분산 처리 환경에서의 기계학습 기반의 뉴스 기사 빅 데이터 분석

*오희빈, 이정철, 김경섭
충남대학교 컴퓨터공학과
e-mail : gmlqls111@naver.com, tree4553@naver.com, sclkim@cnu.ac.kr

News Article Big Data Analysis based on Machine Learning in Distributed Processing Environments

*Corresponding author : Hee-bin Oh, Jeong-cheol Lee, Kyungsup Kim
Dept. of Computer Engineering, Chung-Nam National University

요 약

본 논문에서는 텍스트 형태의 빅 데이터를 분산처리 환경에서 기계학습을 이용하여 분석하고 유의미한 데이터를 만들어내는 시스템에 대해 다루었다. 빅 데이터의 한 종류인 뉴스 기사 빅 데이터를 분산 시스템 환경(Spark) 내에서 기계 학습(Word2Vec)을 이용하여 뉴스 기사의 키워드 간의 연관도를 분석하는 분산 처리 시스템을 설계 및 구현하였고, 사용자가 입력한 검색어와 연관된 키워드들을 한눈에 파악하기 쉽게 만드는 시각화 시스템을 설계하였다.

1. 서론

정보통신 기술의 발전으로 인해 인터넷에는 매우 많은 양의 데이터들이 쏟아져 나온다. 하루에 인터넷에 게재되는 인터넷 기사의 수는 셀 수가 없을 정도이다. 정보 과도로 인하여 구독자가 얻고자 하는 정보를 정확히 찾는 것은 큰 노력이 필요하고 세간의 전체적인 동향을 살피는 것 또한 쉽지 않은 일이다.

본 논문에서는 이러한 구독자의 불편한 사항들을 해결하기 위해서 '분산 처리 환경에서의 기계학습 기반의 뉴스 기사 빅 데이터 분석' 연구를 통해서 사용자가 검색하고자 하는 키워드와 연관된 다른 키워드를 찾아 시각화하여 인터넷 뉴스의 전체적인 동향을 한눈에 파악하기 쉽게 하는 방안을 모색한다.

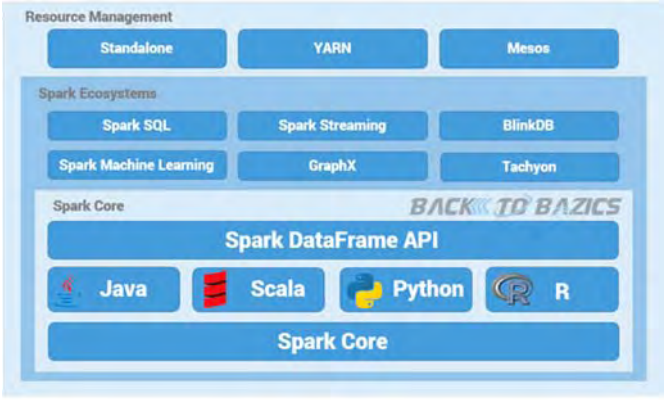
많은 양의 인터넷 뉴스 기사들을 기계학습 시키기 위해서 Hadoop 과 Spark 를 이용한 분산 처리 시스템을 구축하고 뉴스 기사 빅 데이터를 분산 처리 시스템을 이용하여 분석을 진행한다.

분석을 통해 얻어낸 데이터를 사용자가 검색한 키워드와 연관 키워드들로 시각화하여 한눈에 보기 쉽게 하고 인터넷 뉴스 기사들의 전체적인 동향을 파악할 수 있도록 하는 것이 논문에서 다루는 시스템의 목적이다.

논문의 다음과 같은 시나리오로 진행된다. 뉴스 기사를 크롤링하여 기사의 제목, 내용을 시스템에 저장하고 기사의 제목, URL, 날짜를 데이터베이스에 저장한다. 기사의 내용은 Konlpy 를 이용하여 명사화하고 명사화한 기사 내용을 TF-IDF 를 이용하여 불필요하지만, 많이 쓰이는 단어를 걸러낸다. 가공을 거친 데이터를 Word2Vec 에서 사용할 수 있는 txt 파일 형태로 만들어준다. 마지막으로 분산시스템인 Spark 의 Mllib 인 Word2Vec 을 통해 키워드와 연관 키워드의 관계를 분석하고 생성된 Model 을 이용하여 시각화하여 사용자에게 보여준다. 자세한 내용과 과정은 논문의 뒷부분에서 다루겠다.

