SW 프로그래밍 팀 프로젝트

코로나19 관련 데이터 분석 및 예측

아주대학교 대학원 지식정보공학과

000000000 OOO
000000000 OOO
202024704 김한호

목차

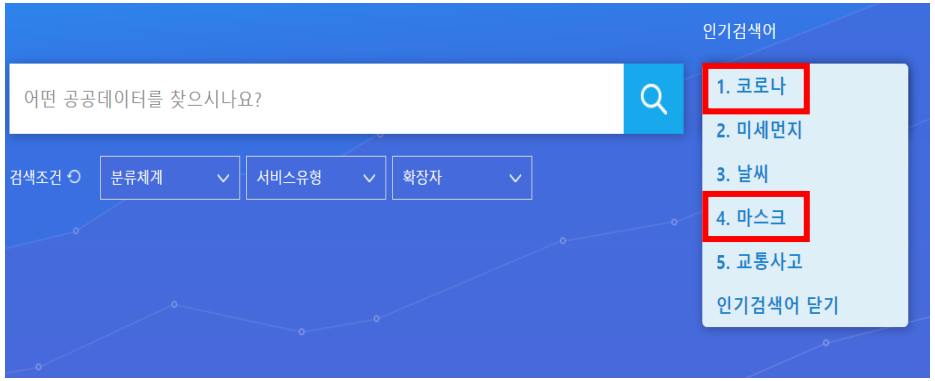
1. 데이터 선정 배경 및 목적

코로나 19 관련 공공 데이터

확진자 수, 지역정보, 확진자 이동 경로, 공적 마스크 판매 정보 등 일반적인 통계나 공지사항

보건의료

코로나바이러스19



공공데이터를 활용한 민간 서비스 개발

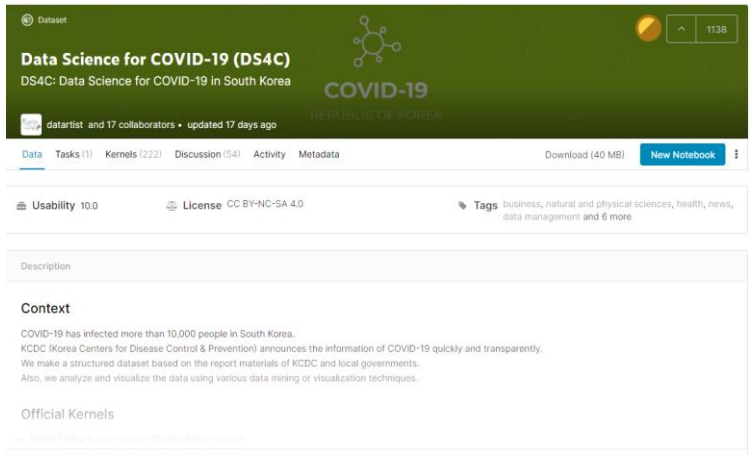
공공데이터를 활용하여 개발된 마스크 알리미, 코로나 19 실시간 상황판 등의
민간 서비스는 국민들에게 도움을 줌

1. 데이터 선정 배경 및 목적

DS4C: Data Science for COVID-19 in South Korea

질병관리본부에서 발표한 코로나19 관련 정보들을 분석 및 모델링하기 적합한 형태로 재가공한 데이터 셋

https://www.kaggle.com/kimjihoo/coronavirusdataset



Context

COVID-19 has infected more than 10,000 people in South Korea.
KCDC (Korea Centers for Disease Control & Prevention) announces the information of COVID-19 quickly and transparently.
We make a structured dataset based on the report materials of KCDC and local governments.
Also, we analyze and visualize the data using various data mining or visualization techniques.

Official Kernels

- [DS4C] What is this dataset (Detailed Description)
- [DS4C] EDA with Floating Population Data
- [DS4C] Who spreads the corona virus?
- [DS4C] time series geospatial EDA using folium.
- [DS4C]Tutorial : All about folium (ing..) + 한국어 설명
- [DS4C] Korea, Wonderland? (Fight against COVID-19)

Update

- We update our dataset every 2 weeks to ensure accuracy and stability of it.
- Last update has been on May 15th, 2020.
  - Up-to-date dataset until 2020-05-14
- Next update is going to be on June 1st, 2020.
  - Up-to-date dataset until 2020-05-31

Acknowledgements

Thanks sincerely to all the members of KCDC and local governments.
Source of data: KCDC (Korea Centers for Disease Control & Prevention)

