분석 설계서

1. 분석 목표 및 환경

1.1. 데이터 분석 목표

코로나 바이러스 감염증 이후 바뀐 우리의 생활, 문화를 분석하여 긍정적 변화와 부정적 변화를 파악하고 이를 토대로 이후에 비슷한 사례에 대비할 수 있도록 한다.

1.2. 분석 환경

개발 언어 : 파이썬(Python)

개발 도구

- Google Colab(Colaboratory)
    - 구글 클라우드 기반의 Jupyter notebook 개발환경.
- Jupyter notebook
    - 문서의 포맷으로 웹 브라우저에서 파이썬 코드를 작성하고 실행할 수 있는 소프트웨어.
- IPykernel, IPython
    - 인터프리터 환경에서 파이썬을 조금 더 유연하게 작업할 수 있도록 사용되는 command shell과 이것을 프론트에서 사용할 수 있도록 지원해주는 코어 기능.

2. 사전 데이터 수집 및 가공

2.1. 파싱(데이터 수집)

파싱 대상 : 네이버 뉴스

수집 기간 : 2020년 01월 01일 ~ 2020년 08월31일

수집 키워드 : “코로나”

대상 언론사 : 네이버 뉴스 포맷을 지원하는 방송 / 신문사 24곳

데이터 저장 형식 : csv 파일로 pandas의 DataFrame 자료구조 형태로 저장.

2.2. 전처리(정제 & 정규화)

저장된 데이터는 이전과 동일하나, 기사 본문(text) 필드만 전처리 후 저장.

뉴스 기사에는 기자의 이름, 이메일, 언론사 이름, 광고 등 분석할 때 필요 없는 데이터들이 섞여있는 경우가 있어 정규표현식을 사용해 수행하고 중복 또는 비어있는(Nan) 기사들을 제거하여 저장.

2.3. 토큰화

파이썬에서는 변수에 담겨있는 데이터를 구조 그대로 저장할 수 있는 pickle 이라는 라이브러리가 제공

전처리된 데이터를 활용하여 기사 본문과 제목을 형태소분석기중 MeCab을 이용해 용도에 따라 사용하기 위해 모든 품사인 경우와 명사만을 이용하는 경우로 나누어 저장함. 이때, 명사는 일반명사, 고유명사(NNP,NNG)만을 사용

3. 시나리오 분석 선정

3.1. 전체 기사 건수 시계열 분석

목표
크롤링한 전체 뉴스 기사를 1주 단위로 취합하여 기사 건수의 변화 추이를 시계열 방식으로 시각화하는 시나리오

기능
기사 건수의 큰 변동이 있을 때 사건, 사고의 발생으로 인한 뉴스량의 변동을 통해 시기를 알고 이를 신규 확진자와 비교

예시)

특정 법안 실행 이후 전세금 변동률

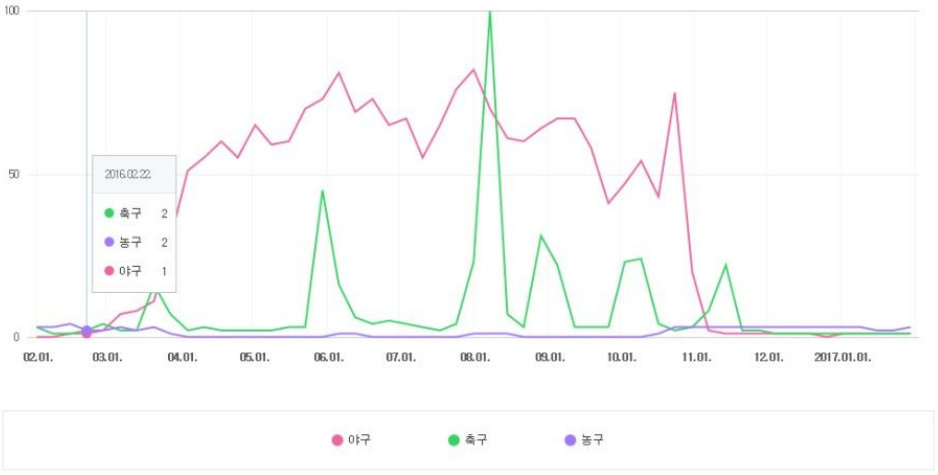
출처 [서울 전세시장 뒤집어졌는데...한국감정원은 이번주도 "안정적" - 땅집고 > 시황·분양 (chosun.com)]