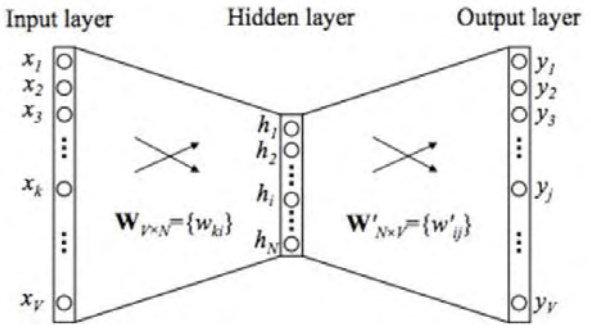
2. 관련 연구 - 텍스트 마이닝

Apache Spark[1]는 그림 1 과 같은 구조로 메모리상에서 동작하는 확장성이 뛰어난 분산 시스템으로, Java, Scala, Python, R 과 같은 여러 언어를 사용하여 Application 을 개발할 수 있는 기능을 제공한다. Mahout 과 같은 Apache 시스템은 이제 MapReduce 를 대신하여 프로세싱 엔진으로 자리매김하고 있으며, Spark Application 은 Hive 를 사용할 수 있으므로 Hive 와 직접 데이터를 주고받을 수 있다.



[그림 1: spark 구조]

구글에서 만들어진 Word2Vec[2]은 word embedding 과 관련된 학습 모델이며, 그림 2 와 같이 Word2Vec 모델은 신경망 구조를 가지며, Deep Learning 이 아닌 Shallow, two layer 구조로 되어 있다. 이는 언어적인 단어 문단을 재구성하고 학습하는데 용이하다.
Word2Vec 은 input 으로 큰 텍스트를 받아 수백 차원의 벡터 공간을 만들어 낸다. 만들어진 벡터 공간에는 단어에 따른 기준 단어와 연관된 단어들이 거리에 따라 분류된다.



[그림 2: Word2Vec 구조]

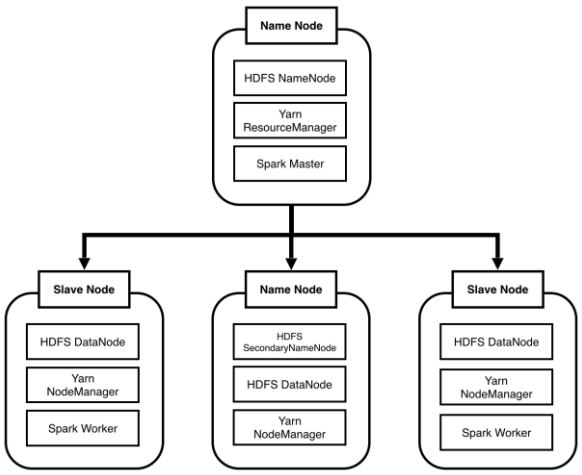
Konlpy[3]는 한국어 정보처리를 위한 Python 패키지이며, 한국어 문장을 입력받아 이를 원하는 형태로 가공할 수 있다. 명사 추출, 형태소 분리 등 여러 가지 기능들을 이용하여 한국어 문장의 형태를 가공할 수 있다.

TF-IDF[4]는 정보 검색과 텍스트 마이닝에서 이용하는 가중치로, TF 는 특정한 단어가 문서 내에 얼마나 자주 등장하는지를 나타내는 값이며, IDF 는 특정한 단어가 몇 개의 문서 안에서 쓰였는지에 대한 DF 의 역수이다. TF-IDF 는 TF 와 IDF 를 곱한 값이다. 여러 문서로 이루어진 문서 군이 있을 때 어떤 단어가 특정 문서 내에서 얼마나 중요한 것인지를 나타내는 통계적 수치이다. 문서의 핵심어를 추출하거나, 검색 엔진에서 검색 결과의 순위를 결정하거나, 문서들 사이의 비슷한 정도를 구하는 등의 용도로 사용할 수 있다.

3. 빅데이터 분석 시스템 설계 및 구현

3.1 분산 시스템 구성

대량의 인터넷 뉴스 기사를 분석하고 학습하기 위해서 분산처리 환경을 Hadoop 과 Spark 를 이용하여 구축했다. 대량의 인터넷 뉴스 기사들을 처리하기 위해서 HDFS 를 기반으로 YARN 을 통해 자원 관리와 스케줄링을 실행했다. 이를 위해 그림 3 과 같이 분산 시스템을 1 대의 NameNode 와 그 NameNode 를 포함한 3 대의 DataNode 를 설정하고 Master Node 에는 추가로 Yarn ResourceManager, SecondaryNameNode, Spark Master 그리고, Yarn NodeManager 를 설정해 준다. 그리고 2 대의 Slave Node 에는 추가로 Spark Slave, Yarn NodeManager 를 설정해서 분산 시스템 환경을 구축했다.



