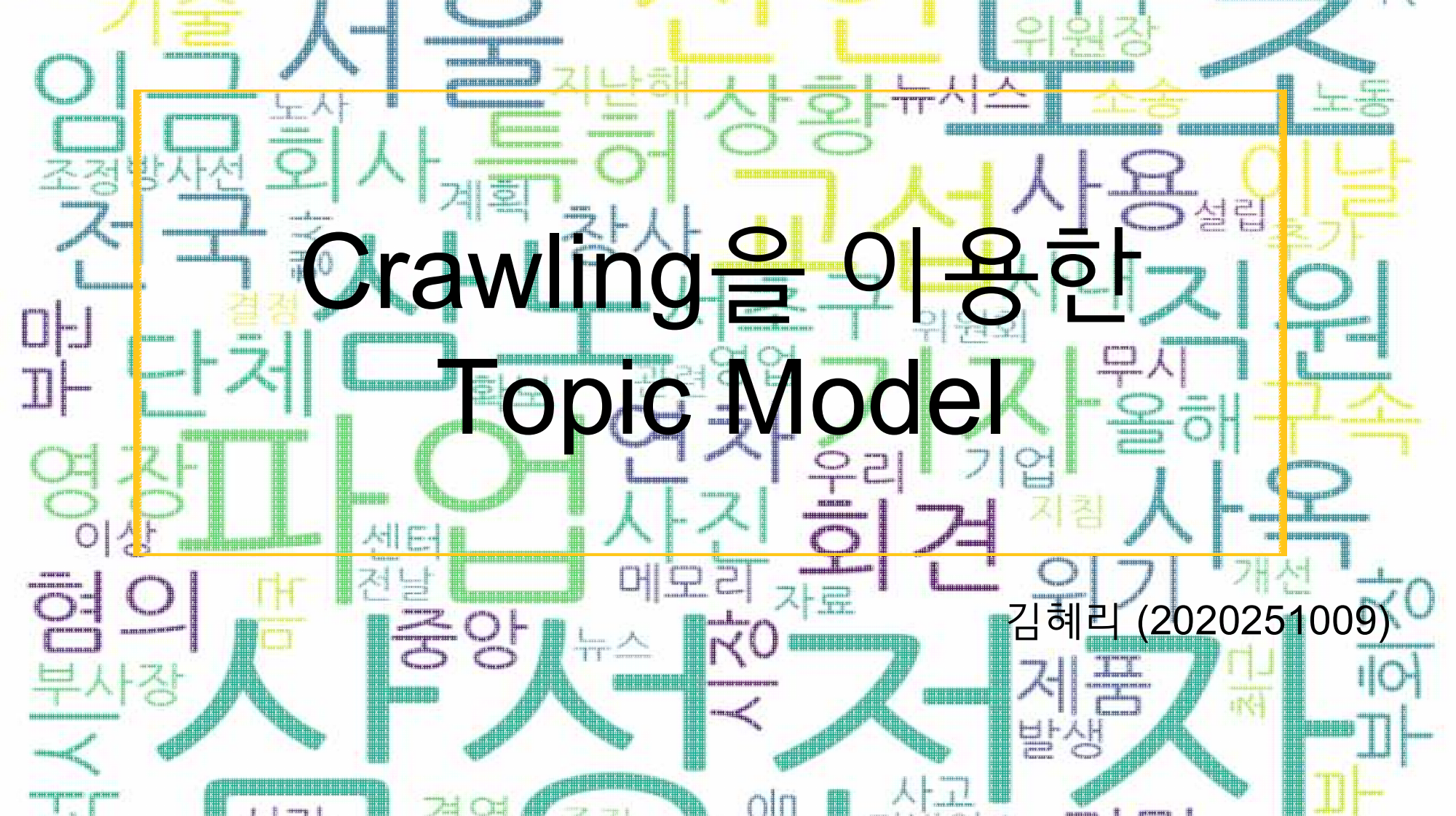


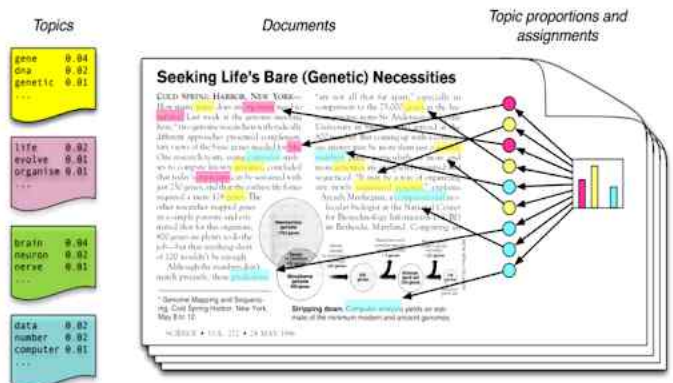
Crawling을 이용한 Topic Model

김혜리 (2020251009)

목차

프로젝트 목적

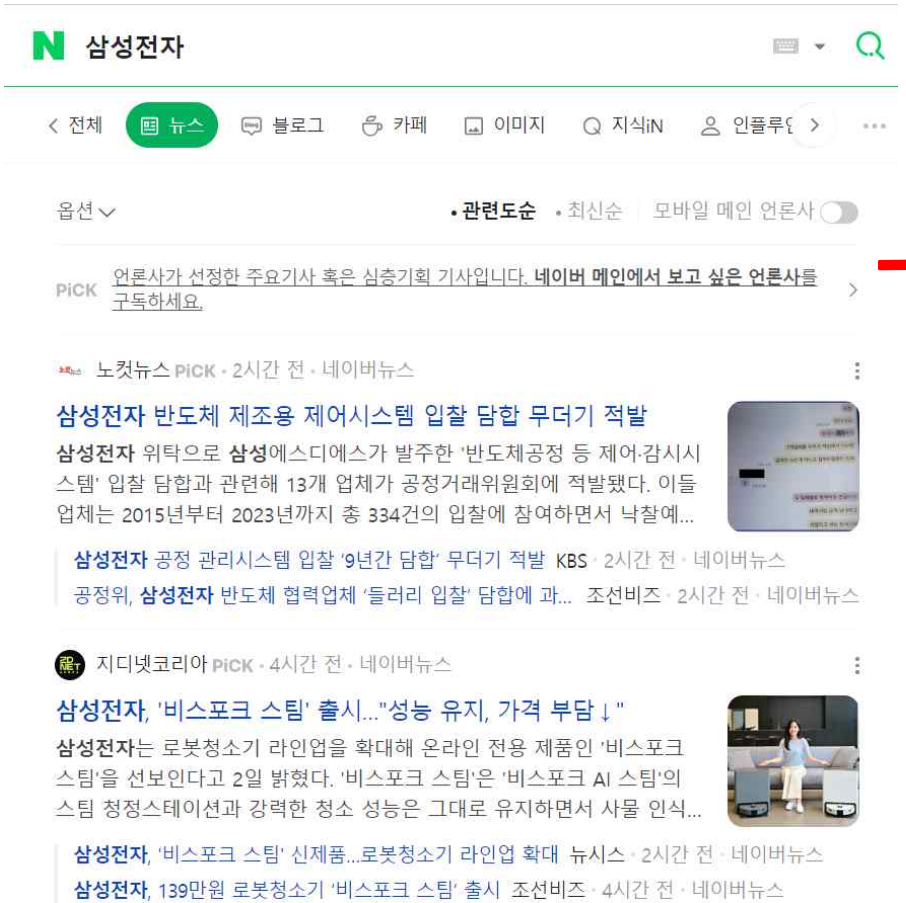
언론사가 제공하는 뉴스는 종종 복잡하고 이해하기 어렵다.
이로 인해 우리가 언론사로 부터 제공받는 정보로부터 자료를 요약하고, 최적의 선택을 하기 힘들게 만든다.
이는 뉴스를 접하는 현대인에게 하여금 혼란을 야기하고, 불필요한 비용 지출을 초래하기도 한다.



-크롤링을 이용해 데이터 수집, 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.

