경 영 학 박 사 학 위 논 문

실시간 크롤링을 통한 웹 마이닝 최적화 분산처리시스템 설계

2017년 8월

부 경 대 학 교 대 학 원

경 영 학 과

이 종 화

경 영 학 박 사 학 위 논 문

실시간 크롤링을 통한 웹 마이닝 최적화 분산처리시스템 설계

지도교수 이 현 규

이 논문을 경영학 박사 학위논문으로 제출함

2017년 8월

부 경 대 학 교 대 학 원

경 영 학 과

이 종 화

이종화의 경영학박사 학위논문을 인준함.

2017년 8월

위원장 공 학 박사 한 재 호 (인)

위 원 이 학 박사 박 만 곤 (인)

위 원 경 영 학 박사 문 용 은 (인)

위 원 경 영 학 박사 김 종 원 (인)

위 원 경 영 학 박사 이 현 규 (인)

< 목 차 >

< 표 목 차 >

< 그 림 목 차 >

Distributed Processing Systems Design For Web Mining Optimization Through Realtime Crawling

JongHwa Lee

Department of Business Administration
Graduate School of Pukyong National University

Abstract

The media such as newspapers, televisions, radios and magazines played the important role in informing people about the facts that have occurred. Nowadays, the development of information and communication technology and the Internet makes it possible to access to various news more easily, and the influence of the media providing information that can form public opinion is larger and larger.

According to the statistics of registration notification for “Internet newspaper" published by the Ministry of Culture, Sports and Tourism, the number of Internet newspapers was 286 in 2005, 2,484 in 2010, and 6,605 in 2015. With the spread of the Internet, the emergence of mobile devices, and the development of networks, the number has increased more than 23 times in 10 years. It has been utilized as a research subject of many researchers as reflecting the use value of users. However, since most article analysis studies do not reflect up-to-date information, it is difficult to predict accurately the direction of public opinions that change over time. Therefore, it is necessary to study the method for quickly analyzing news data generated in real-time with the latest information.

This research constructed a real-time analysis system for

150 Internet newspaper companies in Korea as the objects of the study. Through the web mining process of desired Internet news articles, various visualizations were presented through real-time data collection and text mining analysis, and the Hadoop system, a distributed processing technology for rapid processing, was also applied. The opinion mining analysis of Netizen's comments in real-time collected news and standard dictionaries will be opened to the public, which will be helpful for future researchers.

Listening to customers and analyzing consumer's patterns and needs quickly will help the companies to establish operations and strategies. We expect that this study will contribute to the development of big data processing system that can deal with real-time processing not only simple frequency analysis but also association analysis, cluster analysis, classification analysis, and predictive analysis.

Key Words : Association Analysis, Web Mining, Text Mining, Opinion Mining, Real-time Processing

제 1 장 서 론

제 1 절 연구의 배경 및 목적

1. 연구의 배경

인류는 ICT를 기반으로 한 새로운 산업혁명을 준비하고 있다. 1, 2차 산업혁명을 통해 새로운 동력을 발견하였고, 이는 인류가 빠른 생활패턴을 가지게끔 하였다. 3차 산업혁명은 IT의 보편화와 생산 시스템의 자동화가 산업 혁명을 주도하였고, 인류의 편리함과 글로벌 환경을 가속화 시키는 촉매가 되었다.

3차 산업혁명 이후의 ICT(Information and Communication Technology, 정보통신기술)는 모든 경제, 문화 환경을 연결하는 고리로써 충분히 그 역할을 하고 있으며 단순 네트워크의 기능을 넘어서 산업간 장벽도 허물어주는 초 연결 사회를 만들고 있다. 이와 더불어 곧 다가올 제4차 산업 혁명의 기술인 Big Data, IoT(Internet of Things, 사물인터넷), 3D Printing, AI(Artificial Intelligence, 인공지능)등은 ICT 발전과 기업들의 마케팅전략과 더불어 급속도로 팽창하고 있는 산업 기술들이다(Kemp, 2016). 즉, 컴퓨터가 업무의 도구로 활용되면서 우리의 작은 행동 하나하나가 데이터로 기록되고 있다. 최근 들어 네트워크와 하드웨어 처리 속도가 향상되면서 데이터의 패턴을 찾아 소비자의 행동을 분석하는 연구가 활발히 진행되고 있는 분야가 빅데이터이다(O'Driscoll et al., 2013). 이러한 데이터에 기계학습 시스템을 불어 넣으면 바로 스스로 학습이 가능한 인공지능 시스템이 탄생하는 것이다(Morase et al., 2013).

이와 더불어 컴퓨터만으로 인터넷 접속이 가능했던 시절엔 생각지도 못했던 스마트 폰의 등장과 언제 어디서나 네트워크에 접속할 수 있는 유비쿼터스 환경의 영향으로 모든 사물에 네트워크가 연결되었으며, 사물이 가진 특성은 더욱 지능화되었다. 센서들을 통해 사물과의 소통을 시작해 보다 지능적인 환경의 사물인터넷이 등장한 것이다(UIT, 2005).

또 다른 혁신적 기술은 3D 프린팅 기술이다. 재화를 하나 생산하는 과정에는 많은 공정과 비용이 발생한다. 간단한 플라스틱 제품을 하나 생산하더라도 여전히 복잡한 과정을 거쳐야 소비자 손에 들어갈 수 있다. 하지만 이젠 필요한 제품을 소비자가 설계하고 직접 생산까지 가능한 시대가 되었는데, 이 계기는 바로 3D 프린팅 기술의 등장이라 할 수 있다(Rengier et al., 2010).

모든 사물에 음양의 조화가 필요하듯 우리 산업은 제조업을 기반으로 한 기계적 기술인 하드웨어 성장과 알고리즘과 코딩, 그리고 소프트웨어 성장이 맞물려 완성체가 된다. 또한, 개개의 산업만이 독자적인 기술로 제 4차 산업 혁명을 이끌 수 있는 것이 아니라, ICT 기반의 융・복합 산업발전이 앞으로 글로벌 환경의 새로운 경제 산업 구조로서 미래를 이끌어 나갈 것이다.

사회・인문학 분야도 4차 산업혁명의 중요한 산업 기술과 연관된다. 이에 본 연구는 빠른 네트워크 기술과 데이터 처리 기술을 결합하여 사용자의 니즈를 찾고자 한다. 인류가 생존하는 한 데이터는 계속하여 쌓이기 마련이다. 이렇게 고객의 니즈를 찾기에 강력한 도구이며 새로운 산업에서도 그 중요한 가치를 확인할 수 있는 것이 바로 빅데이터이다. 또한, 일반적인 사회 현상의 표본을 추출하여 설문 조사(Survey) 방식으로 모집단을 확률적으로 예측하는 방법도 현재 공존하는 데이터 분석 방법이다. 하지만 모집단 전체를 연구 대상에 넣고 자료 수집, 빠른 분석, 그리고 패턴을 찾기

위한 연구 방법론 등이 4차 산업혁명의 견인차 역할을 해 줄 것으로 기대된다.

이에 본 연구는 4차 산업혁명의 가장 기본적 토대가 되는 빅데이터 분야를 중심으로 웹 포털 사이트에서 실시간 생성되는 뉴스를 웹 마이닝(Web Mining) 처리를 통해 분석하는 웹 서비스 시스템을 구축하고자 한다. 본 시스템은 웹 브라우저(Web Browser) 환경에서 사용자가 원하는 키워드 중심으로 검색, 크롤링, 전처리 과정, 마이닝 처리, 시각화, 문서 군집, 댓글 분석, 연관 분석 등의 과정을 실시간으로 처리하는 방식으로 구현하고자 한다.

2. 연구의 목적

빅데이터에 대해 많은 연구자들이 관심을 갖고 노력을 기울이고 있다. 하지만 기존 대부분의 빅데이터 관련 연구는 방대한 자료 수집과 자료 분석이 이원화되어 연구어 왔다. 즉, 연구 대상의 자료가 모집단 전체를 대상으로 한다면 자료 수집 시간과 그 자료를 토큰(Token) 단위의 식별 가능한 자료로 분해하여 분석 기법이 적용되는 시간 등 많은 데이터 량과 함께 분석 시간도 늘어나는 비례관계에 있다. 또한, 연구 시간에 비례하여 여론을 예측하기 어려운 현실의 한계점이 많이 기술되고 있다(Lee and Lee, 2015; 이철성 외, 2013; 임좌상・김진만, 2014).

따라서 본 연구의 목적은 첫째, 원하는 인터넷 뉴스 기사의 웹 마이닝 과정을 통하여 실시간 자료 수집과 텍스트 마이닝(Text mining) 분석을 하고자 한다.

둘째, 실시간 수집된 뉴스의 네티즌(Netizen, 누리꾼) 댓글을 통한 오피니언 마이닝(Opinion Mining) 분석을 하고자 한다.

셋째, 방대한 자료를 신속히 처리하기 위한 분산처리기술인 하둡

(Hadoop)을 적용하여 대기 시간 및 처리 시간의 차이를 확인하고자 한다.

마지막으로 실시간 수집, 가공된 분석 결과를 시각화하는 웹 페이지를 구축하여 고객의 니즈 분석에 한 발 더 가까이 다가서는 것을 목적으로 한다.

제 2 절 연구의 범위 및 구성

1. 연구의 범위

본 연구에서는 국내 최대 포털 사이트인 네이버(http://www.naver.com) 사와 제휴된 국내 언론사를 대상으로 실시간 기사 분석 시스템을 구축하고자 한다. 검색 키워드에 관련된 네이버 뉴스 기사를 웹 마이닝으로 추출하여 텍스트 데이터를 자연어 처리 기반으로 문서 분류(Document Classification), 정보 추출(Information Extraction), 문서 군집(Document Clustering) 등의 분석을 하고자 한다. 오픈소스 소프트웨어(Open Source Software)를 활용하여 하둡을 이용한 분산처리시스템 구축과 오픈 소스 통계솔루션인 R프로그램을 활용한다. 문장 단위 분석을 통하여 텍스트 마이닝(Text Mining), 연관분석(Association Analysis), 오피니언 마이닝(Opinion Mining)을 진행하고자 한다. 그 외 리눅스 운영체제를 비롯한 PHP, MySql, HTML, CSS, JavaScript, Python 등 실시간 웹 마이닝을 지원하기 위한 도구와 한글 자연어 처리 실시간 연구라는 사례가 없는 시스템을 구축하여 연구문제를 해결하고자 한다.

2. 연구의 구성

본 연구는 5개의 장으로 구성하였으며 다음과 같다.

제 1 장은 연구의 배경과 논문의 주제에 대한 문제 제기 그리고, 연구의 범위와 구성에 대하여 서술하였다.

제 2 장은 웹 마이닝를 비롯한 텍스트/오피니언 마이닝의 기존 연구자들의 선행연구와 한글 자연어처리 패키지와 비표준어 처리 연구, 분산처리시스템의 선행연구를 살펴본다.

제 3 장에서는 연구의 목적과 이론적 배경을 바탕으로 한 실시간 분석 프레임워크를 제시하고 표준어를 사용하는 뉴스 기사의 분석과 상대적으로 비표준어를 많이 사용하는 뉴스 댓글 분석을 위한 패키지 알고리즘을 제시한다. 또한, 웹 마이닝의 크롤링 방법과 알고리즘을 자세히 설명하고 연구 질문의 해결 과정을 상세히 기술한다.

제 4 장은 현재 시대적 이슈를 실시간 분석을 통하여 시간의 흐름에 따라 변화하는 여론과 네티즌들의 반응을 시각화하여 웹페이지로 결과를 확인한다. 한 대의 컴퓨터로만 처리하는 집중처리(Centrallized Processing)와 n대의 컴퓨터를 한 대처럼 사용하는 분산처리(Distributed Processing)의 처리 시간에 대한 차이를 비교하여 실시간 이슈 분석 시스템의 속도 차이를 살펴본다.

제 5 장은 논문의 요약 및 결론 부분으로 결과 분석을 통해 연구의 의미와 공헌도 그리고, 연구의 한계점과 향후 연구방향에 대해 제시한다.

제 2 장 이론적 배경

본 연구 시스템의 이론적 근거가 되는 웹 마이닝(Web Mining)에 대한 일반적인 고찰과 자연어 처리 기반의 텍스트 마이닝, 오피니언 마이닝 이론과 개념들을 살펴보고자 한다. 또한, 한글 처리의 표준어/비표준어 처리에 관한 선행 연구와 실시간 웹 마이닝에 필요한 오픈 소스 소프트웨어의 웹 언어들에 관한 연구들을 살펴보고자 한다.

제 1 절 웹 마이닝(Web Mining)

1. 웹 마이닝

웹 마이닝은 데이터 마이닝 기술을 활용하여 웹 문서 및 서비스에서 지식 검색에 대한 정보를 자동으로 검색, 추출, 분석 및 유용한 패턴을 찾는 것을 의미한다. 이는 데이터 마이닝 기술을 응용하여 동적인 웹 환경의 텍스트와 콘텐츠에서 정보를 발견하는 기법이다(Kosala and Blockeel, 2000). 웹 데이터는 반정형적 또는 비정형적 텍스트, 시간 흐름에 변화하는 형태로 이루어져 있다. 웹의 다양한 정보는 웹 마이닝의 세 가지 분류에 의하여 연구되고 있다(Zhang and Segall, 2008).

2. 웹 마이닝 분류

웹 마이닝은 데이터 유형에 따라 텍스트, 이미지, 오디오, 비디오를 사용

하는 웹 콘텐츠 마이닝(Web Content Mining)과 웹페이지와 하이퍼 링크를 사용하는 웹 구조 마이닝(Web Structure Mining) 그리고, 사용자 프로파일을 활용한 웹 사용 마이닝(Web Usage Mining)으로 구분한다(Zhang and Segall, 2008). 웹 콘텐츠 마이닝의 텍스트 정보를 활용하여 텍스트 마이닝, 오피니언 마이닝 기법들을 사용한다(Kosala and Blockeel, 2000; Bin and Zhijing, 2003; Chakarbarti, 2003; Henzinger, 2004; Srivastava et al., 2002). 이와 같이 세 가지 분류별 연구를 살펴보면 다음의 [표 2-1]과 같다.

[표 2-1] 웹 마이닝 데이터 유형에 따른 연구

구 분	관련 연구
Web Content Mining (텍스트, 이미지 오디오, 비디오)	Zhang et al., 2003; Lau et al.., 2004; Lihui and Lian, 2005; Liu and Chen-Chuan-Chang, 2004; Darmont et al., 2007; Graves et al., 2007; Pol et al., 2008; Xu et al., 2011; Sivakumar, P. 2015;
Web Structure Mining (웹 페이지)	Lihui and Lian, 2005; Chakrabarti, 2002; Fang and Sheng, 2004; Hay et al., 2004; Guan and McMullen, 2005; Song and Sheppard 2006; Dinucă, Totad and PVGD, 2009; Dinucă, 2011; Yan, 2014; Victor and Rex, 2016;
Web Usage Mining (사용자 프로파일)	Mobasher et al., 2000; Spiliopoulou, 2000; Srivastava et al., 2000; Clendaniel, 2001; Fenstermacher and Ginsburg, 2003; Pierrakos et al. 2003; Song and Shepperd, 2006; Pabarskaite and Raudysv 2007; Tyagi et al., 2010; Bari and Chawan, 2013; Wang et al., 2015; Vellingiri et al., 2015

Kosala and Blockeel(2000)는 웹 마이닝을 통한 기계 학습, 정보 검색, 자연 언어 처리, 정보 추출을 시도하였으며 Han and Chang(2002)는 웹의 링크와 의미 구조를 분석하여 마이닝 웹 검색 엔진에 관한 연구를 진행하

였다. Chau et al.(2004)는 웹 콘텐츠 마이닝을 통하여 다국어에 관한 연구를 진행하였고 Kolari and Joshi(2004)는 시맨틱(Semantic) 웹 마이닝을 연구하였다.

3. 웹 마이닝 사례

웹 환경의 변화와 모바일 시장의 급 팽창은 웹 접근성을 필요로 하며 모바일의 작은 화면에 고객의 욕구를 충족하기 위한 개발자들의 노력은 더욱 필요로 해 보인다. 이러한 환경에서 보안과 손쉬운 UI(User Interface)환경의 개발을 위한 웹 언어 또한, 다양하게 웹 페이지 구현에 사용되고 있다. HTML를 비롯한 JavaScript, CSS, PHP, Ajax, JQuery 등 다양한 언어들의 진화가 웹 개발 환경을 더욱 복잡한 구조로 만들고 있다.

웹 페이지 접근이 점점 복잡한 구조로 이루어져 웹 마이닝 연구자들은 연구과정에서 많은 어려움을 한계점을 지적하고 있다. 강한훈 외(2010)는 HTML, JavsScript, Ajax 등으로 이루어진 다양한 쇼핑몰 웹 페이지를 계층 구조 분석을 통해 연구를 진행하였다. 최승배 · 강창완(2011)의 웹 마이닝 분석은 학교 로그데이터를 지원 받아 방문자 패턴을 연구하여 지역적, 신입생 유치 방안 등을 연구하였다. 정민영(2016)은 포털 사이트인 네이버, 다음, 구글에서 제공하는 인기 키워드들의 생존 분석 연구를 진행하였고 네이버, 다음의 통합적인 주요 이슈들을 분석 및 도출하였다.

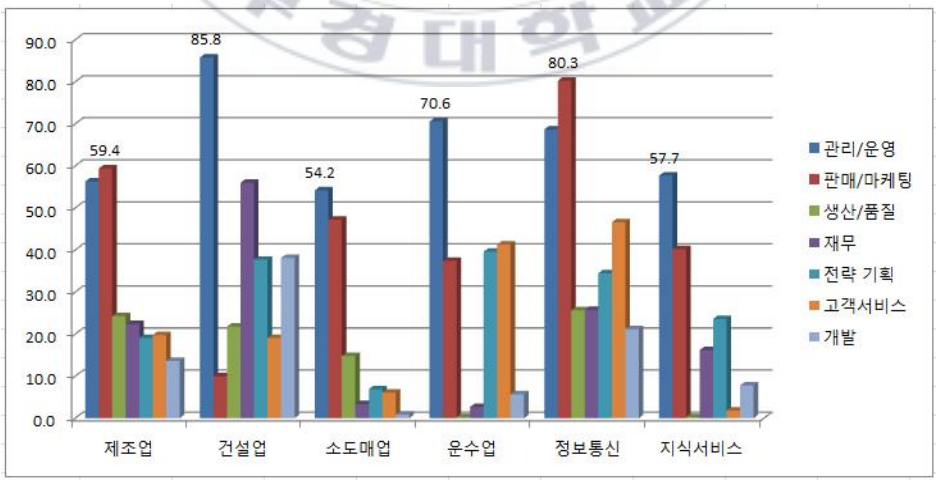
국내외 포털사이트 운영 회사들은 자체적인 DB 자료를 바탕으로 실시간 분석 서비스를 운영하고 있으며 단순 키워드 중심의 빈도를 시계열로 나타내고 있다. 구글(Google)은 구글트랜드, 다음(Daum)은 다음소프트, 네이버는 DataLab를 통하여 서비스하고 있다.

제 2 절 데이터 마이닝(Data Mining)

1. 데이터 마이닝

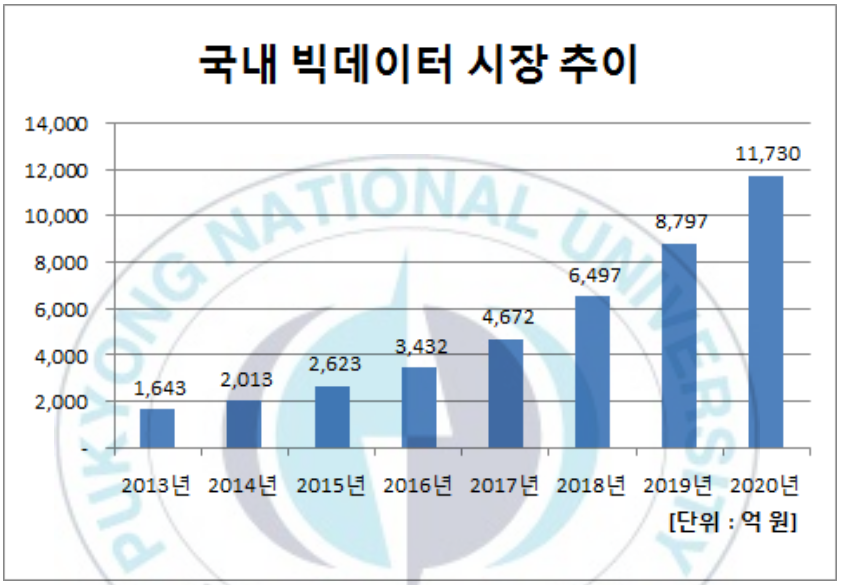
데이터 마이닝은 빅데이터에서 의미 있는 패턴을 찾거나 의사결정을 위한 예측에 빅데이터를 활용하는 기술이다. 가설이나 가정에 따른 분석, 검증인 통계분석과는 다르다는 것이다.

[그림 2-1]은 국가통계포털인 e-나라지표 자료를 분석한 것으로 2015년 3,500여개의 기업을 대상으로 한 빅데이터 산업별 활용분야 조사에 따르면 제조업은 관리/운영분야와 판매/마케팅분야에 빅데이터를 활용하는 것으로 나타났다. 건설업은 관리/운영분야에 86% 가까이 빅데이터를 활용하고 있다고 조사되었다. 소도매업, 운수업, 지식서비스 분야도 관리/운영 업무에 데이터를 사용하고 있고 정보통신 분야는 판매/마케팅 분야에 높은 빅데이터 이용 비율을 보였다(lee et al., 2016).



[그림 2-1] 2015년 산업별 빅데이터 활용 분야(www.index.go.kr)

한국정보화진흥원에서 발표한 국내 빅데이터 시장 추이를 살펴보면 [그림 2-2]와 같다. 2014년에 2,013억 원 규모로 나타났다고 한다. 2015년은 2,623억 원, 2016년은 3,432억 원의 규모로 매년 30% 이상의 성장을 보이고 있으며 2020년 국내 빅데이터 시장을 11,730억 원으로 성장할 것으로 전망된다(www.nia.or.kr).



[그림 2-2] 국내 빅데이터 시장 추이(www.nia.or.kr)

마이닝 도구의 발달로 방대한 양의 데이터를 분석하는 것이 가능해졌으며, 데이터 마이닝은 정형적 자료 형태에서 분석이 가능한 정형 데이터 마이닝(Structured Data Mining)과 정형화 되지 않은 자료 형태를 분석하는 비정형 데이터 마이닝(Unstructured Data Mining)으로 나누어진다. 일반적인 정형 데이터 마이닝은 분류분석(Classification Analysis), 예측분석(Prediction Analysis), 군집분석(Clustering Analysis), 연관분석(Association Analysis)등이 있으며, 비정형 데이터 마이닝은 텍스트 마이닝(Text mining), 사회네트워크분석(Social Network Analysis)으로 구분한다.

2. 텍스트 마이닝

텍스트 마이닝은 수많은 단어들로 구성된 문서 또는 텍스트 미디어에서 중요한 정보를 추출하는 프로세스이다(Lee and Lee, 2015). 텍스트 정보는 문서의 단어 집합에서 추출한 텍스트를 분석하여 서로의 관련성을 파악하고자 한다. Fayyad et al., (1996)이 제시한 텍스트 마이닝 프로세스에 의하면 먼저 연구 대상 텍스트를 선택(Selection), 대상 데이터를 저장 및 전처리과정(Proprocessing and Cleaning)을 수행, 얻고자 하는 정보의 특성을 파악하고 관련 단어를 추출(Feature selection and extraction), 데이터 변환 과정을 통하여 분석 모델 설계와 테스트 과정인 마이닝 처리(Text Mining), 단어들의 패턴을 찾아 분석 결과에 대한 평가(Interpretation Evaluation)순이며, 이 과정을 거친 자료들이 비즈니스에 활용된다. 텍스트 마이닝은 문서 분류(document classification), 문서 군집(document clustering), 메타데이터 추출(metadata extraction), 정보 추출(information extraction) 등으로 구분한다(Written et al., 2016).

기본적으로 텍스트 마이닝은 텍스트를 요인이나 숫자로 변환하여 데이터 마이닝 프로젝트에서 소셜 네트워크 분석, 문서 클러스터링, 상관관계, 예측 등과 같은 분석에 사용될 수 있다. 텍스트 마이닝의 진화된 연구 중 오피니언 마이닝(또는 감정 분석이라고도 함)은 최근 많은 관심을 끌었던 어려운 작업 중 하나이다(Liu, 2013).

많은 연구자들은 텍스트 마이닝 연구의 진행을 위해 오픈소스 소프트웨어이며 통계 패키지인 R을 사용한다(장경애 외, 2015). R은 기존 통계 프로그램(SAS, SPSS, 등)의 기능들을 포함하면서 최신 분석 방법과 R 플랫폼을 통한 마이닝 기능을 제공한다. 다양한 시각화 기법을 통한 사용자 인

터페이스 차원의 대시보드 (Dashboard)를 구현할 수 있는 언어이기도 하다. 그리고 언어에 가까운 문장 형식이므로 기능들의 자동화가 비교적 쉽고, 사용자들이 여러 예시들을 공유하여 많은 패키지들을 매일 생성하고 있다(https://www.r-project.org/).

또한, 개발 환경에서 GUI로 돕기 위한 R스튜디오는 오픈소스로 함께 제공되며 다양한 운영체제를 지원한다(Lee and Lee, 2015). R스튜디오는 프로그램 작업을 편하게 해주는 스크립트 창(Script Window)과 변수 타입이나 변수 값을 확인할 수 있는 워크스페이스(Workspace), 작업 폴더 구조, 차트 결과를 나타내는 Plots 공간과 Console을 통한 명령어를 실행 할 수 있어서 편리하게 작업할 수 있다(https://www.rstudio.com/).

3. 소셜 네트워크 분석

소셜 네트워크 분석(Social Network Analysis)은 노드(Vertex)와 링크로 구성된 그래프 이론과 행렬 구조를 활용하여 개인과 집단들 간의 관계를 계량적으로 분석하는 과정이다. Martino and Spoto(2006)는 행과 열이 만나는 셀에 특정 값을 입력해 행과 열 사이의 관계를 나타내는 행렬 구조를 비정형 데이터 분석에 활용하는 연구를 하였다. 네트워크의 구조를 파악하기 위해 연구자들은 중심성(Centrality), 밀도(Density), 중심화(Centralization) 등 주로 이 세 가지 기법을 이용한다.

중심성(Centrality)은 중앙에 위치한 정도를 나타내며 한 점이 다른 점과 직접 연결된 정도에 따라서 지역 중심성을 나타내며, 한 점이 연결망 전체의 중심자리에 따라 전체 중심성을 나타낸다. 밀도(Density)는 전체에 영향을 주는 관계수중에서 실제로 연결된 관계 수의 비율을 뜻하며 Scott(1992)의 실험에 집단의 크기에 반비례하는 연구 결과가 있다. 전체

연결망의 형태가 어느 정도 중앙에 집중되어 있는지를 나타내는 정도를 중심화(Centralization)라 한다. 중심성이 한 노드가 전체 망에서 얼마나 핵심적 위치를 차지하는가에 초점을 둔다면, 중심화는 전반적으로 얼마나 중심에 집중되어 있는 구조를 띄는지를 측정하는 것이다.

4. 연관 분석(Association Analysis)

대규모의 데이터 항목들 중에서 유용한 연관성을 찾는 기법으로 활동이 이루어진 항목들의 상호 연관성을 찾아내어 항목들 간의 규칙을 발견하기 위해 사용된다. “A상품을 구입한 이후에는 C상품을 구매 한다.”는 구매 패턴을 찾아내는 장바구니 분석(Market Basket Analysis)이라고도 한다. 고객 충성도의 정도에 따라 고객들의 다양한 행동 중 교차구매(Cross Selling)나 상향구매(Up Selling)를 분석할 때 효과적이다.

연관분석의 측도는 지지도(Support), 신뢰도(Confidence), 향상도(Lift)로 나누어진다.

지지도(Support)는 전체 상품 중에서 A상품과 C상품을 동시에 포함하는 거래에 대한 확률이나 비율을 의미한다. 식(1)은 상품 A, C가 동시에 포함하는 구매 거래가 어느 정도인지를 나타내며, 전체 빈도 경향을 파악할 수 있다.

$$\text{Support} = P(A \cap C) = \frac{A\text{와 } C\text{를 포함하는 거래수}}{\text{전체 거래수}(N)} \quad (1)$$

식(2)의 신뢰도(Confidence)는 상품 A를 포함한 거래 중에서 상품 A와 상품 C가 같이 포함될 확률을 의미하며 A를 구입한 경우, 이 중에서 얼마나 C를 구매할 것인지를 의미한다.

$$\text{Confidence} = P(C|A) = \frac{P(A \cap C)}{P(A)} = \frac{A\text{와 } C\text{를 포함하는 거래수}}{A\text{를 포함하는 거래수}} \quad (2)$$

식(3)의 향상도(Lift)는 상품 A를 구매한 경우 그 거래가 C를 포함하는 경우와 C가 임의로 구매되는 경우의 비율을 나타내며 상품 A와 C의 구매 패턴이 독립적인지, 서로 상관이 있는지를 나타낸다. 향상도가 1이면 두 항목이 서로 의존성이 낮은 독립적인 관계이고 1보다 작으면 음(-)의 상관관계, 1보다 크면 서로 양(+)의 상관관계가 있다고 해석된다.

$$\text{Lift} = \frac{P(C|A)}{P(C)} = \frac{P(A \cap C)}{P(A)*P(C)} = \frac{A\text{와 } C\text{를 포함하는 거래수}}{A\text{를 포함하는 거래수} * C\text{를 포함하는 거래수}} \quad (3)$$

본 연구에는 연관 분석에서의 상품이나 서비스 리스트가 아니라 연구 대상에서 자주 등장하는 단어로 그 리스트를 대신하고 전체 거래 수(N)는 연구 대상 문장의 수로 대신하여 연구를 진행하였다. 비정형 데이터를 정형 데이터로 전환하여 분석하는 과정을 구현하고자 한다.

제 3 절 오피니언 마이닝(Opinion Mining)

1. 오피니언 마이닝

사람들의 의견(Opinion), 감성(Sentiment), 평가(Evaluation), 태도(Attitude), 감정(Emotion)을 분석하는 마이닝을 오피니언 마이닝(Opinion Mining) 또는 감성분석(Sentiment Analysis)이라 한다(Bing and Liu, 2012). Pang and Lee(2008)은 소셜 네트워크의 가장 거대한 연구 관심 분야로 특정 문서를 만든 사람의 의견을 추출해 내기 위해 만든 고급기술로

오피니언 마이닝을 정의하기도 하였다(Rao et al., 2014).

오피니언 마이닝 기술은 기계학습 접근 방법(Machine Learning Approach)과 어휘기반 접근 방법(Lexicon-based Approach)으로 크게 나누어진다(Medhat et al., 2014).

기계학습 접근 방법은 자료 수집과 결과를 통해 예측한다는 점에서 텍스트 마이닝과 유사하지만 스스로 자료 수집과 학습이 가능한 시스템이란 점에서 이와 차이가 있다. 이는 인공지능(Artificial Intelligence)의 한 분야로써, 빅데이터 핵심 기술로 사회적 관심을 받고 있다(Qiu et al., 2010; Lu et al., 2010; Lane et al., 2012).

어휘기반 접근 방법은 감성 사전을 기반으로 일치하는 단어들을 분석하는 기법이다. 부정 표현(Negative Opinion)과 긍정 표현(Positive Opinion)을 구분하여 네티즌의 의견 및 감성을 찾아내는 맥락에서 매우 보편적인 방법이다(Neviarouskaya et al., 2010; Heerschop et al., 2011; Moreo et al., 2012; Balahur et al., 2012; Le et al., 2015; 이종화・이현규, 2015).

2. 오피니언 마이닝 분류

어떤 논제에 대하여 네티즌들이 참여한 댓글을 살펴보면 “힘내세요.”, “오늘은 피곤해도 기분 좋네” 등의 긍정적 의견이나 감성을 담아 표현하는 것과 “지긋지긋하다.”, “끔찍한 일입니다.” 등 마음에 들지 않는다는 부정적 의견을 제시한다. 제품의 재질이나 특징, 색상 등 사용 리뷰가 급증하면서 온라인 쇼핑에서도 연구 사례를 많이 찾아볼 수 있다(Choi and Chang, 2011; 서지훈 외, 2015; 윤영선, 2013; Le et al., 2015; 이종화・이현규, 2015).

[그림 2-3]은 다양한 감성이나 감정 분류 기법은 어휘 기반 접근 방법과 기계학습 접근 방법으로 나누어 분류하였다.

[그림 2-3] 감성 분류 기법(Medhat et al., 2014)

기계학습 접근 방법은 지도학습(Supervised Learning)과 비지도학습(Unsupervised Learning)으로 구분되며 지도학습(Supervised Learning)은 훈련 데이터로부터 하나의 함수를 유추해 내기 위한 방법으로 지도학습을 통해 유추된 함수로 연속적인 값을 예측하는 회귀분석(Regression)과, 어떤 요인의 값인지 나타내는 분류(Classification)가 있다. 즉, 미리 알려진 훈련 데이터를 학습시켜서 새로 수집되는 데이터가 속할만한 군집을 찾는 것이다. 훈련 데이터가 커지면서 데이터로부터 나타나지 않던 현상까지도 일반화시켜서 시스템에서 처리할 수 있게 되는 것이며 구글의 “Alpha-Go” 시스템이 그러한 예이다.

비지도학습(Unsupervised learning)은 특성이 유사한 데이터를 합쳐서 군(Group)으로 분류하는 방법으로 우리가 흔히 아는 군집화(Clustering)를 통하여 이루어지며 훈련 데이터 군을 사용하지 않기 때문에 “비지도”라 하며 취미, 관심사에 대한 분류에 해당하는 예이다. 이런 지도 학습과 비지도 학

습을 통해 인공지능(AI) 분야에서 사람의 학습을 모델링하는 것을 기계학습이라 하며 그런 알고리즘의 집합을 딥 러닝(Deep learning)이라고 한다.

어휘기반 접근 방법은 긍정적 용어와 부정적 용어로 분류된 사전을 기반으로 분석하는 방법이며 사전 기반 접근 방법(Dictionary-based Approach)과 말뭉치기반 접근 방법(Corpus-based Approach)으로 나누어진다. 사전 기반 접근 방법은 감성 사전과 준비된 데이터를 비교하여 문서를 분류하는 방법이며 연구 데이터 대상에 따라 사전을 분류하면 정확성을 가지는 장점이 있다. 물론 사전의 범위가 제한적이라는 단점이 존재한다. 말뭉치기반 접근 방법은 연구 대상문서를 하나의 덩어리인 말뭉치로 구현하여 빈도가 높은 문서의 정보를 찾아내는 방법이다.

3. 오피니언 마이닝 분류별 연구 사례

[표 2-2] 오피니언 마이닝 관련 연구

종류	관련 연구
사전 기반 접근 방법 (Dictionary-based Approach)	Neviarouskaya et al., 2010; Heerschop et al., 2011; Moreo et al., 2012; Balahur et al., 2012.
Support Vector Machines (SVM)	Bai, 2011; Yan-Yan et al., 2010; Fan and Chang 2011; Kang et al., 2012; Walker et al., 2012; Balahur et al., 2012; Lane et al., 2012.
단순 베이즈 (Nalve Bayes)	Bai, 2011; Yan-Yan, 2010; Fan and Chang 2011; Walker et al., 2012; Lane et al., 2012.
최대 엔트로피 (Maximum Entropy)	Bai, 2011; Duric and Song, 2012.
규칙기반 분류 (Rule-based Classifiers)	Lu et al., 2010; Qiu et al., 2010; Lane et al., 2012.
말뭉치기반 접근 방법 (Corpus-based Approach)	Neviarouskaya et al., 2010; Cao et al., 2011; Zhou et al., 2011; Zirn et al., 2011; Hu et al., 2012; Robaldo and Di Caro 2013; Zhang et al., 2012; Keshtkar and Inkpen, 2013; Maks and Vossen, 2012.

기계 학습 접근 방식의 지도학습 기법에는 의사결정나무분류(Decision Tree Classifiers), Support Vector Machines, 신경망네트워크(neural Network), 규칙기반 분류(Rule-based Classifiers), 최대 엔트로피(Maximum Entropy), 베이지안 네트워크(Bayesian Network), 단순 베이즈(Naive Bayes) 등으로 나누어 연구가 진행되며 [표 2-2]와 같다.

제 4 절 비표준어-한글(Nonstandard Words-Korean)

1. 비표준어

한 나라에서 공용으로 사용하는 말의 규범을 표준어(標準語)로 칭한다. 표준어 규범에 벗어난 언어를 비표준어(Nonstandard Words)라 한다.(Lee et al., 2016). Web 2.0의 환경이 시작되면서 모바일의 급속한 발전과 빠른 네트워크 환경에서 사용자들의 의견이나 감성이 묻어 있는 댓글이 증가하였으며, 이에 따라 사용후기나 리뷰(Review)에 대한 많은 연구가 진행되었다(Kim et al., 2014; Won and Kim, 2014; Lee and Lee, 2015, Le and Lee, 2014; Le et al., 2015; 장경애 외, 2015). “표준말”은 그 시대 중류 사회에서 사용하는 서울말을 의미하였으며, 이는 1993년에 제정되었다(www.korean.go.kr). “표준말”이 “표준어”로 표현된 것은 “비표준어”가 등장하면서 부터인데, 이는 표준말과 대비나 말결에 맞지 않아 “표준어”로 바꾼 것이다. 이처럼 비표준어는 표준말을 표준어로 바꿀 정도로 많은 관심을 받고 있는 연구대상이다. 또한, 인터넷어, 채팅어, 새롭게 등장하는 신조어 등이 비표준어의 일종으로 다양하게 우리 생활에 사용되고 있다. Lee et al.(2016)의 비표준어 사전을 참고하면 “귀요미”, “긔요미”, “내칭구”, “내

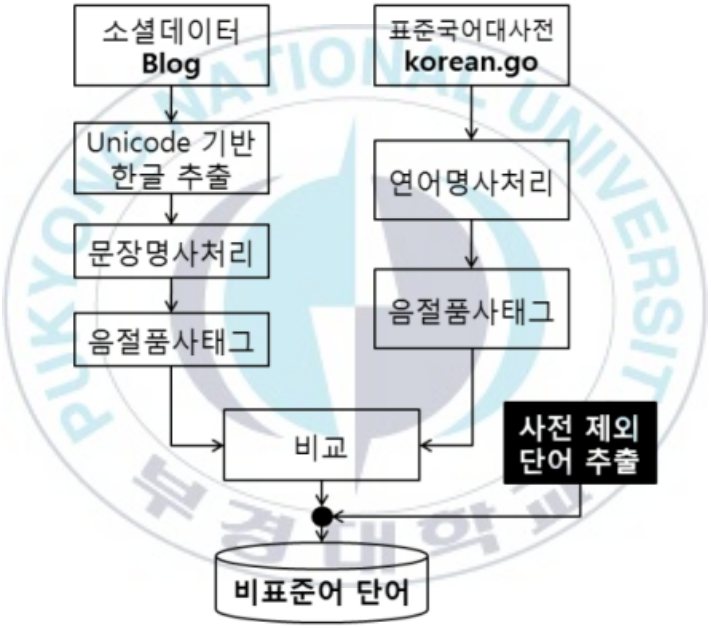
팅구”, “안뇽”, “칭구”, “칭구덜” 등 인터넷에서 사용되는 채팅어, “습니다”를 “스빈다”로 “삽니다”를 “사빈다”로 “합니다”를 “하빈다”와 같이 오타를 그대로 파생시켜 사용하는 파생어, “코드갱어”, “낫닝겐”, “노답”, “노잼”, “딘치” 등과 같이 국어와 외래어를 조합 및 결합하여 사용하는 합성단어, “ㅋㅋ”, “ㅇㅋ”, “ㅎㅎ”, “ㅇㅇ”, “ㅠㅠ”, “ㅡㅡ” 등과 같이 한글의 자음이나 모음 등을 이용한 이모티콘(Emoticon) 등의 비표준어 단어를 연구하였다.