[그림 3: 분산 시스템 구성]

3.2 크롤러

크롤러는 Python 을 이용하여 구현했으며, 네이버 뉴스 웹페이지에서 뉴스 기사를 수집했다. 뉴스 기사를 크롤링하는 과정은 네이버 뉴스 웹페이지의 속보: 정치 부분의 URL 을 확인하여 URL 의 날짜를 입력하는 부분을 파라미터로 받게 한다. 그 후에 파라미터로 날짜를 입력하면 네이버 뉴스 웹페이지의 첫 번째 페이지부터 마지막 페이지까지 순서대로 URL 을 HTTP Request 를 통해 기사 원문을 서버로부터 받는다.

3.3 TF-IDF 단어 가중치 기법

뉴스 기사를 크롤링하는 과정에 의미 없이 자주 사용되는 단어들이 있다. 예를 들면 '본문', '내용', '바로 가기' 등등이 있는데 Word2Vec 의 머신 러닝 과정에 이러한 단어가 포함되게 되면 키워드의 연관 검색어로 등록돼버린다. 이를 막기 위해 TF-IDF 단어 가중치 기법을 이용한다. TF-IDF 값은 단어 빈도와 역문서 빈도의 곱으로 의미 없이 자주 쓰이는 단어는 가중치 값이 작다. 이를 이용하여 기준값을 설정하고 기준값보다 작은 값의 단어를 제거하였다.

3.4 Word2Vec

대량의 기사 내용을 이용해서 뉴스 기사들의 키워드와 연관된 키워드를 찾기 위해 분산 시스템인 Spark 의 Mllib(Machine Learning Library)인 Word2Vec 을 이용하고 프로그래밍 언어는 Scala 를 선택하였다. Python 에서도 Word2Vec 을 제공하지만 분산 처리 환경에서 뉴스 기사 빅데이터를 기계 학습시키기 위해서 Spark 의 Mllib 을 이용하였다. Word2Vec 을 사용하기 위해 Konlpy 와 TF-IDF 를 이용해 명사화와 불필요한 단어를 제거한 기사들을 하나의 txt 파일에 모두 합쳐서 저장한다. 생성된 txt 파일을 HDFS 로 저장시켜 Spark 에서 이용할 수 있게 한다. 그 후 Word2Vec 을 통해 각 키워드의 연관도에 대해 기계학습된 결과값인 Model 이 생성된다. 이 Model 을 통해 키워드와 다른 키워드 간의 연관도를 파악할 수 있다.

3.5 시나리오

본 논문에서 구현된 시스템의 전체 시나리오는 그림 4 와 같이 이루어져 있다.

1) bs4 라이브러리를 이용하여 네이버 뉴스 기사를 크롤링한다.

2) 크롤링한 데이터 중 뉴스 제목, URL, 날짜를 뉴스 기사 데이터베이스에 저장한다.

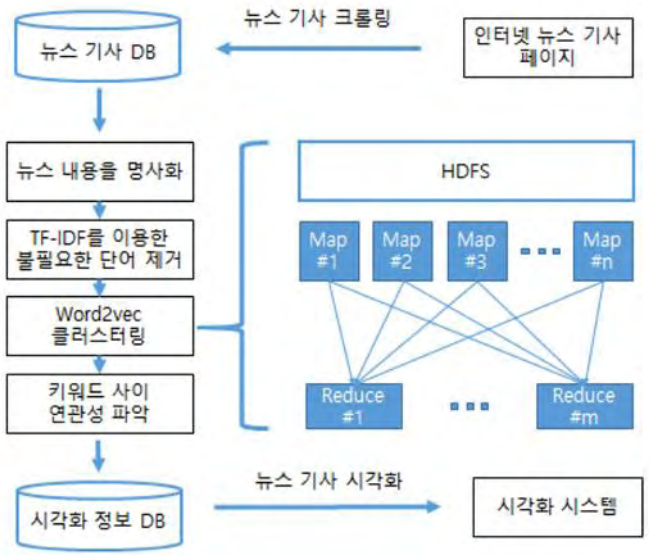
3) 크롤링한 데이터 중 뉴스 기사의 내용을 Konlpy 라이브러리 중 twitter 를 이용하여 명사화하여 날짜별로 기사를 저장한다.