-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

Crawling



<모델링 수행을 위한 csv 파일 추출>

원하는 검색어를 입력하고 검색어와 관련된 뉴스의 본문 데이터를 추출하기 위해 Crawling을 사용하여 실시간으로 변환되어 제공되는 데이터를 수집

Word Cloud

- 단어의 글자 크기가 크면 클수록 해당 단어의 빈도수가 많다는 것을 의미
- 기사 문서에 담긴 키워드를 추출하고 그 빈도수를 추출하여 기사에서 자주 사용하는 단어를 추출
- 기사내 어떤 토픽이 존재하는지 대략적으로 유추가능

Topic Model

- 토픽모델은 문서와 단어로 구성된 행렬(DTM; document-term matrix)을 기반으로 문서에 잠재되어 있다고 가정된 토픽의 등장 확률을 추정하는 통계적 텍스트 처리기법

- 토픽을 찾는 방법 중 가장 많이 알려진 것은 잠재적 디리클레 할당(LDA; latent Dirichlet allocation)모형이 있고, 이를 확장한 상관토픽모형(CTM;correlated topic model), 구조적 토픽모형(STM;structural topic model)이 있음.

- 이 중 가장 대중적으로 많이 사용되는 LDA는 문서 모형 처리 및 분류 등에 기본적으로 사용되는 방법임

