비즈니스 프로그래밍 기말 보고서

202021703

박다현

목차

**장 제목 입력(수준 1)1**

장 제목 입력(수준 2)2

장 제목 입력(수준 3)3

**장 제목 입력(수준 1)4**

장 제목 입력(수준 2)5

장 제목 입력(수준 3)6

1. 주제 및 주제 선정 이유  
   필자는 코로나-19 수혜 기업이라고 불리는 유한킴벌리의 뉴스룸을 크롤링 및 분석하여, 코로나 이전 시기(2018-2019), 코로나팬데믹 발병 시기(2020), 위드 코로나 시기(2021-2022)를 비교해 기업의 핵심 가치 및 비전이 어떻게 변화하였는지, 생산 및 판매가 어떤 제품을 중심으로 이루어졌는지 등 흐름을 알아보고자 한다. 더 나아가 코로나의 영향에 따라 달라진 기사의 성격과 사회 경제 및 일상생활의 변화도 빈도 분석을 통해 발견하고자 하였다. 주제를 선정할 때, 뉴스를 크롤링해 언론사의 정치적인 성향을 찾기 보다는 사회적 배경을 고려하면 좋을 것 같다는 생각이 들었고, 유한킴벌리’는 마스크 등을 생산 및 판매하는 대표적인 기업이므로 코로나-19에 관한 기사가 많을 것이라 생각되어 프로그래밍 대상으로 삼았다.
2. 코드 설명  
   대부분의 코드 설명은 주석에 달아놨지만, 대표적인 코드만 순차적으로 하나씩 소개하겠다.  
   1) 페이지 URL수집  
   텍스트이(가) 표시된 사진

   자동 생성된 설명  
   BeautifulSoup로 html데이터를 lxml형태로 파싱을 해준다. soup.select()를 통해 내가 찾고자 하는 a태그를 반환해 urls에 저장한다. 저장된 urls에서 하나씩 꺼내 링크를 news\_items\_urls1에 저장한다. Orderset은 뉴스 url들을 담게끔 해준다.  
   다른 세 개의 url에도 똑같이

2) 수집한 URL로 각 기사에 들어가 제목, 날짜, 본문 등 크롤링  
텍스트이(가) 표시된 사진

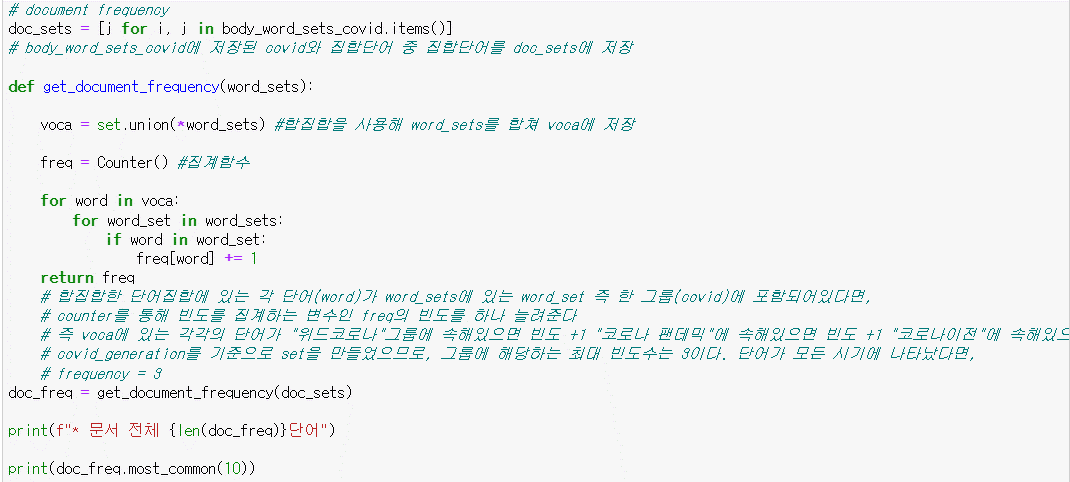
자동 생성된 설명  
기사의 페이지 수집이 완료되었으면, 형태소 분석을 도와주는 kiwi모듈과 시간lapse를 주는 time 모듈을 호출한다. 빈 리스트를 news1에 저장하고, soup를 이용해 각 찾고자 하는 리스트를 반환해 각 변수에 저장한다. 저장된 변수들은 news1에 딕셔너리 형태로 추가한다.