4) 저장한 뉴스 기사 내용을 TF-IDF 알고리즘을 이용해 기사에 불필요한 명사들을 제거하고 Word2Vec 에 이용할 수 있게 하나의 txt 파일로 저장한다.

5) 생성된 txt 파일을 Spark 분산 시스템을 이용해서 Mllib 인 Word2Vec 알고리즘을 통해 각각의 키워드와 연관된 키워드의 Model 을 만들어 낸다.

6) 생성된 Model 에 저장된 키워드 간의 연관도를 시각화 정보 데이터베이스에 저장한다.

7) 시각화 정보 데이터베이스에 저장된 내용을 시각화 시스템을 통해 사용자가 이용할 수 있게 만들어 준다.



[그림 4: 시나리오 구성도]

4. 실험 결과

4.1 크롤링 결과

그림 5 와 같이 기사의 제목과 날짜, URL 을 DataBase 에 저장하고, 그림 6 과 같이 각각 기사 내용을 Konlpy 를 이용해서 명사를 추출하여 txt 파일로 그림 7 과 같이 하나하나 저장하게 해서 기사를 수집했다.

```
| 4402 | http://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=100&oid=081&aid=0002848900
| ['이재용측 "사실 오인" 항소장… 2심은 '묵시적 청탁' 법리전쟁']

                    | 20170829 |
| 4403 | http://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=100&oid=023&aid=0003308795
| ['사드, 내달초 추가배치 끝내기로']

                    | 20170829 |
| 4404 | http://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=100&oid=023&aid=0003308791
| ['北이 쏜 단거리 발사체는 방사포 아닌 탄도미사일']

                    | 20170829 |
| 4405 | http://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=100&oid=023&aid=0003308789
| ['결론형 전자담배 때문에… 한국당 내부 혼선']

                    | 20170829 |
```

[그림 5: DB 에 저장된 데이터]

```
article_20170829_2323.txt  article_20170829_3668.txt  article_20170829_980.txt
article_20170829_2324.txt  article_20170829_3669.txt  article_20170829_981.txt
article_20170829_2325.txt  article_20170829_366.txt   article_20170829_982.txt
article_20170829_2326.txt  article_20170829_3670.txt  article_20170829_983.txt
article_20170829_2327.txt  article_20170829_3671.txt  article_20170829_984.txt
article_20170829_2328.txt  article_20170829_3672.txt  article_20170829_985.txt
article_20170829_2329.txt  article_20170829_3673.txt  article_20170829_986.txt
```

[그림 6: 수집된 기사의 형태]

```
1 본문 내용 플레이어 플레이어 오류 우회 함수 추가 서울 뉴시스 추상 기자 이낙>
  연 국무총리 오후 서울 강동구 강동 경희대 병원 위안부 피해자 상숙 할머니 빈소
  조문 서울 뉴시스 이종희 기자 이낙연 국무총리 위안부 피해자 상숙 할머니 빈소
  총리 서울 강동구 강동 경희대학교 병원 상숙 할머니 빈소 조문 총리 세종시 일>
  정 마치 서울 장례식 총리 조문 과정 유족 고향 중국 국적 고인 해외동포 위해 조
  성 천안시 국립 망향 동산 이야기 총리 즉시 현장 박능후 보건복지부 장관 상숙 >
  할머니 망향 동산 안치 조치 발인 일인 절차 관심 달라 당부 총리 조문 마치 자신
  소셜네트워크서비스 할머니 중국 중국인 결혼 남편 여의고 막내딸 중국 할머니 >
  한국 정부 지원 치료 작년 귀국 병원 어제 별세 세상 아픔 할머니 지난 오전 별세
  발인 오전 할머니 별세 국내 등록 명의 일본군 위안부 피해자 할머니 생존자 뉴>
  시스 빅데이터 주가 시세 바로가기 뉴시스 페이스북 트위터 본문 내용
```

[그림 7: 수집한 기사]

