정보통신공학과 졸업논문

TextRank 알고리즘을 이용한 뉴스 기사 요약 성능 개선

2018-12-17

한국외국어대학교

정보통신공학과

201303036

전 철 민

지도 교수: 한 희일

졸업논문제출 청구 및 승인서

제 목 : TextRank 알고리즘을 이용한 뉴스 기사 요약 성능 개선

대 학 : 한국외국어대학교 학 과 : 정보통신공학과

학 번 : 201303036 성 명 : 전 철 민

이 논문을 제출하오니 승인하여 주십시오.

2018 년 12 월 17 일

성 명 : 전 철 민 (인)

.....

위 학생의 논문 제출을 승인함.

2018 년 12 월 17 일

지 도 교 수 : 한 희 일 (인)

목차

그림 목차

표 목차

1. 서론

1.1 연구 배경 및 목적

현재 인터넷의 발달로 인해 많은 데이터가 매일 생겨나고 있다. 이 모든 정보를 바쁜 현대인들이 일일이 기사를 읽고 정보를 확인하기에는 현실적으로 불가능하다. 따라서 시간을 절약하면서 모든 정보를 효율적으로 습득할 수 있게 관심있는 긴 기사를 요약하여 보여준다면 사용자들에게 원하는 정보를 찾는 데 큰 도움이 될 것이다. 보통 뉴스 검색 후 제목을 보면 대략적으로 어떤 내용일 것이라는 것을 알 수 있다. 이 점을 생각하여 제목 문장을 포함한 뉴스 기사에서 중요한 키워드만을 선별하여 키워드가 많이 들어가 있는 문장을 중요 내용을 포함한다는 근거 하에 많은 사용자들에게 뉴스 요약문을 제공해준다. 여기서 말하는 요약은 한 기사에서 핵심적인 문장들을 추려 사용자에게 중요한 내용을 알려주는 것을 목표로 한다. 따라서 핵심의미를 유지하면서 내용을 효과적으로 줄여 내용을 쉽게 이해할 수 있어야한다.

1.2 관련 기술 동향 및 연구

최근 많은 기업에서는 매일 생겨나는 수많은 데이터를 수집, 분석하여 가치 있는 결과를 추출한다. 이와 같은 기술을 빅데이터라고 한다. 기업에서는 각종 SNS 글을 분석하여 소비자의 니즈를 파악하여 마케팅을 실시하고 있다. 이처럼 데이터를 이용하여 분석하는 기술인 텍스트 마이닝과 오피니언 마이닝이 빅데이터에서 많이 사용되고 있다. 텍스트 마이닝은 말 그대로 글(텍스트)을 캐낸다는 의미로 단어의 출현빈도, 단어 간 관계성 등을 파악하여 유의미한 정보를 추출하는 것이다. 텍스트 마이닝을 이용한 뉴스 기사 헤드라인 작성, 제품에 대한 고객 반응 분석, 각 브랜드에 대한 고객 생각 등을 알아내는 데 많이 활용되어지고 있다. 오피니언 마이닝은 소셜미디어 등의 정형/비정형 텍스트의 긍정, 부정, 중립의 선호도를 분석하고 더 나아가 그 원인을 도출하는 것을 목적으로 한다. 정치, 경제 사회적인 특성 사안들이 발생했을 때, 여론이나 대중의 관심도가 실시간으로 어떻게 변화하는지 확인할 수 있다. 오피니언 마이닝 사용의 예로는 2011년 페이스북에 '우울하다', '불안하다' 등의 부정적인 단어가 증가한 후 미국의 실업률이 증가하였고, 트위터에 부정적인 언급이 다수 등장하는 기업의 주가는 88% 이상 하락했다고 한다. 이뿐만 아니라 기업은 신제품 개발, 마케팅 전략 수립 등 다양한 방면에서 오피니언 마이닝 기술을 사용하고 있다.

NAVER에서도 작년 말부터 문장의 중요도를 분석한 자동추출기술로 기사내용을 요약해주는 서비스를 실시하고 있다. 아직 베타버전이라 문제점이 많지만 사용자들에게 편의성을 제공해 관심을 끌고 있다. DAUM 또한 2016년 다음에서 특허 등록한 기사요약 서비스 서버 및 방법 기술이 적용된 자동요약 서비스 시작하였다.

미국에서나 영미권은 이미 요약형 저널리즘이 많이 나오고 있는 반면 우리나라는 자연어 처리 등에서 뒤처지고 있다. 실제 해외에서는 5년여 전부터 '섬리(Summly)', '와비(Wavii)' 등을 통해 보편화된 서비스를 자리매김하는 추세다. 2015년 출범한 프랑스의 '브리프미(Brief me)'는 그날의 중요한 뉴스 5가지를 선정한 뒤 기사 내용을 요약해 저녁에 독자 메일로 보내준다. 미국의 '아골로 (Agolo)'는 AI 기술을 이용해 복잡한 자료를 빠른 시간에 정리 요약해준다. 또한 최근에는 딥러닝, 자연어 처리 기법 관련해서 문서 요약 연구가 진행되고 있다.

위 내용과 같이 문서요약에는 여러가지 방법들이 있다. 문장 간의 이 논문에서는 Google의 검색엔진에 사용되는 PageRank 알고리즘을 텍스트에 적용한 TextRank 알고리즘을 이용하여 뉴스 기사에서 중요 내용들만 추출하는 3줄 요약 시스템을 구현하고 더 나아가 요약 성능 개선에 대해 논한다.

그림 1 네이버 요약봇, 미국의 Agolo, Yahoo의 Summly

2. 본론

2.1 문서 요약 기법

문서 요약이란 한 문서로부터 특정 사용자나 작업에 적합하게 축약된 형태의 문서를 생성하는 것이다. 즉, 문서의 복잡도를 줄이면서 중요하고 필요한 정보는 유지한다. 문서 요약 기법에는 유형에 따라 생성 요약, 추출 요약, 지시적 요약, 정보적 요약 등 다양한 기법들이 있지만 본론에서는 뉴스 요약과 같은 글 요약 유형에 쓰이는 요약 기법의 두 가지 종류 생성 요약과 추출 요약을 설명한다.

2.1.1 생성 요약

생성 요약은 원문서로부터 중요한 단어들을 선별하여 자연어 처리를 통하여 새로운 문장을 구성한 후에 사용자에게 요약문으로 제공하는 것이다. 즉, 문서의 내용을 압축하여 새로운 문장을 생성한다. 쉽게 설명하면 우리가 글을 읽고 직접 손으로 글을 요약하는 것과 같다고 생각하면 된다. 생성 요약은 중요한 문장을 추출하는 추출요약보다 문서를 더 압축할 수 있지만 한국어를 자연어 처리하는 과정에서 한글의 여러가지 특성 때문에 어려움이 발생한다. 키워드를 선별하기 위해서는 한국어는 형태소 분석 작업을 하여 조사나 어미 등 불용어들을 제거하는 등과 같이 사람과 같은 수준의 자연어 생성은 아직까지 어렵기 때문에 기존 내용을 이용한 키워드 추출 방법을 적용하는 추출 요약을 많이 사용한다.

2.1.2 추출 요약

추출요약은 원 문서로부터 중요하다고 생각되는 문장만을 선별하여 요약문으로 제공한다. 생성 요약과 달리 문장을 새로 만드는 것이 아닌 기존의 문장들을 가지고 분석하여 핵심 문장을 추출하여 전체 기사를 요약해주는 것이라고 생각하면 된다. 문장이 가지는 단어의 빈도수 및 가중치를 통해 문장과 단어 간의 관계를 분석하여 중요 문장을 추출하는 방식으로 이루어진다. 본 연구는 본문의 문장을 그대로 활용하는 추출 요약 기법을 사용한다. 형태소 분석기를 사용해서 문장 내에 명사, 형용사, 동사를 추출하여 각 문장 간 동시에 출현하는 단어의 빈도를 기반으로 문장 간의 유사도를 분석하여 중요문장들을 추출한다. 형태소 분석기, 유사도 분석 등 기술들에 대해서는 뒤에 이어서 설명한다.

2.2 TextRank 알고리즘

TextRank 알고리즘은 Mihalcea와 Tarau가 제안한 알고리즘으로 텍스트에 관한 그래프 기반 랭킹 모델로써 구글의 검색 엔진에 사용된 PageRank를 기반으로 한 알고리즘이다. 즉, PageRank 알고리즘을 텍스트에 적용한 것이 TextRank인 것이다. TextRank의 이해를 위해 먼저 PageRank에 대해 알아본다.

2.2.1 PageRank 알고리즘

PageRank는 월드 와이드 웹과 같은 하이퍼링크 구조를 가지는 문서에 상대적 중요도에 따라 가중치를 부여하는 방식이다. PageRank의 아이디어는 중요한 페이지는 다른 사이트로부터 링크를 받는다는 관찰에 기초하고 있다. PageRank는 각각의 페이지를 Node, 서로의 링크를 Edge로하여 만들어진 그래프로 단순한 계산을 하여 사용자가 검색할 경우 랭크 값을 이용해 가장 유용한 정보 제공하게 해줍니다. 이해를 위해 예를 들어 설명하자면, A, B, C, D 4개의 페이지가 있고 아래의 그림과 같이 연결되어 있다.(설명의 단순화를 위해 방향성이 없다고 가정한다, A->B, B->A) 모든 정점에 1의 가중치를 준 다음 각 정점이 가지고 있는 가중치를 링크되어 있는 정점에게 나눠준다. 즉, A는 B와 D에게 0.5씩 D는 나머지 정점들에게 0.333씩 나눠주게 된다. C의 경우 D에게만 연결되어 있으므로 1 전체를 넘겨준다. 이 과정을 어느정도 반복하면 한 값에 수렴한다. 그 결과가 아래에 있는 마지막 그래프 그림이다. 최종적으로 A, B는 1, C는 0.5, D는 1.5의 가중치를 갖게 됩니다. 따라서 D가 다른 페이지로부터 많은 링크를 받았기 때문에 가장 높은 랭크 값을 가지는 결과가 나온다.

