

Term project

1. 구체적인 프로젝트의 목적
2. 수행방법
3. 진행 결과 및 성과 수준

1. 구체적인 프로젝트의 목적

a. 주제 선정 및 목적

현대사회에서 뉴스는 우리의 일상생활과 비즈니스 활동에 필수적인 요소로 자리 잡았다. 뉴스는 우리에게 세상의 다양한 사건과 이슈들을 알려주며, 우리의 의사결정에 중요한 역할을 한다. 하지만 언론사가 제공하는 뉴스는 종종 복잡하고 이해하기 어렵다. 특히, 우리가 원하는 아티클의 주제를 파악하는데 어려운 요소가 많고, 이로 인해 우리가 언론사로 부터 제공 받는 정보로부터 자료를 요약하고, 최적의 선택을 하기 힘들게 만든다. 이는 뉴스를 접하는 현대인에게 하여금 혼란을 야기하고, 불필요한 비용 지출을 초래하기도 한다.

본 보고서는 이러한 문제를 해결하기 위해 크롤링을 통해 관련된 내용의 뉴스 데이터를 빠르게 수집하며, 이를 LDA(잠재 디리클레 할당)를 통해 각 문서를 이루는 토픽의 비중은 어떻게 되는지를 분석한다. 또한, 각 토픽을 이루는 단어의 비중은 어떤지에 대한 분석을 통해 TOPIC MODEL을 시각화 하여 보다 명확하게 이해할 수 있게 한다.

이 분석으로 각 단어나 문서들의 집합에 대해 숨겨진 주제를 찾아내어 문서나 키워드별로 주제끼리 묶는다. 이후 주어진 문서에 대하여 각 뉴스에 어떠한 주제들이 존재하는지를 서술하는 확률적 토픽 모델링을 통하여, 단어수 분포를 분석한다. 이를 통해 관련 뉴스의 검색어에 관해 어떤 주제들을 함축하고 있는지 파악하며, 뉴스를 더 잘 이해하고, 더 효율적인 의사결정을 할 수 있게한다.

2. 수행방법

2-a) 크롤링

크롤링이란 스크래핑을 통하여 인터넷에서 존재하는 데이터를 컴퓨터 프로그램을 통하여 자동화된 방법으로 웹에서 데이터를 수집하는 모든 작업을 말한다.

크롤링을 통하여 네이버의 뉴스 기사를 관련 기사들로 수집해서 분류하고 저장한다. 본 보고서는 삼성전자에 대한 뉴스를 기반으로 크롤링을 하였으며, 이것을 통해 기사 문서에 담긴 키워드를 추출하고 인덱싱 하였다. 삼성전자의 뉴스 아티클의 수집은 2024년 5월 27일 18

시 21분 부터 2024년 6월 21일까지 진행하였다. 총 136개의 데이터를 수집하였고, 각각 뉴스가 쓰인 시각과, 뉴스의 제목, 뉴스의 본문내용, 네이버 뉴스의 링크들이 feature에 포함되었다.

엑셀로 추출한 데이터의 예시는 다음과 같다.