4.2 TF-IDF 결과

TF-IDF 공식을 이용하여 직접 알고리즘을 구현하고 그 알고리즘을 통해 그림 8 과 같이 각각의 단어의 TF 와 IDF 를 구한다. 이를 이용하여 TF-IDF 의 기준을 설정하여 의미 없이 자주 사용되는 단어를 걸러낼 수 있고, 이 TF-IDF 를 이용해서 제거할 단어를 선택해서 그림 9 와 같이 단어를 제거하여 모든 뉴스 기사들을 하나의 txt 파일로 합쳐 Word2Vec 에서 사용할 수 있게 만들어준다.

```
배포  4310 4279 1.1111778242624633
본문  17291 8646 0.0010640416537256467
무단  4373 4237 1.1287009950740077
전재  3850 3847 1.2619215371828025
우회  8677 8646 0.0009888616333633294
추가  9925 8646 0.0010035133604642732
여러분  2534 2509 1.829861629017969
제보  5343 2772 1.8425840544636614
기자  10336 7373 0.27868760969047823
내용  18574 8646 0.0010718461483837 96
국회  5271 3202 1.6065720142818478
서울  8491 4674 1.0505292758610834
뉴스  6772 4075 1.2524305807882057
오류  8667 8646 0.0009887359063638723
오전  3451 2681 1.8000624983346545
금지  4500 4328 1.0988388505049353
함수  8646 8646 0.0009884714068630587
플레이어  17293 8646 0.0010640542649573209
클릭  3044 3043 1.5807277591262265
```

[그림 8: TF-IDF 를 이용해 제거할 단어 추출]

```
1 활능후 보건복지부 장관 왼쪽 류영진 식품의약품안전처 여의도 법제사 위원회
  전체 예산 심사 순서 코리아
2 다시 탄도미사일 미사일 일본 상공 분간 북태평양 해상 낙하 한반도 유사시 증
  원 병력 출동 주일미군 기지 타격 포위 사격 능력 과시 동시 미국 군사 대응 측
  발 도발 수위 합동 참모 본부 오늘 평양 순안 일대 동해 방향 불상 탄도 미사일
  일본 상공 비행 거리 최대 고도 앞서 사흘 전인 강원도 대령 일대 김책 남단 >
  연안 동해 불상 단거리 수발 당시 단거리 한미 공동 평가 결과 단거리 탄도미사
  일 가능성 일본 중거리 탄도미사일 화성 형일 가능성 해상자위대 함대 사령관 >
  고다 요지 해장 인터뷰 미사일 안팎 비행 보아 미국 주변 화성 형일 공산 탄도>
  미사일 직후 일본 일부 지역 철도 노선 운행 중지 소동 기도 일본 당국 미사일
  전국 순간 경보 시스템 얼럿 발령 하자 도부 철도 측은 이세 사키 닛코 일부 구
  간 철도 운행 중지 철도 운행 재개 이형민 국민일보 홈페이지 페이스북 취재 대
  행 국민일보
3 오늘 평양 순안 일대 동쪽 방향 탄도미사일 탄도미사일 일본 동북지방 상공 북
  태평양 인공위성 용이 장거리 로켓 탄도미사일 일본 상공 이번 처음 합동 참모
  본부 탄도미사일 비행 거리 최대 고도 관계자 미사일 현재 중거리 탄도미사일 >
  계열 고각 관계자 미사일 비행 조각 분리 일본 매체 보도 분석 탄도미사일 지난
  강원도 대령 일대 단거리 탄도미사일 사흘 당시 탄도미사일 킬로미터 동해 합>
  참 탄도미사일 연습 대한 반발 차원 무력 시위 탄도미사일 통해 미국 증원 기지
  타격 능력 과시 전략 여건 조성 목적 장덕수 저작권 복제
```

[그림 9: 단어 제거 후 전체 기사 내용]

4.3 Word2Vec 결과

TF-IDF 를 끝낸 후 생성된 txt 파일을 이용하여 그림 10 과 같은 spark-submit 명령어를 입력하여 구현한 Scala 기반의 Word2Vec 을 실행시키면 그림 11 과 같이 HDFS 파일에 Word2Vec Model 이 저장된다. 생성된 Model 을 이용하여 그림 12 와 같이 입력한 키워드의 연관 키워드들을 얻어낼 수 있다.

```
spark@hadoop2:~/news_visualization$ /home/spark/spark-2.1.1/bin/spark-submit --c
lass spark.word2vec --master spark://hadoop2:7077 --executor-memory 10G --total-
executor-cores 10 /home/spark/news_visualization/scala_test-1.0.jar
```