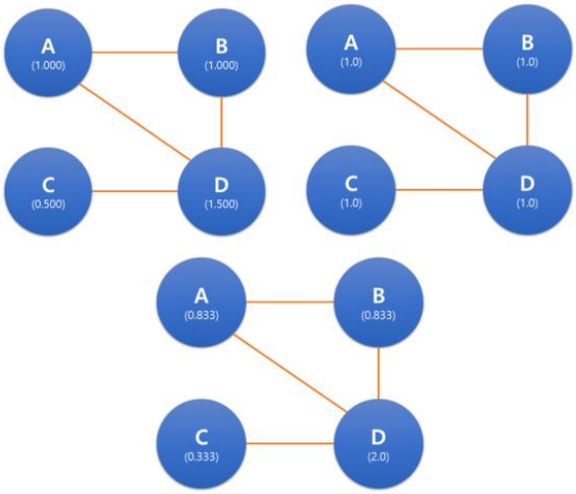

그림 2 PageRank 설명 그래프

$$PR(A) = \frac{(1-d)}{N} + d\left(\frac{PR(T1)}{C(T1)} + \frac{PR(T2)}{C(T2)} + \ldots + \frac{PR(Tn)}{C(Tn)}\right) \quad \text{···· 식 (1)}$$

- PR : PageRank
- PR(A) : A라는 페이지의 PageRank
- T1, T2, … , Tn은 페이지를 가리키는 다른 페이지들
- d : Damping Factor, PageRank 논문에 따르면 damping factor란 어떤 마구잡이로 웹 서핑을 하는 사람이 그 페이지에 만족을 못하고 다른 페이지로 가는 링크를 클릭할 확률이라고 한다. 보통 0과 1사이의 값에서 정해지는데 보통 0.85로 설정해 놓았다고 되어 있다. 여기서도 0.85로 설정하였다.
- N : 전체 페이지 수

위에서 설명한 내용에 텍스트를 적용한 것이 TextRank 알고리즘이다. 이제 한 Node가 페이지가 아닌 한 문장이 되고 Edge는 한 문장 내에서 명사, 형용사, 동사를 추출하여 다른 문장과 유사도 분석을 통해 구한다.

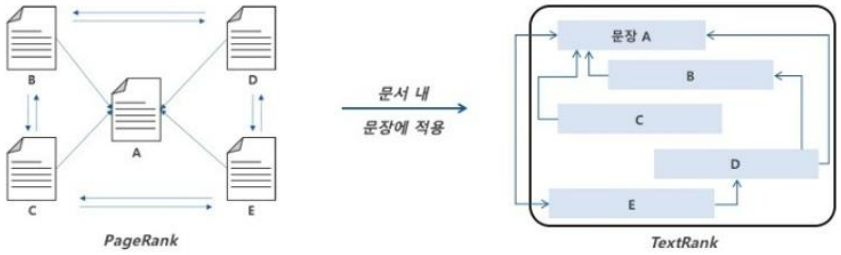

그림 3 PageRank를 Text에 적용한 TextRank

$$TR(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} TR(V_j) \quad \text{···· 식 (2)}$$

- TR(Vi) : 문장 또는 단어 Vi 에 대한 TextRank 값
- Wij : 문장 또는 단어 i와 j 사이의 가중치
- d : damping factor, PageRank에서 한 사람이 웹 서핑을 하는 도중 페이지의 내용에 만족하지 못하고 다른 페이지로 이동할 확률로써 TextRank에서도 같은 값을 사용. (TextRank 논문에서는 0.85로 설정하였기에 같은 값으로 설정)

2.3 뉴스 기사 요약 시스템 흐름도

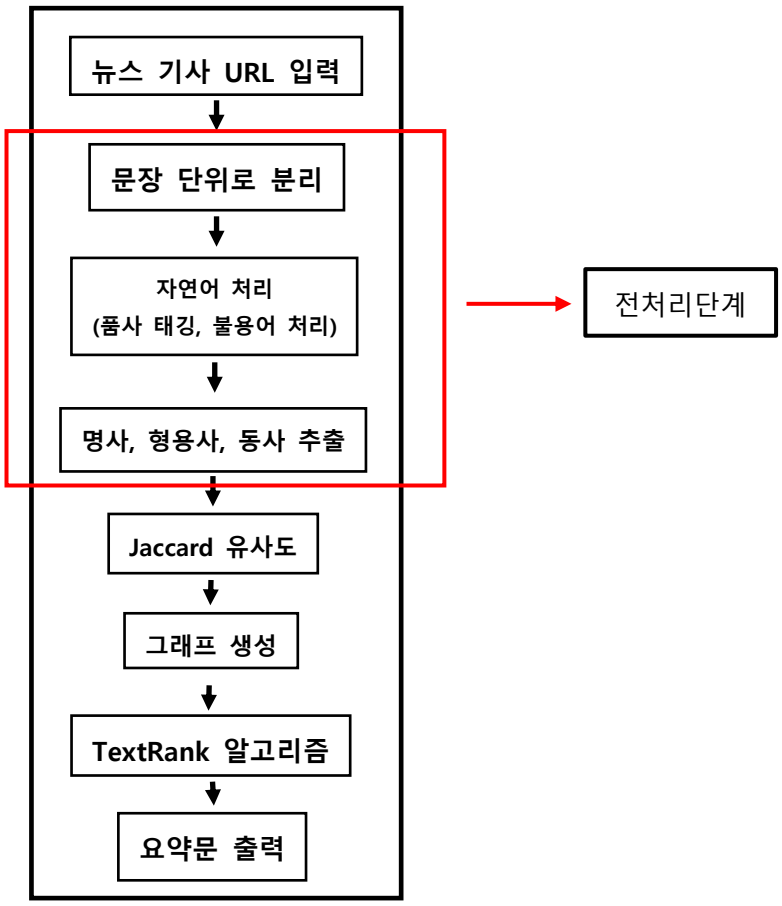

그림 4 시스템 Flow Chart

1. 뉴스 기사의 URL을 넣으면 Python 라이브러리 Newspaper로 기사 내용을 가져온다.
2. ' . ' , '? ', '! ', '₩n', '.₩n' rule-based로 기사 내용을 문장 단위로 분리
3. 형태소 분석기를 통해 품사 태깅, 불용어는 미리 지정해 두어서 품사 태깅 후 불용어가 나오면 추출하지 않고 문장의 의미있는 명사, 형용사, 동사만을 추출
4. Jaccard 유사도를 이용해 PageRank 알고리즘을 돌리기 위한 Edge 값을 구한다.
5. PageRank 알고리즘을 적용하여 기사 내의 문장들의 중요 순위를 정한다.
6. 값이 높을수록 기사의 중요한 문장을 뜻하므로 가장 높은 값부터 3문장 추출하여 사용자에게 보여준다.

2.4 데이터(뉴스 기사) 크롤링

여러 기술을 적용하기 이전에 뉴스 기사를 가져와야 한다. 현재 사용자가 보고싶은 기사의 URL을 입력하면 Python 라이브러리 중 사용자가 지정한 URL에서 Text를 추출해주는 Newspaper을 이용하여 뉴스 기사 제목과 기사 내용을 크롤링해주는 방법을 사용하였다.

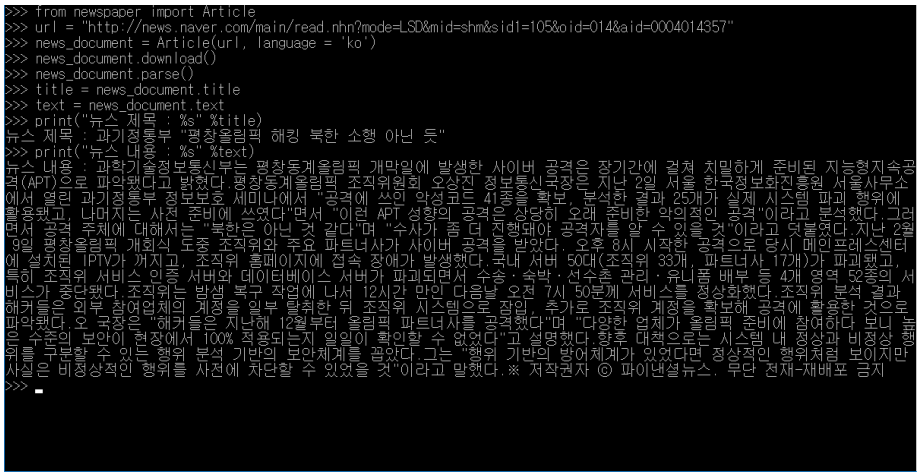

그림 5 Newspaper를 이용한 크롤링 결과

[뉴스 기사 출처 : http://news.naver.com/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=014&aid=0004014357]

2.5 전처리 단계

2.5.1 문장 단위로 분리

뉴스 기사를 TextRank 알고리즘의 각 Node에 적용하기 위해 한 문장 단위로 분리하는 과정이 필요하다. 먼저 사용자가 입력한 URL 기사를 문장 단위로 분리하기 위해서 문장 끝 '. ', '? ', '! ', '₩n ', '.₩n ' rule-based로 문장을 분리한다. 좋은 방법은 아니지만 한국어 형태소 분석기들 중에선 빠르고 원문을 보존하면서 문장을 구분해 주는 기능이 없고 큰 문제가 되지 않기 때문에 이 방법을 이용한다. 참고로 KoNLPy(한국어 자연어 처리를 위한 파이썬 패키지)에 있는 형태소 분석기들 중 Kkma는 문장을 잘라주는 기능이 있지만 이 과정에서 명사 추출을 하는 데 학습되어진 대로 명사가 띄어쓰기로 나눠진 채 반환되어 지기 때문에 원문이 유지되지 않는다. 아래의 예를 보면 전철민이라는 이름을 전철 민으로 띄어쓰기 된 채로 결과가 나오기 아래의 품사 태깅 과정에 적합하지 않기 때문에 rule-based로 문장을 분리한다.

```
kkma.sentences("형태소 분석기를 이용하여 문장을 요약한다. 정보통신공학과 13 학번 전철민.")
['형태소 분석기를 이용하여 문장을 요약한다.', '정보통신공학과 13 학번 전철 민.']
```

2.5.2 각 문장의 중요도 측정을 위한 품사 태깅

이 다음으로는 TextRank 알고리즘에서 설명한 문장 간의 Edge(영향력)을 정의할 차례이다. 여러가지 방법이 있겠지만 두 문장의 공통으로 등장하는 명사, 형용사, 동사에 의한 유사도로 분석을 통해 Edge를 구한다. KoNLPy의 형태소 분석기를 사용해 문장에서 명사, 형용사, 동사를 추출하여 리스트 형식으로 얻는다.

