데이터 분석 정의서

1. 분석 프로젝트 명칭

코로나 이펙트

2. 분석 배경 및 필요성

- 현재 대한민국은 코로나 바이러스 감염증(COVID-19)으로 인한
팬데믹 상태에 빠져있음. 이로 인해 사람들이 살아가는 방식 또한 변화되고 있음

- 코로나 바이러스의 확산을 막기 위한 사람들의 철저한 개인위생,
초, 중, 고, 대학교의 대면 수업을 대신한 온라인 수업,
사회적 거리 두기로 인한 인터넷 쇼핑 사용량의 증가 등 여러 가지 방면의
긍정적인 변화를 보여줌
하지만 가정폭력의 증가, 시장경제의 침체 등 부정적인 변화도 있었음.

- 코로나 바이러스 이전과 이후 생활, 문화 방면에서의 차이점을 토대로 빈도 분석하여 코로나로 인한 문화적 영향력을 확인함.
이를 알아보기 위해서 여러 가지 방법이 존재하는데, 이 중에 우리는 인터넷 뉴스를 기반으로 여론을 조사함.

- 텍스트 마이닝이란 자연어 처리 기술을 활용하여 반정형 및 비정형 데이터를 정형화하고 특징을 추출하기 위한 기술

- 텍스트 마이닝을 활용하여 비 정형화된 뉴스 기사를 다량의 데이터로 가공.
다양한 데이터 분석 방법을 이용하여 코로나로 변화된 문화를 해석

3. 분석 목표

코로나 바이러스 감염증 관련 뉴스 기사와 댓글을 정형화하고 이를 시계열 분석하여 생활, 문화의 변화를 분석하여 사용자가 보기 쉽게 가시화

4. 분석 시나리오

5. 사용 기술 설명

5.1 크롤링

크롤링이란 웹상의 존재하는 여러 데이터들 중 필요한 데이터만 추출하여 분석하고 활용하기 쉽게 데이터를 수집하는 행위
웹 서버에서 API를 지원하면 사용자가 손쉽게 데이터를 가져올 수 있으나, 그렇지 못한 데이터들은 사용자가 직접 수집해야 함. 이때 사용할 수 있는 기술이 웹 크롤링
사용되는 언어는 파이썬(Python)이고, Beautiful Soup 패키지를 사용
검색엔진과 데이터분석 등 다양한 분야에서 사용됨.

5.2 Topic Modeing -LDA(Latent Dirichlet Allocation)

토픽 모델링은 문서의 집합에서 토픽을 찾아내는 프로세스를 말함. 이 중 잠재 디리클레 할당(LDA)은 비지도학습의 일종이고 토픽 모델링의 대표적인 알고리즘
LDA는 문서의 집합으로부터 어떤 토픽이 존재하는지 알아내기 위해 사용함
우리는 토픽의 개수를 직접 지정해 가공된 데이터를 토픽에 할당
우리는 토픽의 주제를 보고 원하는 토픽을 선정 및 사용
사용되는 언어는 파이썬(Python)이고, gensim 패키지를 사용

5.3 WordCloud

워드 클라우드란 한마디로 '핵심 단어를 시각화하는 방법'. 문서의 키워드, 개념 등을 직관적으로 파악할 수 있도록 시각적으로 돋보이게 하는 기법임.
많이 언급될수록 단어를 크게 표시하는 등 여러 가지 기법이 존재.
사용되는 언어는 파이썬(Python)이고, wordcloud 패키지를 사용

5.4 Word to Vector

워드투벡터 word2vec은 word embedding에서 자주 사용되는 방법론 중 하나. 각 단어를 고차원으로 벡터화하는 것인데 이때, CBOW, Skip-gram과 같은 알고리즘을 사용하여 주변 단어와 중심 단어를 맞추어가며 가중치를 적용하는 방법.
결과적으로 유사한 단어들끼리 가까이 그룹이 이루어지게 됨.
word2vec 모델을 구현해둔 곳은 대표적으로 tensorflow와 gensim이 있으며, 이 중 gensim의 word2vc을 사용

5.5 Bokeh

데이터 시각화에서 사용되는 라이브러리 중 하나이며 파이썬을 지원. 다른 라이브러리로는 matplotlib, seaborn가 있으며, 이들과의 차이는 bokeh는 그래프를 이미지가 아닌 JS로 변환하여 렌더링할 수 있음. 원하는 데이터를 골라 동적으로 그래프를 변경하여 사용자에게 그래프를 조금 더 보기 쉽게 제공할 수 있다는 장점으로 장점으로 bokeh를 사용.

6. 사용자 환경

사용자는 웹 환경을 통하여 결과를 볼 수 있음.
웹 환경에는 코드를 숨기고 토글 버튼을 두어 사용자가 원하면 볼 수 있도록 구성.