[그림 10: spark-submit]

```
spark@hadoop2:~/news_visualization$ hdfs dfs -ls
Found 4 items
drwxr-xr-x   - spark supergroup          0 2017-09-05 17:05 .sparkStaging
drwxr-xr-x   - spark supergroup          0 2017-09-11 22:24 model
drwxr-xr-x   - spark supergroup          0 2017-09-05 18:12 myModelPath
drwxr-xr-x   - spark supergroup          0 2017-09-05 14:18 spark-2.1.1
```

[그림 11: 저장된 Model]

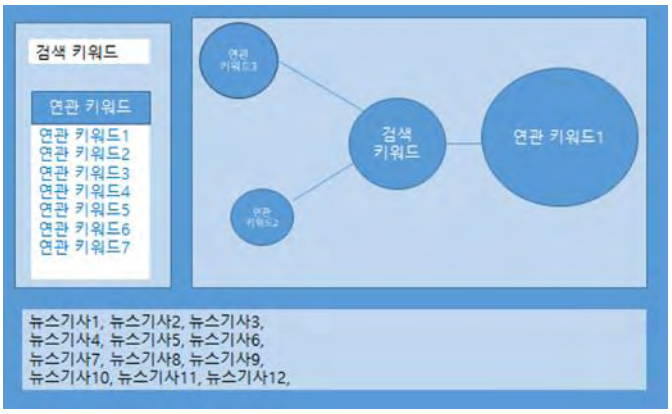
```
spark@hadoop2:~$ python noun.py 북한
[('일본', 0.7079618573188782), ('탄도미사일', 0.6985783576965332), ('관련', 0.69
82226371765137), ('발사', 0.6969292163848877), ('회의', 0.6939438581466675), ('
대표', 0.6852156519889832), ('한국', 0.6736423373222351), ('투쟁', 0.66515713930
13), ('위원회', 0.6637365818023682), ('평양', 0.6495269536972046)]
```

[그림 12: 연관 키워드 검색]

4.4 시각화

사용자가 시각화 시스템을 이용하는 방법은 그림 13 과 같은 홈페이지에서 사용자가 검색어를 입력하고 Enter 를 입력한다. 그 후 그림 14 와 같은 메인 페이지에 검색어가 전해지고 시각화 정보 데이터베이스에 저장된 정보들을 시각화하여 사용자에게 보여 준다. 오른쪽 창에 검색 키워드는 중앙에 위치하고 연관 키워드들이 검색 키워드 주위에 나타나게 된다. 연관 키워드가 검색 키워드와 연관성이 높을수록 원의 크기가 커진다. 왼쪽 창의 검색 키워드에 검색어를 입력하여 재검색을 할 수 있고, 아래의 연관 키워드를 클릭하면 연관 키워드를 검색어로 재검색을 할 수 있다. 아래 창의 뉴스 기사는 사용자가 클릭한 원의 키워드가 갖는 뉴스 기사의 제목을 보여주고 제목을 클릭하면 사용자는 뉴스 기사의 링크로 이동하게 된다.

[그림 13: 홈페이지]



[그림 14: 메인 페이지]

5. 결론

본 논문에서는 매시간 발생하는 대량의 인터넷 뉴스, 즉 뉴스 기사 빅데이터에서 기계학습을 통하여 유의미한 데이터를 추출하는 방법을 다루었다. 또한, 빅데이터의 기계학습을 효율적으로 진행하기 위해 분산처리 환경을 사용하였다. 분산처리 환경에 사용된 PC 수가 그리 많지 않아서 연산에 필요한 시간을 대폭 줄이진 못했지만 많은 성능 향상이 있었다. 분산처리 환경은 확장이 용이하므로 후에 더 많은 PC 를 이용한다면 연산처리 시간을 현저히 낮출 수 있을 것으로 보인다.

본 논문에서는 뉴스 기사에 대해서만 분석을 진행하였지만 뉴스 기사뿐만 아니라 텍스트 형태의 모든 자료를 분석할 수 있다. 논문, 특허, 인터넷 게시판 등의 많은 분야에 검색하고자 하는 키워드가 갖는 연관 키워드를 시각화된 형태로 확인할 수 있을 것이다.

참고문헌

[1] Framton, Mike [Mastering Apache spark: 정보문화사] , 2015

[2] Yoav Goldberg. Omer Levy [word2Vec Explained: deriving Mikolov et al.’s negative-sampling word-embedding method], 2014

[3] 박은정, 조성준, “KoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지”, 제 26 회 한글 및 한국어 정보처리 학술대회 논문집, 2014

[4] https://ko.wikipedia.org/wiki/TF-IDF