KoNLPy에는 총 다섯개의 형태소 분석기를 사용할 수 있다. 그 중 Kkma 형태소 분석기를 사용한 이유는 품사 분석 범주가 가장 많고 정확한 품사 정보 분석을 필요로 할 때 적합한 형태소 분석기이다. 또한 다른 형태소 분석기에 비해 전반적으로 분석 품질이 좋아서 사용하였다.

그림 6 KoNLPy 한국어 자연어 처리를 위한 Python 패키지

```
>>> from konlpy.tag import Kkma
>>> from konlpy.utils import pprint
>>> kkma = Kkma()
>>> pprint(kkma.pos("나는 밥을 먹는다."))
[('나', 'NP'),
 ('는', 'JX'),
 ('밥', 'NNG'),
 ('을', 'JKO'),
 ('먹', 'VV'),
 ('는', 'EPT'),
 ('다', 'EFN'),
 ('.', 'SF')]
>>>
```

그림 7 형태소 분석기를 이용한 문장 분석 결과

(NP : 대명사, JX : 보조사, NNG : 보통명사, JKO : 목적격 조사, VV : 동사, EPT : 시제 선어말 어미, EFN : 평서형 종결 어미, SF : 마침표, 이 중에서 명사, 형용사, 동사만 추출 아래에 품사태그표 참고)

표 1 꼬꼬마 형태소 분석기 품사 분류표

꼬꼬마 형태소 분석기			
태그	설명	태그	설명
NN	명사	VA	형용사
NNG	보통명사	VXA	보조 형용사
NNB	일반 의존 명사	VC	지정사
NNM	단위 의존 명사	VCN	부정지정사 형용사 '아니다'
NNP	고유명사	VCP	긍정지정사, 서술격조사 '이다'
NP	대명사	VV	동사
NR	수사	VXV	보조 동사
OH	한자	VX	보조 용언

2.5.3 불용어 처리

불용어란 조사, 접속사, 어미 등과 같이 문장 분석 시에 문장에서 의미가 없는 단어들을 말한다. 여기에서는 뉴스 기사를 분석하는 데에 있어서 불필요한 단어를 하나의 리스트로 만들어서 문장 간 유사도를 분석할 때 불용어가 포함된 단어는 제외하고 수행하도록 하였다.

아래에는 불용어 리스트 일부분이다. 불용어는 인터넷의 불용어 사전을 참고했고 거기에 추가적으로 기사를 읽으면서 불필요한 단어들을 직접 추가하였다. (불용어가 많아서 일부분만 표로 나타내고 나머지는 부록 부분의 불용어 리스트를 참고.)

표 2 불용어 리스트

연합뉴스	데일리	동아일보	중앙일보	조선일보	기자
을	를	우리	반대로	아이쿠	아
해야한다	여러분	각종	각자	그래서	고로
한	까닭에	때문에	과연	관하여	대하여
약간	말하자면	등	하는	근거하여	첫번째로
조차도	까지도	너희들	네	예	잠시

2.5.4 명사, 형용사, 동사 추출

위에서 기사를 문장 단위로 분리한 후 형태소 분석기를 통해 품사를 태깅하였다. 이제 그 문장 간 유사도 분석을 하기 위해 형태소 분석기를 이용한 품사 태깅 후 명사, 형용사, 동사를 추출한다. 다르게 말하면 문장의 키워드를 추출하는 과정이다. 위에 있는 형태소 분석기 품사 태그 표에 나와있는 품사 태그를 이용해서 명사, 형용사, 동사를 추출하였다. 아래에는 뉴스 기사의 일부분을 가져와 품사 태깅 후 명사, 형용사, 동사를 추출한 결과이다.

- 나노기술은 4차 산업혁명 시대에 미래 사회 변화 대응을 위한 한계극복 기술로 평가받는다.

표 3 형태소 분석기를 이용한 문장1 품사 태깅표

나노	NNG	기술	NNG	은	JX
4	NR	차	NNM	산업	NNG
혁명	NNG	시대	NNG	에	JKM
미래	NNG	사회	NNG	변화	NNG
대응	NNG	을	JKO	위하	VV
ㄴ	ETD	한계	NNG	극복	NNG
기술	NNG	로	JKM	평가	NNG
받	VV	는	EPT	다	EFN

['나노기술', '차', '산업혁명', '시대', '미래', '사회', '변화', '대응', '한계', '극복', '기술', '평가', '위하', '받']

- 혁신성장동력의 기반기술로서 나노분야 기초·원천기술개발과 산업화를 위한 각국 경쟁이 펼쳐지고 있다.

표 4 형태소 분석기를 이용한 문장2 품사 태깅표

혁신	NNG	성장	NNG	동력	NNG
의	JKG	기반	NNG	기술	NNG
로서	JKM	나노	NNG	분야	NNG
기초	NNG	원천	NNG	기술	NNG
개발	NNG	과	JKM	산업화	NNG
를	JKO	위하	VV	각국	NNG
경쟁	NNG	이	JKS	펼쳐지	VV
고	ECE	있	VXV	다	EFN

['혁신', '성장동력', '기반', '기술', '로서', '나노', '분야', '기초', '원천', '기술', '개발', '산업화', '각국', '경쟁', '위하', '펼쳐지', '있']

추가적으로 위에서부터 여러가지 품사 중에 명사, 동사, 형용사만을 추출한 이유는 아래의 그림과 같이 한국어 문장 성분 중에 주성분에 가장 가까우므로 세 가지 품사를 이용했을 경우 가장 문장의 주요한 키워드이자 중요한 내용들을 포함하기 때문에 사용하였다.

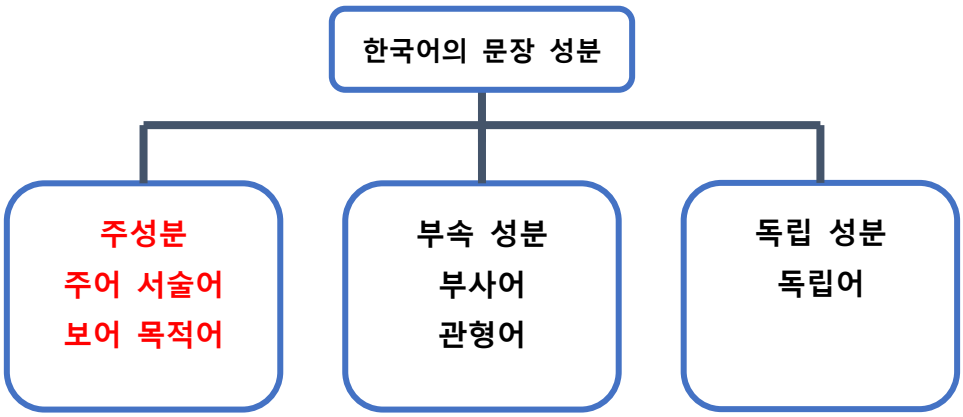

그림 8 한국어 문장 성분 종류

2.6 Jaccard 유사도

이전의 과정에서 추출한 품사들을 가지고 두 문장의 유사도 계산은 Jaccard similarity를 사용한다. Jaccard similarity는 집합 간의 유사도를 검사하는 여러 방법 중 하나로써 두 집합의 유사도는 두 집합의 교집합 크기를 두 집합의 합집합 크기로 나눈 값으로 정의된다. 여기서는 문장 안 명사, 형용사, 동사들을 사용해 두 문장의 유사도를 구하는 것인데 즉, 같은 명사, 형용사, 동사가 있을수록 유사도가 높아진다. 여기서 한 문장마다 안에 있는 명사, 형용사, 동사를 문장마다 각각 서로 다른 리스트로 구성하게 하였다. 그래서 한 집합을 한 리스트라 생각하면 된다.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \qquad \cdots\cdots \text{식 (3)}$$

['나노기술', '차', '산업혁명', '시대', '미래', '사회', '변화', '대응', '한계', '극복', '기술', '평가', '위하', '받', '혁신', '성장동력', '기반', '기술', '로서', '나노', '분야', '기초', '원천', '개발', '산업화', '각국', '경쟁', '펼쳐지', '있']

- ['기술', '기술', '위하']
- 3/29, 두 문장은 10%의 유사도를 가진다.

이 과정에서 교집합, 합집합 계산을 위해 Python 라이브러리 중 Counter를 사용해서 각 문장 내에 명사, 형용사, 동사의 수를 카운트하여 값으로 나타나게 하여 그 값을 통해 유사도 분석 결과 값(두 문장이 얼마나 관계가 있는지를 나타내는 정도)을 구한다.

아래의 식은 Python에서 구현한 Jaccard 유사도 식이다.

$$Similarity = \frac{Sum((bow1 \,\&\, bow2).values())}{Sum((bow1 \,|\, bow2).values())} \quad \cdots\cdots \text{식 (4)}$$

2.7 TextRank 알고리즘 적용을 위한 그래프 생성

이제 문장을 Node로, Edge에 부여할 값을 정하였으니 이제 그래프를 만들고 TextRank을 적용하는 과정만이 남았다. 그래프를 생성하는 것은 Python 라이브러리 NetworkX가 해준다. 각 Node, Edge에 들어가는 값들을 입력해주면 자동으로 그래프를 생성해준다. 아래 그림은 직접 한 뉴스 기사에서 Node, Edge값을 구해 생성한 그래프이다.