2. 비표준어-한글 연구 사례

국가 간의 문화 교류를 통하여 협력증진을 목적으로 설립된 유네스코(Unesco) 한국위원회에 따르면 우리말 “한글”의 근본인 “훈민정음 해례본”을 1997년에 세계유산으로 등록하여 관리하고 있다. 최근 우리 전통놀이인 “줄다리기”를 문화유산에 지정한 바가 있다(www.unesco.or.kr). 한글은 한국인에게 문자 혁명을 가져왔다는 글을 시작으로 한글의 소개와 독창성을 강조하며 설명되어 있다. 영어 문장은 알파벳 나열 방식으로 ‘G’, ‘o’, ‘o’, ‘d’를 붙쳐 “Good”로 표현하지만 한글은 ‘ㅇ’, ‘ㅕ’, ‘ㄹ’, ‘ㅡ’, ‘ㅁ’이라 쓰지 않고 “여름”처럼 묶어서 쓰는 방식을 취하고 있다. 즉, 자음과 모음의 음소로 음절 단위로 묶어 글자로 만들어 쓴다. 네티즌들이 사용하는 사용후기나 댓글들은 표준어 이외의 단어들이 더 많이 등장하고 있어 이러한 비표준어가 많은 한글 처리 연구에 한계점을 제시하고 있다(Kim et al., 2014; Yun, 2008; Hong and Cha, 2013; Sim, 2011; 박소연, 2013; 안정국・김희웅, 2015).

각 나라별 언어를 모두 표현하기 위해 나온 코드 체계 중 유니코드(Unicode)를 활용하여 비표준어 연구가 진행되고 있다(Lee et al., 2016). 수집한 소셜 데이터 중 유니코드 기반의 한글을 추출하는데 한글 초성, 중

성, 종성의 가지 수는 11,172이며 해당되는 범위 즉, 유니코드 내 임계치(아스키코드 기준 : -21504 ~ -10333)에 해당하는 글자를 먼저 선별한다. 국어국립원의 우리말 표준 국어 대사전을 명사화한 결과에 기존 연구 데이터와 비교하여 표준어를 제외시킨 후 나머지 데이터를 활용하여 비표준어를 찾아내는 연구를 진행하였다. 연구 관련 프레임워크는 다음과 [그림 2-4]와 같다.



[그림 2-4] 비표준어 추출 프레임워크
(Lee et al., 2016)

블로그 데이터 4,000건의 자료 수집에서 나온 비표준어 300여개의 비표준어 사전을 [표 2-3]과 같이 연구하였다.

[표 2-3] 비표준 단어 사전의 표제어 목록(Lee et al., 2016)

순번	단어	순번	단어	순번	단어	순번	단어	순번	단어	순번	단어
1	뀰	54	은탱	107	길냥이	160	셀카족	213	훈대딩	266	디질랜드
2	듄	55	저퀄	108	김치녀	161	셀피족	214	훈선수	267	똥꼬발랄
3	쩐	56	존맛	109	까도남	162	쇼루머	215	힐링녀	268	리즈시절
4	가놀	57	졸팅	110	까도녀	163	스빈다	216	ㄸㄹㄹ	269	리터루족
5	감솨	58	즐감	111	깔맞춤	164	씹덕사	217	ㅍㅎㅎ	270	멘탈붕괴
6	감튀	59	직캠	112	꼬돌남	165	악화녀	218	ㅎㄷㄷ	271	모루밍족
7	갠소	60	출책	113	꼬돌남	166	엄친딸	219	돌취생	272	모루밍족
8	갠톡	61	치맥	114	꿀벅지	167	엔디족	220	글설리	273	몰고어족
9	게삭	62	칭구	115	나핑족	168	엔지족	221	넘사벽	274	미담진족
10	겜방	63	칼답	116	낫닝겐	169	여미족	222	극혐오	275	번달번줌
11	겜종	64	컴겜	117	내칭구	170	완벽녀	223	쓸고퀄	276	베이글남
12	겟관	65	국방	118	내팅구	171	완판녀	224	피꺼솟	277	베이글녀
13	겨냄	66	굴녀	119	노노족	172	요섹남	225	솔까말	278	부끄부끄
14	구챗	67	폰겜	120	노모족	173	우쭈쭈	226	사바사	279	불펌금지
15	구플	68	피닝	121	뇌색남	174	운도남	227	고고씽	280	브런치족
16	극험	69	핸폰	122	뇌섹남	175	운도녀	228	김여사	281	브로맨스
17	길막	70	헐ㅋ	123	뇌섹녀	176	울아들	229	품절남	282	블링블링
18	길빵	71	횻팅	124	눔프족	177	울아빠	230	가격딴지	283	빼박캔트
19	꿀잼	72	훈남	125	뉴빠족	178	울언니	231	감성포텐	284	사토리족
20	노답	73	훈녀	126	능욕짤	179	웃겨염	232	견어물녀	285	삼포세대
21	노잼	74	힐빙	127	니트족	180	유느님	233	걸크러쉬	286	세테크족
22	노켐	75	광클	128	답정녀	181	유트족	234	고럼고럼	287	쇼루밍족
23	놀족	76	엄크	129	당글녀	182	응믐응	235	고퀄리티	288	쓰담쓰담
24	뇌섹	77	깜놀	130	대세녀	183	의느님	236	고퀄병맛	289	아리얀족
25	닉추	78	짝남	131	듣보잡	184	자출족	237	골로갈뻔	290	알바추노
26	님선	79	짝여	132	떡실신	185	좋슴닿	238	군대드립	291	야지디족
27	닝겐	80	ㅋㅋ	133	똥남자	186	즐설리	239	군데렐라	292	오랜지족
28	덕질	81	ㅇㅋ	134	똥따떠	187	짜짜시	240	굿쩐문화	293	으랏차차
29	덧답	82	ㅎㅎ	135	띠껍다	188	짱짱맨	241	그래니룩	294	치렐루야
30	득템	83	겨털	136	런피스	189	쩍벌남	242	그루밍족	295	캥거루족
31	딘치	84	팀킬	137	레어템	190	차도녀	243	꺄르르르	296	코드갱어
32	렉방	85	멘붕	138	로엘족	191	출장족	244	꽁냥꽁냥	297	프리타족
33	막방	86	득템	139	로엘족	192	츤데레	245	꿀잼꿀팁	298	프티타족
34	맞팔	87	ㅇㅇ	140	로코남	193	치느님	246	끄어어억	299	힙스터족
35	먹튀	88	호갱님	141	먹튀꾼	194	치맥족	247	낄끼빠빠	300	앵그리맘
36	멍미	89	가싶남	142	몰링족	195	칭구덜	248	나포츠족	301	금사빠녀
37	모객	90	간장녀	143	몰캉족	196	컴터끄	249	나홀로족	302	남성눔프족
38	병맛	91	감쪽녀	144	물계급	197	컴터앞	250	날르가즘	303	매스티지족
39	블그	92	갑툭튀	145	뭐하삼	198	퇴장족	251	낮져밤이	304	베재댜겨쇼
40	생얼	93	개교주	146	반갑남	199	퉁퉁녀	252	내일러들	305	부키싱글족
41	소푸	94	개미족	147	반갑녀	200	펫팸족	253	노마드족	306	썸썸남썸녀
42	스샷	95	개소름	148	방콕남	201	포미족	254	뉘에뉘에	307	역쇼루밍족
43	스압	96	개이득	149	방콕녀	202	포미족	255	뉴시니어	308	찌아찌아족
44	스포	97	검색질	150	방통럭	203	푸어족	256	느므느므	309	커우커우족
45	심남	98	골쥼이	151	버카충	204	푸하핫	257	니에니에	310	코드커터족
46	심쿵	99	구글링	152	벌벌족	205	피딩족	258	달관세대	311	하이타오족
47	썸남	100	굿굿굿	153	뽀샤시	206	하빈다	259	달빛동맹	312	흐규규규뀨
48	썸녀	101	귀요미	154	뽀통령	207	핫핑크	260	대프리카	313	
49	썸맥	102	글램핑	155	삑사리	208	핫핫핫	261	덕페이스	314	
50	썸친	103	금벅지	156	사빈다	209	호갱님	262	도로로로	315	
51	안눙	104	금사빠	157	상오빠	210	호빗족	263	도쩐개쩐	316	
52	안습	105	긔요미	158	세피족	211	후덜덜	264	등쉰색히	317	
53	왤케	106	기승전	159	셀카봉	212	훅와닿	265	디스하다	318	

제 5 절 한글자연어처리(KoNLP)

1. 자연어 처리

인류가 서로의 의사를 표현하기 위하여 사용하는 언어를 IT 디바이스에서도 처리가 가능하게 변환하는 기술을 자연어 처리(Natural Language Processing)라 한다(John and Govilkar, 2016). 인간이 일상적으로 사용하는 언어를 의미 분석이나 형태소 분석을 통해 컴퓨터가 처리할 수 있게 변환한다는 뜻이다. 한글은 음절 단위로 묶어 다시 한 글자로 만들어진 문자이며 이런 한글 자연어 처리에 대해 활발한 연구가 이루어지고 있다(박경미・황규백, 2011; 김승우・김남규, 2014; 안정국・김희웅, 2015; 이종화・이현규, 2016; Le et al., 2015; Lee et al., 2015; 박대민, 2016; 강정배 외, 2013).

자연어 처리를 위한 데이터 과학에서 텍스트 분석 도구로 많은 연구자들이 Python, Java, R등을 사용한다.

Bird et al.(2009)의 연구는 Python의 Natural Language Toolkit(NLTK)을 사용하여 자연어 처리 작업 수행하는 오픈 소스 플랫폼을 토큰, 형태소 분석, POS 태깅, 구문 분석 및 의미론적 추론 등의 인터페이스로 연구를 진행하였다.

Socher et al.(2013)와 Manning et al.(2014)의 연구는 Java 프로그램의 “CoreNLP”를 사용하여 POS 태깅, 기본적인 NLP 작업을 지원하는 프레임워크로 인식(Recognition), 파싱(Parsing), 회귀분석뿐만 아니라 감성 분석(Sentiment Analysis)을 적용하는 연구가 이루어졌다.

Rehurek and Sojka(2010)는 데이터가 가질 수 있는 주제에 대해 정해진

확률인 확률분포와 주제들에서 나타나는 단어들의 확률분포를 찾았을 때, 프로세스에 의해 새로운 문서를 생성할 수 있다고 역설했다. 문서들이 있으면 전 상황과 반대로 확률분포들을 추정하여 주제를 찾는 방법인 Latent Dirichlet Allocation(LDA)를 Python으로 연구하였다.

Che et al.(2010)의 연구는 C++를 사용하여 Language Technology Platform(LTP)에 관한 연구를 하였다. LTP는 단어 세분화, 태그 부착과정을 거쳐 어휘 분석, 구문 분석 및 의미 분석 모듈을 포함하여 중국어를 위한 오픈 소스 NLP 시스템을 설계하였다.

2. KoNLP 패키지 연구 사례

한글 자연어 처리에 대한 연구도 국내 연구자들에 의하여 형태소 분석, 구문분석 처리 모듈 등 언어 처리에 필요한 연구들이 진행되었다(박경미・황규백, 2011; 강정배 외, 2013; 김승우・김남규, 2014; 안정국・김희웅, 2015; 이종화・이현규, 2016; Lee et al., 2015; 박대민, 2016).

이종화・이현규(2016)의 연구에 의하면 기존 개발된 패키지의 명사 추출 신뢰성을 높이기 위한 새로운 패키지를 제안하였다. [그림 2-5]과 같이 기존 R를 활용한 extractNoun 함수의 비정상적 명사 처리과정을 확인하였다.

```
> extractNoun("한글날을 하루 앞두고 개최된 이 행사는 외국인 어린이들의 한글에 대한 흥미를 높이고 마련되었다.")
 [1] "한글날을" "하루"     "개최"     "행사"     "외국"     "어린이"   "들"       "한글에"   "흥미"     "마련"
>
```

[그림 2-5] 1차 명사 분리 작업(이종화・이현규 2016)

“한글날을”은 “한글날/N+을/J”로 “한글날”은 명사, “을”은 조사로 구분되어 “한글날”이 비정상적 명사 추출이 일어나는 것을 보완하기 위하여 [그림 2-6] 같이 명사 추출 후 품사 태그를 재확인 즉, 한 번 더 확인 과정을 거쳐 진행하였다. 표준어를 상대적으로 많이 사용하는 뉴스와 비표준어 단어가 많이 들어가는 블로그를 대상으로 기존 명사 추출 함수와 New Algorithm 으로 명사추출과정을 거쳐 명사 추출율의 개선된 점을 확인할 수 있었다.

[그림 2-6] 품사 태그를 활용한 명사 추출 프레임워크 (이종화・이현규 2016)

이종화・이현규(2016)의 연구에 다르면 뉴스 기사는 23%의 개선 단어로 명사 추출되었고, 블로그는 29%의 개선 단어로 명사 추출이 개선되었다. [표 2-4]와 같은 명사 추출이 불일치된 명사리스트는 다음과 같다.

[표 2-4] extractNoun()과 new_Noun()함수와의 불일치 명사 리스트 (이종화 · 이현규, 2016)

No	extractNoun()	new_Noun()	extractNoun()	new_Noun()	extractNoun()	new_Noun()
1	10선을	10선	맛집이였어여	맛집	책테마파크에	책테마파크
2	13작품이	13작품	먹다보니	먹다보	책테마파크에서	책테마파크
3	2구역은	2구역	먹방은	먹방	초밥들이	초밥들
4	2인석에	2인석	미니우동하나를	미니우동하나	초밥먹을꺼면	초밥먹을꺼
5	2인석은	2인석	부른배에도	부른배	초밥은	초밥
6	3구역은	3구역	성남문화재단은	성남문화재단	초밥이	초밥
7	4구역으로	4구역	시켜먹었어여	시켜먹었어	초밥이랑은	초밥
8	4구역은	4구역	식감이	식감	초밥입니다	초밥
9	갈꺼에요~	갈꺼	식감자체가	식감자체	초밥집이	초밥집
10	고깃집에	고깃집	쌈밥같은	쌈밥같	초밥집입니다	초밥집
11	고깃집임에도	고깃집	쌈채소나	쌈채소	촉촉~해서	촉촉
12	굽느냐에따라	굽느냐	안되서	안되	키즈룸입니다	키즈룸
13	내안에를	내안에	안받고	안받	하시더라구	하시
14	내안에에서는	내안에	없어여	없어	한국콘텐츠진흥원과	한국콘텐츠진흥원
15	네이버가	네이버	예삿놈이	예삿놈	한글날에는	한글날
16	네이버는	네이버	용비어천가를	용비어천가	한글날을	한글날
17	대답하시더라구	대답하시	이히히히~~	이히허히	한글에	한글
18	떡~~~하니~	떡~~~하	잔뜩~	잔뜩	한글은	한글
19	뚫려있는곳은	뚫려있는곳	잘어울린다는	잘어울린다	한글을	한글
20	뜨거운물이	뜨거운물	종류만해도	종류만해	한글이	한글
21	리필을	리필	주문햇습니다	주문햇습니	한글자료도	한글자료
22	말했어여	말했어	쫙~	쫙	한글자료를	한글자료
23	맛있더라구여	맛있더라구	찜갈비도	찜갈비	햇기에	햇기
24	맛집에서는	맛집	찜갈비를	찜갈비	허형만	허형
25	맛집이였습니다	맛집	책테마파크가	책테마파크	활어로	활어

또한, 기존에 명사로 구분되어 명사 빈도에 포함된 단어들 중 비정상적 추출과정에 포함된 조사나 수식어, 독립어 등의 단어는 명사로 배제되어 보다 신뢰도를 높여 명사추출 개선에 도움이 된 제외된 명사 리스트는 [표 2-5]과 같다.

[표 2-5] new_Noun()함수 사용으로 제외된 명사 리스트 (이종화・이현규, 2016)

No	extractNoun()	new_Noun()	extractNoun()	new_Noun()	extractNoun()	new_Noun()
1	각	"각/M"	안	"안/M"	처음	"처음/M"
2	계속	"계속/M"	애	"애/I"	체	"체/I"
3	들이	"들이/M"	약	"약/M"	하더라구	"하/P+더라구/E"
4	떡	"떡/M"	어리둥절	"어리둥절/M"	한	"하/P+ㄴ/E"
5	맛있어	"맛있/P+어/E"	요놈	"요놈/I"	합니다	"하/P+ㅂ니다/E"
6	몇	"몇/M"	위해	"위하/P+어/E"	해	"하/P+어/E"
7	물론	"물론/M"	이	"이/M"	해서	"하/P+어서/E"
8	사왔습니다	"사/P+아/E+오/P+아ㅂ니다/E"	이미	"이미/M"	회	"회/M"
9	석	"석/M"	저	"저/M"	후	"후/I"
10	순	"순/M"	전	"전/M"		
11	시	"시/I"	쪽	"쪽/M"		

M : 수식언(관형사, 부사), I : 독립언(감탄사), P : 용언(동사, 형용사), E : 어미(연결어미)

제 6 절 오픈 소스 소프트웨어(Open Source Software)

소프트웨어를 구성하는 명령어, 소스코드를 무료로 공개하고 배포하는 프로그램을 오픈소스 소프트웨어(Open Source Software)라 한다. 사용자가 소스에 접근할 수 있고 사용과 수정이 가능하며 재배포까지 자유롭게 진행이 가능하다(문장식・김홍기, 2014). 사용의 개방적 특성으로 누구나 쉽게 개발환경을 접할 수 있으며 ICT 산업이 하드웨어 중심에서 소프트웨

어 중심으로 시장도 이동하였다. 하드웨어 제품에 소프트웨어를 융합하여 보다 그 가치를 높이자는 뜻이기도 하다. 즉, 일반적인 신발의 부가가치보다 그 신발에 웨어러블 센싱이 가능한 제품으로 판매한 부가가치가 더 높다는 설명이다(미래부창조과학부_SW혁신전략, 2014).

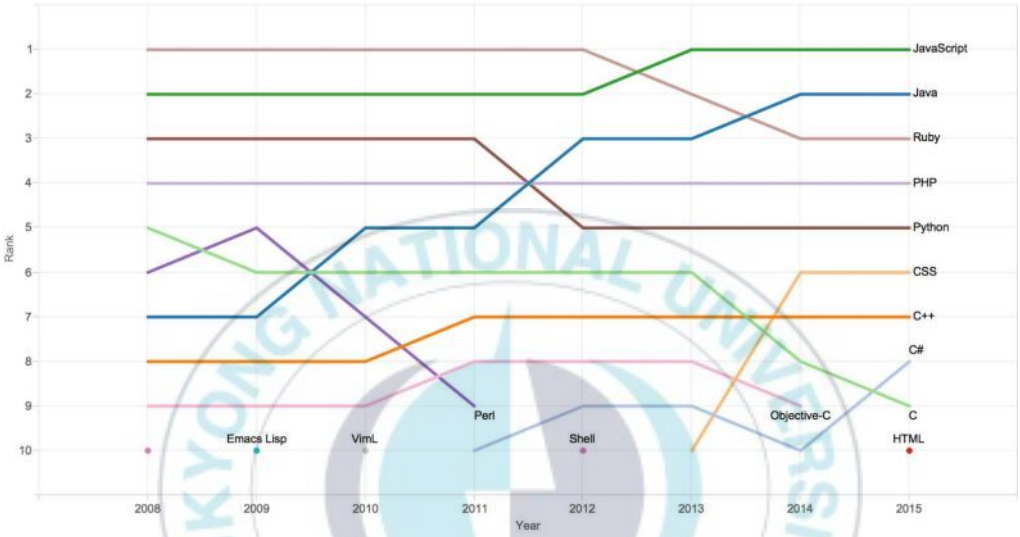
본 연구에 사용되는 모든 도구들은 오픈소스 소프트웨어를 활용하며 개발 환경을 보다 개방(Openness)적이고 개발 툴 간의 호환 및 분석 데이터들의 연동성을 확인하였다. JavaScript, JQuery, HTML, CSS, PHP, MySql, Python 등이 연구에 활용된 예를 보자.

1. 자바스크립트(JavaScript)

웹페이지는 정보 제공자인 Server측과 정보를 요구하는 Client측으로 나누어진다. 웹 페이지의 내용과 모양을 제어하기 위하여 ASP(Active Server Page), JSP(Java Server Page) 등의 서버 페이지 제작 언어들이 존재한다. 현재 대부분의 웹 페이지에서는 HTML(Hyper Text Markup Language), CSS(Cascading Style Sheets), JavaScript등의 언어들이 클라이언트 페이지 제작을 담당하고 있다. 웹 페이지의 큰 틀을 제작할 때는 HTML언어를 사용하며, 페이지 내 글씨체나 색깔과 같은 디자인적 요소들을 표현할 때는 CSS가 담당한다. JavaScript는 웹 페이지의 동작을 담당하는데 '버튼을 선택했을 때 시간을 보여줘'라는 형태의 명령을 내릴 수 있는 풍부한 효과를 넣을 수 있다(정원기 · 문수묵, 2010; 류석영, 2016).

소스 코드 관리 툴인 'Git'을 사용하는 프로젝트에 웹 호스팅 서비스를 제공하는 'GitHub'는 2013년 이후 CSS와 함께 웹 페이지 구축에 많이 사용되고 있다. [그림 2-7]를 살펴보면 객체 지향 언어인 JavaScript, Java, Ruby 등이 개발자들 사이에 많이 사용되고 있다는 것을 알 수 있다. 지난

2008년부터 매년 정기적으로 발표되는 인기 랭킹 집계 내용에 따르면 웹페이지 구축을 위한 언어로 단연 JavaScript가 많이 사용됨을 알 수 있다. PHP와 CSS 그리고 HTML이 그 뒤를 잇는다(김진국 외, 2015; http://github.com/).



[그림 2-7] GitHub 사용자의 소스 코드 언어 순위

2. 제이쿼리(JQuery)

자바스크립트로 작성할 경우 엄청난 소스 코드를 입력해야하는 기능들, 예컨대 디자인을 변경하거나 애니메이션을 삽입하는 등의 효과를 Jquery를 통해 쉽고 빠르게 작성할 수 있다. 이는 웹 페이지를 개발하는 개발자의 생각하는 방법을 바꾼 언어로 자바 스크립트 라이브러리이다(장명현 외, 2011). Jquery 환경에서 사용하는 Ajax 기법은 대화식 웹 페이지 제작을 위한 웹 개발 기법이다. Ajax는 페이지 이동 없이 고속으로 화면 전환이 가능하며 수신하는 데이터의 양을 줄일 수 있고, 클라이언트에게 처리를

위임할 수도 있다. 단점은 Ajax와 호환성에 문제가 있는 브라우저는 Ajax 기법을 사용한 페이지를 볼 수 없다는 것이다(https://wikipedia.org/).

[그림 2-8]은 웹페이지에서 JavaScript로 작성된 소스 코드와 JQuery로 작성된 소스 코드를 비교한 예문이다. wrapper라는 아이디를 가진 div태그 안에 ul태그가 있고 그 자식으로 li태그가 4개가 있고 li태그의 자식으로 span태그가 있는 html 구성에서 세번째 'li'태그의 글자색을 빨간색으로 변경하려고 할 때 두 언어의 기법을 이용하여 표현한 것이다.

Language	Source Code

Java Script	window.onload = function(){
	//페이지가 로드되면
	var wrapper = document.getElementById('wrapper');
	//'warpper' id를 찾아서 wrapper 변수에 대입
	var ul = wrapper.getElementsByTagName('ul');
	//'ul' 태그를 찾아서 ul 변수에 대입
	var li = ul[0].getElementsbyTagName('li');
	//'li' 태그를 찾아서 li 변수에 대입
	
	for(var i=0; i<li.length; i++)
	{
	// li의 개수만큼 반복
	
	var l = li[i];
	// 각 li를 l 변수에 대입
	if(l.className=='three')
	{
	//l의 class가 three인지 비교해서 맞으면
	var s = l.getElementsByTagName('span');
	//l의 자식인 'span'태그를 s 변수에 대입
	s[0].style.color='red';
	//'span'의 0번째요소 즉, 텍스트의 글자색 변경
	}
	}
	}

jQuery	$(document).ready(function(){
	//페이지 로드되면
	$('#wrapper > ul > .three > span').css('color','red');
	//wrapper id의 자식인 'ul' 태그의 자식인 three class의 자식인 span 태그의 글자색을 red로 변경
	});

[그림 2-8] JavaScript와 jQuery 소스 코드 비교

3. 크롤링(Crawling)

웹 페이지를 그대로 가져와서 데이터를 추출해 내는 과정을 의미하며, 이번 연구에서 가장 많은 노력과 시간이 소요된 과정이다. 크롤링 대상 페이지를 웹 클라이언트를 통해 접속하고 접속된 웹 클라이언트를 통해 HTML 파일을 파싱(Parsing)하여 연구자가 원하는 내용을 가져오는 과정이다. 최민석(2015)은 페이스북의 크롤링 과정을 실증 연구 방법을 통하여 연구하였다. 리눅스(Linux) 크론(Cron)을 이용하여 데이터를 수집하며 MySql DB에 결과물을 저장하였다. HTML, CSS, JavaScript 언어를 활용하여 사용자 인터페이스를 구현하였다. 일반적인 웹 서버와 DB를 이용하여 접근성과 호환성을 확보하였고 적은 비용으로도 시스템 구축이 가능하게끔 하였다. 무엇보다 오픈 소스 API를 이용하여 간단한 웹 프로그래밍 작업으로도 자료 수집에 큰 문제가 없었다고 한다(구홍서, 2000; Cachia et al., 2007; 손수아・박석천, 2015; Black, 2008).

본 연구는 방대한 량의 뉴스 기사들을 크롤링하여 텍스트 마이닝, 오피니언 마이닝, 연관 분석 등을 진행 하고자 한다. 또한, 유래 없이 연구 분석 방법으로 자료 수집과 분석과정을 일원화하여 빠른 시간 방대한 양의 분석을 실시간으로 처리하고자 한다.

기존 연구들은 수집과 분석을 일원화하기 위하여 분석 툴을 활용하여 수집 과정을 확대하는 방법을 사용하였다. 분석 툴은 기본적으로 분석 과정에 패키지들이 제공되며 제한적이지만 수집과정도 일부 스크립트를 수정하면 제한적이지만 웹 페이지를 크롤링이 가능하다는 것을 활용하였다. 정민영(2015, 2016)는 통계 분석 도구인 R언어를 활용하여 네이버의 인기 검색어를 1시간단위로 가져오는 크롤링 알고리즘을 공개하였다. 하지만 R언

어로써 여러 겹의 래퍼로 구성된 웹 페이지를 더 깊이 읽어 들이는 한계를 나타내고 있다.

IoT 환경의 연구에서 센싱들의 데이터 수집으로 로그 데이터 분석 또한, 눈에 띄는 연구들이었다. 이들은 크롤링 과정 없이 연구 대상 기업의 로그 데이터를 제공 받아 분석하는 연구 사례들도 많았다. 최승배・강창완(2011)는 학교 홈페이지 분석을 위한 접속자 IP 정보를 가장 선호하는 메뉴나 클릭 페이지 등을 분석하였다.

제 7 절 분산처리시스템(Distributed Processing System)

1. 분산처리시스템

이용자 한명이 하나의 컴퓨터를 사용하는 단일 테스크와는 달리 분산처리시스템은 여러 대의 컴퓨터를 한 대의 컴퓨터처럼 사용할 수 있는 시스템을 말한다. 데이터 처리량이 늘어나면서 빠른 처리 속도로 데이터 처리의 효율성을 찾고 저비용 디바이스들을 활용하여 고속 처리를 하자는 의미이다. 1998년에 창업한 구글(Google Inc.)은 쓸모없는 정보로만 페이지 상단을 채우는 일이 흔했다. 이를 개선하기 위해 구글은 검색 결과의 랭킹(Ranking)을 활용하여 웹 페이지에 점수를 매기고, 점수가 높은 페이지를 검색 결과의 상단으로 가져오는 방법을 채택하였다. 랭킹 기술의 구현을 위해서는 복잡한 계산이 요구되며, 그만큼 많은 컴퓨터가 필요했다(Brin and Page, 1998). 구글의 분산 스토리지는 GFS(Google File System)의 분산 파일시스템과 Bigtable의 분산 스토리지 시스템을 사용하였다(Ghemawat et al., 2003; Chang et al., 2008). 분산처리를 위한 기반 기술

은 MapReduce를 사용하여 대용량 데이터를 분산하여 처리하였다.

오픈 소스 기반의 분산처리 환경을 구축하기 위해서 하둡(Hadoop)은 HDPS(Hadoop Distributed File System)의 파일 처리 시스템을 사용하며 분산 처리 프레임워크는 Hadoop MapReduce를 사용하고 HBase의 저장 장치를 사용한다(Sheela, 2016). 데이터를 하둡 파일 시스템 수집, 저장하는 플룸(Flume), 데이터를 하둡에 로딩하고 처리 결과를 RBase에 연동 및 저장하는 스쿱(Sqoop), 대량의 데이터를 실시간 저장, 조회 가능한 NoSQL, 하둡을 모니터링 및 관리할 수 있는 척와(Chukwa), 자원을 제어하고 관리할 수 있는 분산 코디네이터의 역할을 하는 주키퍼(Zookeeper)로 구성되어 있다(Lakhe, 2014; Lydia and Swarup, 2015).

2. 하둡 기반 연구 사례

하둡 시스템은 다양한 분야에서 연구되고 있었다. Lewis et al. (2012)은 하둡 MapReduce 프레임워크에서 효율적으로 실행되도록 특별히 설계된 단백질 서열 데이터베이스 검색 엔진인 “Hydra” 알고리즘을 사용하였다. Tare et al. (2014)은 감성분석의 지도학습기법 중 단순 베이즈(Naive Bayes) 분류를 활용하여 트윗 분류를 위한 MapReduce 전략을 연구하였다. Bian et al. (2012)은 Support Vector Machines(SVM)를 사용하여 SNS 메시지 내용을 분석하여 마약 사용자를 찾고 잠재적인 역효과를 찾는 방법을 연구하였다. Liu et al. (2013)은 영화 리뷰 데이터를 이용하여 단순 베이즈(Naive Bayes)분류법으로 감성 분류를 연구하였다. Leo et al. (2009)는 초고속 정렬을 달성하기 위해 그래픽 처리 장치에서 멀티 프로세서를 사용하는 짧은 시퀀스 읽기 정렬 알고리즘으로 하둡 연구를 하였다.

3. RHadoop

RHadoop는 R 함수를 통해 하둡 플랫폼으로 데이터 분석을 수행하기 위한 R의 오픈소스 프레임워크이다. RHadoop 프로젝트는 "rhdfs", "rmr", rhdase" 이렇게 세 가지 R 패키지로 구성되어 있으며 하둡의 두 가지 주요 기능인 HDFS와 맵리듀스를 모델로 하여 설계되었다(Oancea and Dragoescu, 2014).

rhdfs는 R에서 하둡의 모든 HDFS 접근을 제공하기 위한 R 패키지이다. R 함수로 모든 분산 파일을 관리할 수 있다. R과 하둡을 연결해주며 분산 데이터 파일을 쉽게 읽고 쓸 수 있다. rmr은 R과 하둡 맵리듀스의 인터페이스 기능을 제공하는 R 패키지이며 맵퍼와 리듀서를 쉽게 개발할 수 있다. 입력, 출력, 맵퍼, 리듀서 등 여러 잡 매개변수로 하둡 스트리밍 맵리듀스 API를 호출하고, R 맵리듀스 잡을 수행할 수 있다. rhbase는 R에서 HBase 분산 데이터베이스에 있는 데이터를 처리하는 패키지이며 테이블 조작 작업, 초기화 및 읽고 쓰는 여러 가지 메소드에 대한 디자인이 가능한 패키지이다(Hafen et al., 2014; Harish et al., 2015).

제 3 장 연구방법

제 1 절 연구 개요 및 개략 프레임워크

1. 연구 개요

인터넷 신문은 2005년부터 시작하여 현재 2,500곳의 인터넷 신문을 운영하고 있다(http://www.index.go.kr). 인터넷의 보급과 모바일 기기의 등장, 네트워크의 발달로 무려 10년 사이 급성장한 산업이기도 하다. 중앙지를 비롯한 언론사들은 사용자들의 사용 가치를 반영하듯 학계에서도 많은 연구자들이 이슈 분석, 선거 분석의 연구 대상으로 활용되어 왔다(Kam and Song, 2012). 하지만 대부분의 기사 분석 연구들은 최신 정보를 반영하지 못하기 때문에 시간이 지남에 따라 실시간으로 바뀌는 여론의 방향을 정확히 예측하기가 어려운 것이 한계점으로 나타났다(이철성 외, 2013; 임좌상・김진만, 2014; Lee and Lee, 2015).

본 연구는 중요한 일들을 발 빠르게 전달하려는 인터넷 뉴스를 실시간 분석하고자 한다. 포털 사이트의 검색 기능처럼 사용자가 궁금해 하는 이슈 단어와 해당 기간을 선택하면 관련 기사의 크롤링 과정을 거쳐 이슈 분석에 필요한 기사만을 추출하여 텍스트 마이닝, 빈도 분석, 군집 분석, 소셜 네트워크 분석, 연관 분석 등의 시각화를 통해 웹 마이닝 설계과정을

연구하고자 한다.

또한, 특정 기사들의 댓글 분석을 통해 국민들의 여론을 빠르게 분석하기 위해 비표준어 처리 과정을 거쳐 오피니언 마이닝 분석을 진행하고자 한다.

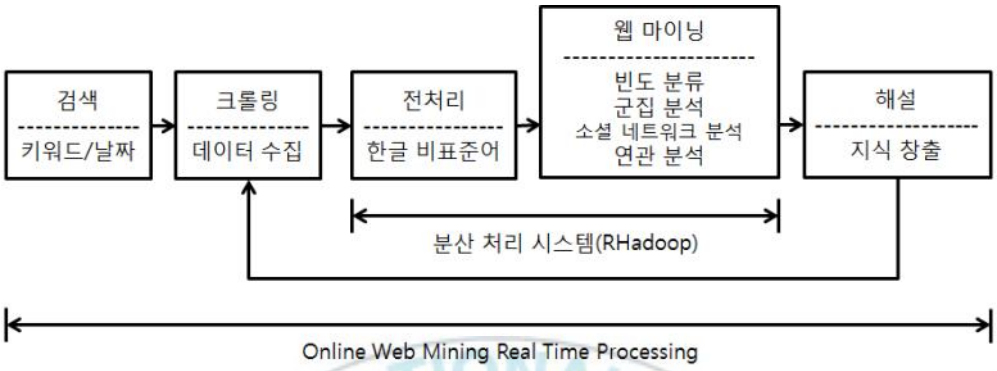
실시간 크롤링 과정과 해당 데이터의 실시간 처리를 위하여 분산처리시스템을 함께 구현하기 위해 Hadoop을 구축하였다. 텍스트 분석을 위해 데이터 마이닝에 많이 사용하는 R에서 제공하는 RHadoop를 함께 탑재하여 마이닝 처리를 분산처리시스템에 활용하여 구축하고자 한다.

그리고 모든 과정을 사용자 유저인터페이스(User Interface)환경으로 구축하여 연구의 결과물을 실생활에 활용할 수 있는 웹 페이지로 구축하고자 한다. 또한, 반응형 웹(Responsive Web)을 적용하여 여러 디스플레이 종류에 따라 화면의 크기가 최적화되도록 하여 보다 시각화 결과를 잘 전달하고자 한다.

2. 연구 프레임워크

하이퍼텍스트 기반의 데이터 분석을 통하여 새로운 의미 있는 결과를 찾거나 새로운 지식을 발견하는 과정을 관련 키워드 입력으로 전 과정을 실시간 자동 처리하는 시스템을 개발하였다.

[그림 3-1]은 본 연구의 분산 처리 시스템을 활용한 웹 마이닝 실시간 분석 처리 시스템 개략 프로세스이다. 전 과정 처리할 수 있는 웹페이지가 제작되며 궁금한 이슈 단어와 기간을 입력하면 웹 마이닝 처리 과정을 거쳐 관련 기사 및 댓글을 DB에 저장한다. 이후 한글 비표준어 처리와 불용어 및 의미 없는 문자 제거 등의 전처리 과정을 거쳐 데이터 마이닝 분석 과정을 시스템에서 진행한다.



[그림 3-1] 본 연구의 프레임워크

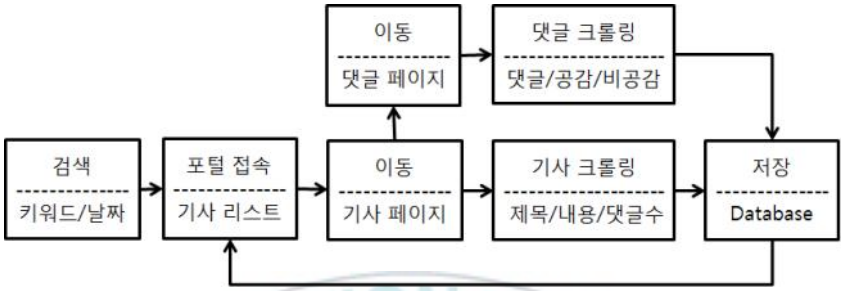
연구의 목적에서도 제시하였듯 뉴스 기사의 웹 마이닝 과정을 통하여 실시간 자료 수집과 텍스트 마이닝 분석 프레임워크 설계와 특정 기사의 댓글 분석을 통한 네티즌 의견을 분석하는 상세 프레임워크를 설계하였다. 또한, 빅데이터 분석 도구인 R 프로그램을 분산처리시스템으로 구축하기 위하여 Hadoop과 함께 결합한 RHadoop 시스템을 구축하여 처리 시간 효율성을 측정하는 프레임워크도 함께 설계하였다.

제 2 절 상세 프레임워크

1. 데이터 크롤링 디자인

개발된 웹 페이지에서 키워드와 날짜 설정 후 검색을 실행하면 포털 사이트의 뉴스 기사가 같은 조건으로 검색이 이루어지며 기사 리스트를 확보한다. 기사별 해당 웹 페이지로 이동하여 기사의 제목, 내용, 댓글 수를 데

이터베이스에 저장한다. 또한, 기사별 댓글 페이지로 이동하여 해당 기사의 댓글과 댓글의 공감 및 비공감 지수를 함께 DB에 저장한다[그림 3-2].



[그림 3-2] 데이터 크롤링 프레임워크

포털에서 뉴스 기사를 검색하면 관련 기사가 DB에 자동으로 저장된다. 물론 기사의 송고 날짜를 구분하여 원하는 기간 동안의 기사에 한하여 크롤링되는 알고리즘을 [그림 3-3]에서 제시한다.