3.2. Google, Naver 검색량 분석

목표
구글의 'trends' 과 네이버 'DataLab'의 검색량을 수집하여 변화 추이를 코로나 사태 이전부터 최근까지의 검색량 변화를 시계열 방식으로 시각화하는 시나리오

기능
검색량은 사람들의 관심의 지표라고 생각하여 두 개의 검색엔진을 통하여 코로나 이전과 이후의 검색량을 통하여 사회적 거리 두기 등 지침을 통한 통제가 이루어졌을 때 사람들의 관심사 및 변화된 생활에 대하여 파악

예시)



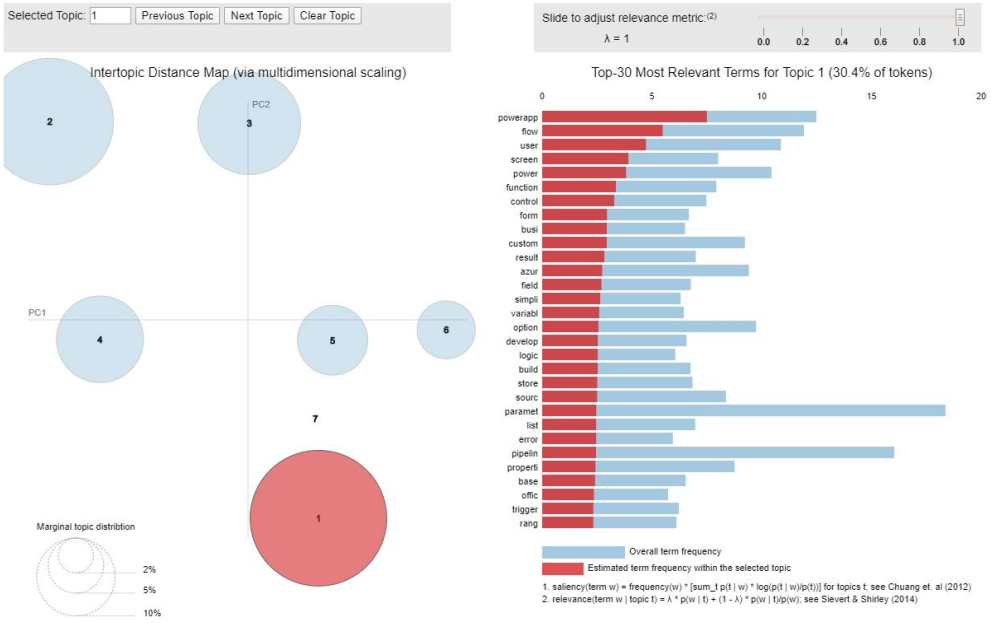
야구, 축구, 농구 3가지의 검색량을 비교한 그래프
출처 [지식 공유 블로그(마케팅, 스포츠, IT정보, 유튜브) : 네이버 블로그 (naver.com)]

3.3. LdaModeling

목표
토픽 모델링 중 하나인 LDA를 사용해 파싱 된 문서의 토픽(주제)을 확인하고 해당 토픽에 속해있는 문서들을 확인하는 시나리오.

기능
우리가 분석하려는 생활, 문화 관련 토픽들을 확인하고 선택하여 이후의 시나리오에 사용하기 위함. 또한 특정 군집이 어떤 토픽으로 이루어져 있는지 등 사용자가 직접 볼 수 있는 기능을 제공

예시)



LDA Topic Modeling을 실시하여 가시화한 결과
출처 [Building an LDA Topic Model with Azure Databricks - Adatis]

3.4. LDA 토픽 모델링을 통해 분류된 키워드 WordCloud

목표

LdaModeling에서 선정한 '생활, 문화' 토픽의 문서를 사용하여 WordCloud 라이브러리를 사용하여 시각화하는 시나리오

기능

'생활, 문화' 토픽의 문서의 키워드를 직관적으로 파악할 수 있도록 가시화하는 기능.