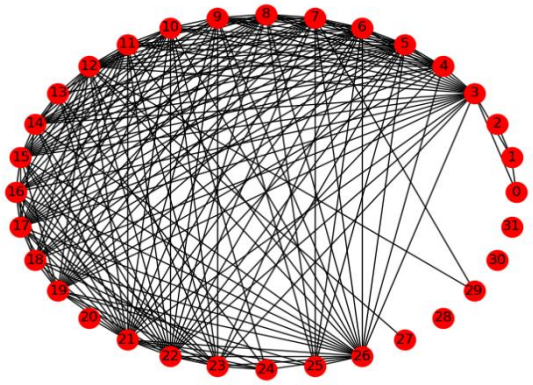

그림 9 NetworkX를 이용한 문장 간의 유사도 그래프

2.8 TextRank 알고리즘 적용

그래프 생성이후 NetworkX 안에 있는 PageRank 함수를 이용해서 서로의 Rank 값을 일정한 값에 수렴할 때까지 반복적으로 계산 후 결과적으로 각각의 문장마다 가중치를 갖게 된다. TextRank와 PageRank 알고리즘은 Node와 Edge를 설정해주는 것만 다를 뿐 계산하는 과정은 동일하여 PageRank 함수를 사용하여도 상관없어서 사용하였다. 계산 후 가중치가 높은 순서대로 3 문장을 출력하면 그것이 뉴스 기사에서 중요한 문장들로 이루어진 요약문이 된다.

3. 실험 데이터 및 결과 분석

실험에는 품사별 실험, [명사], [명사, 형용사], [명사, 형용사, 동사] 뉴스 기사 요약시에 어떤 품사를 넣어야 중요 문장 추출이 더 잘 되는지를 실험하고 이와 동시에 TextRank 알고리즘을 이용한 뉴스 기사 요약 시스템은 각 문장 내의 단어들의 유사도 가중치로 계산하기 때문에 뉴스 기사의 길이에 따른 문장 추출이 달라질 수 있기 때문에 뉴스 기사 길이에 따른 요약의 정도도 살펴본다.

실험에는 2018년 5월 ~ 6월 뉴스 기사 30편을 데이터로 하여 수행하였는데 그 중 품사별, 길이별로 결과를 잘 보여줄 수 있는 두 기사(길이가 짧은 기사, 긴 기사)만을 나타내었다.

3.1 기사 길이가 짧은 뉴스

한겨레

대통령 "유엔, 북 핵실험장 폐쇄 참관을"...IAEA가 갈듯

[한겨레] **구테흐스 사무총장과 통화**

"판문점 선언 지지해달라"

DMZ 평화지대화 검증도 요청

구테흐스 "기꺼이 협력할 것"

문재인 대통령이 1일 안토니우 구테흐스 유엔 사무총장에게 "유엔이 (4·27 남북정상회담에서 합의한) 판문점 선언을 지지하는 선언을 내주고, 북한 풍계리 핵실험장 폐쇄 현장(공개)에 동참해 달라"고 요청했다.

김의겸 청와대 대변인은 브리핑에서 "문 대통령이 오전 11시30분부터 낮 12시까지 구테흐스 사무총장과 통화를 하고 남북정상회담 결과에 대해 의견을 나눴다"며 이렇게 밝혔다. 문 대통령은 구테흐스 사무총장에게 "유엔이 총회나 안전보장이사회를 통해 남북 정상이 합의한 판문점 선언을 환영하고 지지하는 선언을 내주었으면 좋겠다"며 "유엔의 지지는 남북 관계 발전뿐만 아니라 다가오는 북-미 정상회담 성공을 위해서도 큰 힘이 될 것이다"라고 말했다. 또 "김정은 북한 국무위원장이 풍계리 핵실험장 폐쇄 때 한국과 미국은 물론 국제사회에도 투명하게 공개하겠다고 약속했는데, 폐쇄 현장에 유엔이 함께 참가해서 폐기를 확인해주면 좋겠다"고 요청했다. 청와대 관계자는 "유엔이 풍계리 핵실험장 폐쇄 참관을 결정한다면 유엔 산하의 국제원자력기구(IAEA) 쪽이 오게 될 것"이라고 설명했다. 문 대통령은 또 판문점 선언 가운데 비무장지대를 실질적인 평화지대로 전환하는 내용을 언급한 뒤 "그 과정 또한 유엔이 참관하고 이행을 검증해달라"고 부탁했다.

이에 구테흐스 사무총장은 "기꺼이 협력할 용의가 있다"고 답했다. 그는 "문 대통령 요청들이 유엔 안보리의 승인이 필요한 사항들이지만 한반도 평화 정착에 도움이 되도록 노력하겠다"며 "유엔 군축 담당 책임자를 한국과 협력하도록 지정하겠다"고 말했다.

성연철 기자 sychee@hani.co.kr

(출처 : http://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=100&oid=028&aid=0002408334)

그림 10 길이가 짧은 뉴스 기사

3.2 기사 길이가 짧은 뉴스 품사별 요약 결과

[명사만을 이용한 뉴스 요약 결과]

뉴스 제목 : 문 대통령 "유엔, 북 핵실험장 폐쇄 참관을"...IAEA가 갈듯

뉴스 내용

1. 문재인 대통령이 1일 안토니우 구테흐스 유엔 사무총장에게 "유엔이 (4·27 남북정상회담에서 합의한) 판문점 선언을 지지하는 선언을 내주고, 북한 풍계리 핵실험장 폐쇄 현장(공개)에 동참해 달라"고 요청했다.

2. 문 대통령은 구테흐스 사무총장에게 "유엔이 총회나 안전보장이사회를 통해 남북 정상이 합의한 판문점 선언을 환영하고 지지하는 선언을 내주었으면 좋겠다"며 "유엔의 지지는 남북 관계 발전뿐만 아니라 다가오는 북-미 정상회담 성공을 위해서도 큰 힘이 될 것이다"라고 말했다.

3. 이에 구테흐스 사무총장은 "기꺼이 협력할 용의가 있다"고 답했다.

[명사, 형용사를 이용한 뉴스 요약 결과]

뉴스 제목 : 문 대통령 "유엔, 북 핵실험장 폐쇄 참관을"...IAEA가 갈듯

뉴스 내용

1. 문재인 대통령이 1일 안토니우 구테흐스 유엔 사무총장에게 "유엔이 (4·27 남북정상회담에서 합의한) 판문점 선언을 지지하는 선언을 내주고, 북한 풍계리 핵실험장 폐쇄 현장(공개)에 동참해 달라"고 요청했다.

2. 문 대통령은 구테흐스 사무총장에게 "유엔이 총회나 안전보장이사회를 통해 남북 정상이 합의한 판문점 선언을 환영하고 지지하는 선언을 내주었으면 좋겠다"며 "유엔의 지지는 남북 관계 발전뿐만 아니라 다가오는 북-미 정상회담 성공을 위해서도 큰 힘이 될 것이다"라고 말했다.

3. 이에 구테흐스 사무총장은 "기꺼이 협력할 용의가 있다"고 답했다.

[명사, 형용사, 동사를 이용한 뉴스 요약 결과]

뉴스 제목 : 문 대통령 "유엔, 북 핵실험장 폐쇄 참관을"...IAEA가 갈듯

뉴스 내용

1. 문재인 대통령이 1일 안토니우 구테흐스 유엔 사무총장에게 "유엔이 (4·27 남북정상회담에서 합의한) 판문점 선언을 지지하는 선언을 내주고, 북한 풍계리 핵실험장 폐쇄 현장(공개)에 동참해 달라"고 요청했다.

2. 문 대통령은 구테흐스 사무총장에게 "유엔이 총회나 안전보장이사회를 통해 남북 정상이 합의한 판문점 선언을 환영하고 지지하는 선언을 내주었으면 좋겠다"며 "유엔의 지지는 남북 관계 발전뿐만 아니라 다가오는 북-미 정상회담 성공을 위해서도 큰 힘이 될 것이다"라고 말했다.

3. 이에 구테흐스 사무총장은 "기꺼이 협력할 용의가 있다"고 답했다.

3.3 기사 길이가 긴 뉴스

중앙일보

문정인 "평화협정 뒤 미군 주둔 정당화 힘들 것"

미 외교전문지에 철수 가능성 시사
"보수 반발, 문 대통령 딜레마 될 것"
김정은엔 "실용·현실적 인물" 묘사
청와대 "대통령 정책 방향과 달라"
한국당 "김정은의 특보냐" 비판

문정인 대통령 외교안보특보가 지난달 30일(현지시간) "평화협정이 체결된 뒤에는 한반도에서의 주한미군 주둔을 정당화하기 어려울 것"이라고 주장했다. 미 외교전문지인 포린 어페어스에 기고한 '남북 정상회담의 진전과 약속'이라는 제목의 글을 통해서다. 문 특보는 기고문에서 "주한미군의 감축이나 철수에 대해 남한의 보수 야당 세력이 강력히 반대할 것"이라며 "이는 문재인 대통령에게는 중대한 정치적 딜레마가 될 것"이라고 전망했다.

또 문 특보는 "평화롭고 핵 없는 한반도는 문 대통령이 당선 전부터 추구해 온 목표였다"며 "문 대통령은 정권이 바뀐 뒤에도 판문점 선언을 이행하기 위해 국회 비준을 추진하고 있지만 보수파들이 비준을 막고 선언 이행을 어렵게 만들 것"이라고 말했다.

그러면서 "판문점 회담이 문 대통령의 꿈을 실현할 새로운 기회를 열었지만 쉽지는 않을 것이다. 이를 문 대통령은 정확히 알고 있고 신중하게 접근할 것"이라고 전망했다.