본 연구의 실제 웹 페이지의 검색 정보인 키워드, 시작 날짜, 끝 날짜를 PHP 환경으로 이동시킨다. 검색 결과 페이지의 URL 정보를 이용하여 총 기사의 개수와 한 페이지 단위 10개의 기사로 리스트를 확인하고 몇 페이지로 구성되어 있는지 확인한다. 즉 100개의 기사가 검색 되었으면 10페이지로 나누어 기사를 볼 수 있으므로 몇 페이지로 구성 되어 있는지를 확인한다. 크롤링 로봇은 각 기사 링크로 이동하며 언론사, 기사제목, 댓글 개수, 기사 작성 날짜, 기사 내용 등을 각 변수에 저장하게 된다. 이는 MySQL의 Insert query 명령어를 이용하여 이루어지는 과정이다. 또한, 선행연구에서도 확인되었지만 웹 환경의 텍스트 크롤링은 여러 특수 문자들을 함께 가져오게 되어 trim(), strip_tags(), html_entity_decode() 등의 명령어를 결합하여 불용어 처리도 함께 해주었다. 언론사명은 이미지로 제공되어 텍스트 크롤링 환경에서 이미지로 되어 있는 언론사명을 가져오기는

불가능한 일이라 로고 이미지의 파일명이 언론사명으로 된 것을 확인하고 파일명을 크롤링하여 DB내 언론사명 필드에 내장시켰다.

```
$sch_txt = "키워드";
$startDate = "시작 날짜";
$endDate = "끝 날짜";

$html = file_get_html(" 뉴스 기사 URL ");
$total_val = $html->find(" 뉴스 검색 총 기사 개수");
$page_cnt = ceil($total_val/10);  // 페이지 수

$article_list = file_get_html(" 뉴스 기사 URL ");

$i=0;

foreach($article_list->find("건별 기사 링크") as $element)
{
    $article_link = $element->href; // 기사 링크 대입
    $element->href = trim(strip_tags(html_entity_decode($element->href))); // 특수문자 제거

    $img = $article_link->find("언론사 로고 이미지");          //언론사
    $press = $img->title;
    $press = trim(strip_tags(html_entity_decode($press)));   // 웹 특수문자 제거

    $article_title = $article_link->find("기사제목")->plaintext;  // 기사제목
    $article_title = trim(strip_tags(html_entity_decode($article_title)));

    $reply_cnt = $article_link->find("댓글 개수")->plaintext;   // 댓글 개수
    $reply_cnt = trim(strip_tags(html_entity_decode($reply_cnt)));

    $art_date = $article_link->find("기사 작성 날짜")->plaintext;  // 기사 작성 날짜
    $art_date = trim(strip_tags(html_entity_decode($art_date)));

    $contents = $article_link->find("기사 내용")->plaintext; // 기사 내용
    $contents = trim(strip_tags(html_entity_decode($contents)));

    $query = "insert query 명령'";
    $result = mysqli_query($conn, $query);   // DB 저장
}
```

[그림 3-3] 뉴스 기사 크롤링 알고리즘

네티즌들은 인터넷 기사를 읽은 후, 다른 네티즌들의 의견인 댓글을 몇 페이지씩 읽어보고 댓글의 의견에 “공감”이나 “비공감”을 표현한다. 이 과정을 통해 네티즌들의 의견이 계속하여 업데이트되며, 이는 여론의 방향이 된다. 사회적으로 큰 이슈가 되는 기사들의 방문자 수나 리뷰어(Reviewer)들의 엄청난 활동으로 만들어지는 리뷰들은 수만 건의 여론을 읽는데 중요한 자료로 활용가치가 있다. 하지만, 특정 기사 관련 댓글들을 웹 마이닝 데이터로 수집하는 과정은 복잡하다.

먼저, 우리가 기사를 읽고 하단의 최근 댓글 몇 개 읽고 “더보기” 버튼을 이용하여 다음 페이지로 넘기고 읽고 또, 넘기고 읽고 이렇게 반복하듯이 크롤링 로봇이 그 일을 대신 해주어야 한다. 아니면 포털 사이트 내부 데이터베이스의 내용을 마음대로 읽을 수 있는 권한이 있는 연구자 이외는 다른 방법이 없다. Web Driver가 프로그램에 의해서 “더보기” 버튼을 반복하여 클릭하고 마지막 리뷰가 웹 화면에 나타날 때까지 연구자가 할 일을 그대로 대신해주는 API이다.

Web Driver는 범용적인 웹 브라우저에서 제공하는 API로 코드를 통해 실제 이용자가 웹브라우저를 다루는 것처럼 사용이 가능하며 Web Driver를 사용하기 위하여 Python의 Selenium 라이브러리를 활용하였다.

Python의 Selenium Web Driver를 이용하여 실제 사용자가 브라우저를 다루는 것처럼 브라우저 자체를 사용하여 단순 패킷 요청과 JavaScript와 CSS에 의해 만들어진 웹 페이지도 자동으로 테스트를 할 수 있고, 크롤링하기 힘든 포털 사이트에서도 가능하다.

본 연구는 PhantomJS를 이용할 경우, 실제 웹 브라우저를 이용하는 것이 아니라 가상적인 웹 브라우저를 이용해 마치 사용자가 브라우징 하는 것처럼 행동하여 웹페이지 데이터 크롤링이 가능함을 밝혔다. 적용된 알고리즘은 [그림 3-4]과 같다.

```
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
import time

driver = webdriver.PhantomJS(executable_path=' 서버 URL')

line = "뉴스기사 URL"
driver.get(line)

try:
    while True:
        more_button = WebDriverWait(driver, 10).until()  # 더보기
        more_button.click()
except :
    pass
finally:
    file.write("댓글 내용 파일쓰기")
```

[그림 3-4] 댓글 크롤링 알고리즘

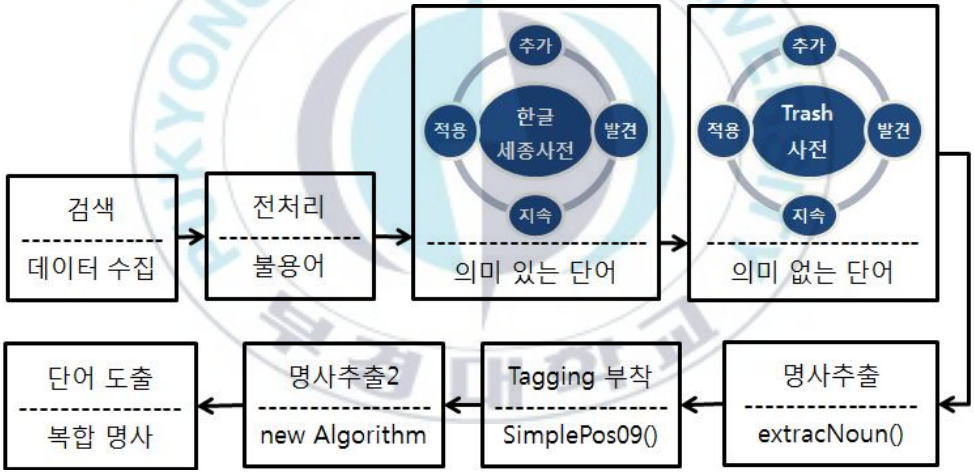
2. 한글 비표준어 처리

Web 2.0으로 인하여 사용자들의 콘텐츠 생성이 자유롭게 이루어지면서 표현 방법도 다양하게 발전하고 있다. 인터넷어, 채팅어, 새롭게 등장하는 신조어 등 모바일 환경의 영향으로 한글 자음과 모음을 이용하여 이모티콘화된 표현과 말 줄임 표현들은 우리 생활에 사용되고 있다(Lee et at., 2016).

따라서 한글은 모음과 자음의 음소(音素)를 음절 단위로 묶어 다시 한

글자로 만들어지며, 한글 자연어처리는 많은 한계점이 있다는 의견이 많다. 네티즌들이 사용하는 사용 후기나 댓글들은 표준어 이외의 단어들이 더 많이 등장하고 있어 한글 자연어처리는 더욱 연구 가치를 더해가고 있다 (Kim et al., 2014; Yun, 2008; Hong and Cha, 2013; Sim, 2011; 박소연, 2013; 안정국・김희웅, 2015).

표준어를 상대적으로 많이 사용하는 뉴스 기사에 비해 관련 기사에 네티즌들의 의견을 표현하는 댓글은 비표준어들이 상대적으로 많이 사용되어 텍스트의 분석의 신뢰성을 높이고자 [그림 3-5]과 같은 프레임워크를 설계하였다.



[그림 3-5] 한글 비표준어 처리 프레임워크

전처리 과정은 불용어나 웹 크롤링 과정의 웹 언어들이 포함되어 있어서 명사 처리 전 제거하였다. “″ ”, “%lsquo;”, “·”, ““”, “””, “’” 등의 웹 문서 상에서 특수문자 코드들이 텍스트 사이에 함께 위치하다보니 관련 특수문자들을 제거한다.

국립국어원 언어정보나눔터에서 제공하는 우리말 사전은 87,007 단어를

기본으로 제공하고 있다. 즉, 87,007 단어는 명사 처리되어 텍스트 마이닝 연구에 사용되는 단어이기도 하다. 하지만 급격히 변화해가는 한글 표현과 산업 발달로 외래어 사용도 점점 많아지면서 표준어에 등록된 명사 말뭉치도 마이닝 연구에 한계를 보이고 있다. 본 연구에서는 실시간 기사 및 댓글분석을 하기 위해 시스템을 개발하고 있다. 따라서 연구자 또는 사용자가 사용하는 과정에 필요에 의해서 새로 등장한 단어를 의미 있는 단어 사전에 추가하여 사용할 수 있는 환경을 만들었다. 다시 말하면 기존 세종사전에 단어를 추가하는 기능을 함께 개발하였다. 의미 있는 단어가 “발견”되면 “추가” 과정을 거쳐 새로이 추가된 단어를 “적용”하여 다시 마이닝 처리한다. 또한, 이후 처리에 “지속”적으로 추가된 단어는 명사 처리 사용이 가능하다. 실시간 분석 시스템이 가동하는 과정에서 발생한 추가 명사 단어 리스트는 본 연구 논문 후미에 부록 처리하여 이후 연구자에게 도움이 되고자 한다.

필요하고 의미 있는 단어를 실시간 추가하는 것도 중요하지만 의미 없는 단어의 리스트를 확보하여 마이닝 처리에서 제외하는 것 또한, 처리 비용을 줄일 수 있는 방법이다. 명사 추출 과정의 결과를 볼 때 의미 없는 단어를 추가하고 DB에 저장하여 지속적인 관리가 중요하다. 의미 없는 단어가 “발견”되면 “추가”과정을 거쳐 새로 추가된 의미 없는 단어를 마이닝 처리함으로서 제외된 것을 확인 및 “적용”하고 이후 연구에도 “지속”적으로 사용하고자 등록 절차를 적용하였다. 또한, 부록에 의미 없는 단어 리스트를 공유할 수 있도록 기재하였다.

선행연구에서 확인했듯이 한글 명사 처리에 비정상적 처리가 발생되는 것을 한 번 더 확인하고 2차 명사 추출과정을 거쳐 비표준어들에 대한 처리에 효율성이 높아진 것을 확인하였다(이종화・이현규, 2016). 기존 연구자들은 한 번의 명사 처리 함수를 이용하여 추출과정을 거쳐 마이닝 처리

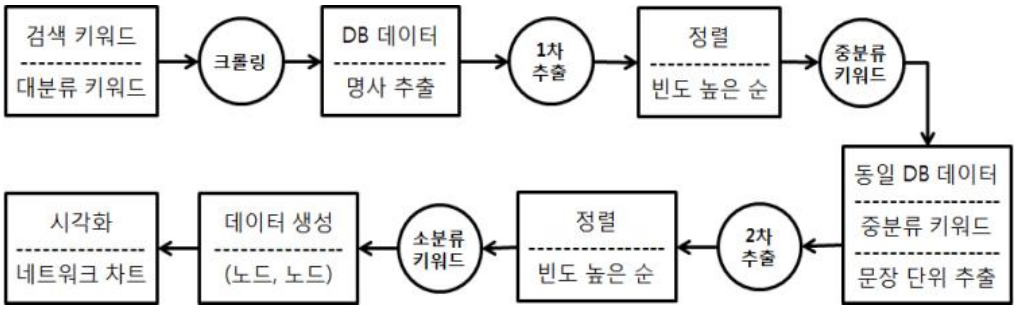
를 하고 있다. 하지만 조사들의 태킹 처리가 비정상적인 부분이 발견되어 한글 자연어 처리의 한계점들을 선행연구에서 기재하고 있다. 이러한 부분들을 해소하고자 새로 개발된 명사 처리 패키지를 적용하였다. 1차 명사들을 다시 한 번 태킹 처리하여 결과를 분석하여 다시 분리 작업을 진행함으로서 그 효율성을 확인하였다.

불용어와 웹 크롤링 과정의 특수문자 표시 Entity 문자 제거, 실시간 업데이트된 한글 세종 사전, 의미 없는 단어들 리스트인 Trash 사전, 명사추출 이후 태킹을 사용하여 한 번 더 명사 추출을 통해서 보다 양질의 명사 추출 과정을 통하여 기사 및 댓글 분석 과정을 진행하고자 한다.

3. 웹 마이닝 디자인

가. 빈도 분석

인터넷 신문사는 지난 10년간 급성장하면서 다양한 분야의 뉴스들을 업로드하고 있다. 수많은 단어들의 조합으로 이루어지는 문장들 사이에서 서로 관련된 것들을 찾고 전달하고자 하는 것들이 등장하는 단어의 빈도와 해당 단어의 2차 연관된 단어 빈도를 함께 보면서 뉴스의 전달력을 한 눈에 볼 수 있을 것이다.



[그림 3-6] 빈도 분석 활용한 네트워크 차트

[그림 3-6]은 크롤링의 기준이 되는 검색 키워드를 중심으로 “대분류”로, 대분류 기준으로 많이 등장한 단어들을 “중분류”로, 다시 중분류 키워드를 기존 크롤링된 데이터에서 다시 재검색하여 각 각의 중분류 기준에 많이 등장하는 단어들을 “소분류”로 구분하여 네트워크 차트화한 빈도 분석 기법이다. 단순 1차원적 검색 데이터에서 텍스트 마이닝 결과의 단어들을 나열한 것이 아니라 동일 데이터에서 1차 마이닝 결과 키워드를 기준으로 다시 마이닝 처리하여 대분류와 관련된 단어들의 패턴 분석에 유의미한 키워드 해석이 보다 편리하도록 처리하였다.

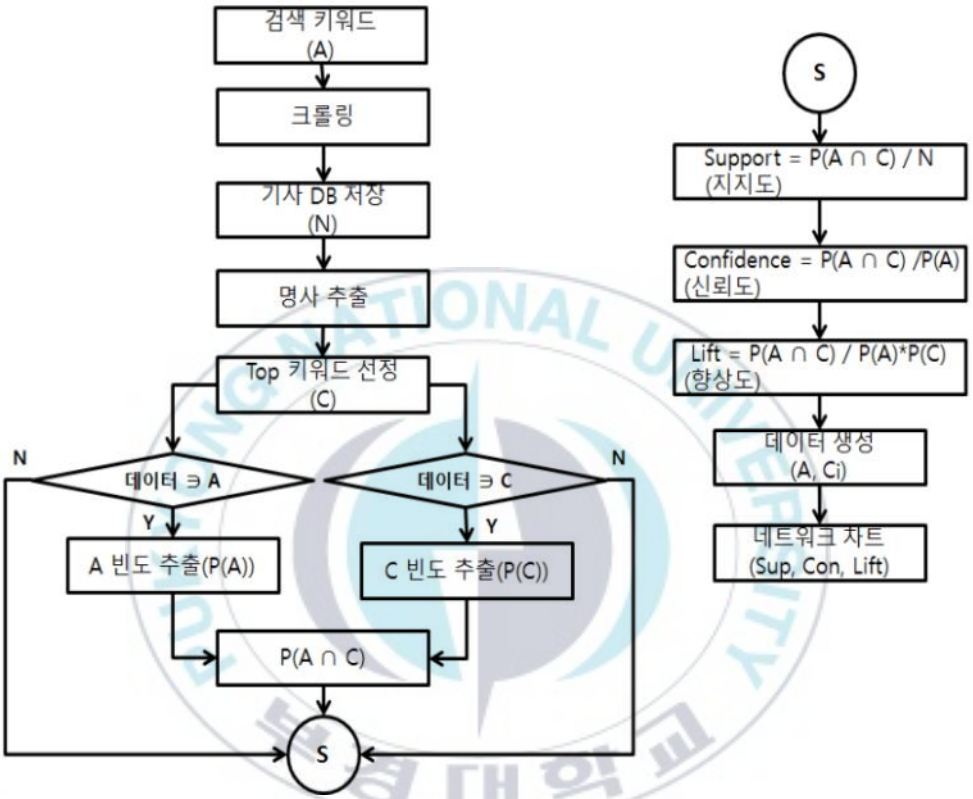
나. 연관 분석

연관성 규칙은 장바구니 분석이라 불린다. 손님의 장바구니에 있는 품목 간의 관계를 알아보는 것처럼 “A제품을 구매한 고객이 C제품을 같이 구매한다.”라는 것이 연관성 분석이다. 이를 통해 우리는 상품간의 관계를 찾아내서 세트메뉴 구성 등 마케팅 전략에 활용한다. 본 연구는 텍스트 기반 뉴스를 연관 분석에 적용하였다. 고객의 장바구니는 인터넷 뉴스 기사로 치환하고 기사들의 단어를 추출하여 장바구니 상품으로 대체하였다.

[그림 3-7]의 연관 분석 프레임워크를 살펴보면 검색 키워드를 중심으로 빈번히 등장하는 키워드들에 대하여 지지도, 신뢰도, 향상도를 측정하였다. 크롤링 데이터를 문장 단위로 분리하여 Top 키워드 빈도가 있을 경우 P(C)값을 증가시키고 해당 문장에 검색키워드의 빈도가 있을 경우는 P(A)값을 증가시킨다. 물론, 해당 문장에 검색키워드와 Top 키워드가 모두 존재하면 P(A∩C)값을 증가시킬 것이다. 선행연구에서도 제시된 지지도, 신뢰도, 향상도를 네트워크 차트에 도식화 하였다.

뉴스 기사 단위 분석이 아니라 문장 단위 분석을 통하여 검색키워드와 Top 키워드들 간의 연관성을 살펴보고자 설계한 연관분석은 다양한 패턴

을 찾아내고 유의미한 키워드들 간의 상관관계를 나타내어 유용한 정보를 제공하고자 설계하였다.



[그림 3-7] 연관 분석 프레임워크

다. 군집 분석

데이터의 특성에 따라 배타적인 여러 가지 집단으로 나누는 것을 의미하며 자료의 개수나 구조에 대한 가정 없이 데이터로부터의 거리 기준에 의해 군집화를 유도한다. 데이터 간 유사성이나 근접성을 측정해 어느 군집으로 묶을 수 있는지 판단해야 하는데 두 점 사이의 거리를 계산하는 유클리드 거리(Euclidean Distance) 유사성 측도를 사용하였다. 관측 값이 서로

얼마나 유사한지 또는 유사하지 않은지를 측정하는 식(4)에 의해 처리하였다.

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \quad (4)$$

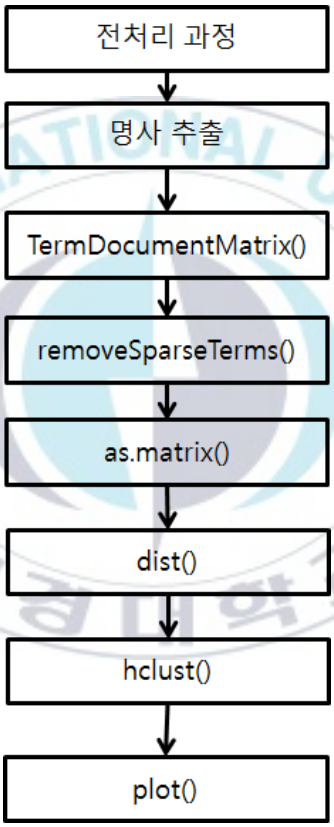
두 데이터 간 거리 측정치를 사용하여 여러 개의 군집으로 시작해 점차 군집의 개수를 줄여나가는 방법인 계층적 군집(Hierarchical Clustering) 방법으로 군집 간 정보의 손실을 최소화하며 편차들의 제곱의 합을 고려하여 군집 내 거리를 최소화하는 와드 연결법(Ward's method)을 사용하였다. 와드 연결법은 군집 분석의 각 단계에서 데이터들을 하나의 군집으로 묶음으로써 생기는 정보의 손실을 군집의 평균과 데이터들 사이의 오차제곱합(SSE)으로 식(5)과 같다.

$$SSE_i = \sum_{j=1}^{N_i}\sum_{k=1}^{p}(X_{ikj} - \overline{X}_{ik})^2 \quad (5)$$
$$SSE = \sum_{i=1}^{g} SSE_i = \sum_{i=1}^{g}\sum_{j=1}^{N_i}\sum_{k=1}^{p}(X_{ikj} - \overline{X}_{ik})^2$$

SSEi는 군집 i의 ESS를 뜻하며 Xijk는 현 단계에 있는 g개의 군집 중 Ni개의 개체를 포함하고 i번째 군집에서 j번째 개체의 k번째 변수에 대한 측정값을 의미한다.

[그림 3-8]은 군집분석의 프레임워크이며 전처리 과정과 명사 추출 과정을 거친 후 단어들의 2차원 테이블 형태로 저장 한다(TerDocumentMatrix). 행렬 구조에서 단어 빈도가 “0”인 데이터들을 불필

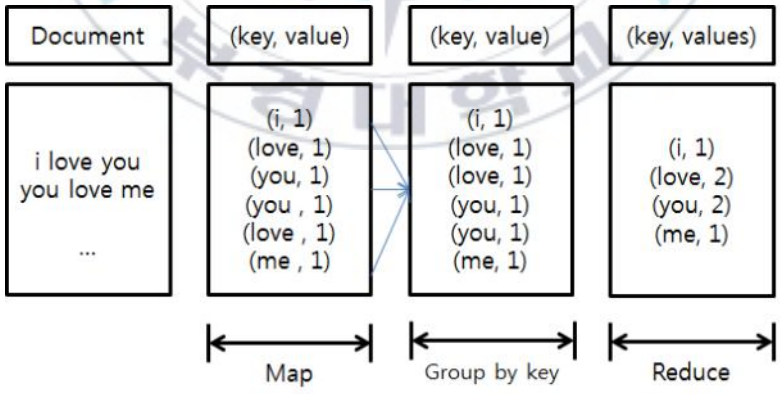
요한 반복을 제어하기 위하여 제거한다. 즉 removeSparseTerms()는 사용되는 않은 항(term)들을 제거하기 위한 함수로 활용하였다. 벡터 데이터를 행렬 구조로 변환하는 작업(as.matrix)을 진행하고 거리를 측정하기 위한 dist()함수를 수행하였다. 유사성 측도로 유클리드(Euclidean distance)거리와 Ward 연결법을 적용하여 군집화과정(hclust)을 거쳤다.



[그림 3-8] 문서 군집 분석 알고리즘

4. RHadoop 디자인

R은 기본적으로 모든 데이터 셋을 메인메모리에 올려놓고 작업하기 때문에 작업 성능이 뛰어나다. 하지만 데이터 크기가 커지면 한계로 작용되고 데이터를 한꺼번에 처리하지 못하는 경우도 발생한다. 맵리듀스(MapReduce) 프레임워크는 대용량 데이터를 분산 처리하기 위한 목적으로 개발된 프로그래밍 모델이다. Google에 의해 고안된 맵리듀스 기술은 대표적인 대용량 데이터 처리를 위한 병렬 처리 기법의 하나로 최근까지 많은 주목을 받고 있다. 맵리듀스는 임의의 순서로 정렬된 데이터를 분산 처리(Map)하고 이를 다시 합치(Reduce)는 과정을 거친다. [그림 3-9]는 텍스트 문서를 맵리듀스 과정을 보여주는 것으로 텍스트 문서를 입력 받아 key-value 쌍으로 생성하는 Map과 키별로 그룹화 하여 해당 key에 속하는 모든 값을 수집하여 출력하는 과정이다.



[그림 3-9] 맵리듀스의 단어 빈도수

RHadoop은 하둡 플랫폼 위에서 R 프로그램을 수행하는 패키지이다. mapreduce()함수를 사용하여 key와 value 형태의 데이터를 입력받고 출력하는 기능이다. [그림 3-10]의 Map의 알고리즘에서 key는 입력 문서이며 value는 문서내의 텍스트의 입력으로 설계되었다. 또한, 단어 빈도수의 결과를 나타내는 Reduce의 알고리즘에서 key는 단어를 나타내고, values는 key의 빈도수를 나타낸다.

```
map(key, value):
    for each word w in value:
        emit(w, 1)
-------------------------------
// key: document name
// value: text of the document
```

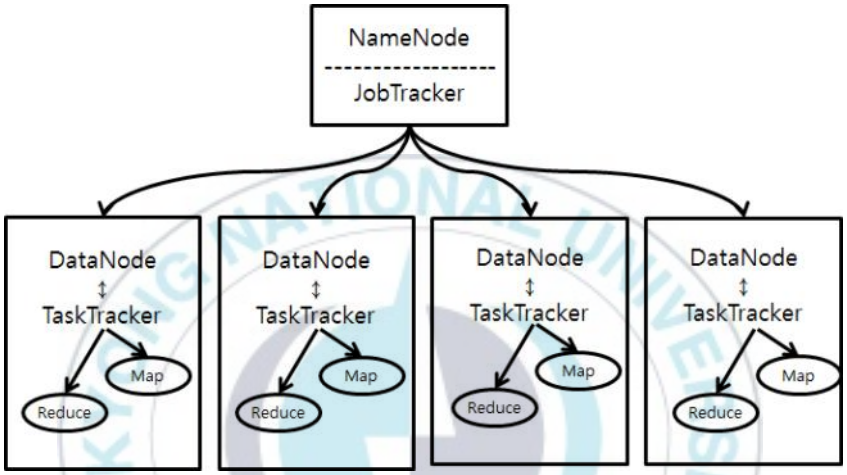
[그림 3-10] Map Algorithm

[그림 3-11]은 Map의 알고리즘에서 key는 입력 문서이며 value는 문서내의 텍스트의 입력으로 설계되었다. 또한, 단어 빈도수의 결과를 나타내는 Reduce의 알고리즘에서 key는 단어를 나타내고, values는 key의 빈도수를 나타낸다.

```
reduce(key, values):
    result = 0
    for each count v in values:
        result += v
    emit(key, result)
-------------------------------
// key: a word
// values: an iterator over counts
```

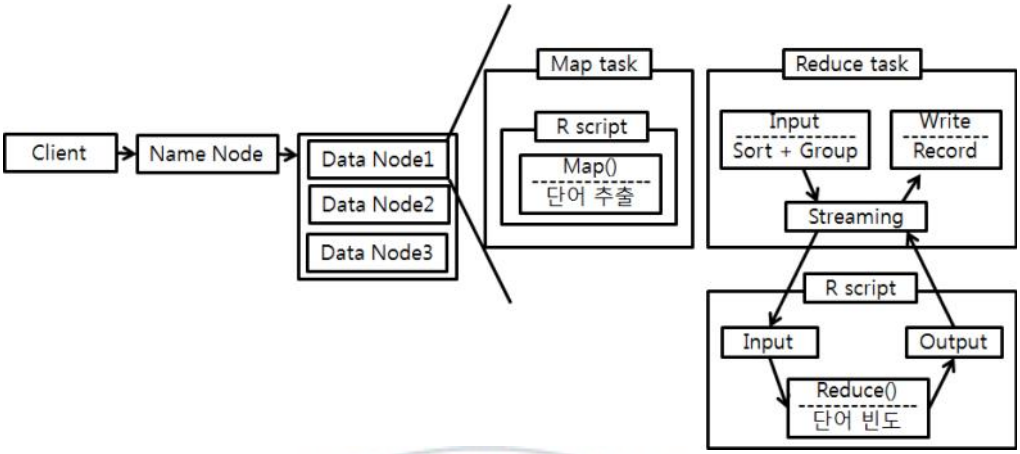
[그림 3-11] Reduce Algorithm

[그림 3-12]은 JobTracker와 TaskTracker의 상호작용으로 사용자가 JobTracker에 데이터 처리를 요청하면, JobTRacker는 해당 작업을 분할하고 서로 다른 map와 reduce 작업을 구성한 뒤 클러스터에 있는 각각의 TaskTracker에 할당하는 구조를 나타낸 것이다.



[그림 3-12] NameNode와 DataNode 구조

[그림 3-13] 은 Hadoop 시스템에 단어 빈도를 계산하는 프레임워크이다. 클라이언트의 단어 빈도 요청에 의하여 네임노드는 데이터노드에 Map Task와 Reduce Task과정을 수행한다. 먼저 Map Task는 [그림 3-9]의 예문처럼 “i love you. you love me.”의 문장에서 각 단어 추출을 목적으로 사용한다. (i, 1), (love, 1), (you, 1), (you, 1), (love, 1), (me, 1)처럼 단어를 추출한다. Reduce Task로 들어온 단어 추출 데이터는 Stream의 정렬 과정을 거쳐 같은 단어별 그룹을 형성하게 한다. R script에 의해서 reduce() 함수는 그룹화 된 단어들을 병합하여 단어의 빈도수를 계산하며 HDFS(Hadoop Distributed File System)에 빈도수를 저장한다.



[그림 3-13] RHadoop 프레임워크

제 4 장 실험과 결과

제 1 절 실험 데이터

본 연구는 국내 최대 인터넷 포털 사이트인 네이버를 대상으로 실험 데이터를 수집하고자 한다. 네이버는 정확한 검색을 통하여 최적의 정보를 전달하고자 다양한 콘텐츠를 제공하며 모바일 환경에서도 새로운 인터넷 경험을 이끌어내고 있다(www.navercorp.com). 1999년 서비스를 시작으로 4,200만명의 회원들이 서비스를 즐기고 있고 모바일 환경에서도 하루 평균 2,600만건의 조회 수를 기록하고 있는 한국의 최대 포털 사이트이다. 또한, 검색뿐만 아니라 메일, 뉴스, 증권, 지식 쇼핑, 지도, 엔터테인먼트 등 생활의 편리함을 제공하여 지속적인 콘텐츠 보급으로 많은 사랑을 받고 있다. 이러한 포털에서 제공하는 뉴스 기사를 대상으로 웹 마이닝을 처리하고자 한다.

네이버 뉴스는 정치, 경제, 사회, 생활/문화, IT/과학, TV연예, 세계, 오피니언, 이슈, 스포츠 등으로 나누어져 다양한 분야별 실시간 뉴스를 제공한다. 또한, 다양한 분야만큼이나 국내 언론 및 방송사들의 뉴스들을 제휴사로 정하고 있다. [표 4-1]은 포털 사이트에 제휴한 언론사 목록이면서 본 연구의 실험 데이터를 제공하는 리스트이기도 하다. 종합 중앙지를 비롯하여 방송사를 포함하여 130여 곳의 국내 인터넷 서비스를 제공하는 언론사의 뉴스 기사를 대상으로 실시간 웹 마이닝 분석을 실시한다.

[표 4-1] 웹 마이닝 실험 데이터 리스트

분 야	언 론 사			
종합(10)	경향신문 서울신문 한겨레	국민일보 세계일보 한국일보	동아일보 조선일보	문화일보 중앙일보
방송/통신(14)	뉴스1 채널A MBC 뉴스 TV조선	뉴시스 한국경제TV MBN YTN	연합뉴스 JTBC SBS CNBC	연합뉴스TV KBS뉴스 SBS뉴스
경제(9)	매일경제 이데일리 헤럴드경제	머니투데이 조선비즈	서울경제 파이낸셜뉴스	아시아경제 한국경제
인터넷(5)	노컷뉴스 프레시안	데일리안	미디어오늘	오마이뉴스
IT(6)	디지털데일리 전자신문	디지털타임스 ZDNetKorea	블로터	아이뉴스24
스포츠/연예(48)	게임메카 데일리e스포츠 마이데일리 스포츠경향 스포츠조선 아이즈ize 엠파이트 일간스포츠 텐아시아 풋볼리스트 KBO OBC TV	골닷컴 디스이즈게임 몬스터짐 스포츠동아 스포츠타임스 앳스타일 윈터뉴스코리아 점프볼 티비데일리 헤럴드POP KBS 연예 OSEN	골프다이제스트 디스패치 베스트일레븐 소프츠서울 스포탈코리아 엑스포츠뉴스 인벤 조이뉴스24 포모스 enews24 MBC연예 SBS funE	뉴스엔 마니아리포트 스타뉴스 스포츠월드 스포츠비뉴스 엠스플뉴스 인터풋볼 테니스코리아 포포투 JTBC GOLF MK스포츠 TV리포트
매거진(16)	뉴스위크한국판 시사IN 이코노미스트 주간조선	레이디경향 신동아 일다 중앙SUNDAY	매경이코노미 씨네21 주간경향 한겨레21	머니S 월간 산 주간동아 한경비즈니스
지역(3)	강원일보	매일신문	부산일보	
전문지(8)	기자협회보 코리아타임스	여성신문 코리아헤럴드	조세일보 코메디닷컴	참세상 헬스조선
포토(4)	신화사 연합뉴스	포토친구	AP연합뉴스	EPA연합뉴스
기타(3)	정책브리핑	코리아넷	성명자료실	

포털 사이트에서 뉴스 카테고리 내에서 원하는 키워드 및 단어들을 검색했을 때 [표 4-1]의 언론사 리스트의 뉴스 기사가 검색되어 나타난다. 검색 결과 내에 기간을 설정하여 방대한 뉴스 기사에서 기사 탐색을 1차 축소하여 검색한다. 키워드 관련 뉴스 기사에 대한 단어 빈도 및 시각화 작업을 진행하여 의미 있는 단어 리스트를 확인할 수 있을 것이다.

또한, 같은 사건 사고 및 이슈를 다양한 방법으로 표현한 뉴스 기사이지만 네티즌들의 많은 관심과 의견들을 표현한 이슈 기사를 확인하고 댓글과 댓글에 참여한 남녀 비율(남/여)과 연령대(10대, 20대, 30대, 40대, 50대) 통계를 함께 연구 대상으로 포함하였다.

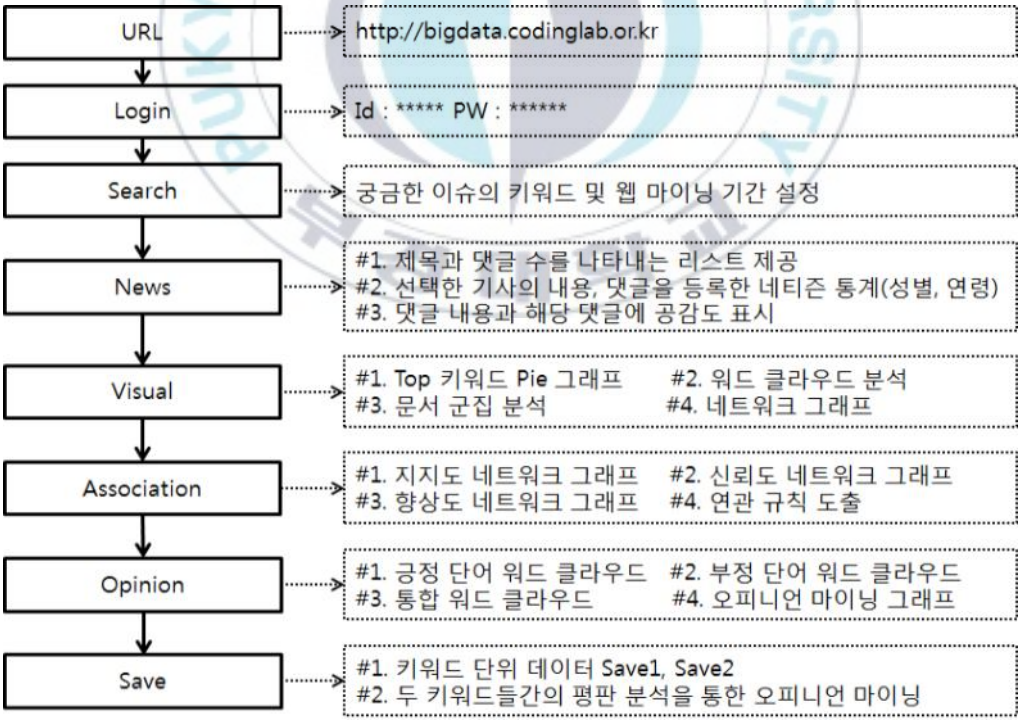
사실이나 사건을 필진의 중립적인 자세로 표현한 뉴스 기사이지만 네티즌들은 글을 읽고 느낀 감정 또는 기사의 내용에 동의하거나 그렇지 않거나를 다양한 의견들을 댓글들로 표현한다. 한 네티즌이 뉴스 기사를 읽고 자유롭게 표현한 글을 이후 네티즌이 이전 네티즌의 댓글에 “공감”이나 “비공감”을 표현하여 호감도를 나타내기도 한다. 댓글과 함께 해당 댓글의 호감도를 이번 연구 대상에 함께 포함시켜 웹 마이닝 처리를 하였다.

매일 24시간 실시간으로 업로드 되는 뉴스 기사, 130여개의 정치, 경제, 사회, 문화, IT, 전문지 등 다양한 언론사들이 경쟁하듯 앞 다투어 핫뉴스를 기고하고 많은 독자들을 확보하기 위한 노력을 진행하고 있다. 이렇듯 본 연구는 유래 없이 웹 환경에 있는 방대한 자료들을 웹 마이닝하여 실시간 자료 수집, 마이닝 처리, 다양한 시각화, 예측 과정까지 연구자의 개입 없이 시스템에서 진행하였다. 수집된 방대한 데이터의 빠른 처리를 위해 분산 처리 시스템인 하둡을 활용하고 마이닝 분석 툴인 R 프로그래밍을 통하여 처리하였다. 무엇보다 연구실에서의 실증연구가 아닌 실제연구로 사용자 인터페이스 및 전체 처리 과정을 현장 연구로 진행하여 보다 다양한 데이터의 패턴과 화려한 시각화를 구현하였다.

제 2 절 실험 설계

본 연구는 130여개의 뉴스 기사를 온라인 실시간 웹 마이닝 처리를 통하여 텍스트 마이닝, 오피니언 마이닝, 연관 규칙 분석 등의 결과를 도출하고자 한다. 또한, 방대한 데이터를 4대의 서버 컴퓨터 네트워크 연결을 통하여 하둡 시스템을 설정하였고 텍스트 마이닝 실시간 처리를 위하여 RHadoop을 추가 탑재하여 데이터 전처리 과정과 웹 마이닝 분석 과정을 진행하였다.

본 실험 과정은 웹 페이지 구축을 통하여 누구나 쉽게 사용할 수 있도록 그래픽 처리를 하였으며 특정 권한이 있는 계정을 이용하여 서비스 제공을 하고자 설계하였다.



[그림 4-1] 실시간 크롤링을 통한 웹 마이닝 프레임워크

[그림 4-1]의 실시간 크롤링을 통한 웹 마이닝 프레임워크를 살펴보면 다음과 같다. 해당 URL 접속을 통하여 특정 계정으로 로그인을 확인한다. 사회적 이슈로 궁금한 사건의 주요 키워드와 검색 기간을 설정 후 검색을 실행하면 해당 뉴스 기사 크롤링 작업을 수행 한다. 뉴스 크롤링 작업은 PHP 프로그램을 활용하여 제작하였다. 작업이 완료되면 주 메뉴는 [Search], [News], [Visual], [Association], [Opinion], [Save]로 구성되어 있다. 웹 페이지에서 메뉴별 구분은 수직 스크롤 단위로 설정하였다. 즉, [Search] 메뉴가 첫 페이지이고 수직 스크롤을 하단으로 하나 내리면 [News] 페이지로 이동하면서 해당 화면을 보여준다. [Visual], [Association], [Opinion], [Save]등도 같은 방법으로 페이지 이동이 가능하며 6개의 수직 스크롤에 모든 메뉴 페이지가 나타난다. [Search] 작업 이후 [News] 페이지로 이동하면 크롤링 결과를 볼 수 있을 것이다. [News]는 3개의 수평 스크롤 단위 화면으로 구성되어 있다. 즉 [News] 페이지는 좌, 우의 이동 버튼이나 하단의 페이지 이동 단추를 사용하여 좌, 우로 페이지를 이동하면서 결과를 확인할 수 있다. [News]페이지의 첫 페이지는 뉴스 기사의 제목과 해당 기사에 누리꾼들의 의견 즉, 댓글 개수를 기록하여 리스트 제공하고 있다. 해당 기사 제목을 읽고 뉴스 내용이 궁금하면 해당 기사를 선택하면 수평 스크롤 두 번째 페이지에서 뉴스 내용과 댓글의 통계가 제공된다. 통계는 남녀 비율과 연령 즉 10대, 20대, 30대, 40대, 50대로 나누어져 통계를 나타낸다. 이후 세 번째 페이지로 이동하면 해당 댓글 리스트가 제공된다. 댓글의 내용을 바로 확인할 수 있으며 해당 댓글에 대한 또 다른 네티즌들의 공감도를 “공감” 또는 “비공감”으로 표시하여 그 개수를 함께 리스트에 추가하였다.