1	A	B	C	D
1	date	title	link	content
	2024.5.27 18:21	검찰, '기밀유출 혐의' 前 삼성전자 부사장 구속영장 재청구	https://n.news.naver.com/mnews/article/003/0012570436?cid=102	<p>지난 1월8일 구속영장 기각 후 재청구상심디스플레이 전 그룹장도 재청구</p> <p>[서울=뉴시스] 서울 서초구 삼성전자 서초사옥(사진=뉴시스DB)(서울=뉴시스) 허종민 기자 = 검찰이 기밀유출 혐의를 받는 전 삼성전자 부사장에 대한 구속영장을 지난 1월에 이어 다시 청구했다.27일 서울중앙지법 정보기술범죄수사부(부장검사 이호는) 전 삼성전자 부사장(尹憲廷) 안오씨에 대해 부장검정 발지 및 영압비밀 보호에 관한 법률 위반(영압비밀 누설 등) 혐의로 구속영장을 청구했다.안씨는 자신의 부하직원이었던 삼성 내 특허담당 직원과 공모해 기밀자료를 유출한 뒤, 이를 자신이 대표로 있는 '시너지IP'와 삼성전자 소송에 활용한 혐의를 받는다.시너지IP는 안씨가 설립한 특허 에이전트 회사로, 특허권자인 '스테인 테카'와 LLC와 함께 삼성전자를 상대로 미국 법원에 우선미러론과 음성인식 관련 특허침해 소송을 제기했다.다만 지난 23일 재판에 등장한 핵심 증거자료는 안씨 등이 개입한 이번 소송이 심각한 불법행위와 부정한 방법으로 제기됐다고 판단해 특허청에 주장에 대해 기각 판결을 내린 바 있다.법원은 판결문에 이들의 불법행위를 '부정적이고, 불공정하며, 기만적이고, 법치주의에 반하는 혐오스러운 행위'라고 명시했다고 한다. 또 이들이 삼성의 기밀정보를 악용해 삼성이 회복할 수 없는 피해를 입었다고 적시했다.판결은 한국, 미국, 중국 특허법인으로 특허 상심디스플레이의 사내 특허 출원 대리인 등 선정 대가로 수년에 걸쳐 함께 약 6억원을 수수한 혐의를 받는 전 삼성디스플레이 출원그룹장 이호씨에 대해서도 배심수재 등 혐의로 구속영장을 재청구했다. 앞서 검찰은 지난달 4일 구속영장을 청구했지만 기각된 바 있다.</p>
2	2024.5.28 11:16	'기밀유출 혐의' 前 삼성전자 부사장, 30일 구속심사	https://n.news.naver.com/mnews/article/003/0012571365?cid=102	<p>특허담당 직원과 공모해 기밀 유출상선 상대로 美법원 특허 소송 제기 배임 혐의 前 디스플레이 그룹장도</p> <p>[서울=뉴시스]서울 서초구 삼성전자 서초사옥(사진=뉴시스DB)(서울=뉴시스)최서진 기자 = 기밀유출 혐의를 받는 전 삼성전자 부사장(尹憲廷) 안오씨가 오는 30일 구속 전 피의자 심문(공판심리)을 받는다.28일 법원에 따르면 서울중앙지법 상판부 영장전담 부장판사는 30일 오전 11시에 안씨에 대한 영장심리심리를 진행할 예정이다.서울중앙지법 정보기술범죄수사부(부장검사 이호는)는 지난 27일 안씨에 대해 부장검정 발지 및 영압비밀 보호에 관한 법률 위반(영압비밀 누설 등) 혐의로 구속영장을 청구했다.안씨는 자신의 부하직원이었던 삼성 내 특허담당 직원과 공모해 기밀자료를 유출한 뒤, 이를 자신이 대표로 있는 '시너지IP'와 삼성전자 소송에 활용한 혐의를 받는다.시너지IP는 안씨가 설립한 특허 에이전트 회사로, 특허권자인 '스테인 테카'와 LLC와 함께 삼성전자를 상대로 미국 법원에 우선미러론과 음성인식 관련 특허침해 소송을 제기했다.다만 지난 23일 재판에 등장한 핵심 증거자료는 안씨 등이 개입한 이번 소송이 심각한 불법행위와 부정한 방법으로 제기됐다고 판단해 특허청에 주장에 대해 기각 판결을 내린 바 있다.법원은 판결문에 이들의 불법행위를 '부정적이고, 불공정하며, 기만적이고, 법치주의에 반하는 혐오스러운 행위'라고 명시했다고 한다. 또 이들이 삼성의 기밀정보를 악용해 삼성이 회복할 수 없는 피해를 입었다고 적시했다.판결은 서울중앙지법은 이날 삼성디스플레이의 사내 특허 출원 대리인 등 선정 대가로 수년에 걸쳐 함께 약 6억원을 수수한 혐의를 받는 전 삼성디스플레이 출원그룹장 이호씨에 대해서도 영장심리심리를 진행할 계획이다.</p>
3				<p>'삼성 레노스 HVAC 북미' 설립·'냉난방공조 시장 경쟁력 강화'</p> <p>[서울=뉴시스]이현주 기자 = 삼성전자가 미국 냉난방공조 기업 '레노스'와 합작법인을 설립하고, 북미 지역에서 수요가 높은 개별 공조 시장 공략에 나선다. 삼성전자는 '레노스(Lennox)'와 합작법인 '삼성 레노스 HVAC 북미(Samsuno Lennox HVAC North America)'</p>

2-b) LDA

LDA란 잠재 디리클레 할당으로, 주제 모델링에서 사용되는 생성 통계 모델이다. 주로 주어진 문서에 대하여 각 문서에 어떤 주제들이 존재하는지에 대한 확률모형으로써 사용된다. 또한 토픽별 단어의 분포나, 문서별 토픽의 분포를 모두 추정한다. 따라서 본 보고서에는 LDA를 통하여 삼성전자에 대한 뉴스아티클 데이터를 갖고 문서 집합에서 얻은 분포로 부터 토픽을 뽑고, 이후 해당 토픽에 해당하는 단어를 뽑아 볼것이다.

위의 크롤링을 통해 수집한 삼성전자에 관한 뉴스아티클 데이터를 갖고 LDA를 통해 토픽모델링을 진행하였고, 시각화를 통해 기사 문서에 담긴 키워드를 추출하고 그 빈도수를 통하여 기사에서 삼성전자에 관해 자주 사용하는 단어를 파악하였다. 기사내의 토픽에 유사한 단어들이 존재하는지 여부를 대략적으로 확인 할 수 있었다.

단어 개수에 따른 다차원의 토픽 구분을 시각화를 위해 차원을 축소하여 여러 단어를 principal component analysis(PCA)를 통해 2차원으로 축소하였다. 이후 Saliency 와 같은 지표를 통해 각 단어의 중요도를 확인해보았다.

다음은 본 연구의 구체적인 수행절차와 방법이다.

```
data = pd.read_excel('/content/drive/MyDrive/시빅주/삼성전자크롤링 (1).xls')
```

5] data

	date	title	link	content
0	2024-05-27 18:21:14	검찰, '기밀 유출 혐의' 前 삼성전자 부사장 구속영장 재청구	https://n.news.naver.com/mnews/article/003/001...	[Wn지난 1월8일 구속영장 기각 후 재청구삼성 디스플레이 전 그룹장도 재청구WnWn...
1	2024-05-30 17:28:12	케이뱅크, 삼성전자 제휴 챗린지박스 6시간 만에 완판	https://n.news.naver.com/mnews/article/366/000...	[WnWnWnWnWn케이뱅크 제공 케이뱅크는 삼성전자와 손잡고 30일...
2	2024-05-31 10:25:11	민트, 삼성전자와 '갤럭시 AI 추가보상' 시행	https://n.news.naver.com/mnews/article/050/000...	[WnWnWnWnWn민트 갤럭시 AI 추가보상 이벤트 이미지. 사진=민트민트가 삼성...
3	2024-05-31 09:25:16	"둘이 반대로 가네"... 삼성전자 '반등' vs SK하이닉스 '위태'	https://n.news.naver.com/mnews/article/015/000...	[WnWnWnWnWn사진=연합뉴스삼성전자와 SK하이닉스의 희비가 엇갈리고 있다. 약...
4	2024-05-30 11:32:20	[포토]삼성전자, 생성형 AI 달고 코파일럿 지원...'갤럭시북4 엣지' 공개	https://n.news.naver.com/mnews/article/018/000...	[WnWnWnWnWn[이데일리 방인권 기자] 박준호 삼성전자 MX사업부 갤럭시 에코...