문 특보는 김정은 국무위원장에 대해서는 "실용적이고 현실적인 인물"이라고 묘사했다. 그 이유로 "(김정은은) 이번 정상회담에서 비핵화의 전제조건으로 주한미군 철수나 감축, 한·미 동맹 등에 관해 언급하지 않았다"고 설명했다.

실제로 최근 미국 내에서도 남북 평화협정 뒤 주한미군 문제가 거론되고 있다. 제임스 매티스 국방장관은 지난달 27일 "평화협정이 체결되면 미군이 한반도에 계속 주둔할 필요가 있는가"라는 기자들의 질문에 "아마도 그것은 먼저 동맹과의 협상에서, 물론 북한과의 협상에서도 우리가 논의할 이슈의 일부"라고 답했다. 또 "그래서 나는 지금 당장 우리가 그 절차에 따라 협상해야 한다고 생각하며 향후 어떻게 될지에 대한 전제나 추정은 하지 말아야 한다. 외교관들이 이제 그 문제를 다뤄야 할 것"이라고 덧붙였다. 주한미군 지위 변화에 대한 가능성을 언급한 것이다.

또 미 NBC 방송은 지난달 30일 도널드 트럼프 대통령이 지난 2월 평창 겨울올림픽 개최 전 주한미군 철수 방안을 검토했지만 존 켈리 백악관 비서실장과의 격렬한 언쟁 끝에 결국 철회했다고 보도했다. 트럼프 대통령이 한·미 무역 불균형 문제를 다룰 카드로 주한미군 철수를 꺼내자 켈리 실장이 저지했다는 내용이다. 백악관은 이 보도를 부인했지만 트럼프 대통령은 실제로 지난 3월 "남북한 사이에 우리 군인 3만2000명이 있는데 무슨 일이 일어나는지 두고 보자"고 말한 적이 있다. 이런 와중에 문 대통령의 외교안보 분야 멘토로 통하는 문 특보가 주한미군 철수 가능성을 시사하는 논의를 제기해 파장이 일고 있다.

청와대는 즉각 진화에 나섰다. 청와대 관계자는 1일 "북한은 주한미군에 대해선 문제 삼지 않겠다는 선대의 유훈이 지금도 유효하며 김정은 국무위원장도 주한미군 철수를 꺼낸 적이 없다"며 "문 특보의 발언은 학자적 주장일 순 있어도 문 대통령의 정책 방향과는 맞지 않는다"고 밝혔다. 이 관계자는 "한국 입장에서 주한미군은 대북 억제력뿐 아니라 중국과 러시아와의 군사적 균형추 역할도 하기 때문에 철수 요청을 할 이유가 없다"며 "북한도 주한미군이나 주일미군이나 위협이 똑같기 때문에 굳이 주한미군만 문제 삼지 않는 것"이라고 덧붙였다. 국방부 관계자도 "현재 한·미 간에 주한미군 철수에 대해 논의한 일도 없다. 향후 논의할 일도 아니다"고 선을 그었다.

보수 진영은 문 특보 기고문에 강력 반발했다. 전희경 자유한국당 대변인은 "문 특보는 우리나라 대통령 특보냐, 북한 김정은의 특보냐"며 "평화협정이 체결되면 주한미군이 필요 없어질지 모른다는 주장은 대한민국 안보를 흔드는 망언이며 논평할 가치조차 못 느낀다"고 비난했다.

최익재·강태화 기자 ijchoi@joongang.co.kr

(출처 : http://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=100&oid=025&aid=0002817703)

그림 11 길이가 긴 뉴스 기사

3.4 기사 길이가 긴 뉴스 품사별 요약 결과

[명사만을 이용한 뉴스 요약 결과]

뉴스 제목 : 문정인 "평화협정 뒤 미군 주둔 정당화 힘들 것"

뉴스 내용

1. 문 특보는 기고문에서 "주한미군의 감축이나 철수에 대해 남한의 보수 야당 세력이 강력히 반대할 것"이라며 "이는 문재인 대통령에게는 중대한 정치적 딜레마가 될 것"이라고 전망했다.

2. 이런 와중에 문 대통령의 외교안보 분야 멘토로 통하는 문 특보가 주한미군 철수 가능성을 시사하는 논의를 제기해 파장이 일고 있다.

3. 청와대 관계자는 1일 "북한은 주한미군에 대해선 문제 삼지 않겠다는 선대의 유훈이 지금도 유효하며 김정은 국무위원장도 주한미군 철수를 꺼낸 적이 없다"며 "문 특보의 발언은 학자적 주장일 순 있어도 문 대통령의 정책 방향과는 맞지 않는다"고 밝혔다.

[명사, 형용사를 이용한 뉴스 요약 결과]

뉴스 제목 : 문정인 "평화협정 뒤 미군 주둔 정당화 힘들 것"

뉴스 내용

1. 문 특보는 기고문에서 "주한미군의 감축이나 철수에 대해 남한의 보수 야당 세력이 강력히 반대할 것"이라며 "이는 문재인 대통령에게는 중대한 정치적 딜레마가 될 것"이라고 전망했다.

2. 트럼프 대통령이 한·미 무역 불균형 문제를 다룰 카드로 주한미군 철수를 꺼내자 켈리 실장이 저지했다는 내용이다.

3. 청와대 관계자는 1일 "북한은 주한미군에 대해선 문제 삼지 않겠다는 선대의 유훈이 지금도 유효하며 김정은 국무위원장도 주한미군 철수를 꺼낸 적이 없다"며 "문 특보의 발언은 학자적 주장일 순 있어도 문 대통령의 정책 방향과는 맞지 않는다"고 밝혔다.

[명사, 형용사, 동사를 이용한 뉴스 요약 결과]

뉴스 제목 : 문정인 "평화협정 뒤 미군 주둔 정당화 힘들 것"

뉴스 내용

1. 문정인 대통령 외교안보특보가 지난달 30일(현지시간) "평화협정이 체결된 뒤에는 한 반도에서의 주한미군 주둔을 정당화하기 어려울 것"이라고 주장했다.

2. 문 특보는 기고문에서 "주한미군의 감축이나 철수에 대해 남한의 보수 야당 세력이 강력히 반대할 것"이라며 "이는 문재인 대통령에게는 중대한 정치적 딜레마가 될 것"이라고 전망했다.

3. 청와대 관계자는 1일 "북한은 주한미군에 대해선 문제 삼지 않겠다는 선대의 유훈이 지금도 유효하며 김정은 국무위원장도 주한미군 철수를 꺼낸 적이 없다"며 "문 특보의 발언은 학자적 주장일 순 있어도 문 대통령의 정책 방향과는 맞지 않는다"고 밝혔다.

3.5 평가 방법

현재 기계의 요약 정도를 평가하는 방법은 사람 요약문과 기계 요약문 비교이다. 그러므로 본 논문에서는 뉴스 기사 문장 하나하나를 읽고 중요도를 측정하여 중요도가 높을수록 중요 문장이므로 기계 요약의 결과에서 중요도 합산이 가장 높은 요약문을 생성해 낸 경우가 성능이 좋다고 생각하여 이 방법으로 평가하였다.

표 5 문장 중요도 평가표

문장	중요도
문재인 대통령이 1일 안토니우 구테흐스 유엔 사무총장에게 "유엔이 (4·27 남북정상회담에서 합의한) 판문점 선언을 지지하는 선언을 내주고, 북한 풍계리 핵실험장 폐쇄 현장(공개)에 동참해 달라"고 요청했다.	88%
김의겸 청와대 대변인은 브리핑에서 "문 대통령이 오전 11시30분부터 낮 12시까지 구테흐스 사무총장과 통화를 하고 남북정상회담 결과에 대해 의견을 나눴다"며 이렇게 밝혔다.	50%
문 대통령은 구테흐스 사무총장에게 "유엔이 총회나 안전보장이사회를 통해 남북 정상이 합의한 판문점 선언을 환영하고 지지하는 선언을 내주었으면 좋겠다"며 "유엔의 지지는 남북 관계 발전뿐만 아니라 다가오는 북-미 정상회담 성공을 위해서도 큰 힘이 될 것이다"라고 말했다.	90%
또 "김정은 북한 국무위원장이 풍계리 핵실험장 폐쇄 때 한국과 미국은 물론 국제사회에도 투명하게 공개하겠다고 약속했는데, 폐쇄 현장에 유엔이 함께 참가해서 폐기를 확인해주면 좋겠다"고 요청했다.	65%
청와대 관계자는 "유엔이 풍계리 핵실험장 폐쇄 참관을 결정한다면 유엔 산하의 국제원자력기구(IAEA) 쪽이 오게 될 것"이라고 설명했다.	40%
문 대통령은 또 판문점 선언 가운데 비무장지대를 실질적인 평화지대로 전환하는 내용을 언급한 뒤 "그 과정 또한 유엔이 참관하고 이행을 검증해달라"고 부탁했다.	55%
이에 구테흐스 사무총장은 "기꺼이 협력할 용의가 있다"고 답했다.	70%