데이터 분석 및 시각화 ▶ 학습 및 확진자 수 예측 ▶ 예측 성능 평가

2. 주요 활용 데이터 (1) DS4C

[표1] 주요 활용 데이터

구분	데이터명	비고
공개 데이터	Case.csv	Data of COVID-19 infection cases in South Korea (확진자 정보 : 거주지, 집단 감염 여부, 위도,경도 등)
	PatientInfo.csv	Epidemiological data of COVID-19 patients in South Korea (확진자의 역학 조사 정보 : 연령, 성별, 주거 지역, 감염 경로 등)
	PatientRoute.csv	Route data of COVID-19 patients in South Korea (확진자의 방문 경로 정보 : 지역, 방문 방법, 위도, 경도 등)
	Policy.csv	Data of the government policy for COVID-19 in South Korea (코로나19 관련 정부 정책에 대한 정보 : 시행 일자 및 종료 일자)
	Region.csv	Location and statistical data of the regions in South Korea (지역별 정보 : 학교 수, 고령인구 수, 위도, 경도 등)
	SearchTrend.csv	Trend data of the keywords searched in NAVER which is one of the largest portals in South Korea (한국의 가장 큰 포털 중 하나인 네이버의 키워드 트렌드 데이터)

2. 주요 활용 데이터 (1) DS4C

구분	데이터명	비고
공개 데이터	SeoulFloating.csv	Data of floating population in Seoul, South Korea from SK Telecom Big Data Hub (SK텔레콤 빅데이터 허브로부터 제공 받은 서울 유동인구 정보)
	Time.csv	Time series data of COVID-19 status in South Korea (시간 경과에 따른 코로나19 검사 수, 확진자 수, 사망자 수)
	TimeAge.csv	Time series data of COVID-19 status in terms of the age in South Korea (연령대를 기준으로 시간 경과에 따른 코로나19 확진자 수, 사망자 수)
	TimeGender.csv	Time series data of COVID-19 status in terms of gender in South Korea (성별을 기준으로 시간 경과에 따른 코로나19 확진자 수, 사망자 수)
	TimeProvince.csv	Time series data of COVID-19 status in terms of the Province in South Korea (지역을 기준으로 시간 경과에 따른 코로나19 확진자 수, 사망자 수)
	Weather.csv	Data of the weather in the regions of South Korea (지역별 평균 기온, 강수량, 풍량, 습도 등 날씨에 대한 정보)

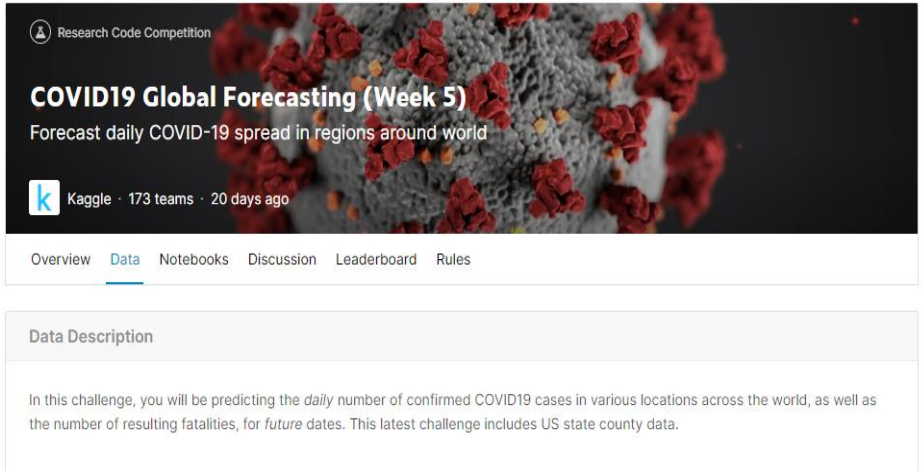
2. 주요 활용 데이터 (2) COVID19 Global Forecasting

COVID19 Global Forecasting (Week 5)

Forecast daily COVID-19 spread in regions around world

전 세계 지역의 코로나19 전파 예측

https://www.kaggle.com/c/covid19-global-forecasting-week-5/data



구분	데이터명	비고
공개 데이터	train.csv	the training data (훈련 데이터)
	test.csv	the dates to predict (테스트 데이터)

3. 데이터 분석 기법

로지스틱 회귀 (Logistic regression)

교차 검증 (Cross Validation)

결정트리 (Decision Trees)

K-평균(Clustering)

스태킹 앙상블 (Stacking)

랜덤 포레스트 (Random Forest)

감사합니다