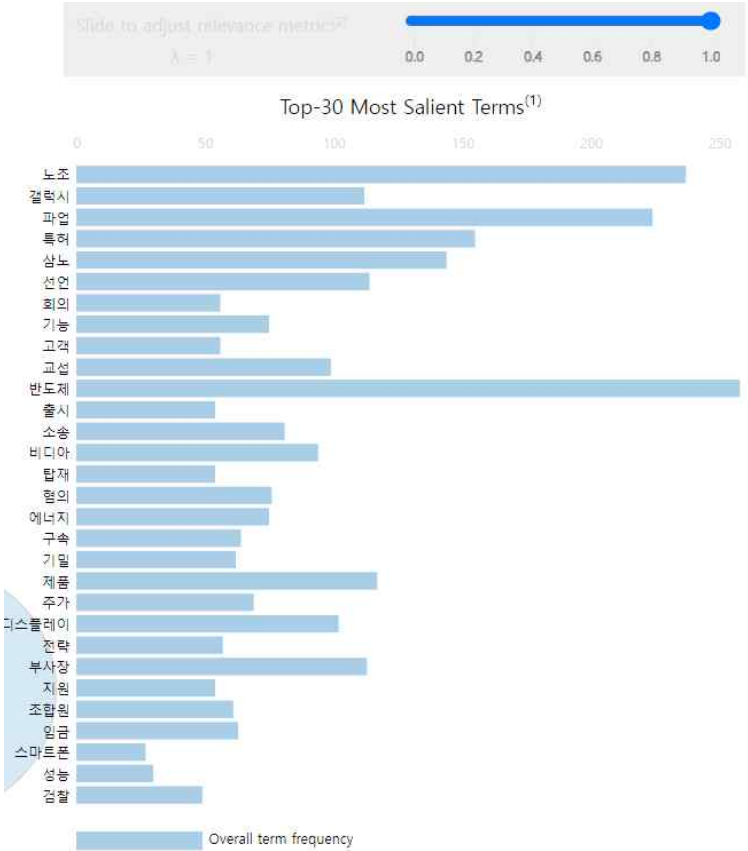
Topic Model

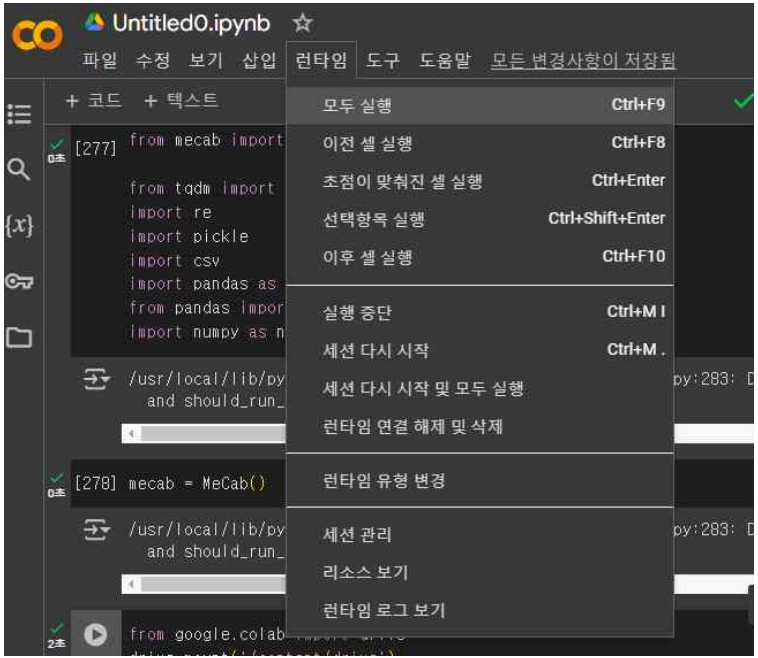
<모델링 수행을 위한 csv 파일>

검색어와 관련된 네이버 뉴스기사의 데이터를
크롤링을 통해 수집

기사 문서에 담긴 명사만 추출해 토큰화

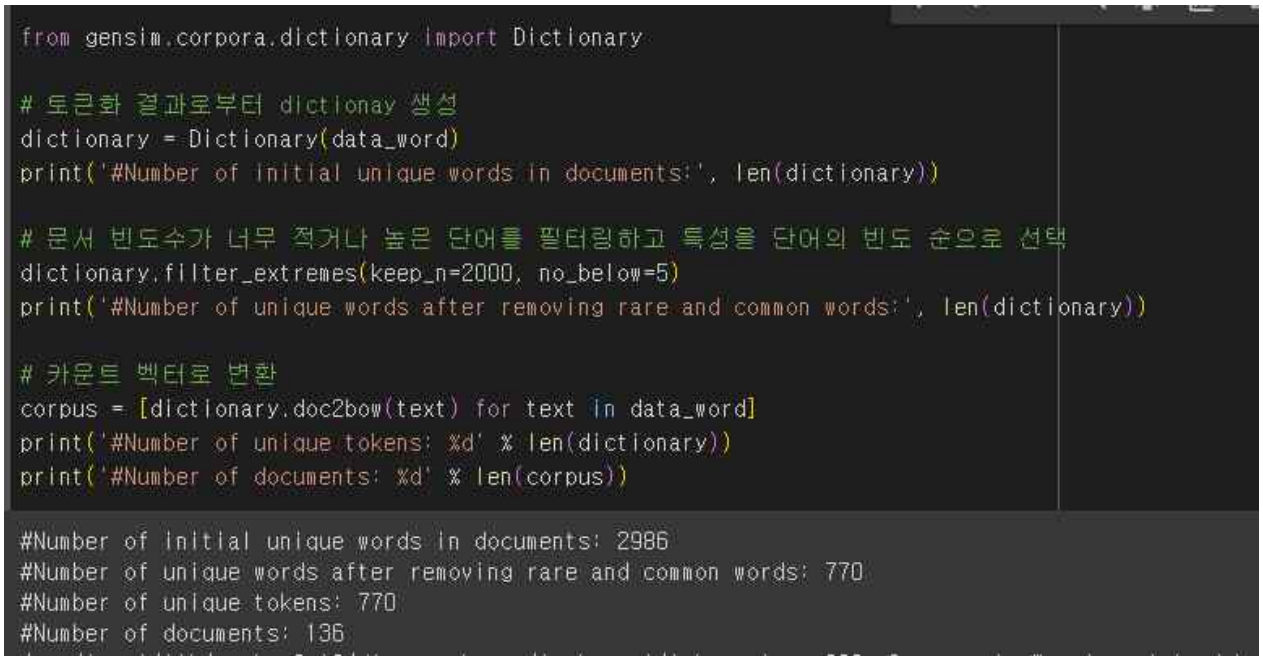
<기사 문서에 담긴 키워드 추출>






```
[280] data = pd.read_excel('/content/drive/MyDrive/지텍추/삼성전자크롤링 (1).xls')
      data
```

해당 프로그램은 google colab에서 작성되었으며, 크롤링된 데이터 csv를 구글 드라이브에 업로드 한 뒤, 데이터를 읽어오는 부분에 해당 데이터셋의 경로를 붙여 넣어준 후 런타임을 실행하면 된다.



```
from gensim.corpora.dictionary import Dictionary

# 토큰화 결과로부터 dictionay 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
```

gensim을 이용해 토픽모델링을 진행하였다.