그는 “문 대통령 요청들이 유엔 안보리의 승인이 필요한 사항들이지만 한반도 평화 정착에 도움이 되도록 노력하겠다”며 “유엔 군축 담당 책임자를 한국과 협력하도록 지정하겠다”고 말했다.	80%
문정인 대통령 외교안보특보가 지난달 30일(현지시간) “평화협정이 체결된 뒤에는 한반도에서의 주한미군 주둔을 정당화하기 어려울 것”이라고 주장했다.	95%
미 외교전문지인 포린 어페어스에 기고한 ‘남북 정상회담의 진전과 약속’이라는 제목의 글을 통해서다.	50%
문 특보는 기고문에서 “주한미군의 감축이나 철수에 대해 남한의 보수 야당 세력이 강력히 반대할 것”이라며 “이는 문재인 대통령에게는 중대한 정치적 딜레마가 될 것”이라고 전망했다.	88%
또 문 특보는 “평화롭고 핵 없는 한반도는 문 대통령이 당선 전부터 추구해 온 목표였다”며 “문 대통령은 정권이 바뀐 뒤에도 판문점 선언을 이행하기 위해 국회 비준을 추진하고 있지만 보수파들이 비준을 막고 선언 이행을 어렵게 만들 것”이라고 말했다.	80%
그러면서 “판문점 회담이 문 대통령의 꿈을 실현할 새로운 기회를 열었지만 쉽지는 않을 것이다.	40%
이를 문 대통령은 정확히 알고 있고 신중하게 접근할 것”이라고 전망했다.	30%
문 특보는 김정은 국무위원장에 대해서는 “실용적이고 현실적인 인물”이라고 묘사했다.	20%
그 이유로 “(김정은은) 이번 정상회담에서 비핵화의 전제조건으로 주한미군 철수나 감축, 한·미 동맹 등에 관해 언급하지 않았다”고 설명했다.	23%
실제로 최근 미국 내에서도 남북 평화협정 뒤 주한미군 문제가 거론되고 있다.	62%
제임스 매티스 국방장관은 지난달 27일 “평화협정이 체결되면 미군이 한반도에 계속 주둔할 필요가 있는가”라는 기자들의 질문에 “아마도 그것은 먼저 동맹과의 협상에서, 물론 북한과의 협상에서도 우리가 논의할 이슈의 일부”라고 답했다.	82%
또 “그래서 나는 지금 당장 우리가 그 절차에 따라 협상해야 한다고 생각하며 향후 어떻게 될지에 대한 전제나 추정은 하지 말아야 한다.	78%
외교관들이 이제 그 문제를 다뤄야 할 것”이라고 덧붙였다. 주한미군 지위 변화에 대한 가능성을 언급한 것이다.	39%

또 미 NBC 방송은 지난달 30일 도널드 트럼프 대통령이 지난 2월 평창 겨울올림픽 개최 전 주한미군 철수 방안을 검토했지만 존 켈리 백악관 비서실장과의 격렬한 언쟁 끝에 결국 철회했다고 보도했다.	45%
트럼프 대통령이 한·미 무역 불균형 문제를 다룰 카드로 주한미군 철수를 꺼내자 켈리 실장이 저지했다는 내용이다.	40%
백악관은 이 보도를 부인했지만 트럼프 대통령은 실제로 지난 3월 “남북한 사이에 우리 군인 3만2000명이 있는데 무슨 일이 일어나는지 두고 보자”고 말한 적이 있다.	55%
이런 와중에 문 대통령의 외교안보 분야 멘토로 통하는 문 특보가 주한미군 철수 가능성을 시사하는 논의를 제기해 파장이 일고 있다.	60%
청와대는 즉각 진화에 나섰다.	15%
청와대 관계자는 1일 “북한은 주한미군에 대해선 문제 삼지 않겠다는 선대의 유훈이 지금도 유효하며 김정은 국무위원장도 주한미군 철수를 꺼낸 적이 없다”며 “문 특보의 발언은 학자적 주장일 순 있어도 문 대통령의 정책 방향과는 맞지 않는다”고 밝혔다.	85%
이 관계자는 “한국 입장에서 주한미군은 대북 억제력뿐 아니라 중국과 러시아와의 군사적 균형추 역할도 하기 때문에 철수 요청을 할 이유가 없다”며 “북한도 주한미군이나 주일미군이나 위협이 똑같기 때문에 굳이 주한미군만 문제 삼지 않는 것”이라고 덧붙였다.	70%
국방부 관계자도 “현재 한·미 간에 주한미군 철수에 대해 논의한 일도 없다.	35%
향후 논의할 일도 아니다”고 선을 그었다.	15%
보수 진영은 문 특보 기고문에 강력 반발했다.	15%
전희경 자유한국당 대변인은 “문 특보는 우리나라 대통령 특보냐, 북한 김정은의 특보냐”며 “평화협정이 체결되면 주한미군이 필요 없어질지 모른다는 주장은 대한민국 안보를 흔드는 망언이며 논평할 가치조차 못 느낀다”고 비난했다.	30%

3.6 결과 분석

여러 뉴스 기사를 테스트 해본 결과, 길이가 짧은 뉴스 기사의 경우 [명사] , [명사, 형용사], [명사, 형용사, 동사] 세 경우 같거나 비슷하게 나왔다. 따라서 길이가 짧은 뉴스 기사의 경우 별 차이는 없었지만 [명사, 형용사, 동사]의 경우가 가장 요약 성능이 좋게 나왔다. 기사의 내용이 긴 경우 [명사], [명사, 형용사], [명사, 형용사, 동사] 중 [명사, 형용사, 동사]를 사용했을 경우 기사의 요약이 잘 되었다. 따라서 명사, 형용사, 동사 세 가지가 들어간 경우가 긴 기사에도 잘 요약되므로 가장 사용하기에 적합하다는 결과가 나왔다.

3.7 문제점 및 한계점

처음에는 문장에서 중요한 품사를 뽑아서 문장 간의 유사도를 측정하므로 기사 요약에 있어서 문장 내 중요한 역할을 하는 품사를 추출하는 것이 중요하다고 생각하였다. 하지만 품사 이외에도 다른 문제점들을 발견하였고 그리고 테스트한 결과 [명사, 형용사, 동사]의 경우가 기사 요약에 있어서 최적의 품사라고 생각하였지만 모든 품사 사용 시에 가장 좋은 요약문이 나와서 이 점에 대해서는 아직까지 명확한 원인을 찾지 못하여 추후에도 조사 및 분석을 하여 내용을 추가할 예정이다.

[모든 품사를 이용한 뉴스 요약 결과]

뉴스 제목 : 문 대통령 "유엔, 북 핵실험장 폐쇄 참관을"...IAEA가 갈듯

뉴스 내용

- 문재인 대통령이 1일 안토니우 구테흐스 유엔 사무총장에게 "유엔이 (4·27 남북정상회담에서 합의한) 판문점 선언을 지지하는 선언을 내주고, 북한 풍계리 핵실험장 폐쇄 현장(공개)에 동참해 달라"고 요청했다.
- 문 대통령은 구테흐스 사무총장에게 "유엔이 총회나 안전보장이사회를 통해 남북 정상이 합의한 판문점 선언을 환영하고 지지하는 선언을 내주었으면 좋겠다"며 "유엔의 지지는 남북 관계 발전뿐만 아니라 다가오는 북-미 정상회담 성공을 위해서도 큰 힘이 될 것이다"라고 말했다.
- 그는 "문 대통령 요청들이 유엔 안보리의 승인이 필요한 사항들이지만 한반도 평화 정착에 도움이 되도록 노력하겠다"며 "유엔 군축 담당 책임자를 한국과 협력하도록 지정하겠다"고 말했다.

또 다른 문제점들은 살펴보면 TextRank 알고리즘을 사용하여 가중치를 계산할 때 단순히 문장 내의 단어들의 빈도수만을 고려해서 계산하기 때문에 두 문장의 의미적 연결을 고려해주지 못하는 문제점이 발생하는 것을 알았다. 예를 들면 동의어의 경우 책/서적, 구입/구매, 가게/상점, 슬픈/우울한 등 빈도수만 고려하면 동의어를 각각 다른 단어로 보고, 또한 복합 명사 또한 같은 의미를 가지지만 논밭을 논과 밭으로 구분해서 나오므로 추가적으로 단어들 간의 의미적 관계를 고려해주는 점이 필요하다고 생각된다. 또 다른 문제점으로는 여기서 형태소 분석기를 사용해서 문장을 분석하는데 이 또한 완벽하지 않다는 점이다. 그래서 이 문제에 대해 분석을 하던 중 한국어가 교착어 성질을 지니는 언어이기 때문에 분석이 힘들다는 점을 알았다. 여기서 한국어의 교착어 성질이라는 것은 어근에 파생 접사나 어미가 붙어서 새 단어를 만든다는 뜻이다. 이해를 돕기 위해 예를 들어 설명을 하자면 "깨뜨리시었겠군요"를 형태소 단위로 분리를 해보면 ['깨' – 어근, '뜨리' – 접사, '시'는 – 높임, '었', '겠', '더' – 시간을 나타내는 어미, '군' – 감탄 어미, '요' – 문장을 끝맺는 어미] 즉, 위 문장을 제대로 분석해보면 8개의 형태소로 분리해야 하지만 형태소 분석기는 아직까지는 교착어 성질로 인해 새로 생겨난 단어를 위처럼 분리하지 못한다. 그리고 평가 방법에 있어서도 객관적인 성향보다는 주관적인 평가인 것 같아 평가 방법에 대해서도 연구하여 보다 객관적인 지표로 평가가 필요하다. 마지막 문제로는 반의어 경우 생기는 문제이다. 만일 제목과 기사의 전체적인 내용이 우울에 대한 내용이라고 가정해보자 그런데 기사의 내용에 행복이라는 단어가 많이 나오면 기사의 전체적인 내용과는 상관없는 행복이라는 문장이 중요 문장으로 요약되어 나올 수 있다. 따라서 유사도 측정시에 이 점을 해결할 수 있는 추가적인 과정이 필요하다.