우선 크롤링을 통해 얻은 136개의 뉴스의 원본에서 뉴스의 본문내용만을 추출하여 분석에 이용하였다.

```
def clean_text(text):
    text = text.replace(".", "").strip()
    text = text.replace("·", " ").strip()
    pattern = '[^ㄱ-ㅣ가-힣0-9]+'
    text = re.sub(pattern=pattern, repl='', string=text)
    return text

/usr/local/lib/python3.10/dist-packages/ipykernel/ipkernel.py:283: DeprecationWarning: `should_run_async`
and should_run_async(code)

clean_list = []
for i in data:
    ct = clean_text(i)
    clean_list.append(ct)

/usr/local/lib/python3.10/dist-packages/ipykernel/ipkernel.py:283: DeprecationWarning: `should_run_async`
and should_run_async(code)

clean_list
```

한 인물이다 그는 2019년 퇴사한 직후 특허관리기업을 설립한 뒤 씨를 통해 불법 취득한 삼성전자 기밀 문건을 이용해 미국에서 삼성전자를 상대로 특허침해소송을 제기한 혐의를 받는다 는 직접 생산 활동은 하지 않은 채 보유한 특허권의 행사만으로 수익을 창출하는 사업자로 일명 특허괴물로 불린다 안 전 부사장은 를 운영하면서 음향기기 업체인 미국 테키야를 대리해 삼성전자와 특허에 대한 라이선스를 협상하던 중 씨가 무단 취득한 2021년 8월 삼성전자의 테키야 특허 관련 분석 보고서를 전달받은 것으로 조사됐다 안 전 부사장은 삼성전자 전 센터 기술분석그룹장 씨와 공모해 보고서에 담긴 기밀정보를 분석한 다음 소송 을 낼 특허를 선별해 2021년 11월 삼성전자를 상대로 9천만 달러의 합의금을 요구하는 소송을 제기했다 삼성전자를 특허 소송으로 위협하면 라이선스 협상에서 유리한 위치를 점할 수 있을 것이라는 판단에서였다 이들은 삼성전자를 효과적으로 위협하기 위해 전략적으로 매출액이 큰 휴대폰 관련 특허 등을 소송 제기 대 상으로 선정하고 중국계 와 삼성전자 내부 보고서를 공유하고 소송 비용을 투자받는 등 기밀 정보를 광범위 하게 활용한 것으로 조사됐다 미국 텍사스 동부지법은 최근 이 소송을 기각하면서 한국 검찰의 수사 결 과를 토대로 안 전 부사장이 삼성전자 내부 자료를 소송에 이용한 사실을 인정하고 부정직하고 불공정하며 법치주의에 반하는 혐오스러운 행위라고 질타했다 검찰은 사건 관련자들을 압수수색해 이들의 범행을 입 증할 물적 증거를 확보해 안 전 부사장과 씨 씨 등 4명을 재판에 넘겼으며 이 과정에서 이들이 내부 보고서

이후 뉴스 본문내용 문자열에 포함된 특수문자(점, 마침표 등)를 처리하여 한글, 숫자, 공백 을 제외한 순수한 문자만을 추출하였다.

```

> # 명사만 추출
data_word_r=[]
for i in range(len(clean_list)):
    try:
        data_word_r.append(mecab.nouns(clean_list[i]))
    except Exception as e:
        continue

# 추출된 단어들 중 한 글자 이상의 명사만 추출
data_word = []
for i in data_word_r:
    noun = []
    for j in i:
        if len(j) > 1:
            noun.append(j)
    data_word.append(noun)

```

mecab 형태소 분석기를 이용하여 전처리를 진행한 데이터 중에서 명사만을 추출하는 과정을 진행하였다.

다음으로, 명사의 길이가 2 이상인 단어들을 추출하여 data_word라는 리스트에 추가하는 토큰화 과정을 진행하였다.

```
data_word
```

```
[['구속',  
  '영장',  
  '기각',  
  '재청',  
  '구삼',  
  '디스플레이',  
  '그룹',  
  '장도',  
  '재청구',  
  '서울',  
  '뉴시스',  
  '서울',  
  '서초구',  
  '삼성전자',  
  '서초',  
  '사옥',  
  '사진',  
  '뉴시스',  
  '서울',  
  '뉴시스',  
  '하종민',  
  '기자',  
  '검찰',  
  '기밀',  
  '유출',  
  '혐의',  
  '삼성전자',  
  '부사장',  
  '구속',  
  '영장',  
  '청구',  
  '서울']
```