[Visual] 페이지는 뉴스 기사의 텍스트 마이닝 결과를 Pie그래프, 워드클라우드, 문서 군집 분석, 네트워크 차트 까지 수평 스크롤 4페이지로 구

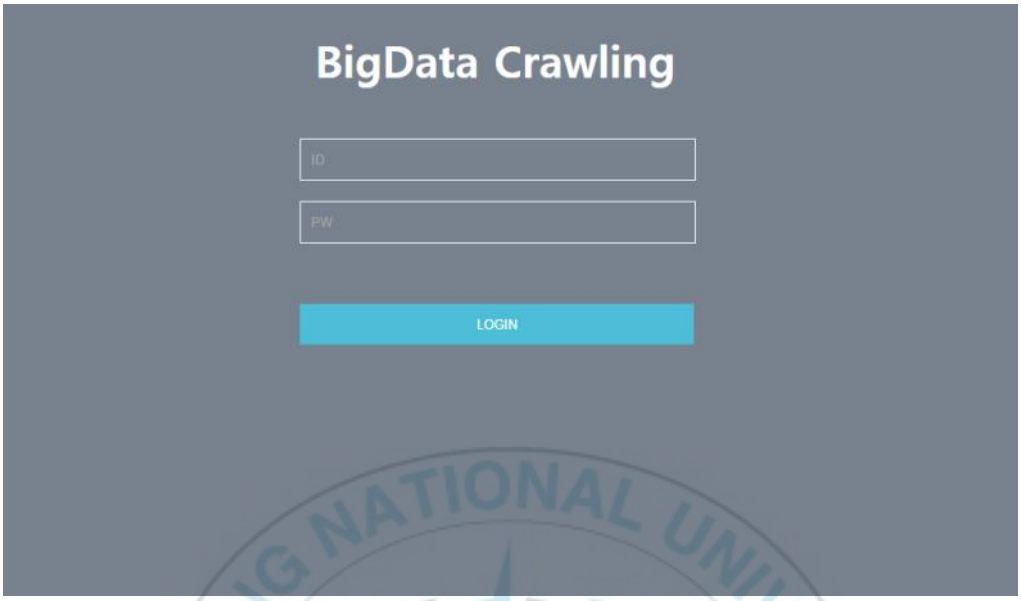
성하였다.

[Association] 페이지는 연관 규칙 분석을 통하여 키워드 간 유용한 연관 관계와 규칙을 발견하고자 한다. 뉴스 기사의 마이닝 결과를 활용하여 진행하였으며 최소 지지 확률, 최소 신뢰 확률, 양의 상관관계를 조정할 수 있는 향상 확률 등을 연구자가 직접 수동으로 수치를 입력하여 결과를 도출하고자 설계하였다. [Association] 페이지는 수평 스크롤을 3페이지로 구성하였다.

[Opinion] 페이지는 뉴스 기사 관련 댓글에 대한 누리꾼들의 평판을 긍정 사전과 부정 사전을 활용하여 오피니언 마이닝 처리를 하였다. 먼저 댓글에 긍정적 단어들을 추출하여 긍정 단어 워드 클라우드를 표시하였고, 부정적 단어 역시 추출하여 부정 단어 워드 클라우드를 표시하고, 긍정적 단어와 부정적 단어를 함께 워드 클라우드 분석으로 표현하여 대조적으로 시각화하였다. 긍정 단어와 부정 단어의 빈도가 얼마나 되는지 수치로 변환하여 Score 차트로 시각화하였다. [Opinion] 페이지는 수평 스크롤 4페이지로 구성하였다.

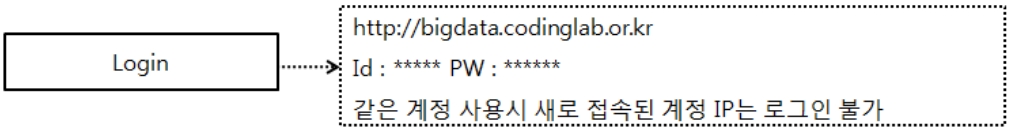
[Save] 페이지는 두 키워드들에 대한 평판 분석을 통하여 서로 비교할 수 있도록 설계하였다. 또한, 두 키워드를 한꺼번에 사용하여야 하기 때문에 자료 수집 시간 또한, 상당 시간 소요될 것으로 판단되었다. 그리하여 기존 수집된 자료를 활용하여 분석하고자 한다. 이전 궁금했던 키워드에 대한 사용자 댓글을 Save1을 통하여 저장한다. 이후 추가로 궁금했던 키워드 댓글 역시 Save2에 저장한다. 자료 수집이 완료되면 두 키워드들 간의 오피니언 마이닝 결과를 도출한다.

웹 페이지에서 구현되는 과정을 다음과 같이 나열한다. 먼저 연구 URL을 통한 웹페이지 접속 페이지는 [그림 4-2]와 같다.



[그림 4-2] LOGIN Web Page

로그인 페이지는 본 연구자가 승인한 계정만 접속 권한을 부여하여 PC 환경, 모바일 환경 등 다양한 디바이스에서 사용할 수 있는 반응형 웹으로 설계하였다. [그림 4-3]은 본 현장 연구 결과 웹 페이지의 프로세스를 설명하고 있다. 한 계정으로 사용자가 사용할 경우 크롤링 뉴스 기사나 검색 결과 값이 섞여서 처리될 수 있으므로 같은 계정으로 이중 로그인한 경우는 자동 로그아웃 되도록 설계하였다.



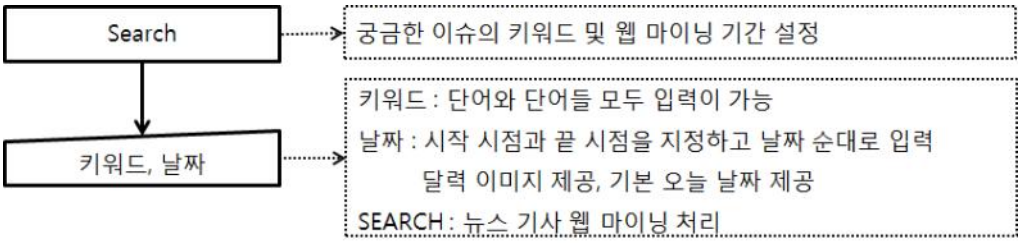
[그림 4-3] Login Web Page 프로세스

2017년 1분기는 국내외적으로 한국의 리더십 부재로 많은 일들이 있었

다. 그 중 가장 외교적으로 큰 마찰은 미국의 “사드”배치로 중국의 보복이 고스란히 경제, 문화산업, 관광산업 등의 다양한 산업에 큰 영향을 끼치고 있는 것이다. [그림 4-4]는 이러한 외교적인 이유로 “사드”라는 키워드와 연구 기간 중 특정일을 기준으로 검색을 실시하였다. 첫 텍스트 박스에는 관련 키워드, 그리고 뉴스 기사 검색 시작일과 종료일을 설정하여 SEARCH를 선택하면 관련 뉴스를 연구용 데이터베이스에 탑재한다.

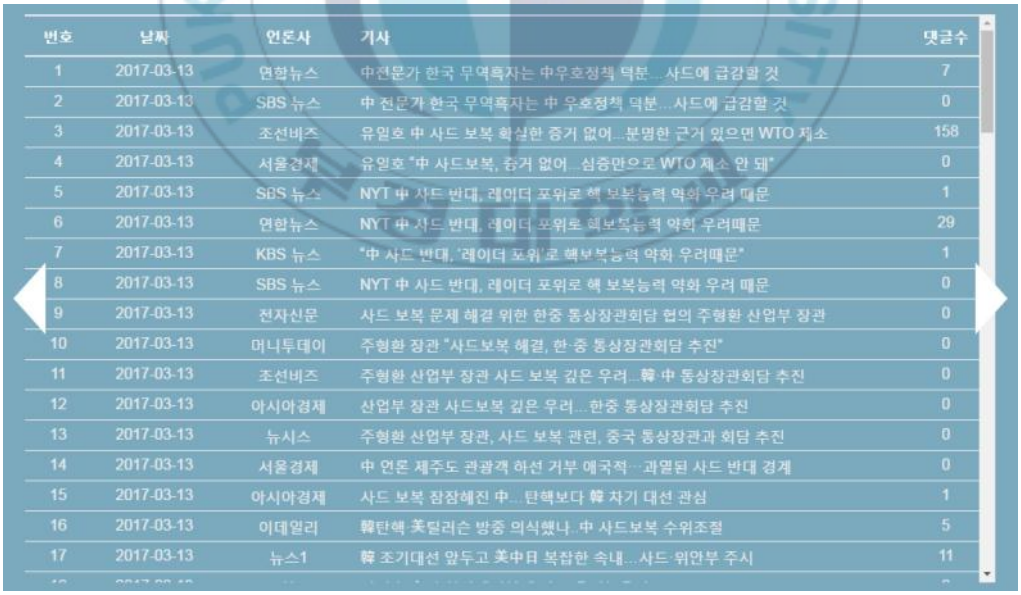
[그림 4-4] Search Web Page

[그림 4-5]는 [Search] 웹 페이지에서 구현되는 과정을 기술한 프로세스이며 키워드 관련 뉴스 기사 크롤링을 시작하는 시작점이기도 하다.



[그림 4-5] Search Web Page 프로세스

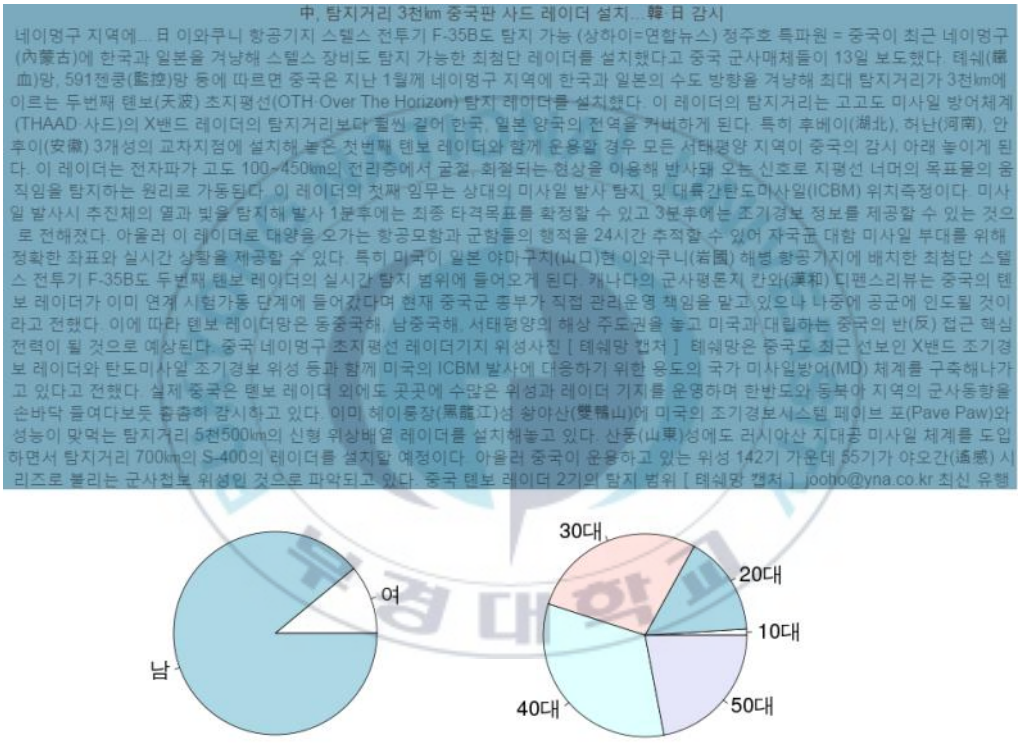
[Search] 웹 페이지에서 검색한 결과는 [News] 페이지로 이동하면 그 결과를 볼 수 있다. [그림 4-6]은 뉴스 기사 크롤링 과정이 종료된 결과 화면 중 일부를 확인하고자 삽입하였다. 뉴스 기사 순번과, 해당 뉴스가 기고된 날짜, 언론사, 기사 제목, 해당 기사에 누리꾼들의 댓글 수 등의 정보를 제공하는 첫 번째 페이지이다.



번호	날짜	언론사	기사	댓글수
1	2017-03-13	연합뉴스	中전문가 한국 무역흑자는 中우호정책 덕분…사드에 급감할 것	7
2	2017-03-13	SBS 뉴스	中 전문가 한국 무역흑자는 中 우호정책 덕분…사드에 급감할 것	0
3	2017-03-13	조선비즈	유일호 中 사드 보복 확실한 증거 없어…분명한 근거 있으면 WTO 제소	158
4	2017-03-13	서울경제	유일호 "中 사드보복, 증거 없어…심증만으로 WTO 제소 안 돼"	0
5	2017-03-13	SBS 뉴스	NYT 中 사드 반대, 레이더 포위로 핵 보복능력 약화 우려 때문	1
6	2017-03-13	연합뉴스	NYT 中 사드 반대, 레이더 포위로 핵보복능력 약화 우려때문	29
7	2017-03-13	KBS 뉴스	"中 사드 반대, '레이더 포위'로 핵보복능력 약화 우려때문"	1
8	2017-03-13	SBS 뉴스	NYT 中 사드 반대, 레이더 포위로 핵 보복능력 약화 우려 때문	0
9	2017-03-13	전자신문	사드 보복 문제 해결 위한 한중 통상장관회담 협의 주형환 산업부 장관	0
10	2017-03-13	머니투데이	주형환 장관 "사드보복 해결, 한·중 통상장관회담 추진"	0
11	2017-03-13	조선비즈	주형환 산업부 장관 사드 보복 깊은 우려…韓·中 통상장관회담 추진	0
12	2017-03-13	아시아경제	산업부 장관 사드보복 깊은 우려…한중 통상장관회담 추진	0
13	2017-03-13	뉴시스	주형환 산업부 장관, 사드 보복 관련, 중국 통상장관과 회담 추진	0
14	2017-03-13	서울경제	中 언론 제주도 관광객 하선 거부 에국적…과열된 사드 반대 경계	0
15	2017-03-13	아시아경제	사드 보복 잠잠해진 中…탄핵보다 韓 차기 대선 관심	1
16	2017-03-13	이데일리	韓탄핵 美달러은 방중 의식했나..中 사드보복 수위조절	5
17	2017-03-13	뉴스1	韓 조기대선 앞두고 美中日 복잡한 속내…사드 위안부 주시	11

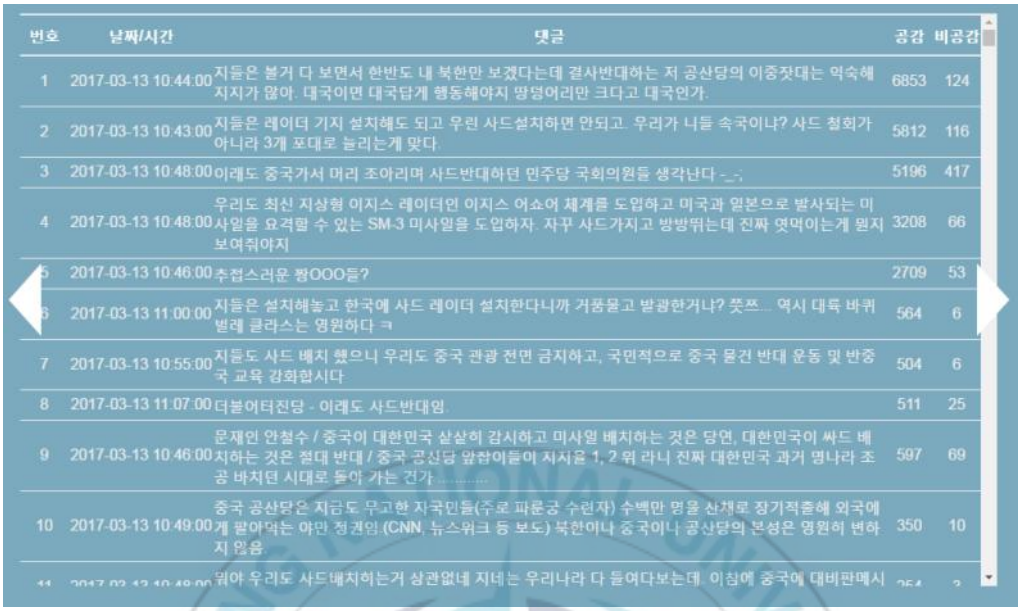
[그림 4-6] News Web Page #1

[그림 4-7]은 “2017-03-13”, “연합뉴스” 기사 중 “中, 탐지거리 3천㎞ 중국판 사드 레이더 설치…韓·日 감시” 제목으로 댓글 2,143개의 정보를 가진 뉴스 기사를 선택하였고 해당 기사의 내용과 댓글 통계 자료를 제시한다. 전쟁 무기에 관련된 뉴스 기사로 댓글을 기록한 성 비율은 남성이 월등하게 많았고, 연령은 40대, 50대가 50% 이상이라는 것을 확인할 수 있다.



[그림 4-7] News Web Page #2

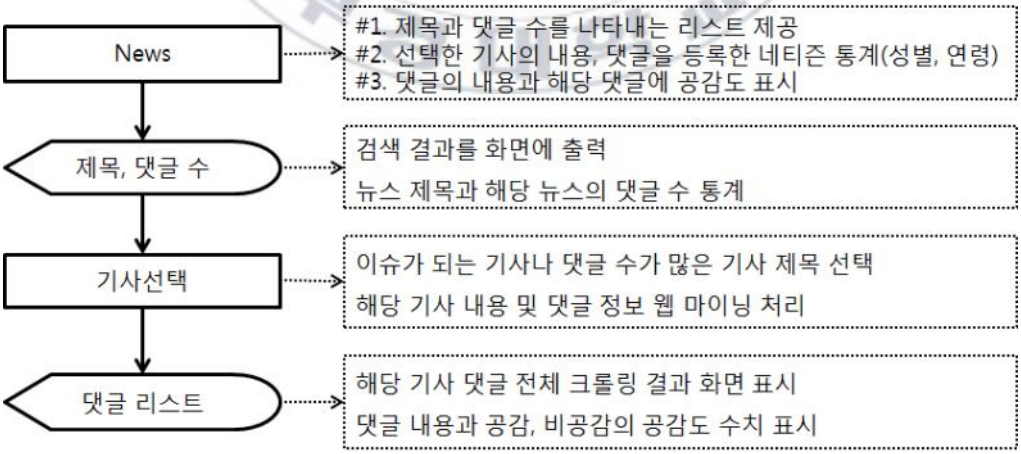
또한, [그림 4-8]은 [News]의 세 번째 페이지로 관련 댓글에 관한 정보가 나타난다. 댓글 번호, 댓글 날짜/시간, 댓글, 댓글에 대한 공감, 비공감 개수를 표현하였다.



번호	날짜/시간	댓글	공감	비공감
1	2017-03-13 10:44:00	지들은 볼거 다 보면서 한반도 내 북한만 보겠다는데 결사반대하는 저 공산당의 이중잣대는 역겨해 지지가 않아. 대국이면 대국답게 행동해야지 땅덩어리만 크다고 대국인가.	6853	124
2	2017-03-13 10:43:00	지들은 레이더 기지 설치해도 되고 우린 사드설치하면 안되고. 우리가 니들 속국이냐? 사드 철회가 아니라 3개 포대로 늘리는게 맞다.	5812	116
3	2017-03-13 10:48:00	이래도 중국가서 머리 조아리며 사드반대하던 민주당 국회의원들 생각난다 -_-;	5196	417
4	2017-03-13 10:48:00	우리도 최신 지상형 이지스 레이더인 이지스 어쇼어 체계를 도입하고 미국과 일본으로 발사되는 미사일을 요격할 수 있는 SM-3 미사일을 도입하자. 자꾸 사드가지고 방방뛰는데 진짜 엿먹이는게 뭔지 보여줘야지	3208	66
5	2017-03-13 10:46:00	추접스러운 짱OOO들?	2709	53
6	2017-03-13 11:00:00	지들은 설치해놓고 한국에 사드 레이더 설치한다니까 거품물고 발광한거냐? 쯧쯔... 역시 대륙 바퀴벌레 클라스는 영원하다 ㅋ	564	6
7	2017-03-13 10:55:00	지들도 사드 배치 했으니 우리도 중국 관광 전면 금지하고, 국민적으로 중국 물건 반대 운동 및 반중국 교육 강화합시다.	504	6
8	2017-03-13 11:07:00	더불어더진당 - 이래도 사드반대임.	511	25
9	2017-03-13 10:46:00	문재인 안철수 / 중국이 대한민국 살살히 감시하고 미사일 배치하는 것은 당연, 대한민국이 싸드 배치하는 것은 절대 반대 / 중국 공산당 앞잡이들이 지지율 1, 2 위 라니 진짜 대한민국 과거 명나라 조공 바치던 시대로 돌아 가는 건가	597	69
10	2017-03-13 10:49:00	중국 공산당은 지금도 무고한 자국민들(주로 파룬궁 수련자) 수백만 명을 산채로 장기적출해 외국에게 팔아먹는 야만 정권임.(CNN, 뉴스위크 등 보도) 북한이나 중국이나 공산당의 본성은 영원히 변하지 않음.	350	10
11	2017-03-13 10:49:00	원아 우리도 사드배치하는거 상관없네 지네는 우리나라 다 들여다보는데. 이참에 중국에 대비판매시	264	3

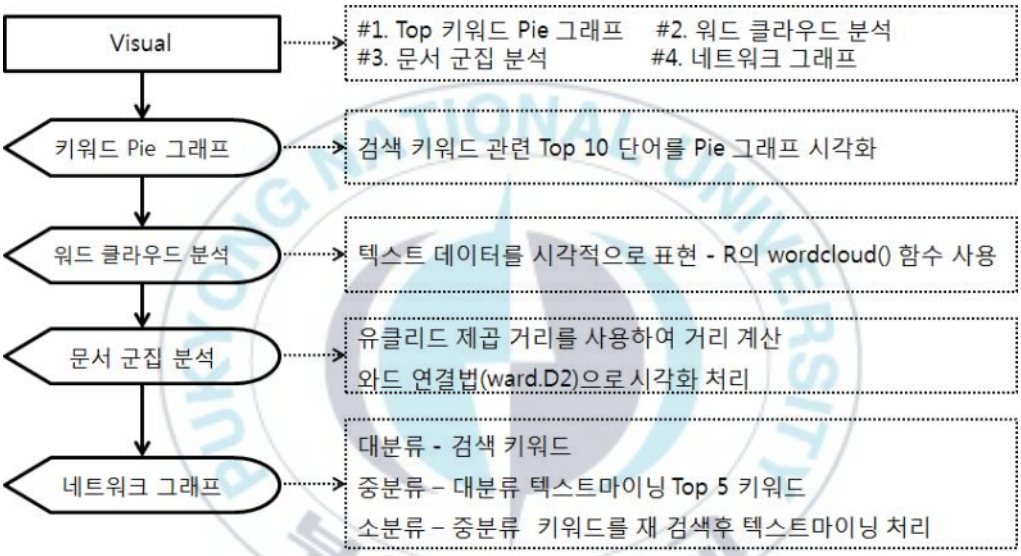
[그림 4-8] News Web Page #3

[News] 메뉴의 프로세스는 [그림 4-9]와 동일하며 관련 기사와 특정 댓글에 대한 자료 수집이 완료된 상태이다.



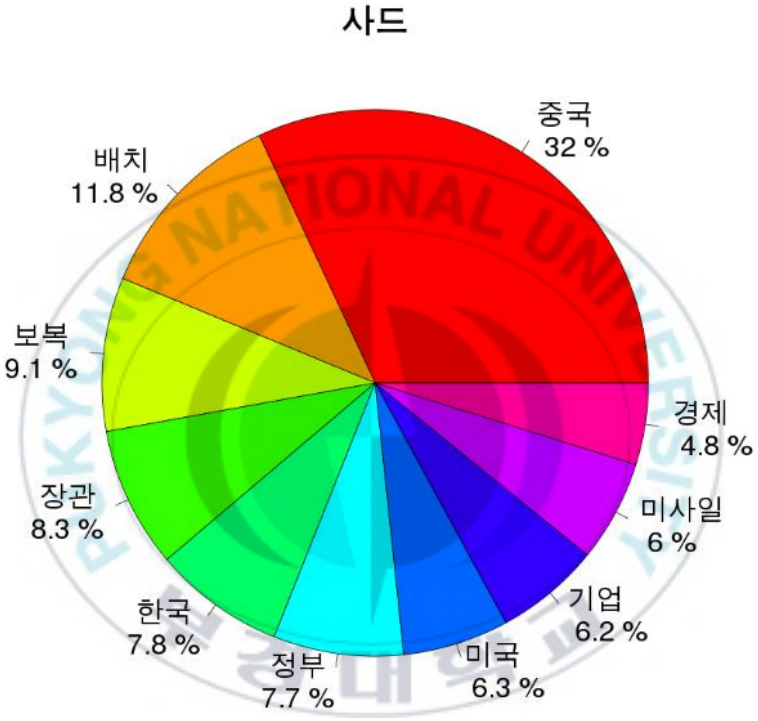
[그림 4-9] News Web Page 프로세스

[Visual] 메뉴는 뉴스 기사들에 대한 시각화 작업이 이루어지는 웹 페이지이다. 전체 4 페이지로 구성되어 있으며 [그림 4-10]의 프로세스 흐름도로 설계하였다. Pie차트, 워드 클라우드 분석, 문서 군집 분석, 네트워크 그래프 등 다양한 시각화를 통하여 키워드 패턴을 확인하고자 한다.



[그림 4-10] Visual Web Page 프로세스

[그림 4-11]은 “사드” 키워드 중심으로 검색된 뉴스 기사들에 대하여 텍스트 마이닝을 처리한 결과를 Top 10 키워드의 빈도를 활용하여 시각화한 Pie 그래프이다.

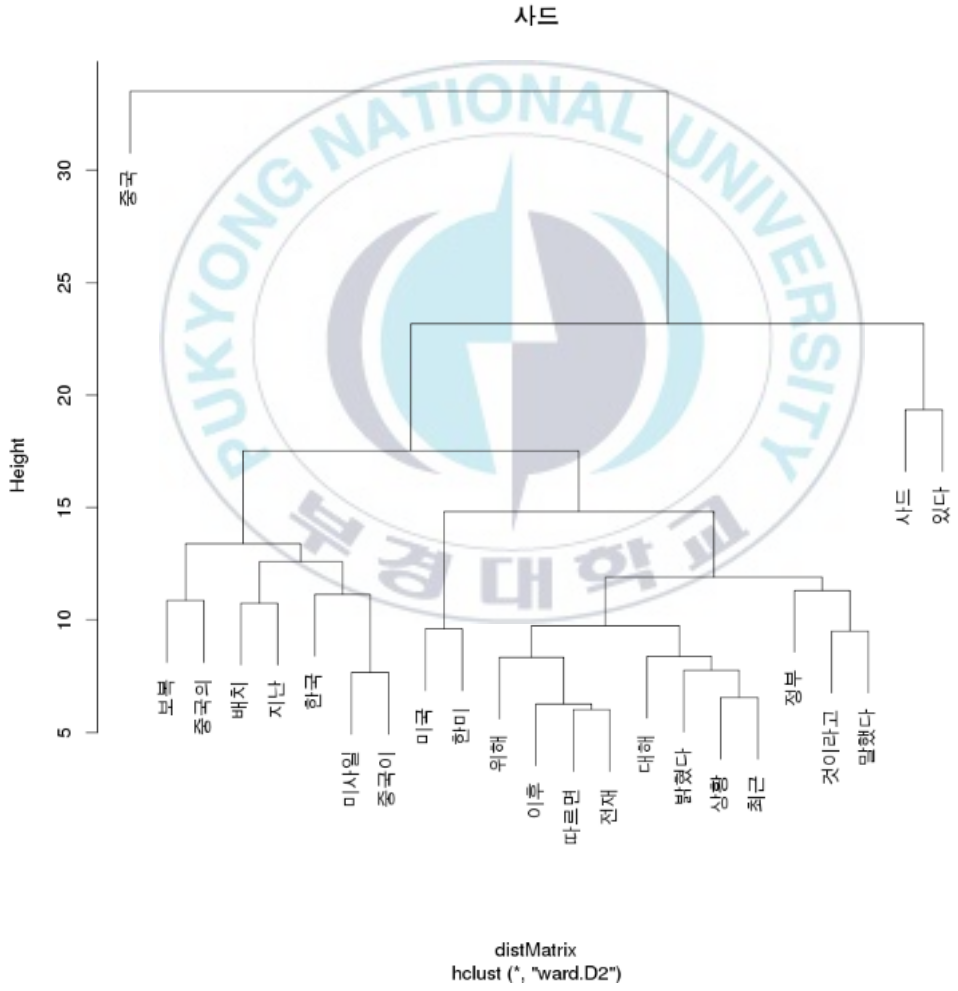


[그림 4-11] Visual Web Page #1

[그림 4-12]은 뉴스 기사에서 등장한 텍스트 단어들을 워드 클라우드 분석을 통하여 시각화한 결과이다. [그림 4-11] Pie 그래프의 Top 10 키워드가 워드 클라우드에서도 눈에 띄는 것을 확인할 수 있다.

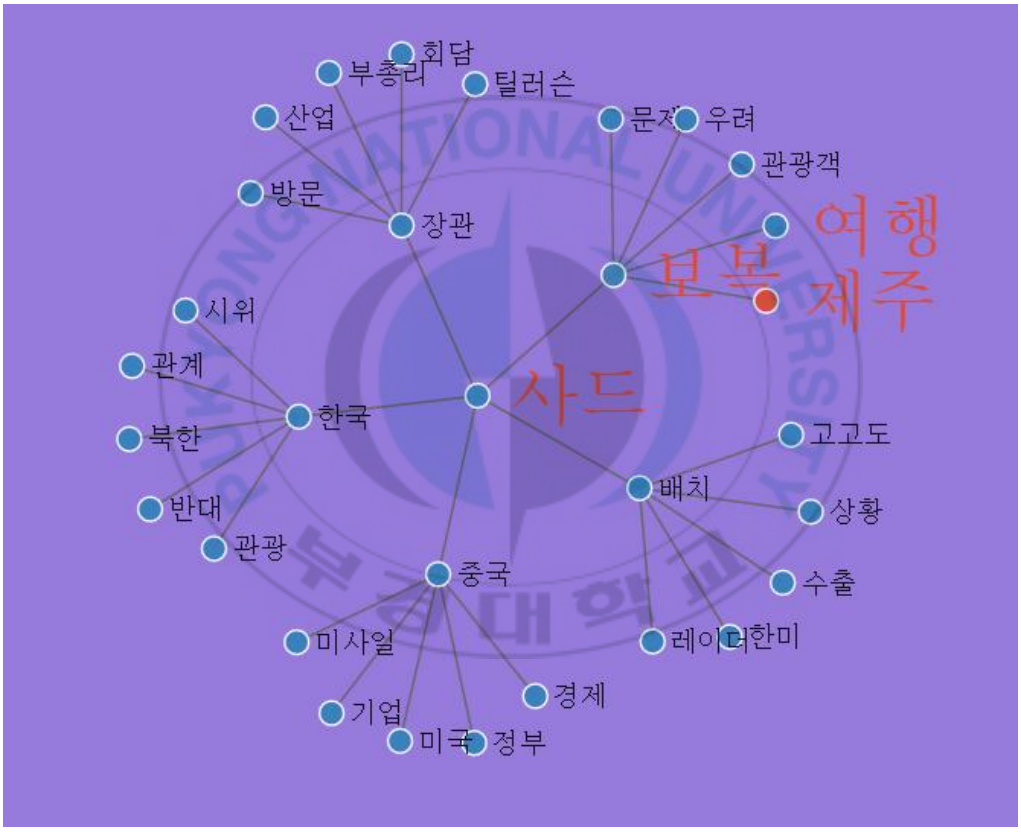
[그림 4-12] Visual Web Page #2

[그림 4-13]은 검색된 뉴스기사들을 문장 단위로 나누어 키워드간의 거리와 한 문장에 얼마나 등장하는지를 나타내는 군집 분석을 하였다. [그림 4-10] Visual Page 프로세스에서 설명하였듯이 유클리드 제곱 거리를 사용한 거리 계산과 와드연결법에 의해서 시각화한 결과인 군집분석을 나타내었다.



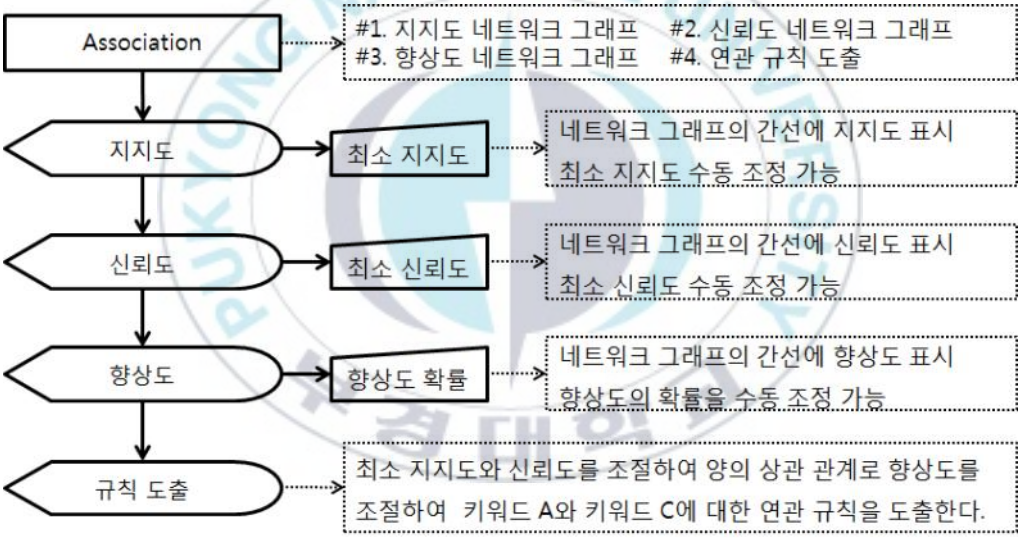
[그림 4-13] Visual Web Page #3

[그림 4-14]은 초기 검색 키워드를 중심으로 Top 키워드를 추출하고 이후 추출된 키워드들 중심으로 각각 재검색하여 다시 추출한 네트워크 그래프 결과이다. 초기 검색 키워드를 “대분류”, 대분류에서 나온 Top 키워드들을 “중분류”, 다시 중분류 기준으로 재검색하여 추출된 중분류 키워드별 Top 단어들을 “소분류”로 하여 이미지화 한 결과이다.



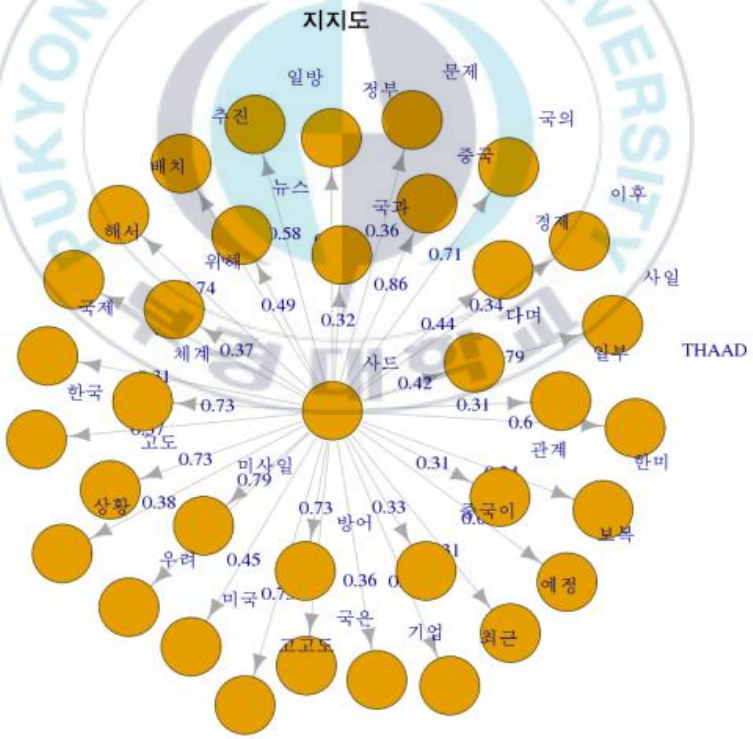
[그림 4-14] Visual Web Page #4

[그림 4-15]는 연관 규칙 분석의 프로세스로 지지도, 신뢰도, 향상도를 사용하여 키워드 A가 등장한 문장은 키워드 C가 함께 사용될 확률적인 통계를 통하여 분석하는 기법이다. 정형적 데이터 분석에서 사용되거나 장바구니 분석에서 사용되는 일반적인 방법을 본 연구에서는 기사 단위가 아니라 기사 내에서 문장 단위로 구분하여 한 문장에서 서로의 키워드 관련성이나 규칙성을 보자는 의미이다. 또한, 연구자가 판단했을 때 최소 지지도 비율을 조절하여 보다 높은 식별성과 효율적인 규칙을 찾기 위함이다.

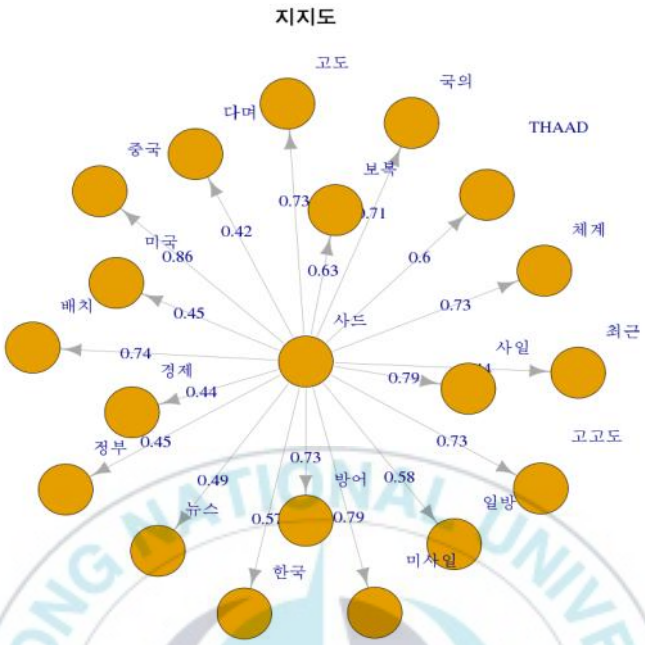


[그림 4-15] Association Web Page 프로세스

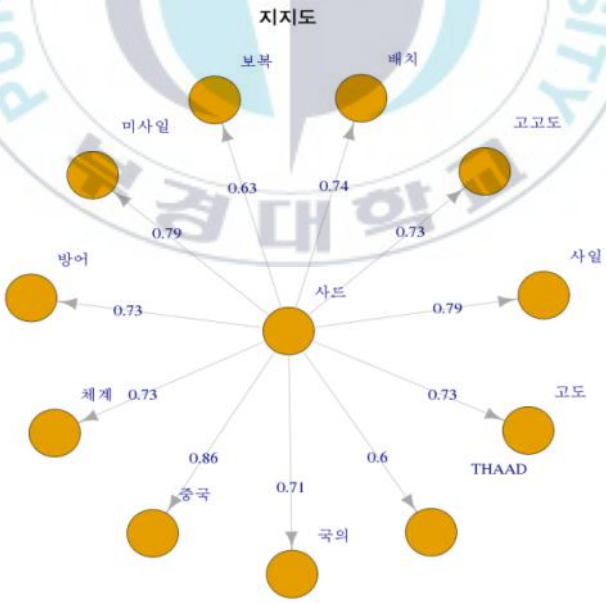
[그림 4-16]은 최소 지지도를 30%로 설정하고 연관 규칙을 확인했을 때의 결과로 검색 키워드를 기준으로 노드에 자리한 키워드와 키워드 링크에 위치한 지지도를 나타낸 결과이다. [그림 4-17]과 [그림 4-18]은 각 40%와 60%의 최소 지지도를 조절했을 때의 결과를 나타낸 그림이다. 이와 같이 연구자가 미리 설정한 기준 이하의 키워드는 제외시키고 임계값 이상에 해당되는 키워드만을 대상으로 신뢰도 및 향상도를 처리하여 세 기준을 모두 만족하는 연관 규칙을 찾는 프로세스이다. 실제 데이터로 실험을 통하여 신뢰도, 향상도 그리고, 최종 연관 규칙을 도출하는 결과는 다음 절에 함께 기술한다.