많은 양의 문서를 사용하므로 WordCloud를 이용하여 데이터의 특징을 알아보고자 사용

예시)

미국의 모양으로 WordCloud를 실행하여 가시화한 결과

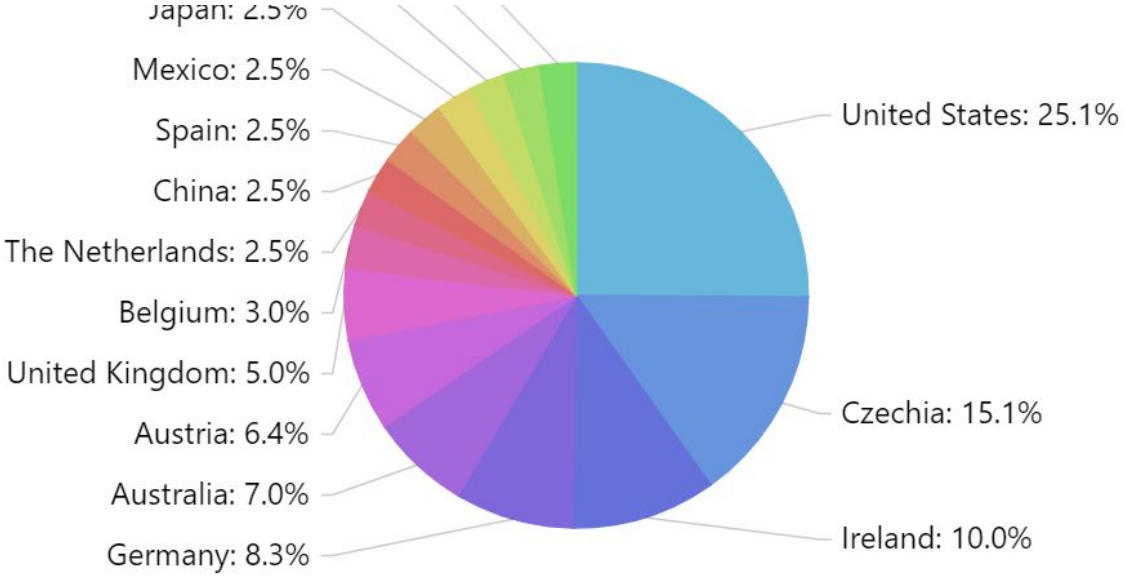
출처 [Create Word Cloud into any Shape you want using Python | by Fahmi Nurfikri | Towards Data Science]

3.5. LDA 토픽 모델링을 통해 분류된 토픽 파이 차트

목표
LdaModeling에서 선정한 '생활, 문화' 토픽의 문서를 사용하여 뉴스 기사 건수의 총 양을 파이 차트로 가시화하는 시나리오

기능
토픽들 간의 기사량의 차이가 있을 것이기 때문에 이를 시각화하여 사람들이 어떤 토픽에 관심이 있었는지, 순서를 보면서 코로나 이후 관심사가 주를 이루었는지 확인

예시)



출처 [PieChart with too many slices – amCharts 4 Documentation]