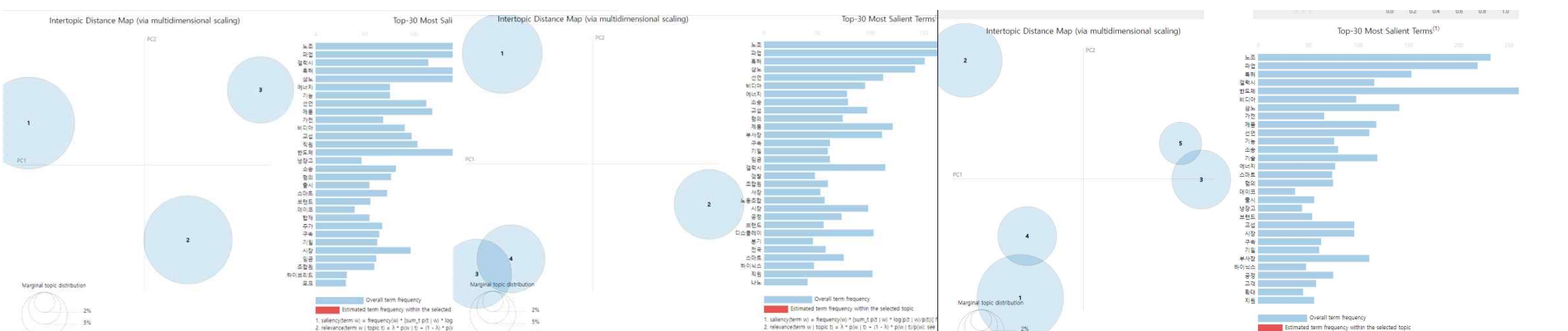
dictionary을 이용해 토큰화된 문서 리스트인 data_word 내의 단어들을 이용해 단어 사전을 생성한 후 문서빈도가 너무 적거나 많은 단어들을 필터링하고, 상위 단어 2000개만 남기고 나머지 단어들을 제거하였다. 추가적으로 단어가 등장하는 문서의 빈도가 5번 이하인 것들을 제거해주었다.