[그림 4-16] Association Web Page #1(0.3)

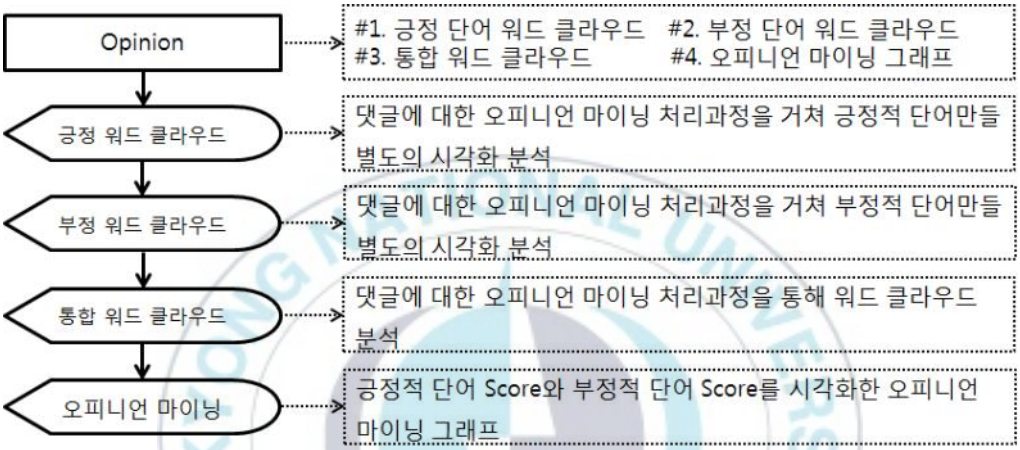


[그림 4-17] Association Web Page #1(0.4)



[그림 4-18] Association Web Page #1(0.6)

네트워크 그래프의 중심 키워드가 문장 중에 사용되었을 때 주변 키워드가 나타날 확률을 나타내는 지지도 결과이다[그림 4-16, 4-17, 4-18]. 연구 대상 문서의 문장 구조나 단어 빈도에 따라 지지도 확률을 조정하여 신속한 결과를 도출할 수 있도록 설계하였고 또한, 신뢰도, 향상도 같은 인터페이스로 설계하였다.



[그림 4-19] Opinion Web Page 프로세스

로그인을 시작으로 이슈 뉴스 키워드 검색, 관련 기사 웹 마이닝 처리, 관련 기사 시각화 처리까지 진행하였다. 또한, 관심 있는 기사의 내용과 관련된 누리꾼들의 댓글을 분석하여 여론의 방향을 보기 위한 프로세스 과정이다. [그림 4-19]는 크롤링 된 댓글의 내용에서 긍정적 단어와 부정적 단어를 비교하여 점수화 과정을 통해 긍정, 부정, 중립의 오피니언 마이닝을 진행 하였다.

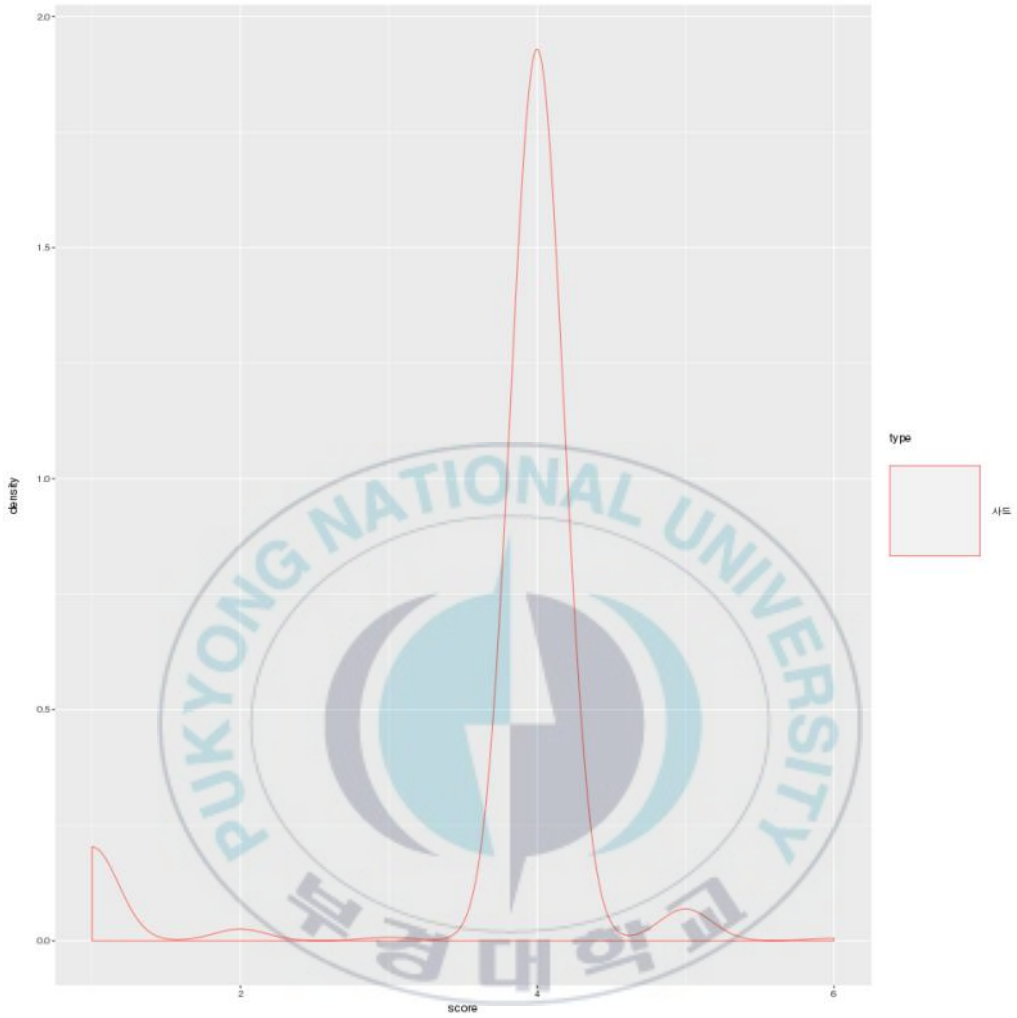
[그림 4-20]은 연구 데이터를 긍정 사전과 비교하여 뉴스 기사에 대한 국민들의 여론을 긍정적 단어들로 이미지화한 워드 클라우드이며 [그림 4-21]은 부정적 단어 사전을 사용하여 관련 단어를 추출 후 빈도 분석을 통해 워드 클라우드 분석 결과를 도출한 예이다.

[그림 4-20] Opinion Web Page #1

[그림 4-21] Opinion Web Page #2

[그림 4-22]는 [그림 4-20]과 [그림 4-21]의 도출 과정에서 추출된 긍정 단어와 부정 단어 양극단어를 두고 시각화 한 워드 클라우드이다. 단순히 두 워드 클라우드를 합병하지 않고 두 개 이상의 문서에서 단어가 사용된 상대 빈도를 비교하는 비교 클라우드라 할 수 있다.

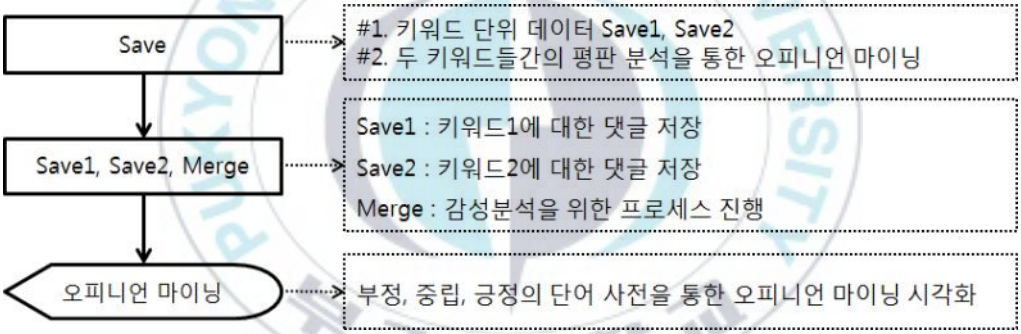
[그림 4-22] Opinion Web Page #3



[그림 4-23] Opinion Web Page #4

[그림 4-23]은 긍정 단어 사전과 부정 단어 사전을 활용하여 연구 댓글을 비교하여 긍정적 단어가 등장하면 '+1', 부정적 단어가 등장하면 '-1'로 계량화하여 나타낸 단어 Score 그래프이다. 긍정, 부정도를 시각화하여 보다 여론을 흐름을 파악하는데 도움이 되며 부록에는 연구에 사용된 긍정 사전과 부정 사전을 함께 공개하였다.

실험 비교를 통한 서로의 평판을 시각적으로 표현하는 프로세스가 [그림 4-24]이다. 두 키워드의 자료 수집과 분석 결과 도출까지의 과정을 처음부터 프로세스를 진행하기에는 많은 시간이 소요된다. 본 연구는 각기 따로 따로 결과물을 검색하여 먼저 검색한 결과물을 Save1에 저장하고 이후 검색한 키워드는 Save2에 검색 결과를 저장하게 된다. Merge를 사용하여 저장된 Save1과 Save2를 비교하여 서로의 키워드 평판을 같은 그래프에 표현하고자 하였다. 실제 연구 데이터는 뉴스 기사의 댓글을 활용한 오피니언 마이닝 프로세스이다. 다음 절의 “실험 결과”의 출력물을 제시하면서 추가 설명하고자 한다.



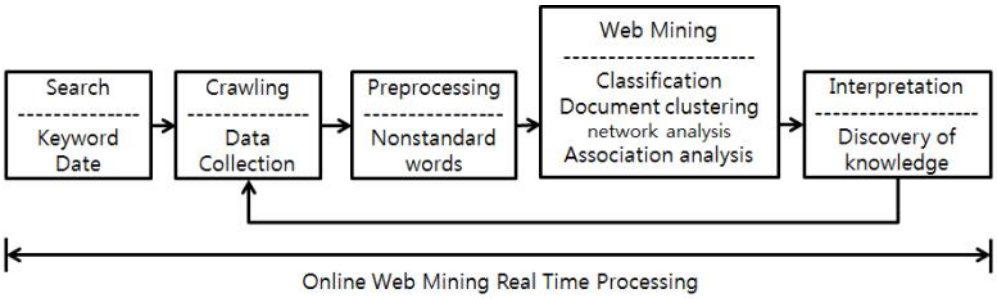
[그림 4-24] Save Web Page 프로세스

제 3 절 실험 결과

본 연구는 국내 인터넷 신문을 대상으로 이슈가 되는 사회 현안들을 키워드 중심의 관련 뉴스 기사 실시간 검색과 해당 뉴스 기사 댓글을 함께 크롤링하여 여론을 보다 빠르게 판단 및 예측하는데 도움이 되고자 설계되었다. 무엇보다 지난 2, 3년간 빅데이터에 대한 연구가 활발히 진행되며 마케팅은 물론, 금융업, 소매업, 통신, 의학, 농업 등 다양한 분야의 비즈니스에 활용되어 왔다(Boss and Mahapatra, 2001).

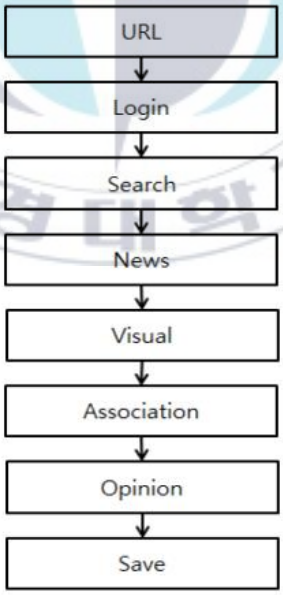
본 연구자의 연구 시점에 사회적 이슈와 정치적 이슈는 대통령 탄핵과 선거에 관한 이슈가 있고, 외교적으로는 한반도 사드(고고도미사일방어체계·THAAD) 배치 결정에 따른 중국과 미국사이의 정치, 경제, 외교적인 이슈가 가장 큰 화두이다. 먼저 중국은 한반도 사드 배치는 미국이 중국을 겨냥한 것으로 판단되어 한국에 사드 보복이란 큰 경제적인 압력을 행사하고 있다. 이런 가운데 중국에 진출한 기업들의 사업장 폐쇄와 수출된 모든 재화들에 대한 턱없이 높은 검열 문턱을 만들어 자국으로 유입되지 않도록 장벽을 치고 있는 것이다. 수출의 30%를 차지하는 주요 수출국인 중국의 사드 보복이란 이슈가 산업 전 분야에 걸쳐서 발생하고 있다. 한편 미국은 테러와의 전쟁 속에서 북한이 매일 같이 시험 발사하는 핵미사일, 장거리 미사일 등 여러 징후에 대응하기 위한 노력을 우리 정부와의 의견을 통해 사드 배치 결정 이후의 수순을 밟고 있는 것이다.

본 연구의 시험 결과에 넣을 실험 이슈는 고고도미사일방어체계 “사드”와 “중국”과 “미국”을 키워드로 선정하였다. 즉 키워드는 “사드 중국”과 “사드 미국”으로 정하고 각 두 차례의 실시간 분석 시스템을 통하여 결과를 도출하고자 한다.



[그림 4-25] 실시간 크롤링을 통한 웹 마이닝 프로세스

본 연구는 [그림 4-25]와 같이 웹 페이지에서 키워드 중심의 관련 기사를 크롤링하여 해당 기사를 텍스트 마이닝 처리와 시각화, 연관 규칙 분석을 통한 예측이 가능하며 네티즌들의 의견을 수렴하는 댓글의 빠른 분석을 통하여 여론을 읽을 수 있는 시스템 프레임워크이다.



[그림 4-26] 실시간 크롤링을 통한 웹 메뉴 프로세스

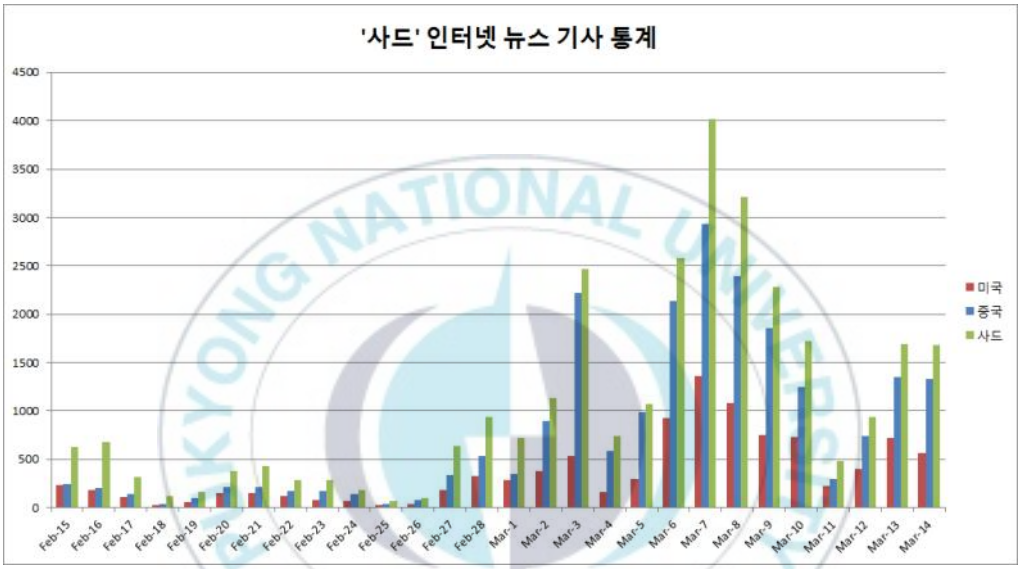
본 연구는 현장 연구의 결과물로 [그림 4-26]은 웹 페이지로 구성된 프로세스 진행 과정을 도식화한 결과이다. 해당 사이트 접속을 시작으로 특정 계정만 시스템 사용을 허용하고 관심 있는 키워드 기간 설정과 검색, 관련 뉴스 크롤링 및 시각화, 특정 댓글 분석을 통한 여론 방향을 키워드 기반으로 긍정, 중립, 부정 등의 결과까지 실시간으로 웹 환경에서 도출된다. 또한, 서로 대립되는 키워드를 선정하여 평판 분석을 통해서 네티즌들의 여론을 도식화한 결과도 제공한다.

[그림 4-27] 실시간 분석 시스템 메인 페이지

[그림 4-27]은 본 연구의 결정체이며 모든 연구 결과가 적재되어 있는 메인 웹 페이지이다. 왼쪽 메뉴는 페이지를 이동하는 메뉴이며 오른쪽 메뉴는 관련 데이터를 분석하는 명령 단추들이 나열되어 있다. 가운데 상단은 검색 키워드와 기간 설정 정보를 제시하였고 상단 중심에 현재 검색 조건을 표시하고 있다.

본 연구의 실험을 통하여 키워드 중심의 검색에서 기간별 이슈의 변화를

살펴보고 마이닝 결과와 비교하여 시스템의 실효를 진단하고자 한다. 먼저 지난 2017년 2월 15일부터 4주간 “사드 중국”, “사드 미국”, “사드”의 키워드로 검색한 뉴스 기사 건별 통계표가 [그림 4-28]과 같이 나타났다. 뉴스 기사 건수가 두드려지게 많이 나타난 2017년 3월 3일(금)과 3월 7일(화)에 어떤 이슈가 있었는지 본 연구 시스템을 통하여 진단하고자 한다.



[그림 4-28] 실험 데이터 통계 자료

1. 2017-03-03 실험 결과

“2017-03-03” 하루 동안 “사드” 키워드 관련 인터넷 뉴스 중 본 연구 시스템에 [그림 4-29]와 같은 결과가 나타났다. 뉴스 기사 제목과 DB내에는 기사 내용이 탑재되었고 네티즌의 댓글 개수 정보까지 리스트에 나타낸 상태이다. 사드 관련 중국의 무역과 관광 등 우리 정부의 미온적으로 대처하는 것을 다루는 기사들이 눈에 띈다. [그림 4-30]은 크롤링 결과 중 79번 연합뉴스의 보도의 “中정부, 韓관광 금지지시하면서 반(反)사드 운동 없다 일축”이란 제목엔 네티즌들의 댓글 반응이 댓글 개수를 보았을 때도 뜨거운 쟁점이 된 것으로 본다. 또한, 90번 연합뉴스 보도에는 “中관광제재 사

드보복 벌써 현실화...예약취소, 홍보단 문전박대"라는 제목의 뉴스 또한, 조회 및 댓글 개수가 높은 것으로 나타났다.

번호	날짜	언론사	기사	댓글수
1	2017-03-03	연합뉴스	中 사드 반발 일환?…서울시 위탁 기관 홈페이지까지 해킹	21
2	2017-03-03	한국경제	한반도 사드 배치 반발에 무작위 해킹 공격…서울시 종합지원센터도 당했다	1
3	2017-03-03	머니투데이	서울시 홈페이지도 당했다…中, 3·1절에 사드반대 해킹 공격	0
4	2017-03-03	디지털타임스	중국 사드 반발 일환?…서울시 위탁 기관 홈페이지까지 해킹	0
5	2017-03-03	뉴시스	유일호 사드 보복, 구두지시 확인되면 적절히 대응	0
6	2017-03-03	뉴스1	유일호 사드 보복, 中 구두지시 확인되면 입장표명	22
7	2017-03-03	스타뉴스	유일호 부총리 사드 보복, 中 구두지시 확인되면 입장표명할 것	1
8	2017-03-03	매일경제	사드 이견…文 다음정부서 安 국론분열 안돼 李 되돌려야	13
9	2017-03-03	연합뉴스	사드 이견…文 다음정부서 安 국론분열 안돼 李 되돌려야	546
10	2017-03-03	중앙일보	이재명 사드, 대통령 되면 중단할 것	27
11	2017-03-03	오마이뉴스	이재명 대통령 되면 사드 바로 중단한다	70
12	2017-03-03	세계일보	대연정·개헌·사드… 민주당 대선주자 첫 합동토론회 어땠나	392
13	2017-03-03	경향신문	[중국 '사드 보복']정부 "중국 설득할 수 있다"더니… 이 정도일 줄 몰랐나	8056
14	2017-03-03	KBS 뉴스	여야, 中 '사드보복' 한목소리 비판…배치엔 시각차	0
15	2017-03-03	아이뉴스24	여야, 中 사드 보복 비판 한 목소리	10
16	2017-03-03	뉴스1	여야 中 도 넘은 사드 보복 비판 일성…정부 대응 촉구(종합)	15
17	2017-03-03	경향신문	[중국 '사드 보복']청치권 "중, 대국답지 않다" 한목소리	27

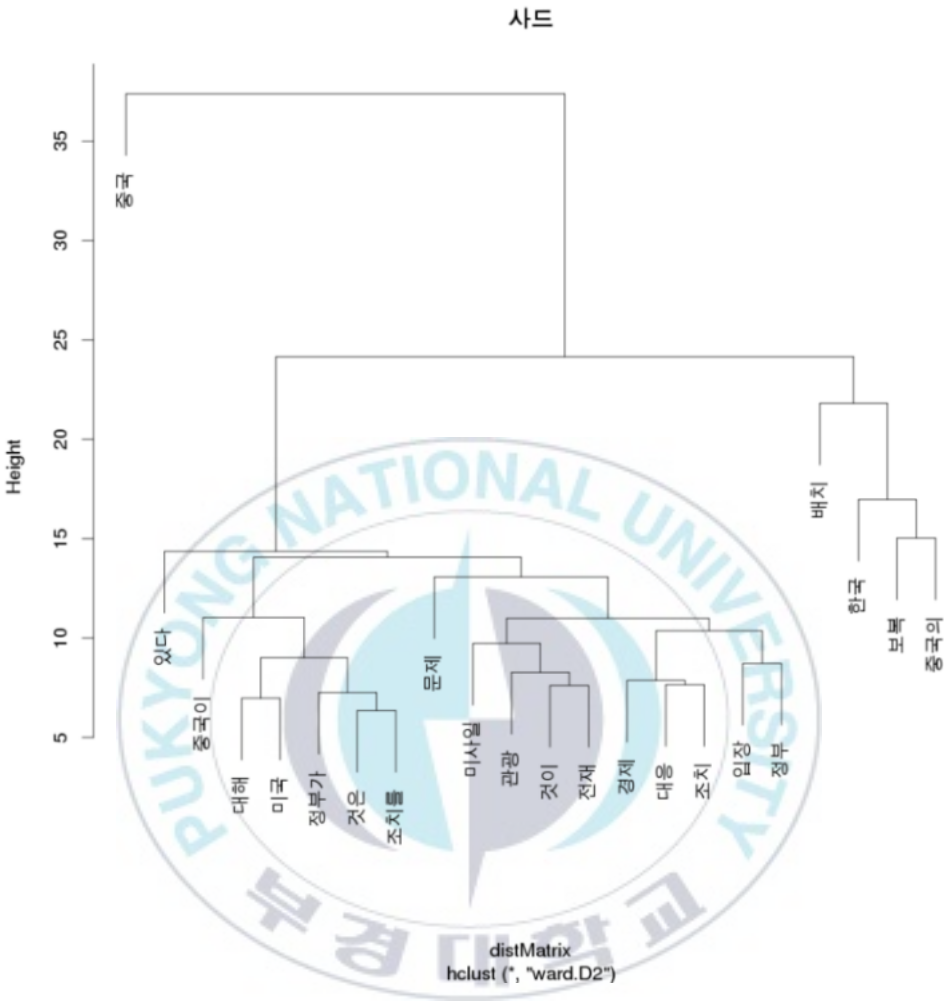
[그림 4-29] 2017-03-03 "사드" 크롤링 결과 #1

74	2017-03-03	SBS 뉴스	美 中 사드 보복, 비이성적이고 부적절 작심 비판	3
75	2017-03-03	연합뉴스TV	미국, 中사드보복 비판 비이성적·부적절…면밀주시	0
76	2017-03-03	MBC 뉴스	美 정부 중국, 사드 보복 비이성적이고 부적절	37
77	2017-03-03	KBS 뉴스	美 "中 사드 보복 조치, 비이성적·부적절"	0
78	2017-03-03	한국경제	문재인·안희정, 中 사드 보복에 나란히 우려 표명	1
79	2017-03-03	연합뉴스	中정부, 韓관광 금지지시하면서 반(反)사드 운동 없다 일축	3513
80	2017-03-03	KBS 뉴스	中정부, "반(反)사드 운동 없다"	5
81	2017-03-03	YTN	中 중국 내에 反 사드 폭력 운동 없어	0
82	2017-03-03	MBN	中정부, 한국 관광 전면금지 해놓고 반(反)사드 운동 없다 일축	0
83	2017-03-03	연합뉴스TV	중국 여행객이 85%…사드 보복에 제주 초비상	2
84	2017-03-03	연합뉴스	[특징주] 中 사드보복 노골화…롯데그룹株 이틀째 급락	0
85	2017-03-03	서울경제	中 사드 보복 본격화…화장품·엔터 여행주 '우수수'	0
86	2017-03-03	연합뉴스	중국 사드 보복 노골화…화장품株 폭삭	2
87	2017-03-03	세계일보	[일착취재] 거세지는 중 사드 보복… 베이징 롯데마트 가보니	92
88	2017-03-03	YTN	여야, 중국 사드 보복 비판…강대국 답지 못해	3
89	2017-03-03	경향신문	[중국 '사드 보복']중 롯데백화점 앞 "철수" 시위…'애국주의 보복' 심상찮다	34
90	2017-03-03	연합뉴스	中관광제재 사드보복 벌써 현실화…예약취소, 홍보단 문전박대	502
91	2017-03-03	연합뉴스	중국 사드보복 부산 크루즈관광 직격탄…유통업계도 당혹	0

[그림 4-30] 2017-03-03 "사드" 크롤링 결과 #2

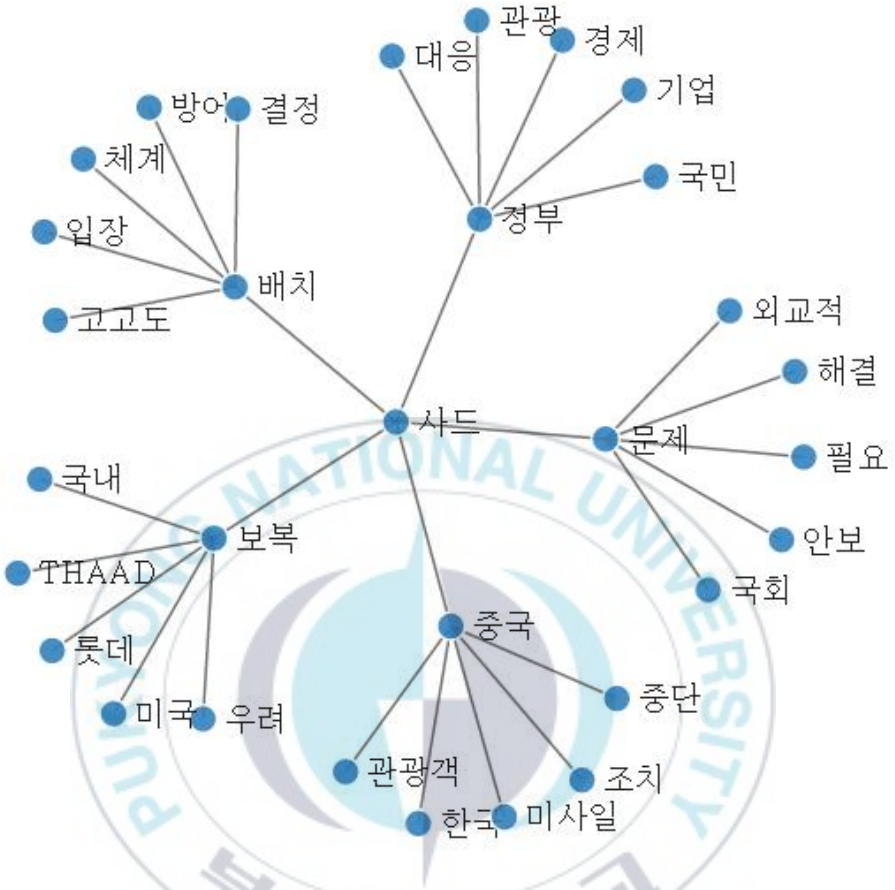
2017년 3월 3일 “사드” 검색어를 통한 뉴스 기사 2,471건을 마이닝한 결과를 살펴보자. 먼저 [그림 4-31]은 워드 클라우드 분석 결과로 사드 “배치”로 인하여 “중국”의 “보복”이 연일 보도되고 우리 “정부”의 “조치”나 “대응”에 대하여 “문제”가 있다는 뉴스 기사들의 관련 단어가 상위 랭킹된 것으로 보인다. “관광객”, “중국인”, “경제”, “미국”, 각 국의 “입장”과 “대응” 등에 관한 뉴스 기사도 언급되어 관련 단어들이 등장하고 있다.

[그림 4-31] 2017-03-03 “사드” 워드 클라우드 분석



[그림 4-32] 2017-03-03 “사드” 문서 군집 분석

가까운 거리의 키워드들 간의 차례로 묶어 나타낸 분석인 [그림 4-32]문서 군집 분석을 살펴보면 “한국의 사드 배치가 중국의 보복으로 나타난다.”는 문장을 유추할 수 있도록 단어들이 서로 가까이 문장을 이루었다는 뜻이다. 우리 정부의 입장에 대한 문제를 지적함과 사드의 정부 조치가 문제가 있음을 지적하는 뉴스 기사가 많은 것으로 보인다.

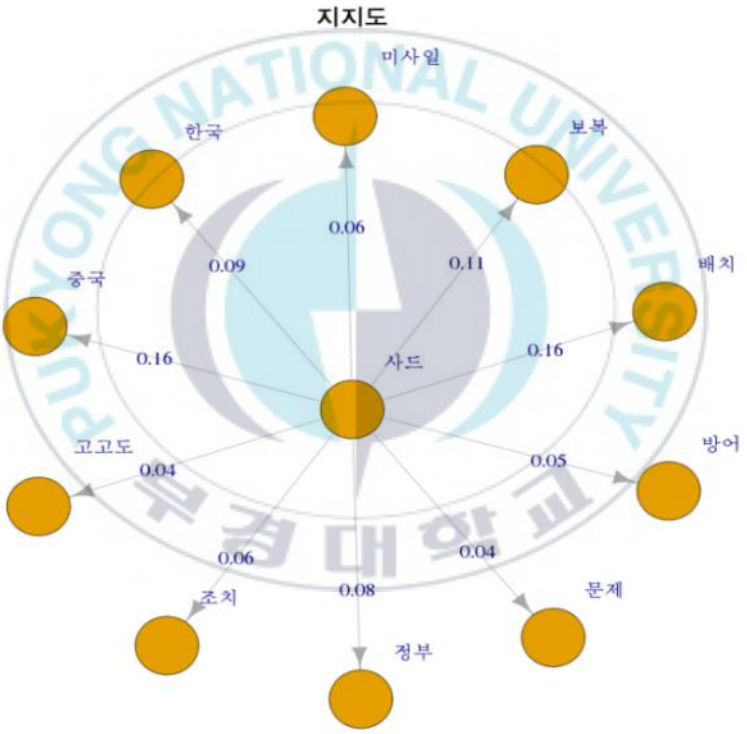


[그림 4-33] 2017-03-03 “사드” 네트워크 그래프

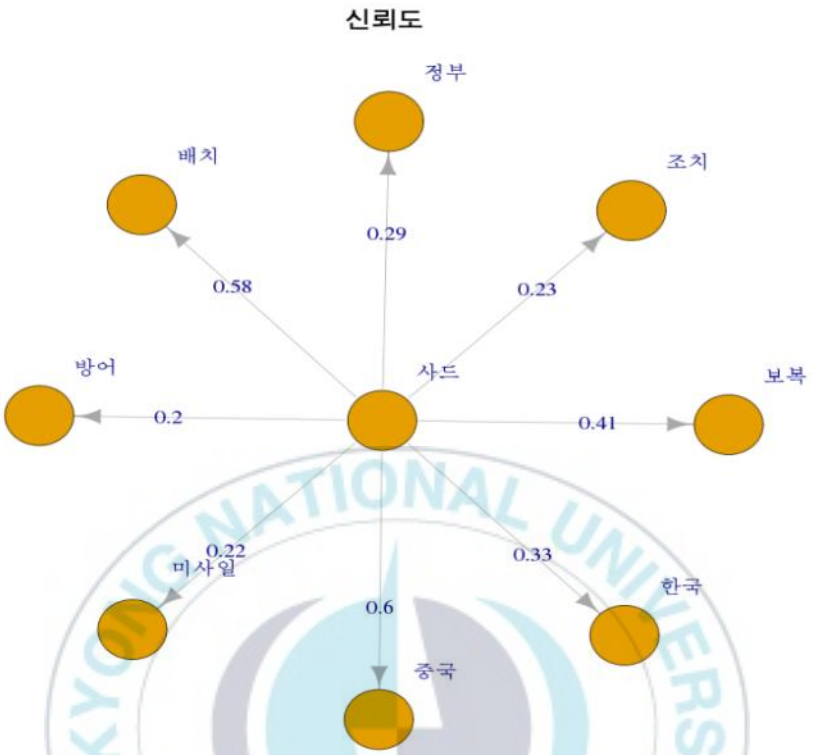
[그림 4-33]은 “사드”의 중심 키워드를 대분류로 보았을 때 크롤링 된 기사는 “중국”, “보복”, “배치”, “정부”, “문제”의 다섯 개의 키워드가 상위 랭킹 되었다. 즉 “사드 보복”, “사드 배치”, “사드 문제”, “사드 중국”, “사드 정부”가 적어도 함께 사용된 문장이 많았고 중분류에 해당되는 키워드 집단이다. 소분류들은 중분류, 대분류 키워드들과 함께 사용된 단어들로서 “사드 보복이 롯데”에 영향을 미치는 뉴스 기사, “사드 배치 결정, 방어, 체계”등 사드 배치에 관한 뉴스 기사, 사드로 인하여 우리 “정부”의 “대응”과 “경제”, 중국 “관광”객에 관한 키워드들이 등장하였다. “문제”와 “중국” 등

중분류 키워드 또한, 다양한 관련 기사들이 있음을 확인할 수 있다.

[그림 4-34]는 키워드간의 연관성과 규칙을 찾기 위한 연관 규칙 분석에 두 키워드가 함께 등장하는 뉴스 기사의 빈도를 나타내는 지지도이다. “사드”와 “배치”나 “중국”이 가장 많이 등장한 것으로 나타났고 “보복”, “한국”, “정부”등이 그 뒤를 이었다. 전체 기사 중에 16%의 문장이 “사드”와 “중국”이 함께 사용되었다고 해석할 수 있다.

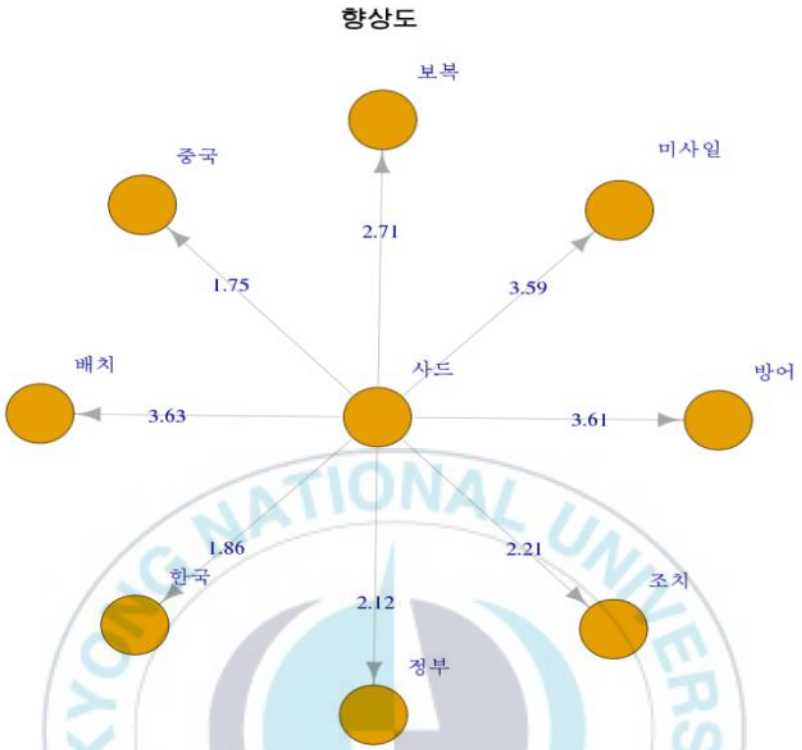


[그림 4-34] 2017-03-03 “사드” 연관 규칙 분석(지지도)



[그림 4-35] 2017-03-03 “사드” 연관 규칙 분석(신뢰도)

[그림 4-35]는 “사드”가 등장한 뉴스 기사 중에서 “중국”이 나타날 확률을 나타낸 신뢰도 분석이다. 60%로 절반 이상의 문장에서 “사드”와 “중국”이 함께 사용되었다고 볼 수 있다. “배치”, “보복” 등의 키워드가 함께 등장할 확률이 상대적으로 높은 것으로 나타났고 최소신뢰도를 20%이상으로 조정하여 처리한 결과이다.

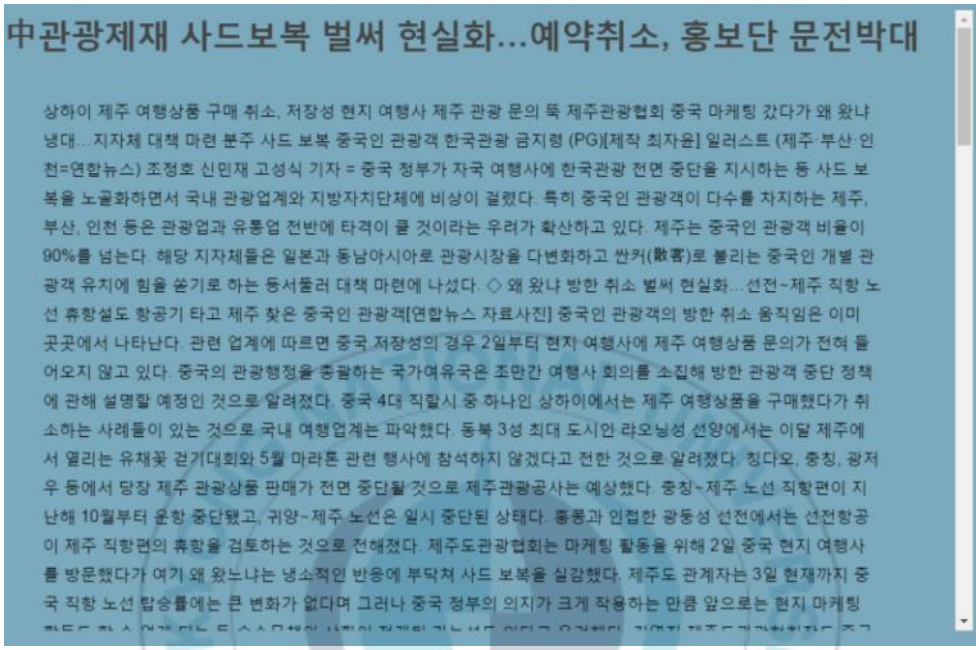


[그림 4-36] 2017-03-03 “사드” 연관 규칙 분석(향상도)

두 단어의 등장 패턴이 서로 독립적인지, 서로 규칙성이 있는지를 나타내는 향상도이다. “사드”를 “배치”라는 키워드와 비교한다면 전체 기사 중 16%의 문장에서 “사드”, “배치”가 함께 사용되었으며[그림 4-34], 두 키워드가 함께 등장할 확률은 58%로 나타났다[그림 4-35]. “사드”와 “배치”를 함께 사용할 비율이 전체 문장에서 “배치”를 사용할 확률보다 3.63배 더 크다고 볼 수 있다[그림 4-36]. 즉 “배치”란 단어 혼자서 사용되는 확률보다 “사드”, “배치”가 함께 등장하는 빈도가 3.63배 높다는 뜻이다.