4. 성능 개선을 위한 추가적인 방법

4.1 TF-IDF(Term Frequency – Inverse Document Frequency)

TF-IDF는 정보 검색과 텍스트 마이닝에서 이용하는 가중치로, 여러 문서로 이루어진 문서군이 있을 때 어떤 단어가 특정 문서 내에서 얼마나 중요한 것인지를 나타내주는 통계적 수치이다. 문서의 핵심어를 추출하거나, 검색 엔진에서 검색 결과의 순위를 결정하거나, 문서들 사이의 비슷한 정도를 구하는 등의 용도로 사용할 수 있다. 여기서 TF는 특정한 단어가 문서(장)내에서 얼마나 자주 등장하는지를 나타내는 값으로, 이 값이 높을수록 문서에서 중요하다고 생각할 수 있다. 하지만 단어 자체가 문서군에서 자주 사용되는 경우 그 단어가 흔한 것을 알 수 있는데 이것을 DF 문서빈도라 하며 이 값의 역수를 IDF라고 한다. IDF를 다시 말하면 한 단어가 문서 집합 전체에서 얼마나 공통적으로 나타나는지를 나타내는 값이다. 즉, TF-IDF는 단어 빈도와 역문서 빈도의 곱이다. 따라서 이 값을 사용하면 모든 문서에 흔하게 나타나는 단어를 걸러내는 효과를 얻을 수 있다.

결론적으로 앞에서 한 차례 불용어를 제거해주었지만 TF-IDF를 이용해 유사도 분석을 한다면 더욱 더 문장에 중요한 키워드를 뽑아주므로 요약의 성능이 더 개선될 것으로 보인다.

$$idf(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|} \qquad \cdots\cdots \text{식 (5)}$$

- |D| : 문서 집합 D의 크기, 또는 전체 문서의 수
- |{d ∈D : t ∈ d}| : 단어 t가 포함된 문서의 수. (즉, tf(t, d) ≠ 0)
- 단어가 전체 말뭉치 안에 존재하지 않을 경우 이는 분모가 0이 되는 결과를 가져온다.
- 이를 방지하기 위해 1+ |{d ∈D : t ∈ d}| 로 쓰는 것이 일반적이다.

하지만 TF-IDF 또한 두 문장 사이의 겹치는 단어가 없다면 유사성을 판단하기 어려운 점이 존재한다. 따라서 TF-IDF를 사용할 때에도 두 문장 간의 의미, 개념적 연결성을 제대로 고려해주지 못한다.

4.2 단어간 의미적 관계 고려 연구

TextRank 알고리즘은 하나의 기사를 문장 단위로 하여 그래프로 변환한 다음, 단어의 빈도수로 문장 간의 유사도를 계산할 때 겹치는 단어의 수가 많을수록 두 문장이 유사하다고 하고 겹치는 단어가 적을수록 유사하지 않다고 본다. 만일 두 문장이 있는데 두 문장 간 겹치는 단어의 수가 아예 존재하지 않아도 서로 연관성이 높은 문장이 될 수 있다. 따라서 두 문장의 의미, 개념적 연결성을 제대로 고려해주지 못하는 점이 있어 이점을 해결하여 요약의 성능을 높이고자 한다.

이 점을 해결하기 위해 동의어, 유의어, 상위어, 하위어 관계 등의 단어 개념을 정리한 시소러스(Thesaurus) 사전을 활용하여 보완하고자 한다. 여러 시소러스 사전 중 기존의 시소러스를 바탕으로 신문에 등장한 용어들을 반영하여 확대, 개편한 한국언론진흥재단이 만든 사전을 이용하였다. 아래와 같이 동의어나 유의어 등을 리스트로 구현하여 같은 단어로 분석되어지게 하여 요약성능을 높이고자 한다. 시소러스 사전에 기재된 총 단어의 수는 27,172개이다.

표 6 시소러스 사전 용어 리스트

가방	의, 골프가방, 배낭, 비치백, 서류가방, 손가방, 여행용가방, 우편가방, 책가방
출근	노동, 출퇴근
최저임금	노동, 최저임금제, 최저임금제도, 근로조건, 임금
체험학습	교육, 견학, 체험, 체험프로그램, 체험학교, 체험활동, 현장체험프로그램, 현장체험학습, 현장학습, 학습, 겨울캠프, 금융체험, 농사체험, 별자리캠프, 봉사체험, 생태체험, 안전체험, 영어캠프, 우주체험, 전통문화체험, 직업체험, 진로체험, 체험행사, 우주체험, 홍보체험관
응원단	스포츠, 서포터즈, 붉은악마, 스포츠관중
선물	금융, 생일선물, 크리스마스선물, 기프트콘, 선물포장

5. 결론

뉴스 기사에서 중요 문장 3문장을 추출하기 위해 기사를 문장 단위로 분리한 후 형태소 분석기를 사용하여 품사를 분리하여 품사에 따른 뉴스 요약 정도를 확인하고 또한 뉴스 길이에 따른 요약 정도를 실험하여 확인하였다. 먼저 뉴스 길이에 따른 요약 결과는 길이가 짧은 기사의 경우 품사의 종류에 따른 결과에 차이가 별로 없었으며 길이가 긴 경우 사용하는 품사가 많을수록 요약이 잘 되었다. 또한 추가적으로 모든 품사를 사용했을 경우도 추가로 확인해본 결과 이 경우가 가장 좋게 요약이 되어서 추가적으로 문장에 영향을 끼치는 품사에 대해 연구를 하여 모든 품사 사용시 어떤 품사가 영향을 많이 끼쳤는 지에 대해 원인을 분석하여 성능을 더 개선할 계획이다. 또한 유사도 분석과정에서 각 문장의 의미적 관계를 고려하는 방법인 시소러스 사전을 이용해서 연구를 하여 성능을 개선할 예정이다. 시소러스 사전을 이용한 방법을 분석 중에서도 약간의 문제점을 발견하였는데 그 점은 만일 사전에 없는 단어나 새로 나온 신조어의 경우는 의미적 관계를 고려해주지 못한다는 점이다. 따라서 계속해서 사전을 업데이트 하는 작업이 필요하다. 그리고 위 문제점에서 언급한 한 뉴스기사의 전체적인 분위기와 키워드가 우울이지만 행복이란 단어가 많이 나올 경우 이 문장을 중요 문장으로 추출할 수 있다는 문제점이 있었는데 이 경우 반의어에 해당하는데 이점은 시소러스 사전에도 없어서 추가적으로 조사하여 해결할 예정이다.

마지막으로는 위에서 기존의 방식에서 대해서 성능 개선 방안을 제시하였지만 이 방법을 아직 실험하지 못하였지만 추후에 동일한 데이터셋을 이용해 실험을 통해 성능이 개선되었음을 더 입증하고자 한다.

부록

꼬꼬마 형태소 분석기 품사 분류표

표 7 꼬꼬마 형태소 분석기 모든 품사 분류표

꼬꼬마 형태소 분석기			
태그	설명	태그	설명
EC	연결어미	IC	감탄사
ECD	의존적 연결어미	JK	조사
ECE	대등 연결어미	JC	접속조사
ECS	보조적 연결어미	JKC	보격조사
EF	종결어미	JKG	관형격조사
EFA	청유형 종결어미	JKI	호격조사
EFI	감탄형 종결어미	JKM	부사격조사
EFN	평서형 종결어미	JKO	목적격조사
EFO	명령형 종결어미	JKQ	인용격조사
EFQ	의문형 종결어미	JKS	주격조사
EFR	존칭형 종결어미	JX	보조사
ET	전성어미	MA	부사
ETD	관형형 전성어미	MAG	일반부사
ETN	명사형 전성어미	MAC	접속부사
EP	선어말어미	MD	관형사
EPH	존칭 선어말어미	MDN	수 관형사
EPP	공손 선어말어미	MDT	일반 관형사
EPT	시제 선어말어미	NN	명사
NNG	보통명사	SP	쉼표, 가운뎃점, 콜론, 빗금
NNB	일반 의존명사	SS	따옴표, 괄호표, 출표
NNM	단위 의존명사	SW	기타기호 (논리/수학기호, 화폐기호)
NNP	고유명사	VA	형용사
NP	대명사	VXA	보조 형용사
NR	수사	VC	지정사
OH	한자	VCN	부정지정사, 형용사 '아니다'
OL	외국어	VCP	긍정지정사, 서술격조사'이다'
ON	숫자	VV	동사
SE	줄임표	VXV	보조 동사
SF	마침표, 물음표, 느낌표	VX	보조 용언

SO	붙임표(물결, 숨김, 빠짐)	XP	접두사
XPN	체언 접두사	XPV	용언 접두사
XSA	형용사 파생접미사	XSN	명사 파생접미사
XSV	동사 파생접미사	XR	어근
UN	명사추정범주		

