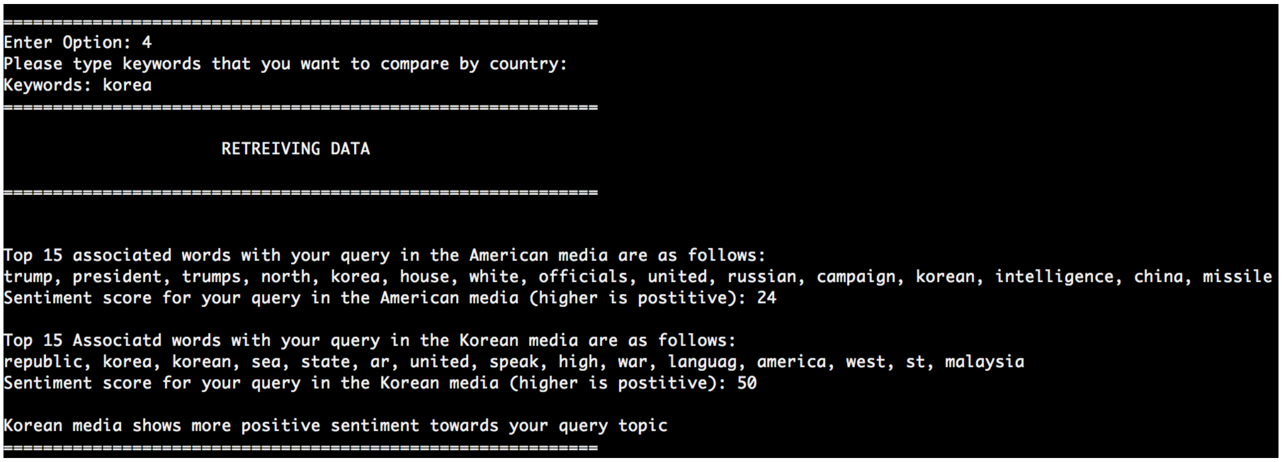
```
============================================================
Enter Option: 3
Please type keywords that you are interested separated by space(e.g. Trump taxes china ...etc:
Keywords: america
============================================================

                      RETREIVING DATA

============================================================

   ************************************************************
        Documents Most Similar 15 Articles to Given Topics
  ************************************************************
  Similarity    Poliarity       Title
  ==========  ==============     ======================================
  0.20300460  Neutral~Positive   사드 배치 서두르더니…정부, 천문학적 청구서 앞 무대책 한치 앞도 못 내다본 정부
  URL: http://www.hani.co.kr/arti/politics/defense/792765.html
  0.18538053  Neutral~Positive   한미FTA 재협상시 향후 5년간 수출손실 최대 170억달러 한국경제연구원 트럼프 재협상 발
  URL: http://www.hani.co.kr/arti/economy/economy_general/792844.html?_fr=st1
  0.17094770  Negative~Neutral   [한겨레 프리즘] 고양이 목에 방울 달기 / 박병수
  URL: http://www.hani.co.kr/arti/opinion/column/790942.html
  0.15740597  Neutral~Positive   칼빈슨호 ‘거짓말’에 중국 “전세계 다 속았나?…미국은 종이호랑이” 칼빈슨호 인도네시아행
  URL: http://www.hani.co.kr/arti/international/china/791409.html
  0.15438709  Neutral~Positive   Republicans decry Trump's defense of Putin, Russia
  URL: http://www.chicagotribune.com/news/nationworld/politics/ct-gop-senators-trump-putin-defense-20170205-story.html#nt=related-content
  0.15438709  Neutral~Positive   Seeing red: Membership triples for the Democratic
  URL: http://www.chicagotribune.com/news/nationworld/politics/ct-democratic-socialists-trump-20170312-story.html
  0.15438709  Positive   Trump returns to 'America first' message in Kenosh
  URL: http://www.chicagotribune.com/news/local/politics/ct-donald-trump-kenosha-visit-met-0419-20170418-story.html
  0.15438709  Negative~Neutral   Clinton calls Trump a 'threat' to democracy
  URL: http://www.chicagotribune.com/news/local/politics/ct-hillary-clinton-springfield-speech-met-0714-20160713-story.html#nt=featured-content
  0.15438709  Neutral~Positive   Does America even want freedom anymore?
  URL: http://www.washingtontimes.com/news/2017/apr/29/does-america-even-want-freedom-anymore/
  0.15438709  Negative~Neutral   Coulter the latest target of liberal attack on fre
  URL: http://www.washingtontimes.com/news/2017/apr/27/ann-coulter-the-latest-target-of-liberal-attack-on/
  0.15438709  Neutral~Positive   Chris Cox: NRA helped save ‘the soul of America’ i
  URL: http://www.washingtontimes.com/news/2017/apr/28/chris-cox-nra-ila-chief-nra-helped-save-soul-ameri/
  0.15438709  Negative~Neutral   The last CIA agent in Saigon bitterly marks its ch
  URL: http://www.washingtontimes.com/news/2017/apr/28/last-cia-agent-saigon-bitterly-marks-chaotic-end/
  0.15438709  Neutral~Positive   Russia can be an ally to patriotic Americans
  URL: http://www.washingtontimes.com/news/2017/apr/6/crosstalk-russia-can-be-ally-patriotic-americans/
  0.15438709  Neutral~Positive   Worries over American ascendancy
  URL: http://www.washingtontimes.com/news/2016/aug/18/americans-worry-over-us-ascendancy-in-new-world-or/
  0.15438709  Neutral~Positive   School rape case raises important questions
  URL: http://www.washingtontimes.com/news/2017/mar/20/much-of-the-heated-discussion-about-america-and-it/
============================================================
```