3) ujson형태로 저장 후 3개의 json을 하나로 통합

4) 통합된 yuhan\_newsroom.jsonl을 데이터 전처리  
텍스트이(가) 표시된 사진

자동 생성된 설명  
세 개의 json을 통합한 json파일을 한 줄씩 읽어와 json\_data에 저장한다. 이후 이전에 데이터 처리할 때와 동일하게, 각 변수에 원하는 리스트를 반환해 저장한다. 형태소는 불용어 제거, 단어제한 옵션을 추가해 단어만 리스트에 추가한다. 이후 news에 딕셔너리 형태로 값을 저장한다.

5) 빈도분석  
빈도분석은 주석에서 충분히 설명했다고 느껴 간단히 짚고 넘어가겠다.  
#term\_frequency  
텍스트이(가) 표시된 사진

자동 생성된 설명  
#document\_frequency  
  
#inverse\_document\_frequency  
텍스트이(가) 표시된 사진

자동 생성된 설명  
#tf-idf  
텍스트이(가) 표시된 사진

자동 생성된 설명

6) 시각화  
텍스트이(가) 표시된 사진

자동 생성된 설명  
워드 클라우드. 최대 단어 수를 정하고, 배경 수, 폰트 경로도 설정. 제목은 covid로 설정하고 사이즈 정해서 시각화  
텍스트이(가) 표시된 사진

자동 생성된 설명  
트리맵. 마찬가지로 필요한 모듈을 호출하고, 최대 단어 수 배경색 폰트 경로 등을 지정해준다. Covid\_generation에 저장되어있는 “위드코로나”, “코로나 팬데믹”, “코로나이전”을 불러와 각 시기에 맞는 단어 중 가장 빈도가 높은 단어 N\_WORDS개를 추출해 괄호 하나 없앤 뒤 word와 count에 할당해 시각화  
텍스트이(가) 표시된 사진

자동 생성된 설명  
막대그래프. 다른 시각화 그래프와 동일하게, 필요한 모듈을 호출하고 폰트, 색깔 글씨 크기 등 결정. Covid\_generation에 저장되어있는 “위드코로나”, “코로나 팬데믹”, “코로나이전”을 불러와 각 시기에 맞는 단어 중 가장 빈도가 높은 단어 20개를 추출해 괄호 하나 없앤 뒤 word와 count에 할당해 시각화 (트리맵과 동일)

1. 해석  
   워드클라우드와 트리맵, 막대그래프의 결과에 따른 분석 결과를 해석하고자 한다.

|  |  |  |
| --- | --- | --- |
| 화이트보드이(가) 표시된 사진  자동 생성된 설명 | 텍스트이(가) 표시된 사진  자동 생성된 설명 | 텍스트이(가) 표시된 사진  자동 생성된 설명 |

일단, 워드 클라우드를 살펴보면, 코로나 팬데믹 이전시기에는 유한킴벌리의 대표 제품인 여성을 위한 생리대에 많은 초점을 맞추었음을 알 수 있다. 시니어를 타겟팅 하는 제품 및 캠페인이 있었을 것을 조심스레 추측할 수 있고, 기저귀와 요실금도 꽤 큰 비중을 차지함을 볼 수 있다.   
 코로나 팬데믹에 들어서자, 생리대에 대한 관심은 조금 줄었으나 유지되었고, ‘마스크’의 비중이 눈에 띄는 것을 볼 수 있다. 코로나 팬데믹 상황으로 마스크에 대한 수요가 폭증함에 따라 소비자의 욕구를 충족시키기 위한 다양한 종류의 마스크 공급에 힘썼을 것을 추측할 수 있다. 하기스와 귀저기로 미루어봤을 때 아기를 위한 제품도 큰 비중을 유지함을 알 수 있다.  
 위드코로나 시기에 들어서자, 생리대와 마스크에 대한 비중은 쏙 들어가고 “환경”에 큰 초점을 맞추고 있음을 볼 수 있다. 점점 환경에 관심을 갖는 똑똑한 소비자가 증가하고, ESG경영이 대두되고, 코로나19에 의한 환경오염이 심각해짐에 따라 기업 핵심 가치를 지속 가능 경영, 친환경 등에 두었을 것으로 예상할 수 있다.