불용어 리스트

["아", "휴", "아이구", "아이쿠", "아이고", "어", "나", "우리", "저희", "따라", "의해", "을", "를", "에", "의", "가", "으로", "로", "에게", "뿐이다", "의거하여", "근거하여", "입각하여", "기준으로", "예하면", "예를", "들면", "들자면", "저", "소인", "소생", "저희", "지말고", "하지마", "하지마라", "다른", "물론", "또한", "그리고", "비길수", "없다", "해서는", "안된다", "뿐만", "아니라", "만이", "아니다", "만은", "아니다", "막론하고", "관계없이", "그치지", "않다", "그러나", "그런데", "하지만", "든간에", "논하지", "않다", "따지지", "않다", "설사", "비록", "더라도", "아니면", "만", "못하다", "하는", "편이", "낫다", "불문하고", "향하여", "향해서", "향하다", "쪽으로", "틈타", "이용하여", "타다", "오르다", "제외하고", "이", "외에", "이", "밖에", "하여야", "비로소", "한다면", "몰라도", "외에도", "이곳", "여기", "부터", "기점으로", "따라서", "할", "생각이다", "하려고하다", "이리하여", "그리하여", "그렇게", "함으로써", "하지만", "일때", "할때", "앞에서", "중에서", "보는데서", "으로써", "로써", "까지", "해야한다", "일것이다", "반드시", "할줄알다", "할수있다", "할수있어", "임에", "틀림없다", "한다면", "등", "등등", "제", "겨우", "단지", "다만", "할뿐", "딩동", "댕그", "대해서", "대하여", "대하면", "훨씬", "얼마나", "얼마만큼", "얼마큼", "남짓", "여", "얼마간", "약간", "다소", "좀", "조금", "다수", "몇", "얼마", "지만", "하물며", "또한", "그러나", "그렇지만", "하지만", "이외에도", "대해", "말하자면", "뿐이다", "다음에", "반대로", "반대로", "말하자면", "이와", "반대로", "바꾸어서", "말하면", "바꾸어서", "한다면", "만약", "그렇지않으면", "까악", "툭", "딱", "삐걱거리다", "보드득", "비걱거리다", "꽈당", "응당", "해야한다", "에", "가서", "각", "각각", "여러분", "각종", "각자", "제각기", "하도록하다", "와", "과", "그러므로", "그래서", "고로", "한", "까닭에", "하기", "때문에", "거니와", "이지만", "대하여", "관하여", "관한", "과연", "실로", "아니나다를가", "생각한대로", "진짜로", "한적이있다", "하곤하였다", "하", "하하", "허허", "아하", "거바", "와", "오", "왜", "어째서", "무엇때문에", "어찌", "하겠는가", "무슨", "어디", "어느곳", "더군다나", "하물며", "더욱이는", "어느때", "언제", "야", "이봐", "어이", "여보시오", "흐흐", "흥", "휴", "헉헉", "헐떡헐떡", "영차", "여차", "어기여차", "끙끙", "아야", "앗", "아야", "콸콸", "졸졸", "좍좍", "뚝뚝", "주룩주룩", "솨", "우르르", "그래도", "또", "그리고", "바꾸어말하면", "바꾸어말하자면", "혹은", "혹시", "답다", "및", "그에", "따르는", "때가", "되어", "즉", "지든지", "설령", "가령", "하더라도", "할지라도", "일지라도", "지든지", "몇", "거의", "하마터면", "인젠", "이젠", "된바에야", "된이상", "만큼 어찌됏든", "그위에", "게다가", "점에서", "보아", "비추어", "보아", "고려하면", "하게될것이다", "일것이다", "비교적", "좀", "보다더", "비하면", "시키다", "하게하다", "할만하다", "의해서", "연이서", "이어서", "잇따라", "뒤따라", "뒤이어", "결국", "의지하여", "기대여", "통하여", "자마자", "더욱더", "불구하고", "얼마든지", "마음대로", "주저하지", "않고", "곧", "즉시", "바로", "당장", "하자마자", "밖에", "안된다", "하면된다", "그래", "그렇지", "요컨대", "다시", "말하자면", "바꿔", "말하면", "즉", "구체적으로", "말하자면", "시작하여", "시초에", "이상", "허", "헉", "허걱", "바와같이", "해도좋다", "해도된다", "게다가", "더구나", "하물며", "와르르", "팍", "퍽", "펄렁", "동안", "이래", "하고있었다", "이었다", "에서", "로부터", "까지", "예하면", "했어요", "해요", "함께", "같이", "더불어", "마저", "마저도", "양자", "모두", "습니다", "가까스로", "하려고하다", "즈음하여", "다른", "다른", "방면으로", "해봐요", "습니까", "했어요", "말할것도", "없고", "무릎쓰고", "개

의치않고", "하는것만", "못하다", "하는것이", "낫다", "매", "매번", "들", "모", "어느것", "어느", "로써", "갖고말하자면", "어디", "어느쪽", "어느것", "어느해", "어느", "년도", "라", "해도", "언젠가", "어떤것", "어느것", "저기", "저쪽", "저것", "그때", "그럼", "그러면", "요만한걸", "그래", "그때", "저것만큼", "그저", "이르기까지", "할", "줄", "안다", "할", "힘이", "있다", "너", "너희", "당신", "어찌", "설마", "차라리", "할지언정", "할지라도", "할망정", "할지언정", "구토하다", "게우다", "토하다", "메쓰겁다", "옆사람", "퉤", "쳇", "의거하여", "근거하여", "의해", "따라", "힘입어", "그", "다음", "버금", "두번째로", "기타", "첫번째로", "나머지는", "그중에서", "견지에서", "형식으로", "쓰여", "입장에서", "위해서", "단지", "의해되다", "하도록시키다", "뿐만아니라", "반대로", "전후", "전자", "앞의것", "잠시", "잠깐", "하면서", "그렇지만", "다음에", "그러한즉", "그런즉", "남들", "아무거나", "어찌하든지", "같다", "비슷하다", "예컨대", "이럴정도로", "어떻게", "만약", "만일", "위에서", "서술한바와같이", "인", "듯하다", "하지", "않는다면", "만약에", "무엇", "무슨", "어느", "어떤", "아래윗", "조차", "한데", "그럼에도", "불구하고", "여전히", "심지어", "까지도", "조차도", "하지", "않도록", "않기", "위하여", "때", "시각", "무렵", "시간", "동안", "어때", "어떠한", "하여금", "네", "예", "우선", "누구", "누가", "알겠는가", "아무도", "줄은모른다", "줄은", "몰랏다", "하는", "김에", "겸사겸사", "하는바", "그런", "까닭에", "한", "이유는", "그러니", "그러니까", "때문에", "그", "너희", "그들", "너희들", "타인", "것", "것들", "너", "위하여", "공동으로", "동시에", "하기", "위하여", "어찌하여", "무엇때문에", "붕붕", "윙윙", "나", "우리", "엉엉", "휘익", "윙윙", "오호", "아하", "어쨌든", "만", "못하다　　하기보다는", "차라리", "하는", "편이", "낫다", "흐흐", "놀라다", "상대적으로", "말하자면", "마치", "아니라면", "쉿", "그렇지", "않으면", "그렇지", "않다면", "안", "그러면", "아니었다면", "하든지", "아니면", "이라면", "좋아", "알았어", "하는것도", "그만이다", "어쩔수", "없다", "하나", "일", "일반적으로", "일단", "한켠으로는", "오자마자", "이렇게되면", "이와같다면", "전부", "한마디", "한항목", "근거로", "하기에", "아울러", "하지", "않도록", "않기", "위해서", "이르기까지", "이", "되다", "로", "인하여", "까닭으로", "이유만으로", "이로", "인하여", "그래서", "이", "때문에", "그러므로", "그런", "까닭에", "알", "수", "있다", "결론을", "낼", "수", "있다", "으로", "인하여", "있다", "어떤것", "관계가", "있다", "관련이", "있다", "연관되다", "어떤것들", "에", "대해", "이리하여", "그리하여", "여부", "하기보다는", "하느니", "하면", "할수록", "운운", "이러이러하다", "하구나", "하도다", "다시말하면", "다음으로", "에", "있다", "에", "달려", "있다", "우리", "우리들", "오히려", "하기는한데", "어떻게", "어떻해", "어찌됐어", "어때", "어째서", "본대로", "자", "이", "이쪽", "여기", "이것", "이번", "이렇게말하자면", "이런", "이러한", "이와", "같은", "요만큼", "요만한", "것", "얼마", "안", "되는", "것", "이만큼", "이", "정도의", "이렇게", "많은", "것", "이와", "같다", "이때", "이렇구나", "것과", "같이", "끼익", "삐걱", "따위", "와", "같은", "사람들", "부류의", "사람들", "왜냐하면", "중의하나", "오직", "오로지", "에", "한하다", "하기만", "하면", "도착하다", "까지", "미치다", "도달하다", "정도에", "이르다", "할", "지경이다", "결과에", "이르다", "관해서는", "여러분", "하고", "있다", "한", "후", "혼자", "자기", "자기집", "자신", "우에", "종합한것과같이", "총적으로", "보면", "총적으로", "말하면", "총적으로", "대로", "하다", "으로서", "참", "그만이다", "할", "따름이다", "쿵", "탕탕", "쾅쾅", "둥둥", "봐", "봐라", "아이야", "아니", "와아", "응", "아이", "참나", "년", "월", "일", "령", "영", "일", "이", "삼", "사", "오", "육", "륙", "

칠", "팔", "구", "이천육", "이천칠", "이천팔", "이천구", "하나", "둘", "셋", "넷", "다섯", "여섯", "일곱", "여덟", "아홉", "령", "영"]

참고 문헌

[1] 안인석, 김현우, 김형주, "텍스트 랭크 알고리즘을 이용한 사용자 타임라인 요약 기법", 정보과학회논문지: 데이터베이스 제 39 권 제 4호, 2012

[2] 홍진표, 차정원, "TextRank 알고리즘을 이용한 한국어 중요 문장 추출". 한국정보과학회 학술발표논문집 36(1C), 2009.6, 311-314 (4 pages)

[3] 설진석, 이상구, "lexrankr: LexRank 기반 한국어 다중 문서 요약", 한국정보과학회 학술발표논문집, 2016.12, 458-460 (3 pages)

[4] 배원식, 차정원, "TextRank 알고리즘을 이용한 문서 범주화", 한국정보과학회 학술발표논문집 36(1A), 2009.6, 80-81 (2 pages)

[5] 조형락, 김성진, 이동호, "의미기반 텍스트 랭크 알고리즘을 이용한 다중 문서 요약", 한국정보과학회 학술발표논문집, 2015.12, 756-758 (3 pages)

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

[7] Rada Mihalcea, Paul Tarau, "TextRank: Bringing Order into Texts", Proceedings of the European Conference on Artificial Intelligence (ECAI 2004)

[8] 조윤성, 안기홍, 김수경, "Enriched TextRank 모델을 활용한 핵심 문장 추출에 관한 연구", 한국정보기술학회, 2013

[9] 권영대, 김누리, 이지형, "문장 수반 관계를 고려한 문서 요약", 한국정보과학회 학술발표논문집, 2016

[10] http://blog.theeluwin.kr/post/146188165713/summariz3

[11] http://konlpy.org/ko/v0.4.3/

[12] http://excelsior-cjh.tistory.com/93

[13] https://ko.wikipedia.org/wiki/Tf-idf

[14] https://www.kinds.or.kr , 한국언론진흥재단, 시소러스(Thesaurus) 사전

[15] https://ratsgo.github.io/from%20frequency%20to%20semantics/2017/05/10/postag/