3.6. LDA 토픽 모델링을 통해 분류된 토픽 시계열 분석

목표
LdaModeling에서 선정한 '생활, 문화' 토픽의 문서를 사용하여 뉴스 기사 건수의 변화 추이를 시계열 방식으로 가시화하는 시나리오

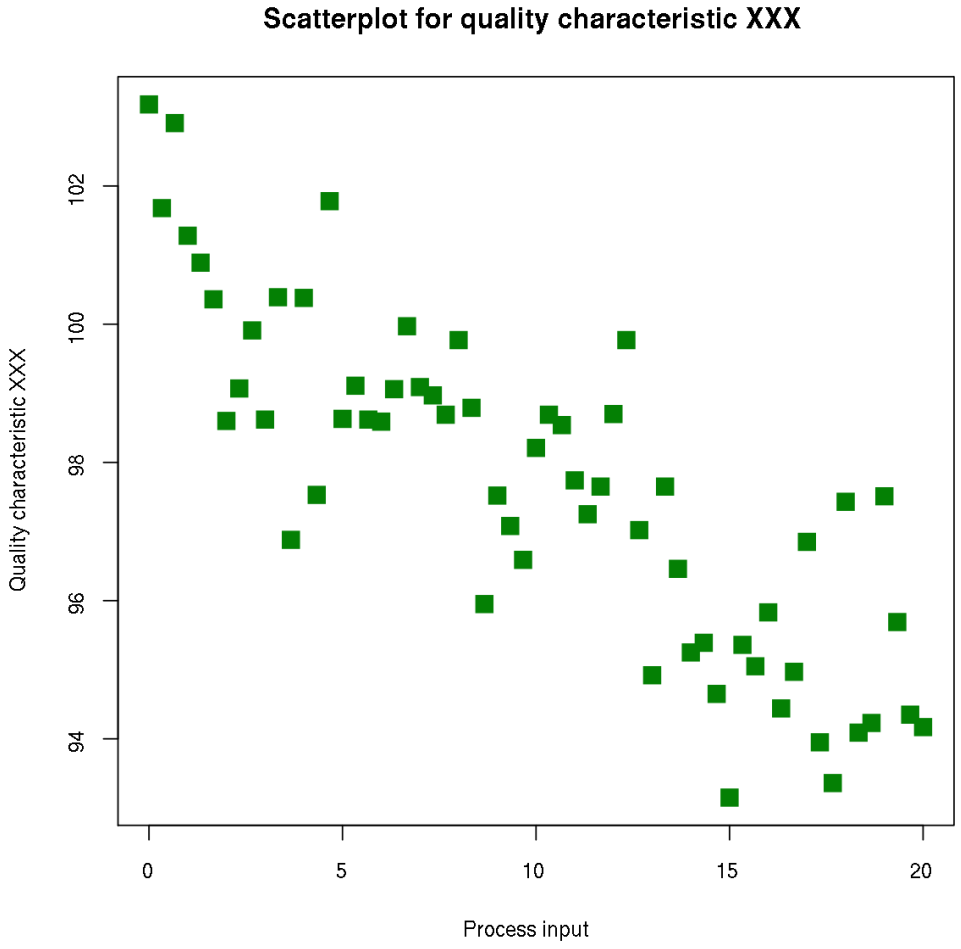
기능
토픽에 따라 특정 사건에 의하여 기사량의 변화가 있을 것이기 때문에 이를 시각화하여 어떤 사건이 일어났을 때 어떤 생활, 문화의 변화가 있었는지를

확인

3.7. Word2Vec을 활용한 연관어 확인 및 시각화

목표
Word2Vec을 이용하여 단어를 벡터화 하고 이를 시각화하여 유사한 단어와 관계의 의미를 분석하는 시나리오

기능
고차원의 단어 벡터를 시각화 하여 유사한 의미의 단어 유사한 맥락에 출현한 단어를 확인



Scatter plot

출처: Scatter Plot

4. 구현상의 고려사항

4.1. 데이터의 수집 및 가공 부분 중 전처리단계에서 추가로 식별되는 불필요한 단어들이 있을 수 있어 정밀한 전처리가 필요함

4.2. 비지도 학습이기에 정확도 측정부분에 어려움이 있을 수 있기에 다양한 방법으로 측정을 해야함

4.3. 시나리오 가시화 부분에서 예상과 다른 결과가 나올 수 있기에 정확도를 높이는 작업이 요구됨

4.4. 개발자가 아닌 사용자도 보고 쉽게 이해할 수 있도록 하고 가시성을 좋게 하여 사용자의 입장을 고려해야함