doc2bow를 이용해 문서를 카운트 벡터로 변환하였고 그결과 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)
```

이후 LdaModel을 이용해 토픽모델을 생성하였다.
토픽수는 3개, 반복 학습 횟수는 15로 지정하였다.

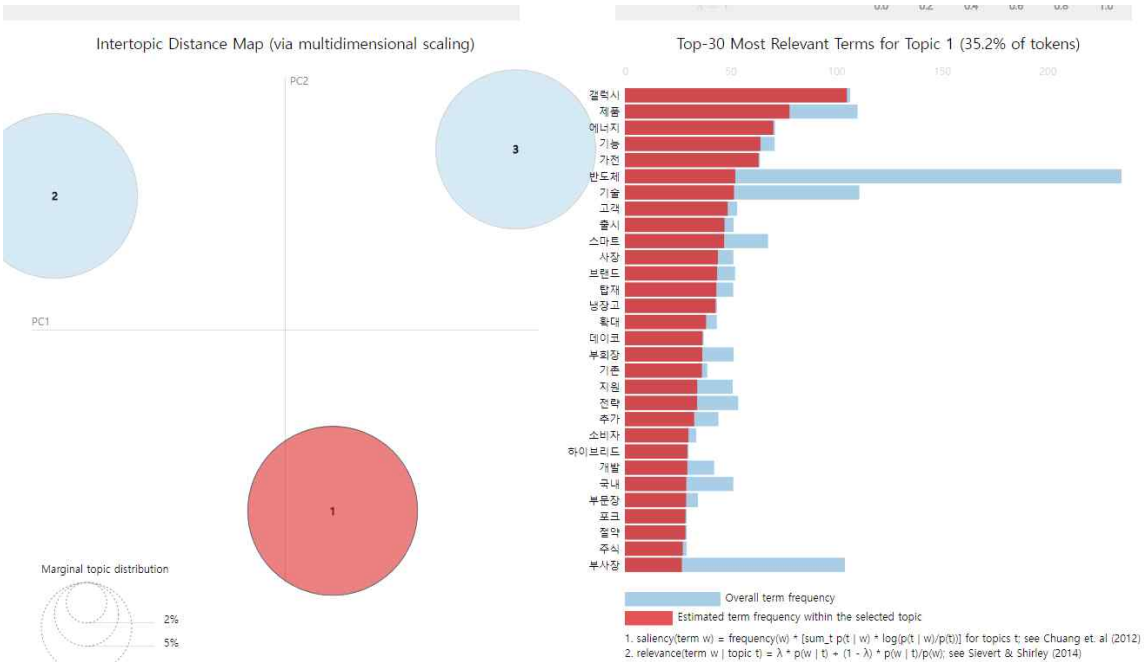


토픽수는 적절한 토픽 수를 정하기 위해 각각 num_topics를 3,4,5로 정한 후
모델을 생성한 결과 토픽 수가 4개가 넘어가는 순간부터 토픽들간 중복되는 부분이 존재하였다.
따라서 각 토픽들이 전체 문서에서 차지하는 비율이 30%로 적절하게 분배되고, 가장 적합한 토픽 수는
토픽들 간 분리가 가장 잘 되어있는 3개라고 결정하였다. 각 토픽에 대해서 주제 단어 관련성에 대한 파라미터인
λ=0.96으로 설정했다.

결과

<1번 토픽>

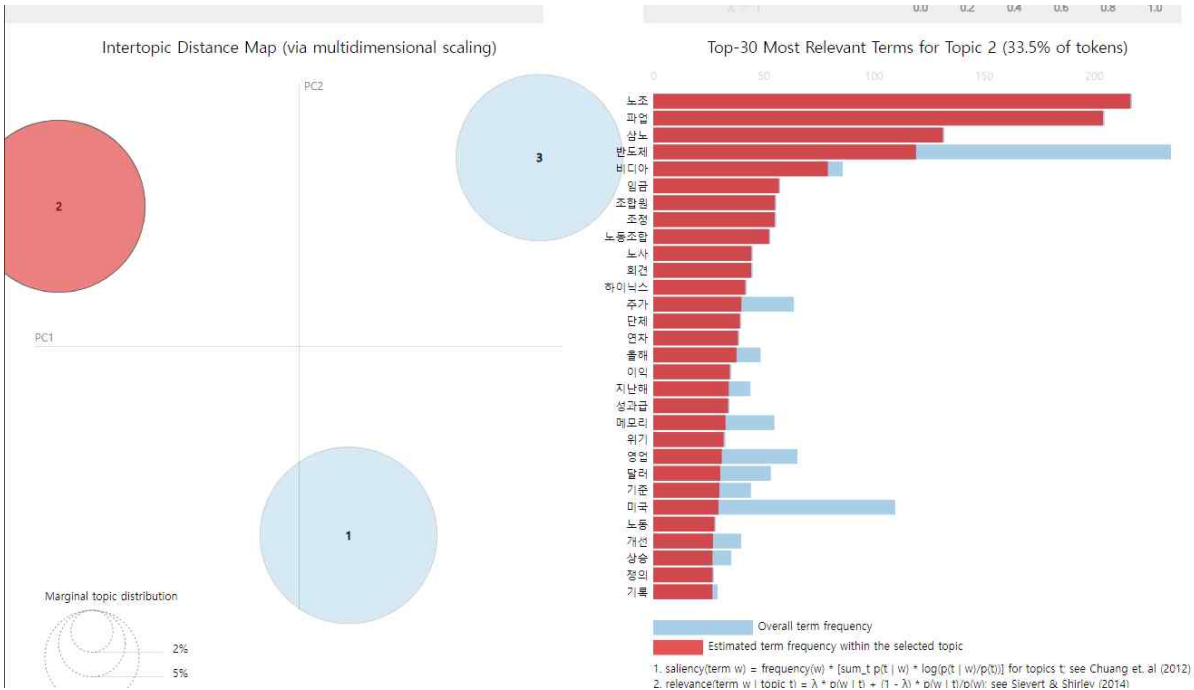
하이퍼파라미터
-num_topics = 3
-passes = 15 (학습 반복 횟수)
- λ =0.96