총 136개의 뉴스데이터를 통해 길이가 2이상인 단어(명사)를 저장한 data_word의 리스트는 위 그림과 같다.

```
stop_words = [
    '가', '가까스로', '가끔', '각', '각각', '각자', '각종', '갖고말하자면', '같다', '같이', '개의치않고', '거니와',
    '거바', '거의', '것', '것과 같이', '것들', '게다가', '게우다', '겨우', '견지에서', '결과에 이르다', '결국',
    '결론을 낼 수 있다', '경사경사', '고려하면', '고로', '곧', '공동으로', '과', '과연', '관계가 있다', '관계없이',
    '관련이 있다', '관하여', '관한', '관해서는', '구', '구체적으로', '구토하다', '그', '그들', '그때', '그래',
    '그래도', '그래서', '그러나', '그러니', '그러니까', '그러면', '그러므로', '그러한즉', '그런 까닭에', '그런데',
    '그런즉', '그럼', '그럼에도 불구하고', '그렇게 함으로써', '그렇지', '그렇지 않다면', '그렇지 않으면', '그렇지만',
    '그렇지않으면', '그리고', '그리하여', '그만이다', '그에 따르는', '그위에', '그저', '그중에서', '그치지 않다',
    '근거로', '근거하여', '기대여', '기점으로', '기준으로', '기타', '까닭으로', '까악', '까지', '까지 미치지',
    '까지도', '과당', '꿈꿈', '까악', '나', '나머지는', '남들', '남짓', '너', '너희', '너희들', '네', '넷', '년',
    '논하지 않다', '놀라다', '누가 알겠는가', '누구', '다른', '다른 방면으로', '다만', '다섯', '다소', '다수',
    '다시 말하자면', '다시말하면', '다음', '다음에', '다음으로', '단지', '답다', '당신', '당장', '대로 하다',
    '대하면', '대하여', '대해 말하자면', '대해서', '덜그', '더구나', '더군다나', '더라도', '더불어', '더욱더',
    '더욱이는', '도알하다', '도착하다', '동시에', '동안', '원바에야', '원이상', '두번째로', '둘', '중중', '위따라',
    '위이어', '등간에', '들', '등', '등등', '딩동', '따라', '따라서', '따위', '따지지 않다', '딱', '때', '때가 되어',
    '때문에', '또', '또한', '윽윽', '라 해도', '령', '로', '로 인하여', '로부터', '로써', '룩', '를', '마음대로',
    '마져', '마저도', '마치', '막론하고', '만 못하다', '만약', '만약에', '만은 아니다', '만이 아니다', '만일',
    '만큼', '말하자면', '말할것도 없고', '매', '매번', '매쓰건다', '몇', '모', '모두', '무렵', '무름쓰고', '무슨',
    '무엇', '무엇때문에', '물론', '및', '바꾸어말하면', '바꾸어말하자면', '바꾸어서 말하면', '바꾸어서 한다면',
    '바꿔 말하면', '바로', '바와같이', '밖에 안된다', '반대로', '반대로 말하자면', '만드시', '버금', '보는데서',
    '보다더', '보도록', '본대로', '봐', '봐라', '부류의 사람들', '부터', '불구하고', '불문하고', '불물',
    '비격거리다', '비교적', '비길수 없다', '비로소', '비록', '비슷하다', '비추어 보아', '비하면', '뿐만 아니라',
    '뿐만아니라', '뿐이다', '배격', '배격거리다', '사', '삼', '상대적으로 말하자면', '생각한대로', '설령', '설마',
    '설사', '셋', '소생', '소인', '좌', '헛', '습니까', '습니다', '시각', '시간', '시작하여', '시초에', '시키다',
    '실로', '실지어', '아', '아니', '아니나다를가', '아니라면', '아니면', '아니었다면', '아래윗', '아무거나',
    '아무도', '아야', '아울러', '아이', '아이고', '아이구', '아이야', '아이쿠', '아하', '아홀', '안 그러면',
    '알기 위하여', '알기 위해서', '알 수 있다', '알았어', '앗', '알에서', '알의것', '야', '약간', '알자', '어',
    '어기여차', '어느', '어느 년도', '어느것', '어느곳', '어느때', '어느쪽', '어느해', '어디', '어때', '어떠한',
    '어떤', '어떤것', '어떤것들', '어떻게', '어떻게', '어이', '어제서', '어젯든', '어쩔수 없다', '어찌',
    '어찌했든', '어찌했어', '어찌하든지', '어찌하여', '언제', '언젠가', '얼마', '얼마 안 되는 것', '얼마간',
    '얼마나', '얼마든지', '얼마만큼', '얼마큼', '얼얼', '에', '에 가서', '에 달려 있다', '에 대해', '에 있다',
    '에 한하다', '에게', '에서', '여', '여기', '여덟', '여러분', '여보시오', '여부', '여섯', '여전히', '여차',
    '연관되다', '연이서', '열', '열차', '열사람', '예', '예를 들면', '예를 들자면', '예컨대', '예하면', '오',
    '오로지', '오르다', '오자마자', '오직', '오호', '오히려', '와', '와 같은 사람들', '와르르', '와아', '왜',
    '왜냐하면', '외에도', '요만큼', '요만한 것', '요만한걸', '요컨대', '우르르', '우리', '우리들', '우선',
    '우에 종할한것과같이', '운운', '윽', '위에서 서술한바와같이', '위하여', '위해서', '윽윽', '윽', '으로',

```

위의 단어들을 stop_word로 지정하였으며, data_word 리스트에 포함된 명사에서 stop_word에 해당되는 단어들을 삭제하였다.

추가적으로 한국어 불용어 사전에서 가져온 단어목록 '서울', '뉴시스', '뉴스', '최근', '이날', '기자', '기업', '선언', '제공', '부문', '사용', '사진', '최대', '사옥', '회사', '이후', '최고', '서초'를 추가하였다.

```
# 제일 많이 나온 단어 |
tags = count.most_common(50)
tags

/usr/local/lib/python3.10/dist-packages/ipykernel/ipkernel.py:344: DeprecationWarning:
and should_run_async(code)
[('삼성전자', 1037),
 ('반도체', 266),
 ('노조', 244),
 ('파업', 230),
 ('삼성', 198),
 ('사업', 185),
 ('특허', 164),
 ('삼노', 148),
 ('기술', 126),
 ('미국', 125),
 ('제품', 124),
 ('부사장', 119),
 ('갤럭시', 119),
 ('디스플레이', 108),
 ('비디아', 96),
 ('소송', 85),
 ('혐의', 80),
 ('기능', 79),
 ('에너지', 79),
 ('스마트', 76),
 ('공정', 75),
 ('영업', 74),
 ('가전', 71),
 ('주가', 71),
 ('구속', 67)]
```

word cloud를 진행하기 위해 data_word에 뉴스 아티클 별로 나뉘져있는 단어들을 noun_list라는 하나의 리스트로 만들었고, 이후 noun_list에 포함된 단어들을 오름차순 기준으로 출현빈도를 확인하였다.