2017년 3월 3일 연합뉴스의 “中관광제재 사드보복 벌써 현실화…예약취소, 홍보단 문전박대”와 관련된 뉴스 기사의 댓글을 분석하였다. [그림 4-37]은 2017년 3월 3일 “사드” 검색어를 통한 뉴스 기사 2,471건 중 임의

의 뉴스 기사를 선택한 데이터이며 관련 기사의 댓글 502개를 분석하고자 한다.

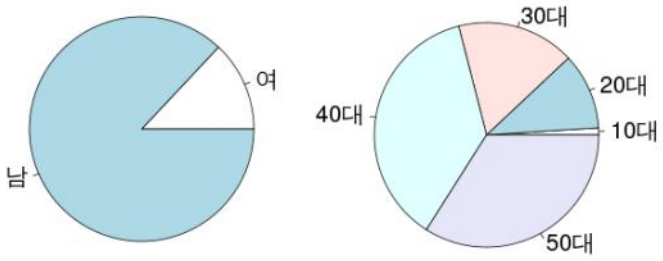


中관광제재 사드보복 벌써 현실화…예약취소, 홍보단 문전박대

상하이 제주 여행상품 구매 취소, 저장성 현지 여행사 제주 관광 문의 뚝 제주관광협회 중국 마케팅 갔다가 왜 왔냐 냉대…지자체 대책 마련 분주 사드 보복 중국인 관광객 한국관광 금지령 (PG)[제작 최자윤] 일러스트 (제주·부산·인천=연합뉴스) 조정호 신민재 고성식 기자 = 중국 정부가 자국 여행사에 한국관광 전면 중단을 지시하는 등 사드 보복을 노골화하면서 국내 관광업계와 지방자치단체에 비상이 걸렸다. 특히 중국인 관광객이 다수를 차지하는 제주, 부산, 인천 등은 관광업과 유통업 전반에 타격이 클 것이라는 우려가 확산하고 있다. 제주는 중국인 관광객 비율이 90%를 넘는다. 해당 지자체들은 일본과 동남아시아로 관광시장을 다변화하고 싼커(散客)로 불리는 중국인 개별 관광객 유치에 힘을 쏟기로 하는 등서둘러 대책 마련에 나섰다. ◇ 왜 왔냐 방한 취소 벌써 현실화…선전~제주 직항 노선 휴항설도 항공기 타고 제주 찾은 중국인 관광객[연합뉴스 자료사진] 중국인 관광객의 방한 취소 움직임은 이미 곳곳에서 나타난다. 관련 업계에 따르면 중국 저장성의 경우 2일부터 현지 여행사에 제주 여행상품 문의가 전혀 들어오지 않고 있다. 중국의 관광행정을 총괄하는 국가여유국은 조만간 여행사 회의를 소집해 방한 관광객 중단 정책에 관해 설명할 예정인 것으로 알려졌다. 중국 4대 직할시 중 하나인 상하이에서는 제주 여행상품을 구매했다가 취소하는 사례들이 있는 것으로 국내 여행업계는 파악했다. 동북 3성 최대 도시인 랴오닝성 선양에서는 이달 제주에서 열리는 유채꽃 걷기대회와 5월 마라톤 관련 행사에 참석하지 않겠다고 전한 것으로 알려졌다. 칭다오, 충칭, 광저우 등에서 당장 제주 관광상품 판매가 전면 중단될 것으로 제주관광공사는 예상했다. 충칭~제주 노선 직항편이 지난해 10월부터 운항 중단됐고, 귀양~제주 노선은 일시 중단된 상태다. 홍콩과 인접한 광둥성 선전에서는 선전항공이 제주 직항편의 휴항을 검토하는 것으로 전해졌다. 제주도관광협회는 마케팅 활동을 위해 2일 중국 현지 여행사를 방문했다가 여기 왜 왔느냐는 냉소적인 반응에 부닥쳐 사드 보복을 실감했다. 제주도 관계자는 3일 현재까지 중국 직항 노선 탑승률에는 큰 변화가 없다며 그러나 중국 정부의 의지가 크게 작용하는 만큼 앞으로는 현지 마케팅

[그림 4-37] 2017-03-03 댓글 분석을 위한 뉴스 기사 전문

댓글 분석 데이터인 “中관광제재 사드보복 벌써 현실화…예약취소, 홍보단 문전박대” 뉴스 기사의 댓글은 [그림 4-38]과 같이 남성 87%, 여성 13%가 댓글에 참여 하였고 10대 1%, 20대 11%, 30대 17%, 40대 37%, 50대 34% 참여율을 보였고, 3, 40대 남성의 참여율이 높은 것으로 나타났다.



[그림 4-38] 2017-03-03 댓글 네티즌 통계

2017년 3월 3일 연합뉴스의 “中관광제재 사드보복 벌써 현실화…예약취소, 홍보단 문전박대”와 관련된 뉴스 기사의 댓글 중 시스템에 [그림 4-39]와 같은 결과로 나타났다. 네티즌이 작성한 댓글 날짜와 시간, 댓글 내용, 해당 댓글의 공감과 비공감 횟수까지 리스트에 나타낸 상태이다. “중국과 단교해라”, “여행 우리도 가지 말자” 등의 부정적이고 험한 단어들을 쉽게 볼 수 있다. [그림 4-40]은 크롤링 결과 중 355번 네티즌은 현 정부의 국내외 현황에 대한 불만을 강하게 표현하였고 358번 네티즌은 현 뉴스 기사와 무관한 대통령 탄핵을 주장하는 의견도 있었다.

번호	날짜/시간	댓글	공감	비공감
1	2017-03-03 21:21:00	사드를 떠나서 짱개와 사회주의는 지금부터 최소한의 교류만하는 쪽으로 전면 재검토해야한다	131	20
2	2017-03-03 21:20:00	멘날당하지만 말고 우린 합법적으로 안산 ,구로동 등 중국불법체류자 3개월만 탈탈 털어버라 수만명은 나올거다 아마 불법체류자가 중국가서 그만하라고 대모할거다	106	18
3	2017-03-03 21:21:00	방정떠는 언론 너거가 더 문제다 알겠나? 너거들이 못난짓을 하고 벌벌 떨고 하니까 ..중국넘들이 더 신이 나서 그딴 짓을 해 대는거다 중국 관광객 많와도 나라 망하지는 않으니까 ..언론사 너희들 걱정이나 해라 ! 너거 회사 망하믄 어쩔 것인지 그거 걱정이나 하라고 ..!	91	13
4	2017-03-03 21:18:00	서서히 중국에서 철수 준비해라. 그넘들한테 언젠가는 발목잡힌다.우리나라 생산물건 모두 망하게 해놓고 어느 시점에 가서는 저넘들 마음대로 가격조정한다. 동남아로 oem 옮겨라	121	24
5	2017-03-03 22:22:00	짱개들한테 굽실 거리지 좀 마 미친 거 같아 진짜	50	6
6	2017-03-03 21:46:00	우리 누나는 영철 좋아함 이대 다니는데 이제 짱개애들 안봐서 좋다고 사진찍지 말라는 캠퍼스 안까지 와서 관광 분위기 망치기 일수 인데 이제 조용히 학교만 다닐수 있을것 같다고 좋아함	56	10
7	2017-03-03 21:36:00	중국 가지마라 깔맞는다	45	8
8	2017-03-03 23:30:00	언제부터 중공이 우방이었나? 다들 정신차려고 지금부터라도 제대로 해라..중공땜에 경제손실..관광손실..어찌고 듣기싫다..옛날에 중공 없이도 잘먹고 살았다..	31	4
9	2017-03-03 23:43:00	크기는 대국 마인드는 소국. 평균으로 내서 중국	11	0
10	2017-03-04 00:04:00	단교해라..핵개발하자.	11	1

[그림 4-39] 2017-03-03 댓글 크롤링 결과 #1

355	2017-03-03 23:01:00	황교활이는 대통령 병에 걸려 구제역 ai 사드문제는 나몰라라하고 나라꼴 개판이구나 ㅉㅉ	5	3
356	2017-03-04 00:10:00	수도공급 중단한다니까 즈그집 수도꼭지 망치로 부수는 놈들 예라 이놈들아 그런 편협한 생각으로 대들어대니까 일본 중국에들한테 비웃음 받는 거다.국내외 정치와 외교 안보를 합리적으로 잘 했으면 이런 개같은 창피는 안 당했을 거 아니냐.앉아서 개오줌 싸는 것과 그 시다바리들....하긴 어쩔 수가 없네...	2	2
357	2017-03-03 22:31:00	중국 여행객들은 실질적으로 면세점 이득이 없다는 선동도 하던데... 말도 안되는 소리 마라... 니들 가족들중 면세점 종사자들이 있다면 물어봐라... 타격이 얼마나 심할지 심장출이고 있을거다... 이 나라 뭣대가리들의 탁상공정이나 니들의 책상머리 선동이나 어찌그리 동일할까... ㅉㅉ	2	2
358	2017-03-03 22:27:00	아무리 중국의 압박수위가 올라가도 닭은 탄핵시키고 보자.	2	2
359	2017-03-03 22:22:00	미국과 유럽에서는 중국을 자유시장국 인정안하고 있다. 한국만 인정해주고 있는거 취소해야지. 중국도 상상도 못할 피해를 입을거다. 다만 공산당이 피해를 통제해서 없는것처럼 보이지만 시진핑도 날아갈수 있는 중대사건이다. 중국 인민들은 눈과 귀가 없는줄아나? 텅텅빈 아파트와 경제불황이 여러곳에서 보이는데 언론서 감춘다고 터질게 안터지냐? 짱깨들 망하는것도 시간문제다. 걱정마라. 중국산 수입 안한다고 죽는거 아니고, 중국 통한 수출 안한다고 업체가 없는것도 아	2	2
360	2017-03-03 21:20:00	쪽팔리게 무슨 중국가서 한국와주세요 하고 있냐. 전면 철수하고 맞대응해라. 재네들은 똑같이 대해줘야돼.	2	2
361	2017-03-03 21:19:00	wto에 제소해야 한다. 중국을 망하게 하려면 고정환율을 완전변동환율제로 안하면 미국은 무역보복해야 한다.	2	2
362	2017-03-03 21:54:00	중국 버리고 동남아 관광객을 유치해야 한다.물론 우리 나라에 쓰레기 관광 상품 파는 놈 단속하고. 덤태기 씌우는 놈들도 싹 단속해서 없애야 하는게 우선되어야 한다.	10	5
		애초에 중국과의 의존도가 너무 컸던게 문제다. 닭정권 3년간 전승절이니 뭐니 해서 서방국에		

[그림 4-40] 2017-03-03 댓글 크롤링 결과 #2

2017년 3월 3일 “사드” 검색어를 통한 뉴스 기사 2,471건 중 임의의 뉴스 기사를 선택한 데이터인 “中관광제재 사드보복 벌써 현실화...예약취소, 홍보단 문전박대” 뉴스 기사의 댓글 502개를 분석해보면 다음과 같다.

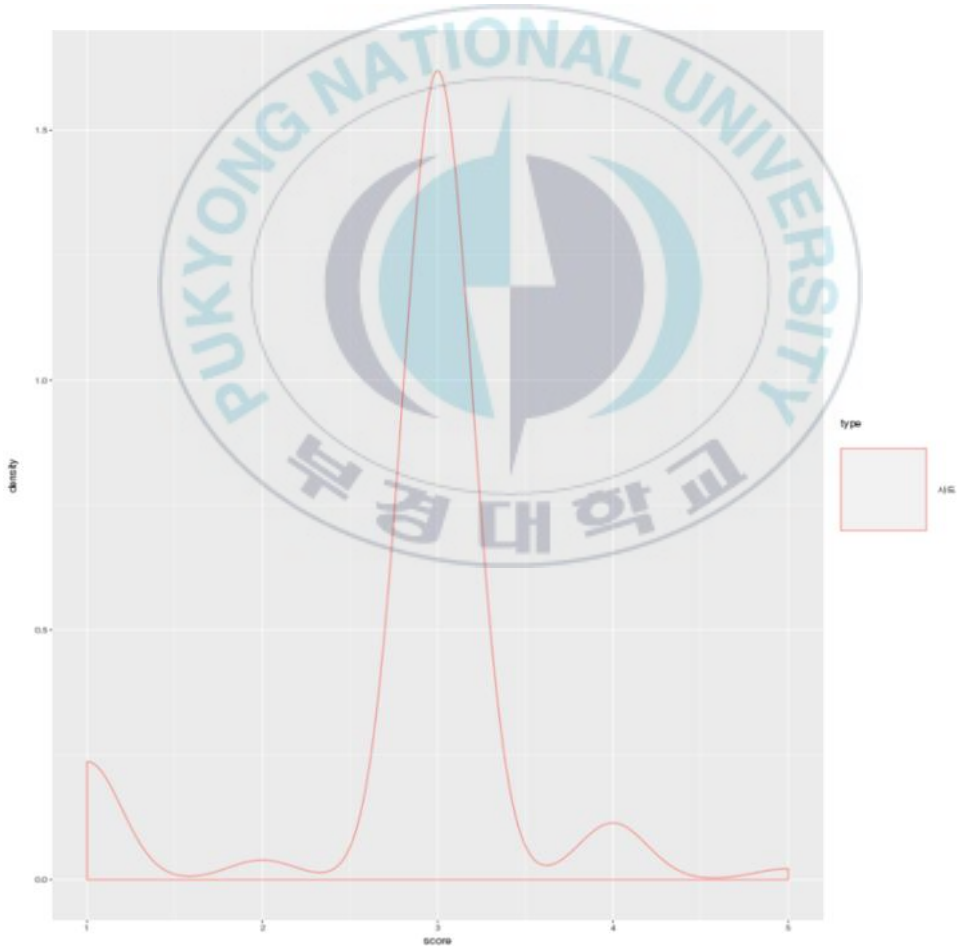
[그림 4-41] 2017-03-03 “사드” 긍정적 단어 워드 클라우드

[그림 4-42] 2017-03-03 “사드” 부정적 단어 워드 클라우드

오피니언 마이닝으로 긍정적 사전의 키워드를 실험 데이터와 비교하여 등장 빈도를 시각화한 [그림 4-41]이다. 실험 주제를 환영하는 네티즌의 자유로운 글들을 긍정적 단어 사전과 비교하여 시각화한 결과 이미지이다. [그림 4-41]과 같이 [그림 4-42]는 부정적 단어 사전과 연구 데이터와의 비교를 통하여 부정적 단어 빈도를 시각화한 결과이다. 비방하는 단어들이 대부분이며 “반대”라는 단어가 네티즌들의 의견을 반영하듯 높은 빈도로 나타났다. 네티즌꾼들의 개개인의 생각을 표현하는 댓글을 분석한 것으로 비표준어나 인터넷어, 함축적인 언어들이 상대적으로 많이 사용되었다.

[그림 4-43] 2017-03-03 “사드” 비교 클라우드

두 단어 집단이 뚜렷하게 구분되며 연구 데이터에서 추출된 단어를 부정적 단어와 긍정적 단어로 구분하여 시각화한 것으로 부정적 단어의 빈도수가 높은 것으로 나타났다. 상대적으로 긍정적 단어의 키워드 개수가 적은 것을 확인할 수 있다. 이는 네티즌들의 반응을 쉽게 엿볼 수 있는 결과물인 것이다[그림 4-43].



[그림 4-44] 2017-03-03 “사드” 오피니언 마이닝

실험 데이터의 수많은 데이터를 대시보드(Dashboard) 하나를 통해 표현할 수 있는 결과물이기도 하다. [그림 4-44]는 오피니언 마이닝의 결과물로 긍정적 사전과 부정적 사전을 이용하여 해당 단어의 빈도가 발생할 때마다 카운트하여 계량화한 그래프이다. 해당 댓글의 내용이 전체적으로 부정적 사전의 단어 빈도보다 긍정적 사전의 단어 빈도가 더 높다는 것을 알 수 있다. 또한, “3”의 빈도가 높은 것은 한 문장에 3번의 긍정적 단어 빈도가 있는 문장이 그 만큼 많이 등장했다는 의미를 나타낸다.

2. 2017-03-07 실험 결과

“2017-03-07”, “사드” 키워드가 관련 인터넷 뉴스 주요 언론사를 포함하여 4,016건으로 본 연구의 실험 기간을 설정한 기간 중 가장 높은 빈도의 뉴스 기사건수가 나타났다. [그림 4-45, 4-46]를 살펴보면 미국으로부터 사드 발사대의 반입 등의 핫뉴스가 등장하였고 북한의 핵미사일의 심각한 상황을 알리기 위한 기사도 눈에 들어 왔다. 또한, 다음 정권으로 사드 문제를 넘겨야 한다는 기사도 있었다. 사드 배치 문제가 국민을 위한 올바른 결정인지 아니면 국민을 무시한 결정인지에 대한 정치권 발언이 뉴스 기사로 나타났으며 한반도 사드 배치를 앞당긴 이유를 취재한 SBS뉴스가 네티즌들이 많은 관심을 보였다는 것을 댓글을 통해서 알 수 있었다.

번호	날짜	언론사	기사	댓글수
1	2017-03-07	KBS 뉴스	사드 발사대 전격 반입…배치 작업 본격화	7
2	2017-03-07	KBS 뉴스	사드 발사대 전격 반입…배치 작업 본격화	1
3	2017-03-07	연합뉴스	러 의회, 사드 한국 배치착수에 반발…러 안보에 직접 위협(종합)	18
4	2017-03-07	KBS 뉴스	러 "군사 대외 활동서 사드 한국 배치 고려할 것"	4
5	2017-03-07	YTN	러 의회, 사드 한국 배치착수에 반발	0
6	2017-03-07	연합뉴스	러 외무부 군사 대외 활동서 사드 한국 배치 고려할 것 경고(종합2보)	39
7	2017-03-07	국민일보	반기문, "사드 배치를 차기 정부에 넘기자는 주장이 중국에 빌미 제공해"	4
8	2017-03-07	연합뉴스	반기문 사드, 차기 정부 넘기자고 해 中이 더 압박(종합)	201
9	2017-03-07	서울신문	주한미군 사드배치 시작…반기문 "사드, 차기 정부 넘기자고 해 중국이 더 압박"	24
10	2017-03-07	YTN	반기문 사드, 차기 정부 넘기자고 해 中이 더 압박	0
11	2017-03-07	세계일보	"문재인 사드 차기정부 이양 주장, 중국에 압박의 빌미 제공하는 것"	2
12	2017-03-07	TV조선	[단독] 반기문 정치권, 사드 배치 이견 한심	72
13	2017-03-07	조선일보	[TV조선 단독] 반기문 정치권, 사드 배치 이견 한심	4
14	2017-03-07	JTBC	사드 전개 첫날, 일단 말 아낀 중국…8일 왕이 회견 주목	9
15	2017-03-07	경향신문	中 사드 보복, 절정은 멀었다…이달 15일이 고비 될듯	20
16	2017-03-07	부산일보	[사드 전격 배치] 군, 사드 장비 전격 반입 배경은	0
17	2017-03-07	KBS 뉴스	"북핵·미사일 심각"…예상 넘는 사드 신속 전개	95

[그림 4-45] 2017-03-07 “사드” 크롤링 결과 #1

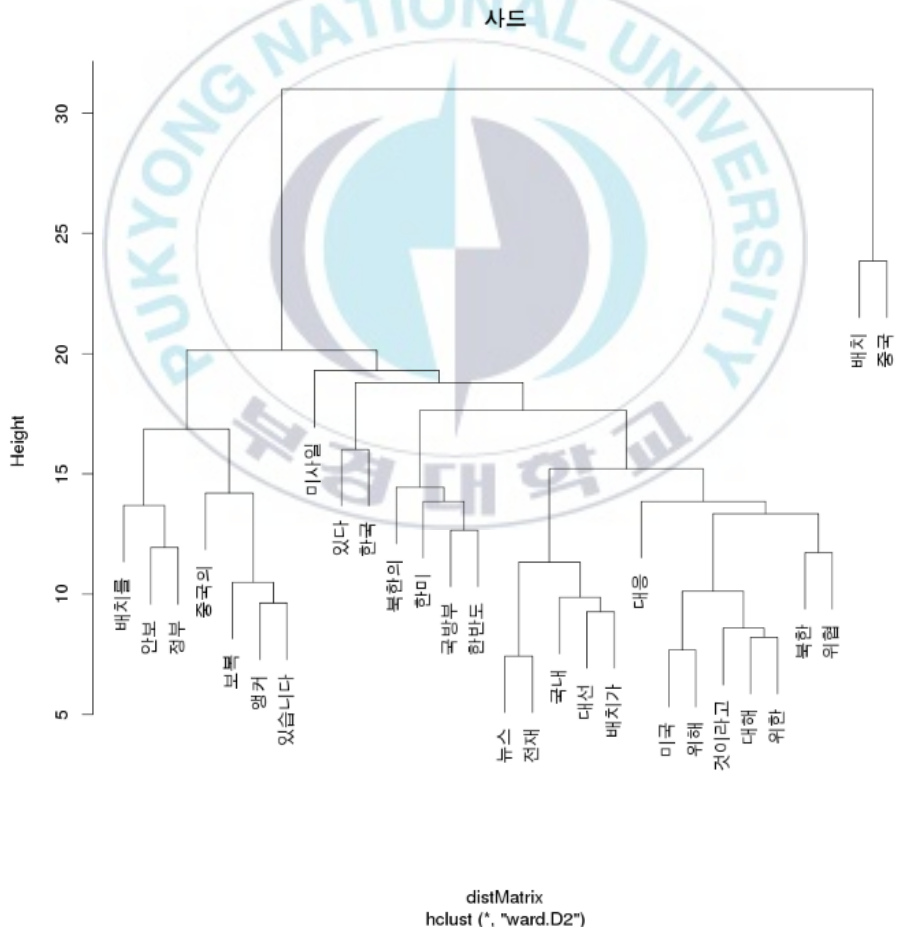
59	2017-03-07	JTBC	속도내는 사드 배치…발사대 2기 등 일부 한국 도착	0
60	2017-03-07	SBS 뉴스	논란 속 사드 배치 시작…어젯밤 발사대 2기 도착	13
61	2017-03-07	서울신문	[영상] 국내 들어온 사드 장비	0
62	2017-03-07	부산일보	[사드 전격 배치] 롯데, 경영권 분쟁에 중국발 불똥까지…	0
63	2017-03-07	세계일보	中 사드 보복에… 금융권도 '비상'	0
64	2017-03-07	연합뉴스	외교부 한미 공조해 중국에 사드 입장 전달	148
65	2017-03-07	SBS 뉴스	외교부 한미 공조해 중국에 사드 입장 전달	0
66	2017-03-07	서울신문	한국 도착한 사드…외교부 "한미 공조해 중국에 사드입장 전달"	1
67	2017-03-07	KBS 뉴스	외교부 "한미 공조해 중국에 사드 입장 전달"	0
68	2017-03-07	뉴시스	청주공항 中노선 이용객 한 달 새 1만명 감소 …중국 사드 보복 영향	0
69	2017-03-07	SBS 뉴스	사드 배치에 엇갈린 정치권…옳은 결정 vs 국민 무시	601
70	2017-03-07	SBS 뉴스	어차피 맞을 매라면…한반도 사드 배치 앞당긴 이유	987
71	2017-03-07	SBS 뉴스	변수 차단 속전속결 사드 배치…트럼프식 힘의 외교	110
72	2017-03-07	연합뉴스	우 원내대표 중국 사드 제재 5단계까지 준비하고 있어	227
73	2017-03-07	SBS 뉴스	대선 전 사드 배치 완료…부지 조성 작업 관건	6
74	2017-03-07	연합뉴스	中 베이징에 사드반대 광고차량 등장…교민사회 불안 고조	76
75	2017-03-07	서울경제	中베이징 시내에 '사드반대' 광고차량 등장	6
76	2017-03-07	SBS 뉴스	中 베이징에 사드 반대 광고 차량 등장…교민사회 불안 고조	0
77	2017-03-07	연합뉴스	베이징에 사드반대 광고차량 등장…교민사회 불안 고조(종합)	1

[그림 4-46] 2017-03-07 “사드” 크롤링 결과 #2

2017년 3월 7일 “사드” 검색어를 통한 뉴스 기사 4,017건을 마이닝한 결과를 살펴보자. 먼저 [그림 4-47]은 워드 클라우드 분석 결과로 국내 사드 발사대 “배치”로 인하여 “중국”의 “보복”에 관한 뉴스는 여전이 높은 빈도를 보여 왔고 “북한”의 “미사일” 관련 키워드가 상위 랭킹된 것을 확인할 수 있다. 또 다른 특징은 사드 “발사대”의 배치가 “시작”되었다는 핫뉴스가 등장한 것이 눈길을 끌었다. 발사대 배치와 관련하여 “오산”기지, “전격”배치 등 관련 키워드가 나타났다. 3월 3일의 등장한 “관광객”, “중국인”, “경제”, 등의 키워드는 상대적으로 빈도가 낮아졌고 “미국”은 “한미”로 대체되어 사용되었다. 또한, 중국의 사드 대응 레이더 요격 미사일이 이미 실전 배치되었다는 뉴스가 나타나면서 “배치”라는 키워드가 높은 빈도로 나타났다.

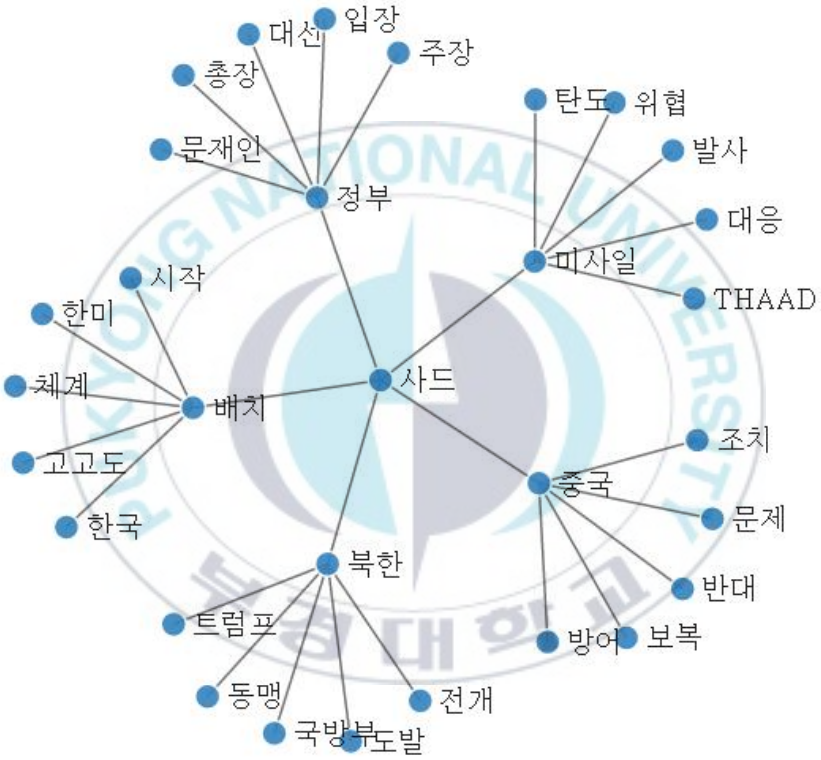
[그림 4-47] 2017-03-07 “사드” 워드 클라우드 분석

[그림 4-48] 문서 군집을 살펴보면 3월 3일 문서 군집 분석과는 다르게 “북한”의 핵 “미사일” “위협”에 “대응”하기 위하여 “미국”과 공동으로 “한미” “대응”으로 “한반도” 사드 “배치”가 진행된 것을 엿볼 수 있다. “중국”과 “배치”가 상위 랭킹 된 것으로 볼 수 있으며 3월 3일 뉴스는 경제나 관광에 보복이 있을 것으로 보는 예측관련 단어들을 많이 사용한 반면에 시간이 흐름에 따라 본격화 되어가는 사드 배치와 북한의 미사일 실험, 중국의 자국 압박의 수위가 높아지는 것을 볼 수 있다. 또한, 국내 대선 판도에도 많은 영향을 줄 것으로 보이는 뉴스들이 나타났다.



[그림 4-48] 2017-03-07 “사드” 문서 군집 분석

“03-07” 뉴스 기사의 “사드”의 중심 키워드를 살펴보면 “03-03” 데이터에서 없던 “북한”과 “미사일” 키워드가 상위 랭킹에 위치하면서 “탄도”, “위협”, “발사”, “대응”, “도발”, “국방부” 등의 키워드가 연관 검색으로 등장했다. 이는 같은 날 북측의 미사일 실험이 있었던 것으로 확인되었고 대부분의 언론사들이 앞 다투어 기사화한 것으로 보인다[그림 4-49].



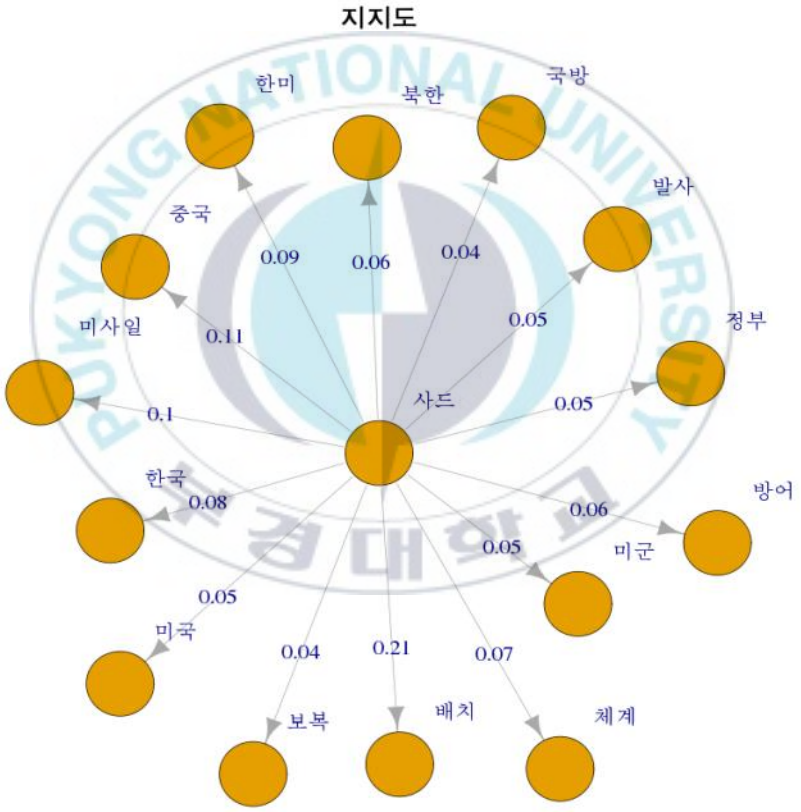
[그림 4-49] 2017-03-07 “사드” 네트워크 그래프

“03-03” 결과와의 비교 [표 4-2]를 살펴보면 “03-03” 결과의 중분류 “정부” 관한 소분류는 “대응”, “경제”, “관광”, “기업” 등 정부의 기업에 대한 대응 즉, 경제적인 사드 보복에 관한 키워드들이 대부분인 반면 “03-07” 결과에서는 “문재인”, “대선”, “입장”, “주장” 등을 보면 여러 국내 정치적 입장과 대안들에 대한 주장을 뉴스로 표현하고 있다.

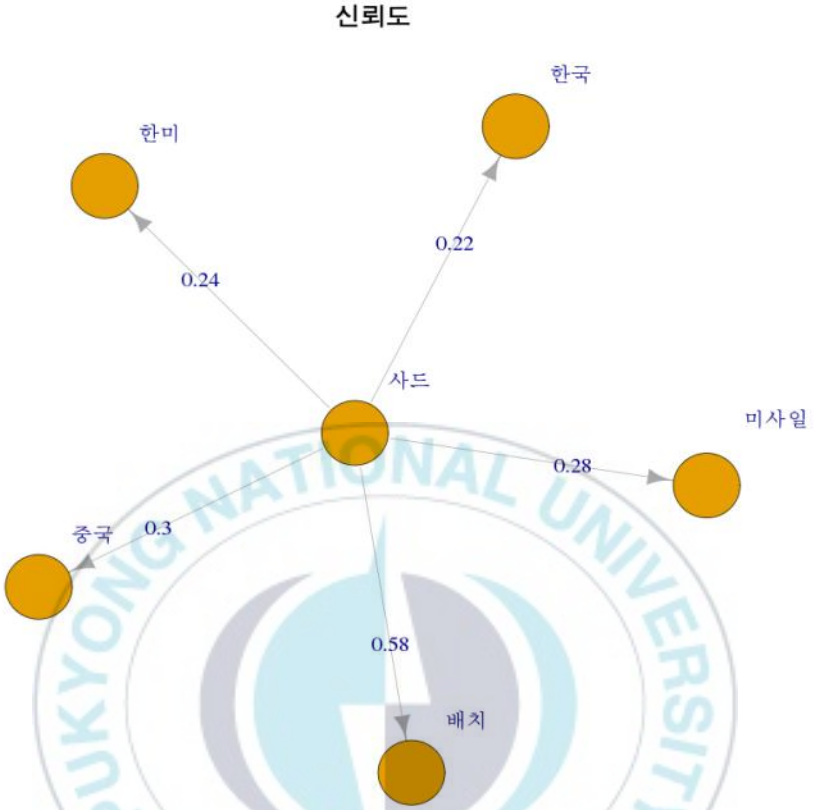
[표 4-2] “03-03”과 “03-07” 키워드 비교

3월	대분류	중분류	소분류
3일	사드	배치	결정 방어 체계 입장 고고도
		정부	대응 관광 경제 기업 국민
		중국	관관객 한국 미사일 조치 중단
		보복	국내 THAAD 롯데 미국 우려
		문제	외교적 해결 필요 안보 국회
7일	사드	배치	시작 한미 차계 고고도 한국
		정부	문재인 총장 대선 입장 주장
		중국	조치 문제 반대 보복 방어
		미사일	탄도 위협 발사 대응 THADD
		북한	트럼프 동맹 국방부 도발 전개

연관 관련 규칙의 지지도인 [그림 4-50]을 살펴보면 “03-03” 결과와 같이 “사드”와 “배치”나 “중국”이 높은 지지도를 보이며 “미사일”, “한미” 등의 새로운 키워드가 10%의 “사드”와 함께 등장한 키워드로 나타났다. 전체 기사 중에 21%의 문장이 “사드”와 “배치”가 함께 사용되었고 “보복” 관련된 키워드는 11%에서 4%로 줄어든 것으로 나타났다.

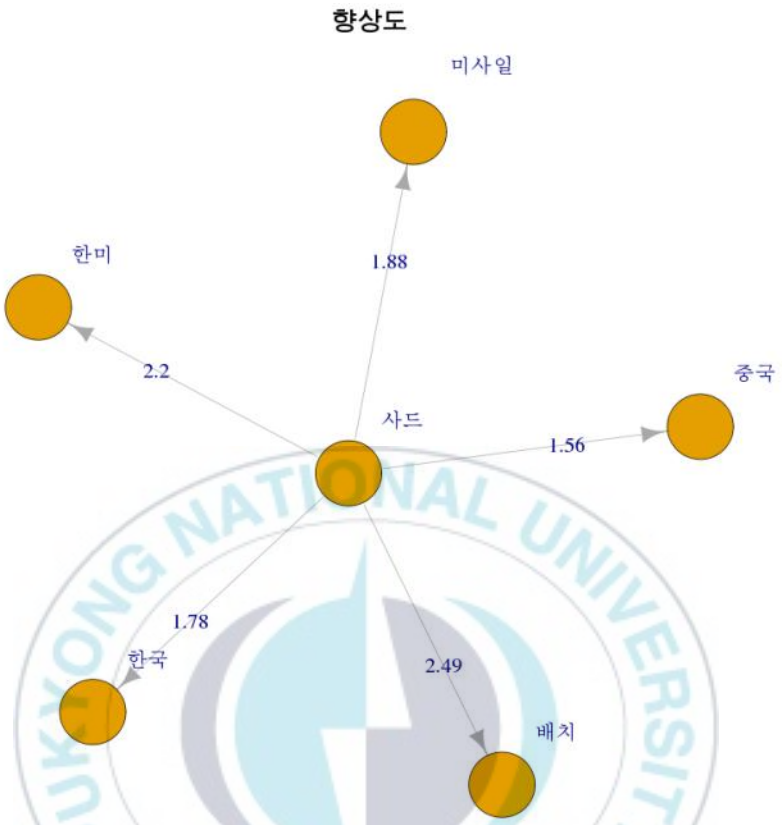


[그림 4-50] 2017-03-07 “사드” 연관 규칙 분석(지지도)



[그림 4-51] 2017-03-07 “사드” 연관 규칙 분석(신뢰도)

[그림 4-51]는 “사드”가 등장한 뉴스 기사 중에서 “배치”가 나타날 확률을 나타낸 신뢰도 분석이다. 58%로 절반 이상의 문장에서 “사드”와 “배치”가 함께 사용되었다고 볼 수 있다. 미국에서 사드 발사대의 배치가 진행되는 이슈가 위와 같은 결과로 이어진 것으로 보인다. 또한, “미사일”, “한미” 등의 키워드가 함께 등장할 확률이 상대적으로 높은 것으로 나타났고 “03-03” 처리 과정과 같이 최소신뢰도를 20%이상으로 조정하여 처리한 결과이다.



[그림 4-52] 2017-03-07 “사드” 연관 규칙 분석(향상도)

[그림 4-52]는 두 단어의 등장 패턴이 서로 독립적인지, 서로 규칙성이 있는지를 나타내는 향상도로 “03-03” 결과와의 큰 차이는 양의 상관관계를 보이는 1이상의 향상도의 키워드 수가 8개에서 5개로 축소된 것과 “사드”와 함께 등장할 확률 또한 3.63에서 2.49로 낮아진 것을 확인할 수 있었다. 이는 선행연구의 “2.3 연관 분석”의 식(1)에서 전체 거래 수(N)에 해당하는 “03-07” 뉴스 기사의 문장 수가 그 만큼 많아졌다는 의미도 될 수 있다. 같은 이슈가 장기간 지속될 경우 다양한 예측과 정보들이 기사화되는 과정이라 볼 수 있다[그림 4-52].

“03-07” 결과는 전체적으로 ‘사드 발사대가 한국에 배치되었다“는 핫뉴스와 북한의 미사일 시험 발사에 관한 뉴스의 관련 키워드가 중심이 되었다. 본 연구의 실시간 뉴스 기사 분석 시스템을 가동한 결과 검색 키워드와 관련된 다양한 연관 키워드들을 표출하였고 수많은 뉴스 기사들을 종합하여 유추할 수 있는 키워드를 추출하고 있는 것을 확인하였다.

3. 오피니언 마이닝 실험 결과

“4.3.1 2017-03-03 실험 결과”에서 나타난 오피니언 마이닝의 결과는 단일 뉴스 기사를 대상으로 댓글을 분석할 경우를 설명하였다. 본 실험은 2개의 뉴스를 대상으로 비교하여 두 이슈간의 차이가 있는지를 나타내고자 한다.