상위 단어들이 갤럭시,제품,에너지,기능,가전 등 인 것으로 보아 토픽1은 삼성의 기술적인 부분과 관련된 토픽인 것을 확인 할 수 있다.

결과

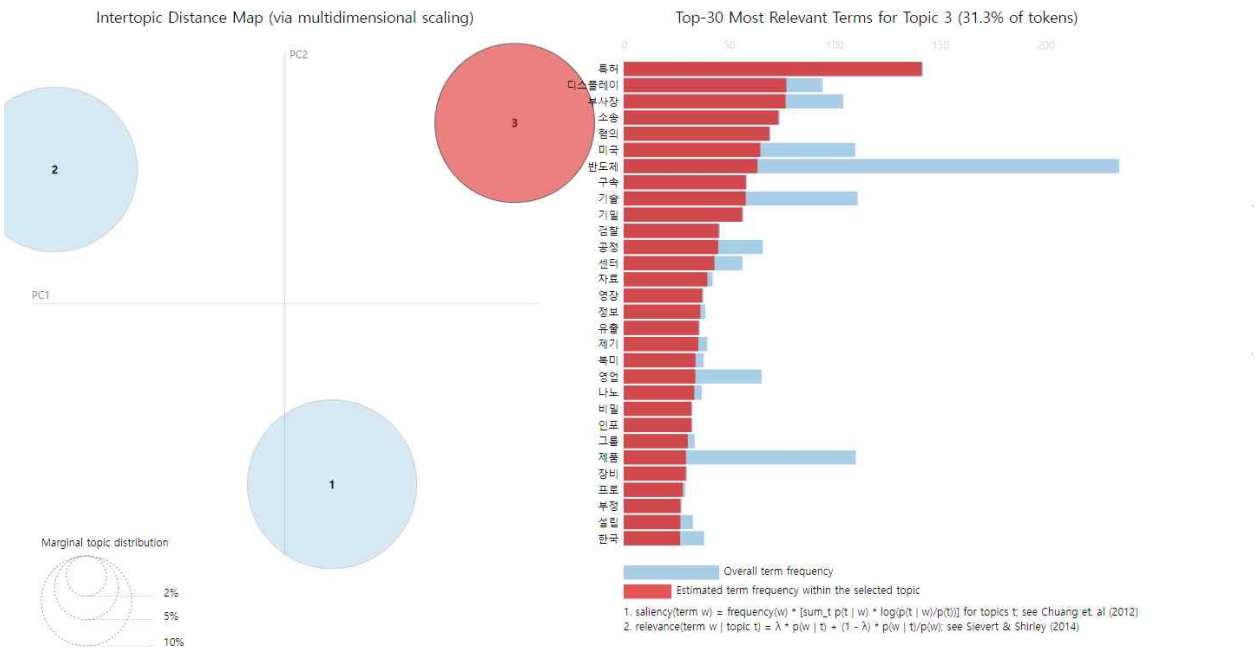
<2번 토픽>



상위 단어들이
노조,파업,삼노,임금,조합원등으로
삼성의 노조 파업과 관련된
토픽임을 알 수 있다.

결과

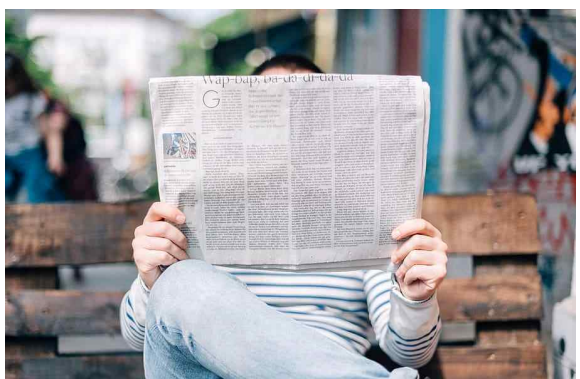
<3번 토픽>



상위 단어들이 특허, 부사장,디스플레이,소송,기밀,미국 등과 관련된 것을 보아

최근 삼성의 특허기술이 유출된 것과 관련하여 소송중인 것을 파악 할 수 있고 이사건에 부사장과 미국등이 관련된 것으로 유추된다.

기대효과



- 뉴스 기사를 활용해 현재 해당 주제에 대한 이슈를 파악할 수 있다
- 바쁜 현대인들에게 관련 뉴스 아티클이 다루는 내용을 쉽고 빠르게 제공할 수 있다

Reference

https://news.samsungdisplay.com/22907
https://gist.github.com/spikeekips/40eea22ef4a89f629abd87eed535ac6a