이것을 시각화하기 위해 워드클라우드 라이브러리를 사용하였고, 만들어진 워드클라우드를 matplotlib을 이용하여 시각화를 진행하였다.


```

from gensim.corpora.dictionary import Dictionary

# 토큰화 결과로부터 dictionary 생성
dictionary = Dictionary(data_word)
print('#Number of initial unique words in documents:', len(dictionary))

# 문서 빈도수가 너무 적거나 높은 단어를 필터링하고 특성을 단어의 빈도 순으로 선택
dictionary.filter_extremes(keep_n=2000, no_below=5)
print('#Number of unique words after removing rare and common words:', len(dictionary))

# 카운트 벡터로 변환
corpus = [dictionary.doc2bow(text) for text in data_word]
print('#Number of unique tokens: %d' % len(dictionary))
print('#Number of documents: %d' % len(corpus))

#Number of initial unique words in documents: 2986
#Number of unique words after removing rare and common words: 770
#Number of unique tokens: 770
#Number of documents: 136

```

위 그림은 gensim라이브러리를 통해 토픽모델링을 진행한 코드이다.

dictionary 라이브러리를 사용하여 토큰화된 문서 리스트 data_word 내의 단어들을 이용하여 단어사전을 생성하였다. 이후 문서빈도가 5 이하인 너무 적게 분포된 단어나 너무 많이 나온 단어들을 필터링하고, 상위 단어 2000개만을 남겼다.

doc2bow를 이용하여 문서를 카운트 벡터를 변환하였고 그 결과 각 문서를 단어 ID와 빈도를 쌍으로 나타내었다.

```

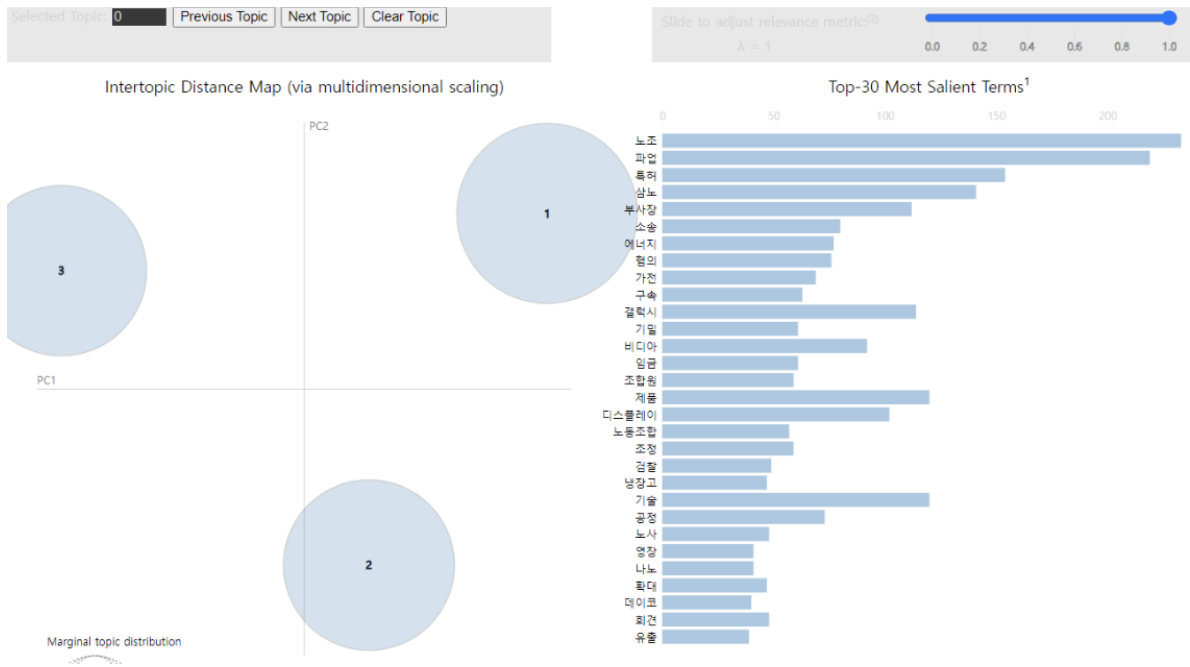
from gensim.models import LdaModel

num_topics = 3
passes = 15 # 반복 횟수
model = LdaModel(corpus=corpus, id2word=dictionary, passes=passes, num_topics=num_topics)

```

이와 같은 과정을 바탕으로 LDAModel을 이용하여 TopicModel을 생성하였다. 본 연구에서는 토픽수를 3개, 반복학습 횟수는 15로 지정하였다.