[그림 4-28] 실험의 통계자료에 따르면 “사드”의 키워드 관련 인터넷 뉴스 주요 언론사를 포함하여 29,920건으로 집계되었다. 실험 초기 “사드” 관련 기사의 네티즌들의 여론과 실험 말기의 “사드”관련 기사의 네티즌들의 여론을 비교하여 실험을 하였다.

[표 4-3] 오피니언 마이닝 데이터 통계

검색 날짜	언론사	뉴 스	댓글 수	전체 뉴스 개수
2017-02-20	한국경제	사드 보복은 핑계…한한령 활용해 자국산업 키우는 중국	612	376건
2017-03-14	연합뉴스	사드 보복 본격화…중국인 한국 단체관광 전면 금지	879	1,682건

[표 4-3]는 “사드” 키워드 검색어로 2월 20일 376건의 검색 뉴스 중 “한국경제” 언론사의 “사드 보복은 핑계…한한령 활용해 자국산업 키우는 중국”관련 뉴스의 댓글을 대상으로 분석하였다. 관련 기사는 한국에서 느끼는 사드 보복과 중국 현지의 기업인들이 느끼는 것이 다소 차이가 있음을 시사 하고 있다.

또 다른 뉴스는 3월 14일 1,682건의 “사드” 관련 뉴스 중 “연합뉴스”의 “사드 보복 본격화…중국인 한국 단체관광 전면 금지” 관련 뉴스의 댓글을 대상으로 함께 분석하였다. 관련 기사는 대형 여행사는 물론 중소여행사까지도 입향 취소와 비자 대행도 불가능한 여러 관광 업계의 사드 보복의 실태를 뉴스 기사로 보여주고 있다.

“사드” 관련 뉴스 건수를 확인해보면 연구 초기의 이슈 즉 뉴스 개수는 376건으로 집계되지만 연구 후반은 1,682개로 5배 정도 높은 이슈를 보이고 있다.

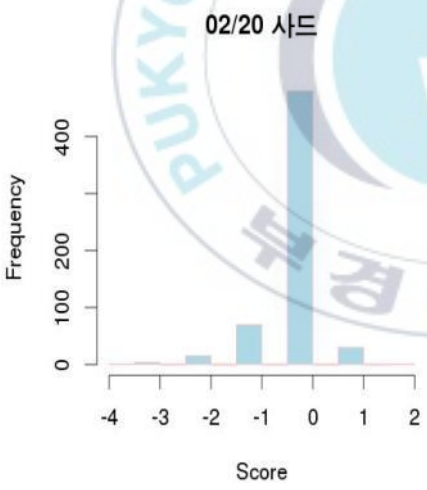
[그림 4-53]과 [그림 4-54]는 두 뉴스 기사를 부정적 단어 사전을 이용하여 관련 단어를 추출해놓은 결과물이다. 연구 초반의 네티즌들은 “거지”, “쓰레기”, “문제”, “꼴”, “반대” 등 일반적인 욕설에 부정적 단어들을 구사하였다. 하지만 연구 후반에 댓글들은 “금지”, “추방”, “반대”, “대가리”, “불법” 등 점점 강한 단어를 사용하고 있는 것을 확인하였다.

[그림 4-53] 2017-02-20 부정적 단어 추출(워드 클라우드)

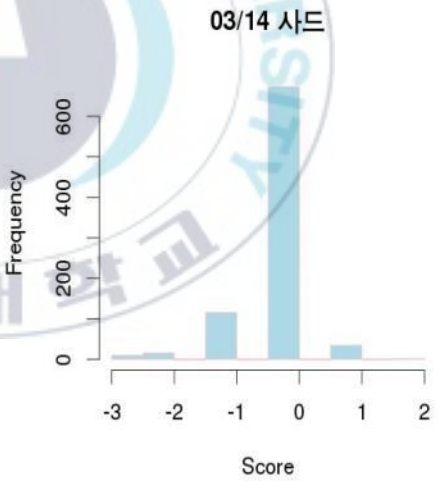
[그림 4-54] 2017-03-14 부정적 단어 추출(워드 클라우드)

오피니언 마이닝의 결과 그래프는 댓글 하나를 한 문장으로 구분하여 문

장내 긍정적 단어 키워드 점수와 부정적 단어 키워드 점수를 합산하여 한 문장의 의견치로 설정하였다. 두 뉴스 기사의 댓글 수는 [표 4-2]를 참고하면 다소 차이는 나지만 “-3”, “-2”에 해당하는 부정적 단어가 [그림 4-55]에 비하여 [그림 4-56]의 결과가 더 높은 것으로 나타났다. “-3”는 한 문장 내 부정적 단어가 3개로 이루어져 있다는 댓글이라 볼 수 있다. “-1”에 해당 되는 문장은 “조선족 불법체류자 추방하면 일자리 30만개는 더 생긴다. 정부는 뭐하나.”로 “추방”이란 단어가 부정적 단어로 처리되었다. “-2”에 해당되는 문장은 “언제는 중국때문에 먹고 살아냐....염병하지말고 오지마”로 “염병”, “오지마”가 부정적 단어 사전으로 등록된 예이다. “-3”은 “이거지 새끼들이 멸종할때 까지 싸우자.”로 “거지”, “새끼”, “싸우자”로 부정적 단어가 세 개로 시스템에서 “-3”의 데이터로 그래프에 반영된다.



[그림 4-55] 2017-02-20
오피니언 마이닝 결과



[그림 4-56] 2017-03-14
오피니언 마이닝 결과

4. RHadoop 기반 텍스트 마이닝 실험 결과

분산처리시스템은 여러 대의 컴퓨터를 한 대의 컴퓨터처럼 사용할 수 있는 시스템이며 데이터 처리량이 늘어나면서 빠른 처리 속도로 데이터 처리의 효율성을 찾고 저비용 디바이스들을 활용하여 고속 처리를 구현 하는 것을 의미한다. 또한, 고속의 처리가 가능한 고가의 하드웨어 시스템을 구입하면 많은 개발 비용을 줄일 수 있지만 비용적인 면이 만만하진 않다. 저가의 일반 PC들을 한 네트워크에 연결하여 마치 하나의 PC처럼 서버가 운영된다면 보다 고속으로 처리될 수 있다. 비용의 효율성도 해결되며 PC의 확장성이 가능하므로 추가적인 네트워크 확장이 처리 속도를 대변해준다.

본 실험은 리눅스 서버 환경에서 뉴스 검색, 분석, 시각화까지의 전 과정의 프로세스가 진행된다. 많은 양의 뉴스 기사를 실시간 처리하기 위해 오픈 소스 소프트웨어인 Hadoop을 사용하였다. Hadoop은 분산 저장기술과 분산 처리 기술의 프레임워크를 제공하며 빅데이터 분석 도구인 R 프로그램과의 확장성이 가능한 소프트웨어이다. 즉, Hadoop 시스템에 R을 다시 연결하여 RHadoop환경이 가능하게 된 것이다. 연구 시스템의 전 처리과정을 분산처리 과정으로 처리한 것이 아니라 데이터 수집 이후 분석 과정을 분산처리시스템을 적용하여 처리하였다.

첫 번째 실험은 네임 노드에서 3개의 데이터 노드를 연결하여 각 데이터 노드의 연결 개수에 따른 데이터 처리량의 변화를 먼저 살펴보았다.

[표 4-4] 데이터 노드에 따른 처리 비교

데이터 용량(MB)	노드 수	완료시간	CPU 사용시간(ms)
37	3대	14	6,300
	2대	21	7,250
	1대	25	7,330

[표 4-4]는 37MB의 한글 데이터를 활용하여 단어 빈도 분석 통계를 나타낸 표이다. 네임 노드에서 데이터를 받은 데이터 노드가 1대만 가동했을 때 25초, 2대의 가동은 21초, 3대의 가동은 14초로 네트워크에 연결하는 데이터 노드 수가 많을수록 처리 시간은 반비례하여 줄어드는 것을 확인할 수 있다. 빅데이터를 분석하는 도구들이다 보니 37MB의 저용량의 데이터로는 그 처리 시간 및 대기 시간의 혁신적인 처리는 아닌 것으로 판단되었다.

두 번째 실험은 네임 노드에서 3개의 데이터 노드를 연결한 상태에서 데이터의 양을 증가 시켰을 때 처리 시간의 변화를 살펴보았다.

[표 4-5] 데이터 용량에 따른 처리 비교

데이터 용량(MB)	노드 수	Map-Task	완료시간	CPU 사용시간(ms)
37	3대	1	15	7,240
111	3대	2	18	15,850
557	3대	6	22	72,500
1,114	3대	12	31	141,700

하둡은 하나의 처리해야할 일이 들어오면 여러 개의 Task로 나누어져 실행된다. 본 연구의 Task는 100MB 단위로 나누어져 Task를 여러 개 만들어진다. [표 4-5]를 참고하여 보면 37MB는 1개의 Map-Task, 100MB는 2개의 Map_task가 만들어지게 된다. 즉 [표 4-4]는 100MB 미만의 데이터 처리를 실험하였기 때문에 Map-Task는 당연히 한 개로 처리된 것이다. [표 4-5]의 1GB 이상의 데이터 처리 시간이 31초로 기록된 것을 확인

할 수 있다. Map-Task의 개수가 12개로 처리되었음을 확인할 수 있다. 데이터 크기가 증가 할수록 Map-Task가 많아지므로 동시에 처리할 수 있는 Task가 증가한 것이다. 빅데이터 분석 도구는 하나의 처리해야 할 일의 크기가 클수록 처리 시간의 효율성이 좋아지는 것을 확인하였다. 작은 데이터 처리를 할 경우 큰 효과를 볼 수 없지만 그 데이터 양이 증가할수록 분산처리량이 증가하므로 빠른 속도를 보였다.

제 5 장 결 론

제 1 절 연구결과의 요약

본 연구는 인터넷 뉴스 기사를 이용한 실시간 이슈분석 시스템을 설계하였다. 대량의 뉴스 기사를 데이터로 설정하여 이슈 관련 키워드의 검색으로 다양한 웹 마이닝 과정을 웹 페이지로 구현하였고 실험까지 진행하였다. 그 결과를 종합하여 정리해보면 다음과 같다.

최신 정보를 업데이트하여 실시간으로 생성되는 데이터를 빠르게 분석할 수 있는 웹 페이지를 구현하였다. 포털 사이트 검색 기능처럼 사회 이슈에 해당되는 키워드와 기간을 설정하면 관련된 뉴스 기사를 크롤링하여 서버 메모리로 적재 후 빅데이터 분석 프로그램인 R 프로그램을 통하여 다양한 통계기법을 사용하여 실험을 전개하였다.

대량의 뉴스 기사의 명사 추출 과정을 간단히 정리하면, 이슈 관련 키워드를 실험 과정에서 키워드 등록과 삭제가 가능하여 마이닝 처리 결과에 연구자의 연구 방향에 초점을 높일 수 있는 편의성을 제공하였다. 자료 수집과 분석 과정, 처리 결과까지 연구자의 개입 없이 실시간으로 이루어지는 시스템이라 결과에 불필요한 키워드나 필요한 키워드가 명사 사전에 미등록되었을 때 연구 결과에도 반영이 안 되는 경우를 보완한 것이다. 사용자 인터페이스 환경을 지원하여 보다 쉽게 키워드 삽입과 삭제가 가능하다. 또한, 비정상적 명사 추출 키워드들을 연구 대상에 반영하기 위하여 일반적 명사 추출 과정에 품사 태킹 처리를 삽입하여 명사 추출에 보다 높은 신뢰성을 확보하였다. 본 연구자는 이와 같이 명사 추출에 많은 과정을 삽입한 이유는 분석의 기초 데이터이며 재료가 되는 복합 명사가 제대로 이

루어져야 이후 모든 빅데이터 분석 기술들이 신뢰할 수 있는 결과를 도출할 수 있기 때문이다.

분석 자료의 정제과정을 통하여 명사 추출 이후 다양한 시각화 및 분석 과정을 설계하였다. 또한, 단어 단위 기반 분석과 문장 단위 기반 분석을 통하여 단조로운 빈도 분석을 보다 다양한 연구방법으로 접근하였다.

먼저 단어 단위 기반 분석을 통하여 Pie 그래프, 워드 클라우드 분석, 네트워크 그래프 등을 시각화하였다. 네트워크 그래프는 초기 검색 키워드를 기준으로 1차 Top 키워드를 선정하고 1차 Top 키워드를 별도 재검색하여 관련 뉴스들에서 2차 Top 키워드를 추출하여 네트워크 그래프를 완성하였다. 링크로 연결된 노드들을 활용하여 문장을 유추하고 관련 이슈에 대한 전체적인 언론의 방향을 읽을 수 있었다.

문장 단위 분석의 문서 군집 분석은 유클리드 제곱 거리를 사용한 거리 계산과 와드연결법에 의해서 시각화한 분석 방법을 선택하였으며 계층적 구조의 키워드의 연결로 관련 키워드의 기사 분석에 시각화가 보다 두드러지게 나타났다. 정형적 텍스트 마이닝에서 활용되는 연관 규칙 분석 즉, 장바구니 분석이라는 제한된 분석 방법을 광범위한 비정형적 텍스트 마이닝에서 시도한 결과를 제시하였다. 이 방법은 고객이 구입한 상품이나 서비스를 장바구니 단위 분석을 통하여 같이 구입한 빈도와 확률, A상품을 구매한 고객이 C상품을 구매하게 되는 비율을 예측할 수 있는 연구 방법이다. 장바구니 단위 분석을 응용한 본 연구에서는 뉴스 기사를 문장 단위로 분해하여 문장을 이루는 단어 간 연관 규칙을 활용하여 결과를 실시간으로 분석하여 제시하였다. 또한, 연구자의 분석 초점에 더 근접하기 위하여 지지도, 신뢰도, 향상도 등 비율을 직접 조율할 수 있는 UI 환경을 함께 제공하였다.

본 연구에서는 사회적 이슈가 어떤 지지를 받는지를 확인하기 위해 댓글

을 분석하는 방법으로 연구하였다. 사회적 이슈를 다루는 뉴스 기사에 대한 네티즌들의 다양한 의견을 오피니언 마이닝 분석을 통해 하나의 이미지인 대시보드로 시각화 하였다. 의견의 대부분은 2가지 원의(原義)로 나타나며 긍정적 단어 사전과 부정적 단어 사전을 활용하여 의견 분석을 해결해 나갔고 단일 이슈의 여론의 방향을 넘어서 두 가지 이슈간의 평판들을 같은 기준으로 분석하여 하나의 대시보드를 작성하므로 두 이슈의 네티즌 의견들을 한 눈에 확인할 수 있었다.

본 연구 과정에는 누구나 자유롭게 소프트웨어를 코딩하여 유용한 기술을 공유하며 사용자들이 서로의 기술을 함께 업데이트 가능한 오픈 소프트웨어를 활용하였으며 리눅스(Linux), MySql DB, HTML, CSS, JavaScript, jQuery, R program, Python 등이 사용되었다.

웹 페이지는 Linux CentOS 서버 환경에 Java 환경의 언어를 동적으로 실행하며 PHP(Hypertext Preprocessor), HTML, JavaScript, jQuery, Ajax 등의 스크립트 언어를 사용하여 사용자 인터페이스를 만들었으며 신문 기사 텍스트 분석엔 R 프로그램을 사용하였다. 또한, 웹 페이지 로딩 시간을 줄이기 위하여 각 페이지를 불러오는 모든 작업은 ajax를 이용하여 실시간으로 페이지 이동 없이 설계되었다. 누리꾼들의 의견을 크롤링하는 과정은 Python을 활용하여 데이터 수집에 보다 빠른 결과를 얻어내었다.

실험은 연구자의 연구 시점에 사회적 이슈인 한반도 사드(고고도미사일 방어체계·THAAD) 배치 결정에 따른 중국과 미국사이의 정치, 경제, 외교적인 이슈를 중심으로 진행하였다. 2017년 2월 15일부터 3월 14일 사이 “사드” 키워드 관련 뉴스 기사들을 관찰하고 뉴스 기사 건수를 분석하여 큰 이슈 시점을 찾아 3월 3일과 3월 7일의 뉴스 기사를 대상으로 진행되었다. 자료 수집 과정인 크롤링 과정과 분석 과정인 시각화 과정이 순조롭게 진행되었으며 편리한 사용자인터페이스로 누구나 편하게 사용 가능한 웹

페이지로 설계되었다.

제 2 절 연구의 시사점

본 연구의 웹 마이닝 분석은 방대한 데이터 내에서 패턴을 찾아 예측을 하는 것이다. 기존 빅데이터 관련 연구는 방대한 자료 수집과 자료 분석이 이원화되어 연구되고 있었다. 연구 대상의 자료가 모집단 전체를 대상으로 한다면 자료 수집 시간과 그 자료를 분석 단위의 식별 가능한 자료로 분해하여 분석 기법이 적용되는 시간이 굉장히 오래 소요되며, 이는 많은 데이터 량과 함께 분석 시간도 비례관계를 보여준다. 이렇듯 연구 시간에 비례하여 여론을 예측하기 어려운 현실의 한계점이 많이 기술되고 있다(Lee and Lee, 2015; 이철성 외, 2013; 임좌상・김진만, 2014).

자료 수집에서 처리와 결과까지 일괄 처리과정을 웹 페이지로 구현하는 연구는 이례적이다. 실시간 처리를 위하여 사무용 PC급들을 활용한 분산처리시스템 구축과 현장 연구의 실제 웹 환경에서의 시스템 사용 등 프로세스 전 과정을 사용자가 확인하며 연구 초점에 가까이 접근할 수 있도록 다양한 사용자 인터페이스 환경을 제공하는 시스템이라 생각한다.

비즈니스 관점의 본 연구는 최종 사용자가 다차원 정보를 직접 검색하여 대화식으로 정보를 분석하고 의사결정에 활용이 가능한 시스템이라 할 수 있다. 빅데이터에서 쉽게 나타나지 않지만 존재하고 있는 의미 있는 정보를 찾을 수 있다. 데이터간의 관계, 패턴, 규칙 등을 찾아내고 실시간 분석을 통하여 기업의 경쟁력 확보를 위한 의사결정에 직접적인 도움이 되는 그래프나 지식으로 변환하는 시스템이라 볼 수 있다. 더 나아가 인공 지능적인 요소를 추가하여 특정 변수나 사건을 예측하게 하고 비즈니스 규칙을 세우기 위한 변수들 간의 규칙을 파악하도록 해주는 연구로 발전하길 기대

한다. 의사결정 지원 시스템인 OLAP(On-Line Analytical Processing)는 수집된 정보를 다양하게 분석할 수 있도록 제공함으로써 기업의 의사결정과 운영을 지원하는 역할을 수행한다. 기업의 운영을 지원하는 데이터에서 온라인 분석 처리를 통한 의사결정의 효과적인 지원을 하는 OLAP와는 차이가 있다고 볼 수 있다. OLAP는 기업의 데이터웨어하우스에서 데이터를 제공받아 CRM 전략을 위해 필요한 다양한 분석 활동을 지원하는 반면 본 시스템은 기업 밖에서의 고객 중심의 방대한 데이터를 수집 가공 처리하여 CRM 전략에 필요한 분석 활동에 기여한다고 볼 수 있다.

본 연구의 목적에서도 제시하였듯이 첫째, 원하는 인터넷 뉴스 기사의 웹 마이닝 과정을 통하여 실시간 자료 수집과 텍스트 마이닝(Text mining) 분석을 통하여 다양한 시각화를 제시하였다. 크롤링 과정을 PHP와 Python을 활용하여 검색 결과의 뉴스 기사와 특정 기사의 댓글 내용, 댓글에 관한 통계자료를 함께 제공하였다. 또한, 워드 클라우드, 군집분석, 네트워크 그래프, 연관 규칙 분석을 활용하여 언론의 방향을 읽을 수 있는 시각화 과정을 통하여 여론을 예측하는데 도움이 될 것으로 본다.

둘째, 실시간 수집된 뉴스의 네티즌(Netizen, 누리꾼) 댓글을 통한 오피니언 마이닝(Opinion Mining) 분석 과정에 연구자의 연구 초점에 더 근접할 수 있도록 표준 사전의 단어 등록과 삭제 등 사용자 인터페이스를 제공하였다. 개발 환경의 스크립트에서 단어 삽입과 삭제가 아니라 실제 웹 페이지에서 연구자의 판단에 따라 적용할 수 있도록 설계하였다. 또한, 단일 여론분석도 가능하지만 두 대립되는 의견들을 같은 조건으로 분석하여 서로의 평판을 확인할 수 있어 평판 분석도 함께 탑재하여 유용하게 사용할 수 있다.

셋째, 방대한 자료를 신속히 처리하기 위한 분산처리기술인 하둡(Hadoop)을 사용하여 연구 프로세스에 적용하였다. 통계학과 공학에서 저

비용으로 시스템 효율을 높이는 연구가 많이 진행되고 있다. 경영학에서 이러한 융합적인 시도를 통하여 고객의 니즈를 판단하고 활용하는 것이 이후 연구자들의 새로운 기초 자료가 될 수 있을 것이다.

넷째, 본 연구에서 사용된 표준 사전 추가 키워드와 불필요한 단어 제거 사전 키워드 등 시스템에서 사용된 키워드를 공개하므로 이후 연구자들의 많은 도움이 될 것으로 판단된다. 또한, 오픈 소스 프로그램을 활용한 연구에 알고리즘을 공개하여 비전공자들의 빅데이터 연구에 보다 많은 힘을 보태는데 공헌할 것으로 기대한다.

또한, 많은 연구자들이 고객의 소리에 귀 기울이며 소비자의 패턴과 니즈 분석, 의미 기반 분석 등의 기능을 향후 추가하여 신뢰성 있는 개발을 기대한다. 단순 빈도 분석을 넘어 연관 분석, 군집 분석, 분류 분석, 예측 분석까지 실시간 처리가 가능한 시스템 개발에 사용자의 대기 시간을 줄이는데 또 다른 지표가 될 것으로 기대한다.

제 3 절 연구의 한계점 및 향후 연구

본 연구는 실시간 크롤링을 통한 웹 마이닝 최적화 분산처리시스템을 설계하고 현장연구를 통하여 실제 연구 결과를 웹 페이지로 제작하였다. 개발 환경에서 적용하지 못한 몇 가지 과제를 제시하고자 한다.

첫째, 빅데이터 분석 기술을 웹 환경에 적용하여 최적화한 본 연구를 다양한 사회 이슈를 분석하는 곳에 적용하는 문제이다. 뉴스 기사 이외 상품이나 서비스 구매 이후 구매 후기 분석을 통하여 실시간 고객의 니즈분석의 연구가 되어야 한다. 오픈 마켓의 다양한 상품의 구매 후기를 비롯하여 맛집 후기, 영화 후기 등 고객들의 의견 분석 범위를 확장할 필요가 있다고 본다.

둘째, 텍스트 중심의 데이터 수집을 사용하여 연구가 진행되었다. 보다 다양한 미디어를 분석 데이터로 활용하는 노력이 절실히 필요해 보인다. 웹 환경은 편리한 UI 환경으로 변화해 가면서 의미 있는 이미지 표현들이 증가하기 때문이다. 사회 인문학, 경영학 등 공학 및 공간 지리 등 다양한 학문들의 융합된 연구들이 앞으로 더욱 활발히 진행되길 바란다.

셋째, 하드웨어 발달과 초고속 네트워크의 실현으로 고객들의 표현이 다양해지고 있다. 본 연구에서 일부 하둡을 활용한 분석을 진행하였다. 하지만 분석 과정에 대하여 국한적으로 적용하면서 전체 프로세스 과정에서 지대한 영향을 주기엔 미진한 부분이 발견되었다. 물론 크롤링 과정이 단일 컴퓨터에서 처리되어야 하는 현 시스템상황이다 보니 가장 많은 시간이 소요되고 있다. 이러한 부분을 해결하기 위하여 시스템 과정이 분산 처리 시스템에서 이루어진다면 보다 빠른 결과를 기대할 수 있는 대기 시간이 현저히 줄어들 것으로 본다. 자료 수집에 관한 연구가 더욱 진행되어야 할 것이다.

참 고 문 헌

< 국 내 문 헌 >

강정배, 최은영, 나운환, “자연어 처리 기술을 활용한 문제행동 유형 분석 연구,” 특수교육재활과학연구, 제52권, 제2호, 2013, pp. 209-237.

강한훈, 유성준, 한동일, “다양한 계층 트리 구조를 갖는 쇼핑몰 상에서의 상품평 수집을 위한 웹 크롤러 래퍼의 설계 및 구현,” 한국지능시스템학회 논문지, 제20권, 제3호, 2010, pp. 318-325.

구흥서, “WWW과 데이터베이스 연동기술의 조사분석,” 정보과학회지, 제18권, 제4호, 2000, pp. 32-40.

김승우, 김남규, “오피니언 분류의 감성사전 활용효과에 대한 연구,” 지능정보연구, 제20권, 제1호, 2014, pp. 133-148.

김진국, 최성수, 지수영, 류관희, “데이터베이스 연동을 통한 빅데이터 분석결과 가시화,” 한국빅데이터서비스학회 논문지2, 2015, pp. 1-10.

류석영, “자바스크립트 웹 앱 분석과 결함 검출,” 정보과학회지, 제34권, 제3호, 2016, pp. 10-14.

문상식, 김기홍, “IT 환경 변화에 따른 한국의 오픈소스 소프트웨어의 정책방향 연구,” 인터넷전자상거래연구, 제14권, 제1호, 2014, pp. 203-221.

박경미, 황규백, “자연어처리 기반 바이오 텍스트 마이닝 시스템,” 정보과학회논문지: 컴퓨팅의 실제 및 레터, 제17권, 제4호, 2011, pp.

205-213.
박대민, “장기 시계열 내용 분석을 위한 뉴스 빅데이터 분석의 활용 가능성,” 한국언론학보, 제60권, 제5호, 2016, pp. 353-407.
박소연, “검색 포털들의 검색어 추천 서비스 분석 평가: 네이버와 구글의 연관 검색어 서비스를 중심으로,” 정보관리학회지, 제30권, 제2호, 2013, pp. 297-315.
서지훈, 조혜진, 최진탁, “한국 문법의 반의어 규칙을 적용한 오피니언 감성사전 설계,” 한국정보기술학회논문지, 제13권, 제2호, 2015, pp. 109-117.
손수아, 박석천, “IoT 기반 실시간 시각화 알고리즘을 이용한 스마트가드닝 시스템 설계 및 구현,” 인터넷정보학회논문지, 제16권, 제6호, 2015, pp. 31-37.
안정국, 김희웅, “Building a Korean Sentiment Lexicon Using Collective Intelligence,” 지능정보연구, 제21권, 제2호, 2015, pp. 49-67.
윤영선, “온라인 리뷰가 온라인 쇼핑행동에 미치는 영향,” 국제회계연구, 제52권, 2013, pp. 139-156.
이종화, 이현규, “오픈소스 소프트웨어를 활용한 자연어 처리 패키지 제작에 관한 연구,” 정보시스템연구, 제25권, 제4호, 2016, pp. 121-139.
이종화, 이현규, “오피니언 마이닝을 통한 국내와 수입 의류 제품에 대한 고객 평판 연구,” 인터넷전자상거래연구, 2015, 제15권, 제3호, pp. 223-234.
이철성, 최동희, 김성순, 강재우, “한글 마이크로블로그 텍스트의 감정 분류 및 분석,” 정보과학회논문지: 데이타베이스, 제40권, 제3호,

2013, pp. 159-167.

임좌상, 김진만, “한국어 트위터의 감정 분류를 위한 기계학습의 실증적 비교,” 멀티미디어학회논문지, 제17권, 제2호, 2014, pp. 232-239.

장경애, 박상현, 김우제, “인터넷 감정기호를 이용한 긍정/부정 말뭉치구축 및 감정분류 자동화,” Journal of KIISE, 제42권, 제4호, 2015, pp. 512-521.

장명현, 박대우, 이예원, “Live site 개념을 도입한 웹사이트 저작도구의 실시간 자동화 모델에 관한 연구,” 한국컴퓨터정보학회, 제19권, 제2호, 2011, pp. 175-177.

정민영, “실시간 검색어 연관 분석을 통한 핵심 이슈 선정,” Journal of Digital Convergence, 제13권, 제12호, 2015, pp. 161-169.

정민영, “포털사이트 실시간 검색키워드의 주간 핵심 이슈 선정 및 차이 분석,” 디지털융복합연구, 제14권, 제12호, 2016, pp. 237-243.

정원기, 문수묵, “웹 페이지 자바스크립트 분석과 자바스크립트 엔진의 성능 평가,” 한국정보과학회, 제37권, 제2호, 2010, pp. 181-182.

최민석, “홍보 효과 증진을 위한 페이스북 팬페이지 분석 시스템 개발,” 디지털융복합연구, 제13권, 제12호, 2015, pp. 135-142.

최승배, 강창완, “웹마이닝 기법을 이용한 학과 홈페이지 분석,” Journal of The Korean Data Analysis, 제13권, 제1호, 2011, pp. 317-329.

Bose, I. and Mahapatra, R. K., “Business data mining - a machine learning perspective,” Information & management, Vol. 39, No. 3, 2001, pp. 211-225.

Chang, C. Y., Jang, J. H., Kim, S, H., Lee, H. K. and Lee, C.

H., "A Study on the Efficient Patent Search Process using Big Data Analysis Tool R," Journal of Korea Safety Management & Science, Vol. 15, No. 4, 2013, pp. 289-294.

Hong, J. P. and Cha, J. W., "Error Correction of Sejong Morphological Annotation Corpora using Part-of-Speech Tagger andFrequency Information," Journal of KISS : Software and Applications, Vol. 40, No. 7, 2013, pp. 417-428.

Kam M. and Song M., "A Study on Differences of Contents and Tones of Arguments among Newspapers Using Text Mining Analysis," Journal of intelligence and information systems, Vol. 8, No. 3, 2012, pp. 53-77.

Kim, W. S., Lee, J. H., Park, j. W. and Choi, j. H.,"A Technique of the Approval Rating Analysis for Political Party Using Opinion Mining," Journal of Korean Institute of Information Technology, Vol. 12, No. 10, 2014, pp. 133-141.

Le, H., and Lee, H. K., "Exploring Relationship Between Social ICT Issues And Academic Research Interests Through Text Mining Analysis," The Journal of Internet Electronic Commerce Research, Vol. 14, No. 5, 2014, pp. 161-180.

Le, H., Lee, J. H. and Lee, H. K., "Purchase Process Aspect-based Opinion Mining : An Application for Online Shopping Mall," The Journal of Internet Electronic

Commerce Research, Vol. 15, No. 2, 2015, pp. 15-28.

Lee, J. H. and Lee, H. K., "A Study on Unstructured Text Mining Algorithm through R Programming based on Data Dictionary," Journal of the Korea Society Industrial Information System, Vol. 20, No. 2, 2015, pp. 113-124.

Lee, J. H., Le, H. S., and Lee, H. K., "Research on Methods for Processing Nonstandard Korean Words on Social Network Services," Journal of the Korea Industrial Information Systems Research, Vol. 21, No. 3, 2016, pp. 35-46.

Sim, K. S., "Syllable-based POS Tagging without Korean Morphological Analysis," Korean Journal of Cognitive Science, Vol. 22, No. 3, 2011, pp. 327-345.

Won, J. Y. and Kim, D. G., "Deduction of Social Risk Issues Using Text Mining," Journal of safety and crisis management, Vol. 10, No. 7, 2014, pp. 33-52.

Yun, B. H., "Natural Language Processing based Information Extraction for Newspapers," Journal of Korean Institute of Information Technology, Vol. 6, No. 4, 2008, pp. 188-195.

< 국 외 문 헌 >

Bai, X., "Predicting consumer sentiments from online text," *Decision Support Systems*, Vol. 50, No. 4, 2011, pp. 732-742.

Balahur, A., Hermida, J. M. and Montoyo, A., "Detecting implicit expressions of emotion in text: A comparative analysis," *Decision Support Systems*, Vol. 53, No. 4, 2012, pp. 742-753.

Bari, P. and Chawan, P. M., "Web usage mining," *Journal of Engineering, Computers & Applied Sciences (JEC&AS)*, Vol. 2, No. 6, 2013, pp. 34-38.

Bian, J., Topaloglu, U. and Yu, F., "Towards large-scale twitter mining for drug-related adverse events," *In Proceedings of the 2012 international workshop on Smart health and wellbeing*, 2012, pp. 25-32.

Bin, W. and Zhijing, L., "Web mining research," *In Computational Intelligence and Multimedia Applications 2003*, 2003, pp. 84-89.

Bird, S., Klein, E., and Loper, E., Natural language processing with Python: analyzing text with the natural language toolkit, O'Reilly Media, Inc, 2009.

Black, E. W., "Wikipedia and academic peer review: Wikipedia as a recognised medium for scholarly publication?," *Online*

Information Revie, Vol. 32, No. 1, 2008, pp. 73~88.

Brin, S. and Page, L., "The anatomy of a large-scale hypertextual web search engine," *Computer networks and ISDN systems*, Vol. 30, No. 1, 1998, pp. 107-117.

Cachia, R., R. Compano and O. D. Costa, "Grasping the potential of online social networks for foresight," *Technological Forecasting and Social Change*, Vol. 74, No. 8, 2007, pp. 1179~1203.

Cao, Q., Duan, W. and Gan, Q., "Exploring determinants of voting for the 'helpfulness' of online user reviews: A text mining approach," *Decision Support Systems*, Vol. 50, No. 2, 2011, pp. 511-521.

Chakarbarti S., Mining the Web: Discovering knowledge from hypertext data. Morgan Kaufmann Publisher, San Francisco, 2003.

Chakrabarti, S., Mining the Web: Discovering knowledge from hypertext data, Elsevier, 2002.

Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M. and Gruber, R. E., "Bigtable: A distributed storage system for structured data," *ACM Transactions on Computer Systems(TOCS)*, Vol. 26, No.2, 2008, pp. 1-14.

Chau, R., Yeh, C. H. and Smith, K. A., "Personalized multilingual web content mining," *In International Conference on Knowledge-Based and Intelligent*

Information and Engineering Systems, 2004, pp. 155-163.

Che, W., Li, Z. and Liu, T., "Ltp: A chinese language technology platform," *In Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, 2010, pp. 13-16.

Choi, J. I. and Chang, J. H., "Implementation of a smart ticketing system using mobile devices," *Journal of the Korea Industrial Information Systems Research*, Vol. 16, No. 5, 2011, pp. 63-71.

Clendaniel, S., "Profitablility and Mining Web Data: Avoiding the Path to Red Ink-Scott Clendaniel discusses how a well-intentioned focus on customer response rates and similar dependent variables may cause," *PC AI*, Vol. 15, No. 5, 2001, pp. 31-35.

Darmont, J., Boussaid, O. and Bentayeb, F., "Warehousing web data," arXiv preprint arXiv:0705.1456.pdf, 2007.

Dinucă, C. E., "Web Structure Mining," *Annals of the University of Petrosani, Economics*, Vol. 11, No. 4, 2011, pp. 73-84.

Duric, A. and Song, F., "Feature selection for sentiment analysis based on content and syntax models," *Decision Support Systems*, Vol. 53, No. 4, 2012, pp. 704-711.

Fan, T. K. and Chang, C. H., "Blogger-centric contextual advertising," *Expert systems with applications*, Vol. 38, No. 3, 2011, pp. 1777-1788.

Fang, X. and Sheng, O. R. L., "LinkSelector: A Web mining approach to hyperlink selection for Web portals," *ACM Transactions on Internet Technology*, Vol. 4, No. 2, 2004, pp. 209-237.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P., "From data mining to knowledge discovery in databases," *AI magazine*, Vol. 17, No. 3, 1996, pp. 37-54.

Fenstermacher, K. D. and Ginsburg, M., "Client-side monitoring for Web mining," *Journal of the American Society for Information Science and Technology*, Vol. 54, No. 7, 2003, pp. 625-637.

Ghemawat, S., Gobioff, H. and Leung, S. T., "The Google file system," *In ACM SIGOPS operating systems review*, Vol. 37, No. 5, 2003, pp. 29-43.

Graves, S., Ramachandran, R., Keiser, K., Maskey, M., Lynnes, C. and Pham, L., "Deployable suite of data mining web services for online science data repositories," *In 23rd Conference on IIPS*, 2007, pp. 1-8.

Guan, S. U. and McMullen, P., "Organizing information on the next generation web-design and implementation of a new bookmark structure," *International Journal of Information Technology & Decision Making*, Vol. 4, No. 1, 2005, pp. 97-115.

Hafen, R., Gibson, T., van Dam, K. K. and Critchlow, T., "Power Grid Data Analysis with R and Hadoop," In Data

Mining Applications with R, 2014, pp. 1-34.

Han, J. and Chang, K. C., "Data mining for web intelligence," *Computer*, Vol. 35, No. 11, 2002, pp. 64-70.

Harish, D., Anusha, M. and Daya, S. K., "BIG DATA ANALYSIS USING RHADOOP," *International Journal of Innovative Research in Advanced Engineering(IJIRAE)*, Vol. 2, Issue. 4, 2015, pp. 180-185.

Hay, B., Wets, G. and Vanhoof, K., "Mining navigation patterns using a sequence alignment method," *Knowledge and information systems*, Vol. 6, No. 2, 2004, pp. 150-163.

Heerschop, B., Goossen, F., Hogenboom, A., Frasincar, F., Kaymak, U. and de Jong, F., "Polarity analysis of texts using discourse structure," *In Proceedings of the 20th ACM international conference on Information and knowledge management*, 2011, pp. 1061-1070.

Henzinger, M., "The Past, Present, and Future of Web Search Engines," *Proceedings of 31st International Colloquium*, 2004, pp. 3-3.

Hu, N., Bose, I., Koh, N. S. and Liu, L., "Manipulation of online reviews: An analysis of ratings, readability, and sentiments," *Decision Support Systems*, Vol. 52, No. 3, 2012, pp. 674-684.

International Telecommunication Union(UIT), "The Internet of Things," 2005

Kang, H., Yoo, S. J. and Han, D., "Senti-lexicon and improved Naive Bayes algorithms for sentiment analysis of restaurant reviews," *Expert Systems with Applications*, Vol. 39, No. 5, 2012, pp. 6000-6010.

Kemp, R., "Fourth industrial revolution," *The Lawyer*, Vol. 31, No. 21, 2016, pp. 12.

Keshtkar, F. and Inkpen, D., "A bootstrapping method for extracting paraphrases of emotion expressions from texts," *Computational Intelligence*, Vol. 29, No. 3, 2013, pp. 417-435.

Kolari, P. and Joshi, A., "Web mining: Research and practice," *Computing in science & engineering*, Vol. 6, No. 4, 2004, pp. 49-53.

Kosala, R. and Blockeel, H., "Web mining research: A survey," *ACM Sigkdd Explorations Newsletter*, Vol. 2, No. 1, 2000, pp. 1-15.

Lakhe, B., In Practical Hadoop Security - Monitoring in hadoop(Ch7), Apress, .2014.

Lane, P. C., Clarke, D. and Hender, P., "On developing robust models for favourability analysis: Model choice, feature sets and imbalanced data," *Decision Support Systems*, Vol. 53, No. 4, 2012, pp. 712-718.

Lau, K. N., Lee, K. H., Ho, Y. and Lam, P. Y., "Mining the web for business intelligence: Homepage analysis in the internet era," *Journal of Database Marketing & Customer*

Strategy Management, Vol. 12, No. 1, 2004, pp. 32-54.

Leo, S., Santoni, F. and Zanetti, G., "Biodoop: Bioinformatics on hadoop," *2009 International Conference on Parallel Processing Workshops*, 2009, pp. 415-422.

Lewis, S., Csordas, A., Killcoyne, S., Hermjakob, H., Hoopmann, M. R., Moritz, R. L. and Boyle, J., "Hydra: a scalable proteomic search engine which utilizes the Hadoop distributed computing framework," *BMC bioinformatics*, Vol. 13, No. 1, 2012, pp. 324.

Lihui, C. and Lian, C. W., "Using Web structure and summarisation techniques for Web content mining," *Information processing & management*, Vol. 41, No. 5, 2005, pp. 1225-1242.