다음으로, 트리맵과 막대그래프를 살펴보고자 한다.

|  |  |
| --- | --- |
| 테이블이(가) 표시된 사진  자동 생성된 설명 |  |
|  |  |
|  |  |

이전 분석과 비슷하게, 코로나 이전에는 생리대, 여성, 기저귀, 화이트, 하기스, 마스크 등 제품 중심이었다면 코로나 팬데믹부터 가능, 환경, 지속, 혁신 등 기업의 핵심가치가 좀 더 잘 나타남을 알 수 있다. 이는 위드코로나 때 정점을 찍었음이 보인다. 위드코로나 시기에는 지속, 환경, 가능, 경영 등의 단어뿐만 아니라, 지구, 탄소 등의 단어도 높은 빈도를 지님을 보여 유한킴벌리가 최근 환경에 큰 중점을 두고 있음을 확실시할 수 있다.

단어의 빈도 분석 결과를 해석하고자 한다.  
텍스트이(가) 표시된 사진

자동 생성된 설명  
----------------------------------------------------------------------------------------------------  
텍스트이(가) 표시된 사진

자동 생성된 설명  
 위는 TF-IDF를 적용하지 않은 순 단어 빈도이고, 아래는 TF-IDF를 적용한 문서 안 단어의 특별성을 고려한 빈도이다. 최대 빈도 30개의 단어를 각 시기별로 뽑았다.  
위드코로나 시기에는 예상과 같이 ‘환경’ 중심 단어 빈도가 문서의 특정성과 관계없이 큰 비중을 차지함을 볼 수 있다. 코로나 팬데믹 시기에는 문서의 특정성을 고려하였더니 코로나가 높은 단어 빈도를 보여주는 단어임을 알 수 있었다. 이외 환경단어 제품단어 모두 골고루 포함됨을 볼 수 있다. 코로나 이전 시기에는 상대적으로 제품 중심의 단어가 주를 이룸을 볼 수 있다. 물론, 환경 단어인 미세먼지의 단어 빈도는 높으나, 이는 유한킴벌리의 기업 핵심 가치를 보여주는 것이 아니라, 그 당시 두드러진 환경 문제라 해석되어 제품 중심으로 보았다.  
 각 시기에 따른 단어 빈도를 시각화함으로써 유한킴벌리의 주된 기업 핵심가치, 각 시기별 제품의 특정성을 한 눈에 파악할 수 있었다.

1. 한계 및 소감  
    유한킴벌리 뉴스룸을 크롤링 및 데이터분석을 하는 과정에서 크롤링은 몹일뿐 분석이 보스라는 것을 깨닫게 되었다. 한 코드가 진행되지 않으면 다음으로 넘어갈 수 없다는 점에서 전체적인 과정은 힘들고 진이 빠졌지만, 코드가 하나 하나 진행될 때마다 성취감이 남달랐다. 빈도 분석에서는 교수님이 올려주신 샘플의 영향을 많이 받았지만, 차근차근 따라하니 윤곽이 어느정도 보였고, 이내 선명해졌다. 개인적으로, 크롤링 할 때, “위드코로나”, “코로나 팬데믹”, “코로나 이전”이라는 기준을 따로 설정하지 않고, 기준을 게시날짜(연도)로 삼았다면 더 정확한 데이터를 갖고 분석할 수 있지 않았을까 생각이 들어 아쉬웠다. 필자는 페이지별로 크롤링했기 때문에, 각 시기와 연도를 정확히 대응시키지 못해 한 두 달의 오차가 필연적으로 발생했기 때문이다. 다음에는 ‘연도’를 기준으로 삼아 다시 한번 분석해보고 싶다.