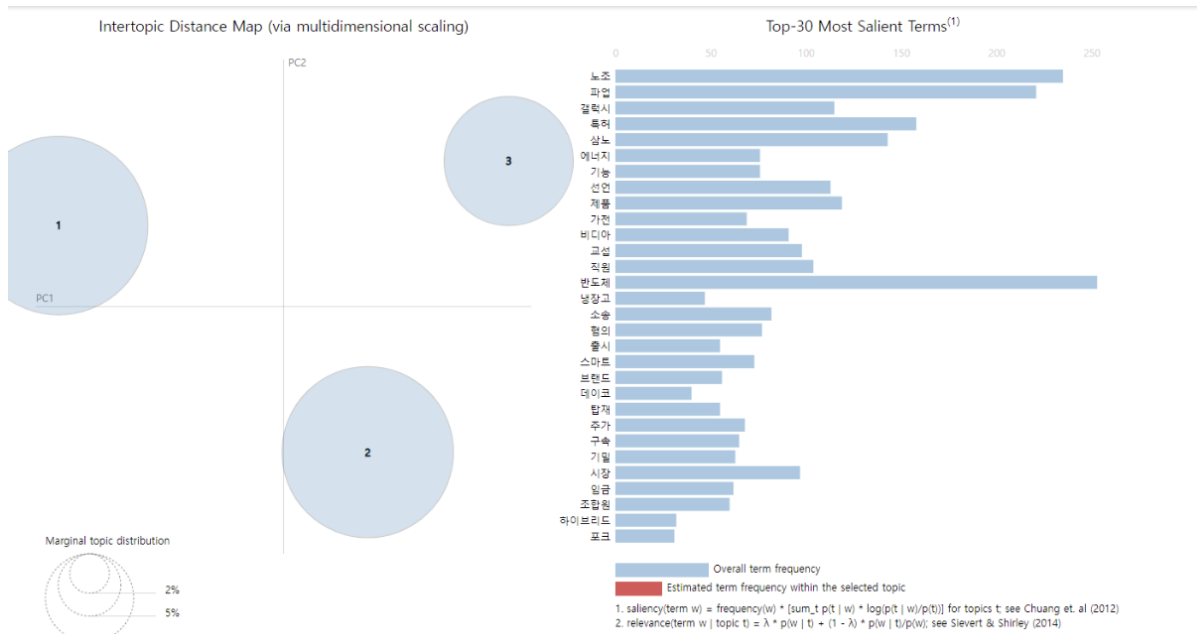
최종적으로 시각화 한 결과는 다음과 같다. 시각화 라이브러리로는 pyLDAvis를 사용하였다.



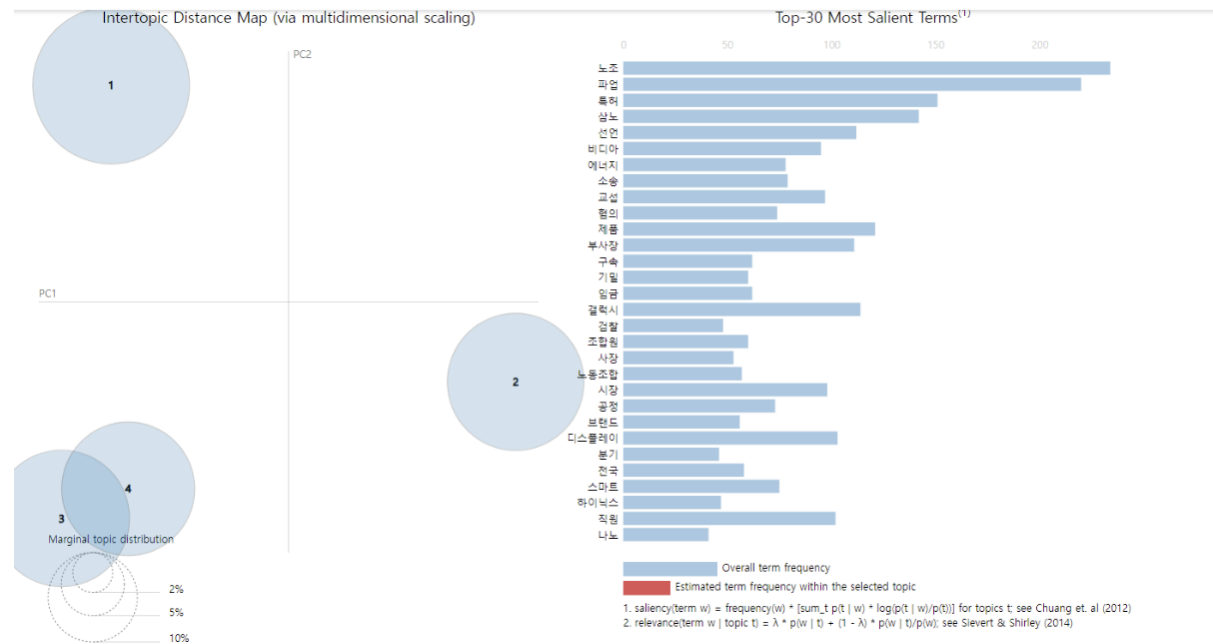
3. 진행결과 및 성과수준

3-a) word cloud

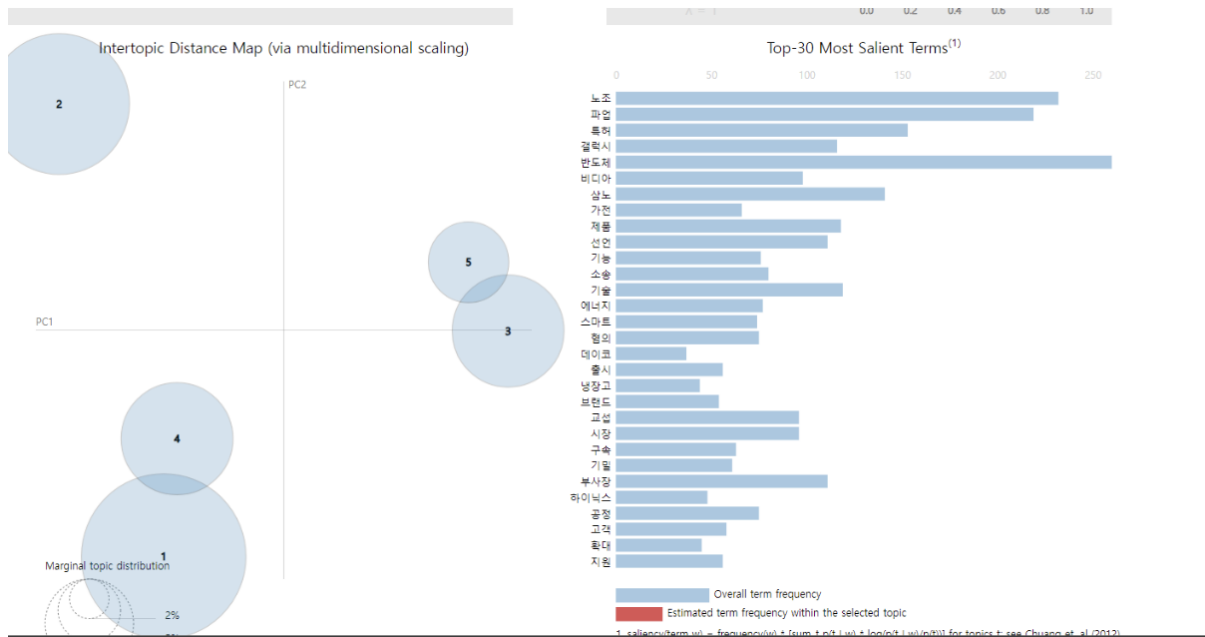
선정된 136개의 뉴스 데이터에서 텍스트 전처리와 불용어 제거 후, 가장 빈도가 높은 단어들을 시각적으로 표현하고 전체적인 주제를 파악하기 위해 빈도수가 가장 많은 상위 50개의 단어를 이용해 word cloud를 진행하였다.