Liu, B., "Sentiment analysis and opinion mining," *Synthesis lectures on human language technologies*, Vol. 5, No. 1, 2012, pp. 1-167.

Liu, B. and Chen-Chuan-Chang, K., "Editorial: special issue on web content mining," *Acm Sigkdd explorations newsletter*, Vol. 6, No. 2, 2004, pp. 1-4.

Liu, B., Blasch, E., Chen, Y., Shen, D. and Chen, G., "Scalable sentiment classification for big data analysis using naive bayes classifier," *In Big Data, 2013 IEEE International Conference on*, 2013, pp. 99-104.

Liu, B., Web Data Mining, 2nd edition, Springer, 2013.

Lu, C. Y., Lin, S. H., Liu, J. C., Cruz-Lara, S. and Hong, J.

S., "Automatic event-level textual emotion sensing using mutual action histogram between entities," *Expert systems with applications*, Vol. 37, No. 2, 2010, pp. 1643-1653.

Lydia, E. L., and Swarup, M. B., "Big Data Analysis using Hadoop components like Flume, MapReduce, Pig and Hive," *International Journal of Science, Engineering and Computer Technology*, Vol. 5, Issue. 11, 2015, 390-394.

Maks, I. and Vossen, P., "A lexicon model for deep sentiment analysis and opinion mining applications," *Decision Support Systems*, Vol. 53, No. 4, 2012, pp. 680-688.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S. and McClosky, D., "The stanford corenlp natural language processing toolkit," *In ACL*, 2014, pp. 55-60.

Martino, F. and Spoto, A., "Social network analysis: A brief theoretical review and further perspectives in the study of information technology," *PsychNology Journal*, Vol. 4, No. 1, 2006, pp. 53-86.

Medhat, W., Hassan, A. and Korashy, H., "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, Vol. 5, No. 4, 2014, pp. 1093-1113.

Mobasher, B., Cooley, R. and Srivastava, J., "Automatic personalization based on web usage mining," *Communications of the ACM*, Vol. 43, No. 8, 2000, pp. 142-151.

Moraes, R., Valiati, J. F. and Neto, W. P. G., "Document-level Sentiment Classification: An empirical Comparison between SVM and ANN," *Expert Systems with Applications*, Vol. 40, No. 2, 2013, pp. 621-633.

Moreo, A., Romero, M., Castro, J. L. and Zurita, J. M., "Lexicon-based comments-oriented news sentiment analyzer system," *Expert Systems with Applications*, Vol. 39, No. 10, 2012, pp. 9166-9180.

Neviarouskaya, A., Prendinger, H. and Ishizuka, M., "Recognition of affect, judgment, and appreciation in text." *In Proceedings of the 23rd international conference on computational linguistics*, 2010, pp. 806-814.

O'Driscoll, A., Daugelaite, J. and Sleator, R. D., "'Big data', Hadoop and cloud computing in genomics," *Journal of biomedical informatics*, Vol. 46, No. 5, 2013, pp. 774-781.

Oancea, B. and Dragoescu, R. M., "Integrating R and hadoop for big data analysis," *Romanian Statistical Review*, 2014, pp. 83-94.

Pabarskaite, Z. and Raudys, A., "A process of knowledge discovery from web log data: Systematization and critical review," *Journal of Intelligent Information Systems*, Vol. 28, No. 1, 2007, pp. 79-104.

Pang, B. and Lee, L., "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, Vol. 2, No. 1 - 2, 2008, pp. 1-135.

Pierrakos, D., Paliouras, G., Papatheodorou, C. and Spyropoulos, C. D., "Web usage mining as a tool for personalization: A survey," *User modeling and user-adapted interaction*, Vol. 13, No. 4, 2003, pp. 311-372.

Pol, K., Patil, N., Patankar, S. and Das, C., "A Survey on Web Content Mining and extraction of Structured and Semistructured data," *In Emerging Trends in Engineering and Technology 2008*, 2008, pp. 543-546.

Qiu, G., He, X., Zhang, F., Shi, Y., Bu, J. and Chen, C., "DASA: dissatisfaction-oriented advertising based on sentiment analysis," *Expert Systems with Applications*, Vol. 37, No. 9, 2010, pp. 6182-6191.

Rao, Y., Li, Q., Mao, X. and Wenyin, L., "Sentiment topic models for social emotion mining," *Information Sciences*, Vol. 266, 2014, pp. 90-100.

Rehurek, R. and Sojka, P., "Software framework for topic modelling with large corpora," *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2010, pp. 46-50.

Rengier, F., Mehndiratta, A., von Tengg-Kobligk, H., Zechmann, C. M., Unterhinninghofen, R., Kauczor, H. U. and Giesel, F. L., "3D printing based on imaging data: review of medical applications," *International journal of computer assisted radiology and surgery*, Vol. 5, No. 4,

2010, pp. 335-341.

Robaldo, L. and Di Caro, L., "Opinionmining-ml," *Computer Standards & Interfaces*, Vol. 35, No. 5, 2013, pp. 454-469.

Scott, J., "Social Network Analysis, 2nd edition," SagePublications, 2000

Sivakumar, P., "Effectual Web Content Mining using Noise Removal from Web Pages," *Wireless Personal Communications*, Vol. 84, No. 1, 2015, pp. 99-121.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y. and Potts, C., "Recursive deep models for semantic compositionality over a sentiment treebank," *In Proceedings of the conference on empirical methods in natural language processing(EMNLP)*, Vol. 1631, 2013, p. 1642-1653.

Song, Q. and Shepperd, M., "Mining web browsing patterns for E-commerce," *Computers in Industry*, Vol. 57, No. 7, 2006, pp. 622-630.

Song, Q. and Shepperd, M., "Mining web browsing patterns for E-commerce," *Computers in Industry*, Vol. 57, No. 7, 2006, pp. 622-630.

Spiliopoulou, M., "Web usage mining for web site evaluation," *Communications of the ACM*, Vol. 43, No. 8, 2000, pp. 127-134.

Srivastava, J., Cooley, R., Deshpande, M. and Tan, P. N.,

"Web usage mining: Discovery and applications of usage patterns from web data," *Acm Sigkdd Explorations Newsletter*, Vol. 1, No. 2, 2000, pp. 12-23.

Srivastava, J., Desikan, P. and Kumar, V., "Web Mining-Accomplishments and Future directions," *Proceedings of National Science Foundation Workshop on Next Generation Data Mining(NGDM'02)*, 2002, pp. 51-56.

Tare, M., Gohokar, I., Sable, J., Paratwar, D. and Wajgi, R., "Multi-class tweet categorization using map reduce paradigm," *International Journal of Computer Trends and Technology(IJCTT)*, Vol. 9, No. 2, 2014, pp. 78-81.

Totad, S. G. and PVGD, P. R., "Amalgamation of web usage mining and web structure mining." *Int. J. of Recent Trends in Engineering and Technology*, Vol. 1, No. 2. 2009, pp. 279-281.

Tyagi, N. K., Solanki, A. K. and Tyagi, S., "An algorithmic approach to data preprocessing in web usage mining," *International journal of information technology and knowledge management*, Vol. 2, No. 2, 2010, pp. 279-283.

Vellingiri, J., Kaliraj, S., Satheeshkumar, S. and Parthiban, T., "A novel approach for user navigation pattern discovery and analysis for web usage mining," *Journal of Computer Science*, Vol. 11, No. 2, 2015, pp. 372-382.

Victor, S. P. and Rex, M. M. X., "Analytical implementation of

web structure mining using data analysis in educational domain," *International Journal of Applied Engineering Research*, Vol. 11, No. 4, 2016, pp. 2552-2556.

Walker, M. A., Anand, P., Abbott, R., Tree, J. E. F., Martell, C. and King, J., "That is your evidence?: Classifying stance in online political debate," *Decision Support Systems*, Vol. 53, No. 4, 2012, pp. 719-729.

Wang, H., Yang, C. and Zeng, H., "Design and implementation of a web usage mining model based on upgrowth and preflxspan," *Communications of the IIMA*, Vol. 6, No. 2, 2015, pp. 69-84.

Witten, I. H., Frank, E., Hall, M. A. and Pal, C. J., Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann, 2016.

Xu, G., Zhang, Y. and Li, L., "Web content mining," *In Web Mining and Social Networking*, 2011, pp. 71-87.

Yan, W., "Block matrix-based marpreduce pagerank algorithm web structure mining applied effect research," *BioTechnology: An Indian Journal*, Vol. 10, No. 5, 2014, pp. 1345-1351.

Yan-Yan, Z., Bing, Q. and Ting, L., "Integrating intra-and inter-document evidences for improving sentence sentiment classification," *Acta Automatica Sinica*, Vol. 36, No. 10, 2010, pp. 1417-1425.

Zhang, H., Chen, Z., Li, M. and Su, Z., "Relevance feedback

and learning in content-based image search," *World Wide Web*, Vol. 6, No. 2, 2003, pp. 131-155.

Zhang, Q. and Segall, R. S., "Web mining: a survey of current research, techniques, and software," *International Journal of Information Technology & Decision Making*, Vol. 7, No. 4, 2008, pp. 683-720.

Zhang, W., Xu, H. and Wan, W., "Weakness Finder: Find product weakness from Chinese reviews by using aspects based sentiment analysis," *Expert Systems with Applications*, Vol. 39, No. 11, 2012, pp. 10283-10291.

Zhou, L., Li, B., Gao, W., Wei, Z. and Wong, K. F., "Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities," *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2011, pp. 162-171.

Zirn, C., Niepert, M., Stuckenschmidt, H. and Strube, M., "Fine-Grained Sentiment Analysis with Structural Features," *In IJCNLP*, 2011, pp. 336-344.

〈부록 A〉 명사 추가 단어 리스트

순번	단어	순번	단어	순번	단어	순번	단어	순번	단어	순번	단어
1	가정사	53		105		157		209		261	
2	경기지사	54		106		158		210		262	
3	고영태	55		107		159		211		263	
4	국민의당	56		108		160		212		264	
5	김정은	57		109		161		213		265	
6	노무현	58		110		162		214		266	
7	대선주자	59		111		163		215		267	
8	대통령	60		112		164		216		268	
9	대한민국	61		113		165		217		269	
10	문재인	62		114		166		218		270	
11	박근혜	63		115		167		219		271	
12	비대위	64		116		168		220		272	
13	비판	65		117		169		221		273	
14	새누리당	66		118		170		222		274	
15	서울신문	67		119		171		223		275	
16	심상정	68		120		172		224		276	
17	안철수	69		121		173		225		277	
18	안희정	70		122		174		226		278	
19	이재용	71		123		175		227		279	
20	입건	72		124		176		228		280	
21	자유한국당	73		125		177		229		281	
22	정의당	74		126		178		230		282	
23	지시	75		127		179		231		283	
24	충남지사	76		128		180		232		284	
25	태극기	77		129		181		233		285	
26	페이스북	78		130		182		234		286	
27		79		131		183		235		287	
28		80		132		184		236		288	
29		81		133		185		237		289	
30		82		134		186		238		290	
31		83		135		187		239		291	
32		84		136		188		240		292	
33		85		137		189		241		293	
34		86		138		190		242		294	
35		87		139		191		243		295	
36		88		140		192		244		296	
37		89		141		193		245		297	
38		90		142		194		246		298	
39		91		143		195		247		299	
40		92		144		196		248		300	
41		93		145		197		249		301	
42		94		146		198		250		302	
43		95		147		199		251		303	
44		96		148		200		252		304	
45		97		149		201		253		305	
46		98		150		202		254		306	
47		99		151		203		255		307	
48		100		152		204		256		308	
49		101		153		205		257		309	
50		102		154		206		258		310	
51		103		155		207		259		311	
52		104		156		208		260		312	

〈부록 B〉 의미 없는 단어 리스트

순번	단어	순번	단어	순번	단어	순번	단어	순번	단어	순번	단어
1	at	53	만큼	105	출시	157		209		261	
2	back	54	말했	106	측은	158		210		262	
3	br	55	모르	107	측이	159		211		263	
4	call	56	무단	108	클릭	160		212		264	
5	co	57	밝혔	109	필요	161		213		265	
6	feedid	58	배포	110	하게	162		214		266	
7	feedname	59	부근	111	하기	163		215		267	
8	gpu	60	부터	112	하면	164		216		268	
9	ionhref	61	분기	113	하지	165		217		269	
10	is	62	분야	114	핫클릭	166		218		270	
11	kr	63	사람들	115	해서	167		219		271	
12	leeyoo	64	사실	116	활용	168		220		272	
13	loc	65	사이	117		169		221		273	
14	msi	66	상황	118		170		222		274	
15	news	67	생각	119		171		223		275	
16	op	68	선전했	120		172		224		276	
17	pf	69	세상	121		173		225		277	
18	PLUS	70	수준	122		174		226		278	
19	seoul	71	시장을	123		175		227		279	
20	sports	72	아침	124		176		228		280	
21	the	73	얘기	125		177		229		281	
22	url	74	어쇼	126		178		230		282	
23	uwg	75	여러분	127		179		231		283	
24	var	76	였다고	128		180		232		284	
25	vs	77	오늘	129		181		233		285	
26	yna	78	오른쪽	130		182		234		286	
27	가운데	79	오전	131		183		235		287	
28	강조했	80	오후	132		184		236		288	
29	건지라고	81	왼쪽	133		185		237		289	
30	경우	82	우리	134		186		238		290	
31	공개	83	으로	135		187		239		291	
32	관련	84	이날	136		188		240		292	
33	금지	85	이번	137		189		241		293	
34	기록했	86	이상	138		190		242		294	
35	기온	87	이유	139		191		243		295	
36	기자	88	이젠	140		192		244		296	
37	까지	89	이후	141		193		245		297	
38	끊임없	90	있다고	142		194		246		298	
39	나가겠다	91	있습니	143		195		247		299	
40	다고	92	저작권자	144		196		248		300	
41	다음	93	전문	145		197		249		301	
42	다하겠	94	전했	146		198		250		302	
43	단지	95	제공	147		199		251		303	
44	대부분	96	제목	148		200		252		304	
45	대표	97	조금	149		201		253		305	
46	대한	98	주장했	150		202		254		306	
47	돌이	99	지방	151		203		255		307	
48	동안	100	지역	152		204		256		308	
49	되기	101	지원	153		205		257		309	
50	됩니	102	직후	154		206		258		310	
51	들이	103	참석	155		207		259		311	
52	때문	104	최고	156		208		260		312	

〈부록 C〉 긍정적 단어 사전 리스트

순번	단어	순번	단어	순번	단어	순번	단어	순번	단어	순번	단어
1	^ㅋ	53	관대	105	매혹	157	섬세	209	연인	261	이점
2	^ㅎ	54	굉장	106	맵시	158	섭리	210	열광	262	익살
3	가능	55	교양	107	맹목	159	성공	211	열렬	263	인격
4	가치	56	구애	108	명랑	160	성숙	212	열망	264	인기
5	간결	57	권위	109	명백	161	성실	213	열성	265	인내
6	간단	58	권유	110	명성	162	성인	214	열심	266	인도
7	감동	59	권장	111	명예	163	성자	215	열의	267	인상
8	감사	60	근면	112	명작	164	성취	216	열정	268	인재
9	감상	61	긍정	113	명확	165	세련	217	엽기	269	인정
10	감질	62	기부	114	모범	166	섹시	218	영감	270	인품
11	감탄	63	기쁨	115	묘미	167	소름	219	영광	271	일관
12	강력	64	기억	116	무료	168	소신	220	영리	272	일류
13	강화	65	기적	117	무해	169	소중	221	영양	273	일치
14	개량	66	깔끔	118	미덕	170	솔직	222	영예	274	자격
15	개방	67	깨끗	119	미래	171	솜씨	223	영웅	275	자랑
16	개별	68	끈기	120	미소	172	숙달	224	영원	276	자비
17	개선	69	나긋	121	미화	173	숙련	225	영향	277	자유
18	개혁	70	낙관	122	민감	174	순결	226	예리	278	자진
19	갱신	71	냥만	123	민첩	175	순수	227	예의	279	장관
20	거대	72	녁녁	124	박수	176	순종	228	예절	280	장점
21	거룩	73	널찍	125	박식	177	숭고	229	온화	281	재능
22	건강	74	노력	126	박애	178	숭배	230	옹호	282	재미
23	건설	75	능가	127	받침	179	스릴	231	용감	283	재치
24	건장	76	능동	128	발굴	180	승리	232	용기	284	저금
25	건전	77	능력	129	발명	181	승인	233	용맹	285	저렴
26	걸작	78	다산	130	번영	182	시원	234	용서	286	적당
27	걸출	79	다양	131	보람	183	신기	235	우상	287	적응
28	격려	80	다작	132	보물	184	신념	236	우수	288	적절
29	결정	81	다재	133	보상	185	신동	237	우아	289	전문
30	겸손	82	다정	134	보석	186	신뢰	238	우월	290	전설
31	경건	83	다행	135	보완	187	신선	239	우정	291	전심
32	경계	84	단순	136	보장	188	신성	240	우호	292	절약
33	경례	85	단언	137	보조	189	신속	241	원활	293	절정
34	경사	86	달성	138	보충	190	신중	242	웰빙	294	접근
35	경외	87	당당	139	보호	191	실제	243	위안	295	정교
36	경의	88	당연	140	복구	192	실현	244	위엄	296	정력
37	경작	89	대단	141	복리	193	씩씩	245	유능	297	정류
38	경탄	90	대접	142	부유	194	안심	246	유리	298	정밀
39	계몽	91	도덕	143	부자	195	안전	247	유망	299	정상
40	계발	92	도량	144	분리	196	안정	248	유명	300	정제
41	고귀	93	도약	145	불꽃	197	압도	249	유용	301	정중
42	고급	94	도움	146	사교	198	애국	250	유익	302	정직
43	고명	95	독창	147	사랑	199	애정	251	유창	303	정착
44	고무	96	따뜻	148	사려	200	애호	252	유쾌	304	정통
45	고상	97	리드	149	상냥	201	양심	253	유행	305	정품
46	공감	98	마법	150	상쾌	202	양질	254	은혜	306	정화
47	공개	99	마음	151	선도	203	엄청	255	응원	307	정확
48	공상	100	막대	152	선량	204	업적	256	의인	308	제왕
49	공손	101	만병	153	선명	205	여유	257	의지	309	조화
50	공정	102	만족	154	선의	206	역시	258	이득	310	존경
51	공짜	103	맞춤	155	선호	207	연대	259	이상	311	존엄
52	공평	104	매력	156	설득	208	연상	260	이익	312	존중

순번	단어	순번	단어	순번	단어	순번	단어	순번	단어	순번	단어
313	좋아	365	쾌활	417	호환	469	상상력	521	사랑사는		
314	주도	366	쿨한	418	화려	470	생존자	522	센세이션		
315	준법	367	탁월	419	화사	471	선명도	523	아름다운		
316	준비	368	태평	420	화해	472	설득력	524	아름다움		
317	준수	369	통일	421	확보	473	손재주	525	오아시스		
318	중요	370	통합	422	환상	474	수상자	526	유머러스		
319	지지	371	특권	423	환심	475	수익자	527	제멋대로		
320	지혜	372	특별	424	환영	476	숭배자	528	즐겨찾기		
321	진보	373	특유	425	환호	477	스마트	529	카리스마		
322	진실	374	튼튼	426	환희	478	스타일	530	엄지손가락		
323	진심	375	판독	427	활기	479	신뢰성	531	응원합니다		
324	진정	376	편안	428	활발	480	아이콘	532	좋은사람들		
325	진지	377	평안	429	활용	481	애국자				
326	진취	378	평온	430	횟팅	482	에너지				
327	진행	379	평판	431	황금	483	엘리트				
328	질서	380	평화	432	황홀	484	연속성				
329	찬사	381	포부	433	회복	485	옹호자				
330	찬성	382	포옹	434	획득	486	위대한				
331	찬양	383	품위	435	횡재	487	유선형				
332	참담	384	품행	436	효율	488	유연성				
333	창조	385	풍부	437	훌륭	489	융통성				
334	창출	386	풍성	438	흥미	490	자발적				
335	천국	387	풍요	439	흥분	491	자비심				
336	천사	388	풍족	440	흥행	492	자신감				
337	천재	389	프로	441	희망	493	자유로				
338	첨단	390	ㅎㅎ	442	희열	494	자존심				
339	청결	391	학식	443	결합력	495	적극성				
340	촉진	392	합리	444	경쟁력	496	적임자				
341	총명	393	합법	445	경제력	497	정당성				
342	최강	394	합승	446	고화질	498	좋았구				
343	최고	395	해방	447	급성장	499	죄가없				
344	최상	396	해학	448	기동성	500	즐거운				
345	최신	397	핸섬	449	기사도	501	즐거움				
346	최적	398	행보	450	능력자	502	참을성				
347	최초	399	행복	451	다목적	503	책임감				
348	추월	400	행운	452	다용도	504	챔피언				
349	추정	401	향기	453	단순화	505	최고점				
350	추천	402	향상	454	달콤한	506	최첨단				
351	축복	403	향유	455	대용량	507	충실도				
352	축제	404	허용	456	대인기	508	통찰력				
353	축하	405	헌신	457	독창력	509	판단력				
354	충분	406	혁명	458	동정심	510	행복감				
355	충성	407	혁신	459	로맨틱	511	화이팅				
356	충실	408	현대	460	마음속	512	활성화				
357	치유	409	현명	461	무제한	513	획기적				
358	친목	410	현실	462	무조건	514	휴대용				
359	친밀	411	현저	463	베스트	515	경이로운				
360	친선	412	혈통	464	보너스	516	기념비적				
361	친절	413	협력	465	붙임성	517	낙천주의				
362	침착	414	혜택	466	비폭력	518	많이있다				
363	칭찬	415	호소	467	사교계	519	변치말고				
364	쾌적	416	호화	468	사용자	520	불가사의				

〈부록 D〉 부정적 단어 사전 리스트

순번	단어	순번	단어	순번	단어	순번	단어	순번	단어	순번	단어
1	ㅜㅜ	53	경련	105	괴짜	157	낙심	209	대폭	261	무력
2	가난	54	경멸	106	교란	158	낙인	210	덤프	262	무례
3	가래	55	경보	107	교묘	159	난리	211	덤핑	263	무모
4	가망	56	경솔	108	교살	160	난색	212	도끼	264	무법
5	가뭄	57	곁눈	109	교전	161	난처	213	도당	265	무시
6	가스	58	계략	110	교활	162	난치	214	도망	266	무식
7	가시	59	계승	111	구걸	163	난파	215	도주	267	무심
8	가증	60	고가	112	구속	164	난폭	216	독단	268	무작
9	가짜	61	고난	113	구제	165	남용	217	독살	269	무장
10	가책	62	고뇌	114	구토	166	남풍	218	독설	270	무지
11	간과	63	고독	115	굴복	167	납빛	219	독재	271	무효
12	간섭	64	고립	116	굴욕	168	낭비	220	돌발	272	묵시
13	간지	65	고문	117	궁금	169	낭패	221	동결	273	문제
14	갈기	66	고물	118	궁리	170	냄새	222	돼지	274	미끼
15	갈등	67	고민	119	궁상	171	냉담	223	두껍	275	미숙
16	갈증	68	고발	120	궁핍	172	냉소	224	두통	276	미움
17	감소	69	고생	121	권세	173	냉정	225	둔감	277	미치
18	감시	70	고소	122	궤변	174	냉혹	226	둔화	278	밀어
19	감염	71	고아	123	균열	175	너무	227	뒷발	279	밀행
20	감옥	72	고약	124	그릇	176	너절	228	땡땡	280	바보
21	감하	73	고정	125	그립	177	노곤	229	레일	281	박대
22	강간	74	고집	126	극단	178	노새	230	레흐	282	박살
23	강력	75	고통	127	극성	179	노쇠	231	마르	283	박쥐
24	강박	76	곤두	128	극심	180	노예	232	마비	284	박탈
25	강제	77	곤란	129	근절	181	노인	233	마약	285	반감
26	강직	78	곤혹	130	금기	182	노크	234	마찰	286	반대
27	강타	79	골수	131	금지	183	논리	235	만료	287	반동
28	강화	80	골절	132	급하	184	논쟁	236	만발	288	반란
29	개뿔	81	공갈	133	급류	185	놋쇠	237	만성	289	반미
30	개판	82	공격	134	급습	186	농담	238	만취	290	반박
31	거만	83	공모	135	기가	187	누출	239	만행	291	반어
32	거부	84	공범	136	기겁	188	눈길	240	말문	292	반역
33	거북	85	공상	137	기괴	189	눈꼴	241	말썽	293	반쪽
34	거세	86	공포	138	기근	190	눈살	242	망상	294	반칙
35	거절	87	공황	139	기만	191	느슨	243	망신	295	반항
36	거지	88	과다	140	기묘	192	단순	244	망치	296	발광
37	거짓	89	과대	141	기분	193	단장	245	매복	297	발진
38	거치	90	과도	142	기소	194	단점	246	매질	298	발화
39	걱정	91	과시	143	기질	195	담배	247	맹렬	299	방랑
40	걸레	92	과실	144	기피	196	담즙	248	맹세	300	방종
41	검댕	93	과열	145	기한	197	답답	249	맹열	301	방탕
42	격노	94	과잉	146	기행	198	답지	250	멍청	302	방해
43	격렬	95	과장	147	기형	199	당혹	251	면목	303	배반
44	격리	96	광기	148	긴요	200	당황	252	면직	304	배신
45	격분	97	광란	149	긴장	201	대결	253	멸망	305	백치
46	격통	98	광포	150	깜박	202	대담	254	모순	306	버릇
47	결여	99	괴기	151	꼬박	203	대량	255	모욕	307	벌금
48	결점	100	괴롭	152	꼭두	204	대상	256	모호	308	벌레
49	결핍	101	괴물	153	꽥꽥	205	대수	257	몰락	309	범죄
50	결함	102	괴벽	154	끽끽	206	대조	258	무감	310	변덕
51	경계	103	괴상	155	나태	207	대중	259	무관	311	변명
52	경고	104	괴이	156	낙담	208	대책	260	무능	312	변칙

순번	단어	순번	단어	순번	단어	순번	단어	순번	단어	순번	단어
313	변태	365	불편	417	성가	469	쓸모	521	역경	573	울적
314	별로	366	불평	418	성급	470	쓸쓸	522	역설	574	원망
315	병나	367	불행	419	성미	471	씨바	523	역성	575	원수
316	병폐	368	불화	420	성실	472	씨발	524	역시	576	원한
317	보복	369	붕괴	421	성질	473	아래	525	역행	577	위기
318	복수	370	비겁	422	세력	474	아첨	526	연기	578	위로
319	복잡	371	비관	423	섹스	475	아프	527	연루	579	위반
320	복종	372	비굴	424	소란	476	아픔	528	연마	580	위법
321	복통	373	비극	425	소름	477	악당	529	연설	581	위선
322	부과	374	비난	426	소모	478	악마	530	연약	582	위압
323	부담	375	비등	427	소문	479	악명	531	연장	583	위조
324	부당	376	비명	428	소박	480	악몽	532	열광	584	위증
325	부동	377	비밀	429	소설	481	악의	533	열변	585	위축
326	부상	378	비방	430	소심	482	악인	534	열병	586	위태
327	부식	379	비상	431	소외	483	악취	535	열성	587	위헌
328	부인	380	비소	432	소유	484	악한	536	열화	588	위험
329	부적	381	비싸	433	소음	485	악화	537	염병	589	위협
330	부정	382	비열	434	소탕	486	안달	538	염증	590	유감
331	부조	383	비위	435	소환	487	안좋	539	에기	591	유독
332	부족	384	비참	436	속물	488	안티	540	에민	592	유령
333	부주	385	비탄	437	속죄	489	암내	541	에속	593	유인
334	부채	386	비통	438	손상	490	암살	542	에측	594	유죄
335	부패	387	비판	439	손실	491	암캐	543	오류	595	유행
336	부하	388	비평	440	손해	492	압박	544	오만	596	유혈
337	부활	389	빈곤	441	솔직	493	애로	545	오명	597	유혹
338	분개	390	빈약	442	쇠약	494	애매	546	오물	598	유황
339	분규	391	빈혈	443	수감	495	애처	547	오버	599	음란
340	분노	392	빙어	444	수다	496	야만	548	오산	600	음모
341	분열	393	뻔뻔	445	수박	497	야비	549	오싹	601	음침
342	분쟁	394	뻐걱	446	수배	498	야유	550	오염	602	음탕
343	분출	395	사기	447	수법	499	약점	551	오용	603	음흉
344	분통	396	사려	448	수척	500	약탈	552	오인	604	응징
345	분투	397	사망	449	수치	501	약화	553	오입	605	의문
346	분한	398	사악	450	스틸	502	얄팍	554	오줌	606	의심
347	분할	399	사임	451	슬픔	503	양립	555	오해	607	의혹
348	분해	400	사직	452	습격	504	어긋	556	올무	608	이간
349	분화	401	사취	453	시늉	505	어둠	557	옴폭	609	이별
350	불결	402	사치	454	시위	506	어색	558	와전	610	익살
351	불경	403	사하	455	식상	507	어찌	559	완강	611	인색
352	불굴	404	산만	456	신고	508	어휴	560	왕따	612	인성
353	불길	405	살인	457	신음	509	억압	561	왜곡	613	인질
354	불량	406	살해	458	실례	510	억울	562	외상	614	일당
355	불리	407	상처	459	실망	511	억제	563	외설	615	일축
356	불만	408	상해	460	실속	512	억지	564	요동	616	일탈
357	불법	409	새기	461	실수	513	언쟁	565	요술	617	잃은
358	불순	410	새끼	462	실패	514	얼룩	566	욕심	618	자갈
359	불신	411	생색	463	싫어	515	엄중	567	욕지	619	자객
360	불쌍	412	서리	464	싫증	516	엄청	568	용자	620	자격
361	불안	413	선동	465	심각	517	엄포	569	우려	621	자극
362	불운	414	설교	466	심연	518	없다	570	우묵	622	자멸
363	불쾌	415	섬뜩	467	ㅆㅂ	519	엉망	571	우미	623	자백
364	불통	416	섬망	468	싸움	520	여분	572	우울	624	자살

순번	단어	순번	단어	순번	단어	순번	단어	순번	단어	순번	단어
625	잔인	677	중단	729	최후	781	파열	833	헌것	885	공모자
626	잔학	678	중독	730	추돌	782	패자	834	험담	886	공수병
627	잔혹	679	중상	731	추락	783	패주	835	험악	887	공짜냐
628	잘못	680	중죄	732	추문	784	편견	836	현세	888	공포증
629	잠식	681	중지	733	추방	785	편협	837	현혹	889	과부하
630	잡다	682	중퇴	734	추위	786	평판	838	혐오	890	광신적
631	잡담	683	증오	735	추잡	787	폐기	839	형벌	891	괴로움
632	잡색	684	지각	736	출혈	788	폐지	840	호구	892	교수형
633	장난	685	지긋	737	충격	789	폐차	841	호색	893	교전국
634	장력	686	지독	738	충돌	790	포기	842	혼돈	894	굶주림
635	재난	687	지루	739	충동	791	포악	843	혼동	895	그림자
636	재발	688	지연	740	충만	792	포위	844	혼란	896	극빈자
637	재앙	689	지옥	741	충혈	793	폭군	845	혼수	897	극악한
638	재잘	690	지체	742	취소	794	폭발	846	혼잡	898	근들거
639	저능	691	지치	743	치명	795	폭탄	847	혼전	899	근시안
640	저속	692	지칠	744	치열	796	폭파	848	홈통	900	기면증
641	저주	693	지탱	745	침략	797	폭포	849	화상	901	기생충
642	저하	694	지터	746	침몰	798	폭행	850	화형	902	깍쟁이
643	저항	695	지하	747	침상	799	표절	851	환각	903	깔따구
644	저해	696	진다	748	침수	800	풍자	852	환멸	904	껑충한
645	적극	697	진압	749	침울	801	피곤	853	환상	905	꼬꼬댁
646	전멸	698	진정	750	침입	802	피로	854	황급	906	꾸지람
647	전복	699	진흙	751	침착	803	피상	855	황달	907	끝난다
648	절규	700	질병	752	침체	804	하강	856	황폐	908	나막신
649	절단	701	질식	753	침해	805	하급	857	회피	909	낙오자
650	절도	702	질책	754	크랩	806	하등	858	횡포	910	난봉꾼
651	절망	703	질투	755	타도	807	하위	859	효험	911	난장판
652	절박	704	짐승	756	타락	808	하인	860	후퇴	912	내리뜬
653	절제	705	징벌	757	타협	809	하품	861	후회	913	너무큼
654	접착	706	징후	758	탄식	810	학대	862	훈계	914	노처녀
655	정복	707	찌네	759	탄핵	811	학살	863	훼방	915	논쟁점
656	정지	708	쩔쩔	760	탈선	812	한탄	864	흉상	916	놀래키
657	정크	709	차별	761	탈수	813	한통	865	흉작	917	농땡이
658	제거	710	차분	762	탐욕	814	함정	866	흉터	918	눈사태
659	제동	711	착취	763	태만	815	항변	867	흉포	919	눈속임
660	제외	712	참견	764	탱크	816	항복	868	흐릿	920	답없다
661	제지	713	창녀	765	테러	817	항의	869	흥분	921	대가리
662	제한	714	창백	766	통렬	818	해고	870	희롱	922	대학살
663	조롱	715	창피	767	통증	819	해골	871	ㅠㅠ	923	대혼란
664	조병	716	채찍	768	통탄	820	해산	872	강제력	924	대홍수
665	조작	717	책망	769	퇴각	821	해충	873	개호구	925	도둑놈
666	조잡	718	처벌	770	퇴보	822	해킹	874	갱스터	926	도둑질
667	존나	719	처지	771	퇴짜	823	행위	875	거드름	927	도망자
668	종북	720	처참	772	퇴폐	824	허가	876	거머리	928	도살장
669	좌절	721	천적	773	퇴행	825	허구	877	거짓말	929	독재자
670	좌초	722	철회	774	투옥	826	허밍	878	건망증	930	돌팔이
671	죄값	723	청산	775	투쟁	827	허세	879	걸림돌	931	동성애
672	죄송	724	체포	776	투정	828	허약	880	겁많음	932	동정심
673	죄수	725	초과	777	퉁명	829	허용	881	겁쟁이	933	두려움
674	죄인	726	초라	778	파괴	830	허위	882	게으름	934	드라콘
675	주색	727	초조	779	파멸	831	허점	883	경쟁자	935	드래그
676	죽음	728	촛불	780	파산	832	허풍	884	곰팡내	936	들창코

순번	단어	순번	단어	순번	단어	순번	단어	순번	단어	순번	단어
937	딜레마	989	버릴거	1041	비평가	1093	인간성	1145	구부러진	1197	인종차별
938	때려잡	990	범좌자	1042	빈민가	1094	자만심	1146	권위주의	1198	적폐청산
939	떨어뜨	991	범죄자	1043	빨갱이	1095	잔소리	1147	귀머거리	1199	절름발이
940	똥줄타	992	변명자	1044	사기꾼	1096	잘못된	1148	그릇되성	1200	주름살지
941	라이벌	993	병뚜껑	1045	사생아	1097	잠꼬대	1149	극단주의	1201	탈탈털린
942	마구간	994	보이콧	1046	살인자	1098	장난감	1150	근질근질	1202	택도없다
943	만지작	995	보잘것	1047	선동자	1099	장애물	1151	꼭두각시	1203	파시스트
944	말꼬리	996	부도덕	1048	소환장	1100	장애인	1152	꼴사나운	1204	편가르기
945	말다툼	997	부랑자	1049	속임수	1101	적대자	1153	꾸불꾸불	1205	피비린내
946	말대꾸	998	부역자	1050	술고래	1102	전리품	1154	나라망친	1206	하품을하
947	말라붙	999	부재자	1051	숨겨주	1103	전염성	1155	나몰라라	1207	허수아비
948	말이없	1000	부적당	1052	스캔들	1104	절름거	1156	나무늘보	1208	헤게모니
949	말장난	1001	부적절	1053	슬럼프	1105	정나미	1157	내리막길	1209	혼란시키
950	매국노	1002	부정성	1054	슬리퍼	1106	정반대	1158	네거티브	1210	화끈하게
951	매춘부	1003	부정직	1055	시기심	1107	제정신	1159	단조로움	1211	회의론자
952	맹공격	1004	부정확	1056	식인종	1108	조바심	1160	뒷걸음질	1212	횡설수설
953	머저리	1005	부조리	1057	신경증	1109	좌절감	1161	떨들썩한	1213	히스테리
954	멍때리	1006	부조화	1058	신경질	1110	죄의식	1162	막무가내	1214	거짓말쟁이
955	멍청이	1007	부주의	1059	실패자	1111	주둥이	1163	만큼주의	1215	멜로드라마
956	명렬한	1008	불가능	1060	싸우자	1112	중독자	1164	말더듬이	1216	무정부주의
957	모조품	1009	불경기	1061	쓰레기	1113	중상자	1165	망연자실	1217	불편한진실
958	목덜미	1010	불공정	1062	아니죠	1114	증후군	1166	무시무시	1218	아이러니한
959	몰래하	1011	불구자	1063	악마의	1115	지저분	1167	무용지물	1219	울통불통한
960	몰이해	1012	불규칙	1064	애새끼	1116	진풍경	1168	미련퉁이	1220	장난꾸러기
961	몰인정	1013	불규형	1065	야만성	1117	징계하	1169	미치광이		
962	몸부림	1014	불균형	1066	야만인	1118	찬바람	1170	바람둥이		
963	몽둥이	1015	불만족	1067	어려움	1119	참을성	1171	바람잽이		
964	무감각	1016	불명예	1068	얼간이	1120	초토화	1172	반사회적		
965	무경험	1017	불복종	1069	엽그레	1121	치찰음	1173	베껴쓰기		
966	무관심	1018	불성실	1070	없을텐	1122	침략자	1174	보기만해		
967	무기력	1019	불안정	1071	엉터리	1123	침력자	1175	부끄러움		
968	무능력	1020	불완전	1072	여드름	1124	타박상	1176	불협화음		
969	무분별	1021	불유쾌	1073	역효과	1125	탐탁찮	1177	브레이크		
970	무섭다	1022	불이익	1074	연고자	1126	퇴박하	1178	비논리적		
971	무의미	1023	불일치	1075	열광자	1127	파괴자	1179	뾰족뒤쥐		
972	무일푼	1024	불충분	1076	오염물	1128	파시즘	1180	성의가없		
973	무자비	1025	불쾌감	1077	오지마	1129	패시브	1181	수다쟁이		
974	무작위	1026	불특정	1078	외고집	1130	편집병	1182	스트레스		
975	무정부	1027	불평등	1079	욕지기	1131	편집증	1183	시기상조		
976	무질서	1028	불필요	1080	용의자	1132	폭발물	1184	쓰잘데없		
977	무차별	1029	불합리	1081	우울증	1133	폭풍우	1185	아이러니		
978	물장구	1030	불협화	1082	욱신거	1134	풋내기	1186	알레르기		
979	미온적	1031	불확정	1083	움푹한	1135	하수구	1187	야단법석		
980	미친짓	1032	비공식	1084	위조품	1136	헛소리	1188	어리석음		
981	바가지	1033	비능률	1085	유행병	1137	현기증	1189	얼어죽을		
982	바보냐	1034	비싸게	1086	이교도	1138	호색적	1190	엉망진창		
983	바보다	1035	비싼건	1087	이기심	1139	회의론	1191	염세주의		
984	반대자	1036	비웃음	1088	이따위	1140	후진성	1192	웃음거리		
985	반체제	1037	비윤리	1089	이방인	1141	휴지통	1193	의미없다		
986	방해물	1038	비인간	1090	이상해	1142	가로채기	1194	의미없어		
987	방화범	1039	비정상	1091	이시끼	1143	가슴앓이	1195	이기주의		
988	배신자	1040	비조직	1092	익살맞	1144	과대평가	1196	이눔들아		