It appears that for the same given query, media companies from different countries provided articles with different focus and perspectives. Further analysis on this comparison is performed in section 2.4.

2.4. Compare Related Keywords by Country

This implementation offers a deeper cross lingual analysis of the given query. We compared associated keywords and sentiment score of the same query between countries. We could observe a clear difference of focus of media depending on the country.

The example below shows the results for the query “korea.” The difference in results is very evident for this input query.



```
============================================================
Enter Option: 4
Please type keywords that you want to compare by country:
Keywords: korea
============================================================

                      RETREIVING DATA

============================================================


Top 15 associated words with your query in the American media are as follows:
trump, president, trumps, north, korea, house, white, officials, united, russian, campaign, korean, intelligence, china, missile
Sentiment score for your query in the American media (higher is postive): 24

Top 15 Associatd words with your query in the Korean media are as follows:
republic, korea, korean, sea, state, ar, united, speak, high, war, languag, america, west, st, malaysia
Sentiment score for your query in the Korean media (higher is postive): 50

Korean media shows more positive sentiment towards your query topic
============================================================
```

The American media associated the given query with keywords such as “north,” “korea,” “missile,” and “russia.” American news media focused more on North Korea and its dictatorship and threats.

On the other hand, the Korean media associated the query with more generic terms such as “republic,” “language,” “united,” “state,” and “america.” Korean News focused more on South Korea and affiliated the query with pro-American culture.

Sentiment score for the query “korea” that the American news media attributed was 24, whereas the Korean media gave the score of 50. As the higher sentiment score indicates, we can clearly see that the Korean media has a much more positive sentiment toward the query “korea,” than its American counterparts do, which are more focused more on North Korea and wars.

3. Limitations / Improvements

3.1. More Robust Translation Methods

Instead of using preexisting translation API's, we developed our own translation methods which used the python regex function to match Korean with English. Our dictionary lacks some proper nouns, acronyms, and neologisms. As news articles could be about a specific person and the user could have searched for a specific neologism, a more robust translation methods could have greatly increased the functionality of the cross lingual information retrieval.