<토픽수 = 4>

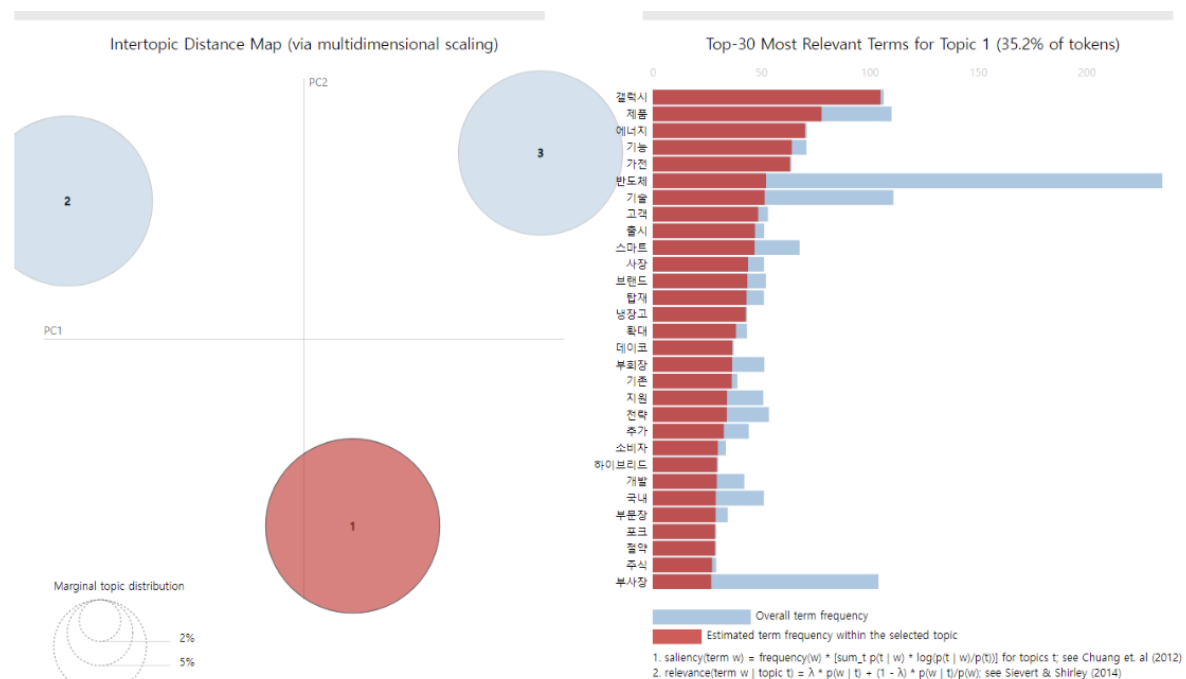


<토픽수 = 5>



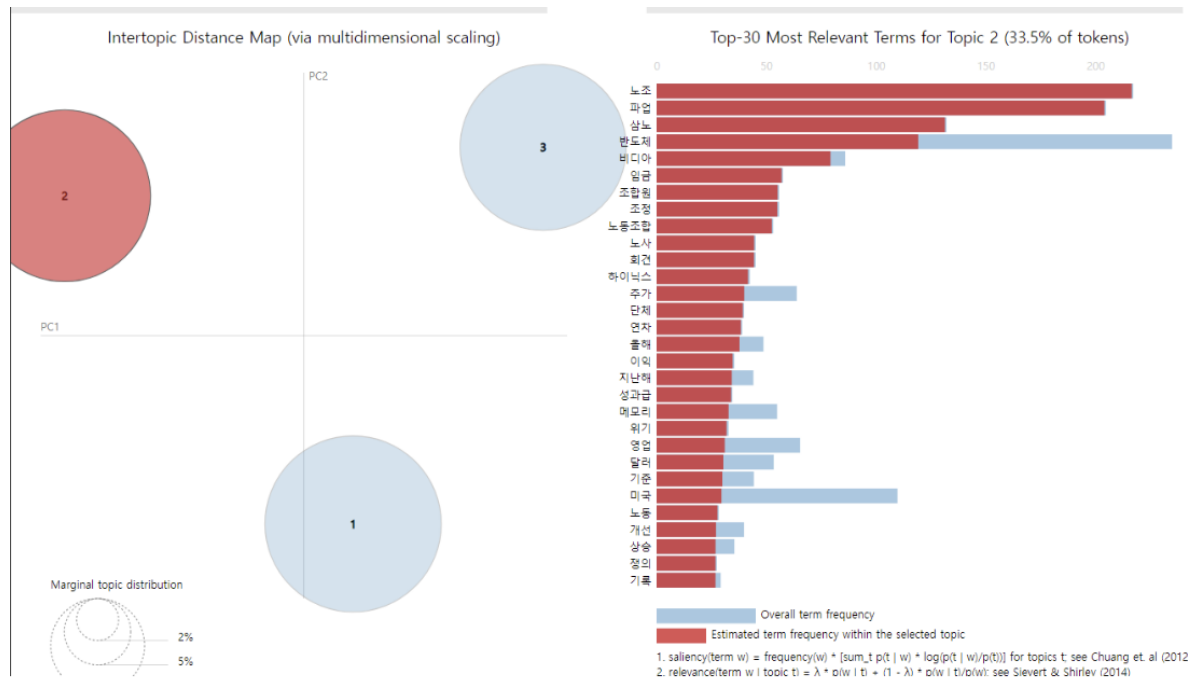
하이퍼파라미터중 하나인 passes = 15로 설정하였다. 즉 모델 학습과정에서 데이터셋 (data_word)를 전체 15번 반복해서 모델을 학습하였다. 적절한 토픽의 개수를 결정하기 위하여 각각 num_topic수를 3,4,5로 정한 후 모델을 생성하였다. 그 결과 토픽 수가 4개가 넘어가는 순간부터 토픽들간의 중복되는 부분이 존재하는것을 확인하였고, 따라서 각 토픽들이 전체 문서에서 차지하는 비율이 30%로 적절히 분배되고, 토픽들간 분리정도가 명확한 토픽수 = 3을 채택하였다. 추가적으로 각 토픽에 대해 주제 단어 관련성에 대한 파라미터 $\lambda=0.96$ 으로 설정하였다.

<토픽 1>



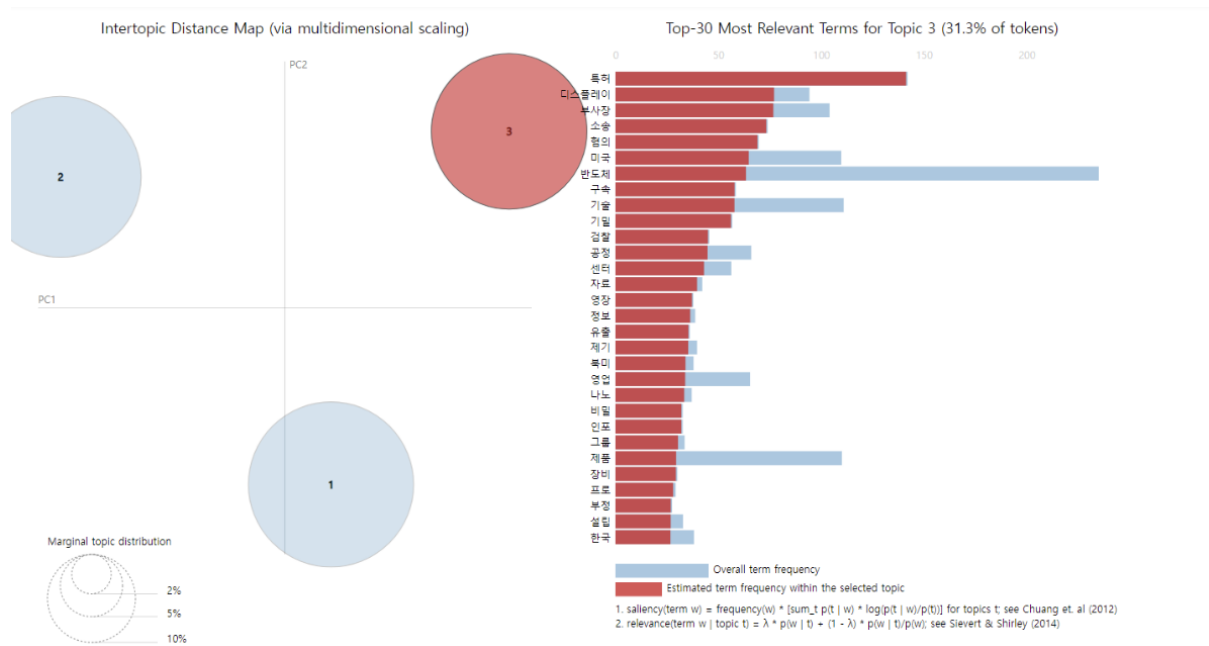
토픽 1의 경우 상위 단어들이 갤럭시, 제품, 에너지, 기능, 가전등인것으로 보아 토픽1은 삼성의 기술적인 부분을 다루는 토픽인것으로 유추할 수 있다.

<토픽2>



토픽2의 경우 상위 단어들이 노조,파업,삼노,임금,조합원들로 나타났다. 이로 유추해보았을 때 삼성의 노조 파업과 관련된 주제를 담고있다는것을 확인하였다.

<토픽3>




토픽3의 경우는 상위 단어들이 특허, 부사장, 디스플레이, 소송, 기밀 미국등과 같이 분포하였는데, 최근 삼성의 특허기술이 유출된것과 관련하여 소송중인 내용을 담고있음을 파악할 수 있었고, 이 사건에 부사장과 미국등이 관련된것으로 유추된다.

Reference

원하는 정보만 수집한다! 크롤링과 빅데이터 분석 활용

빅데이터 분석의 사회적 필요성 현대사회에서의 빅데이터에 대한 지속적인 관심과 실험적인 시도들은 다변화된 현대 사회를 보다 정교하게 예측하고 효율적으로 작동하도록 정보를 제공하며, 개인화된 사회 구

 <https://news.samsungdisplay.com/22907>



<https://gist.github.com/spikeekips/40eea22ef4a89f629abd87eed535ac6a>