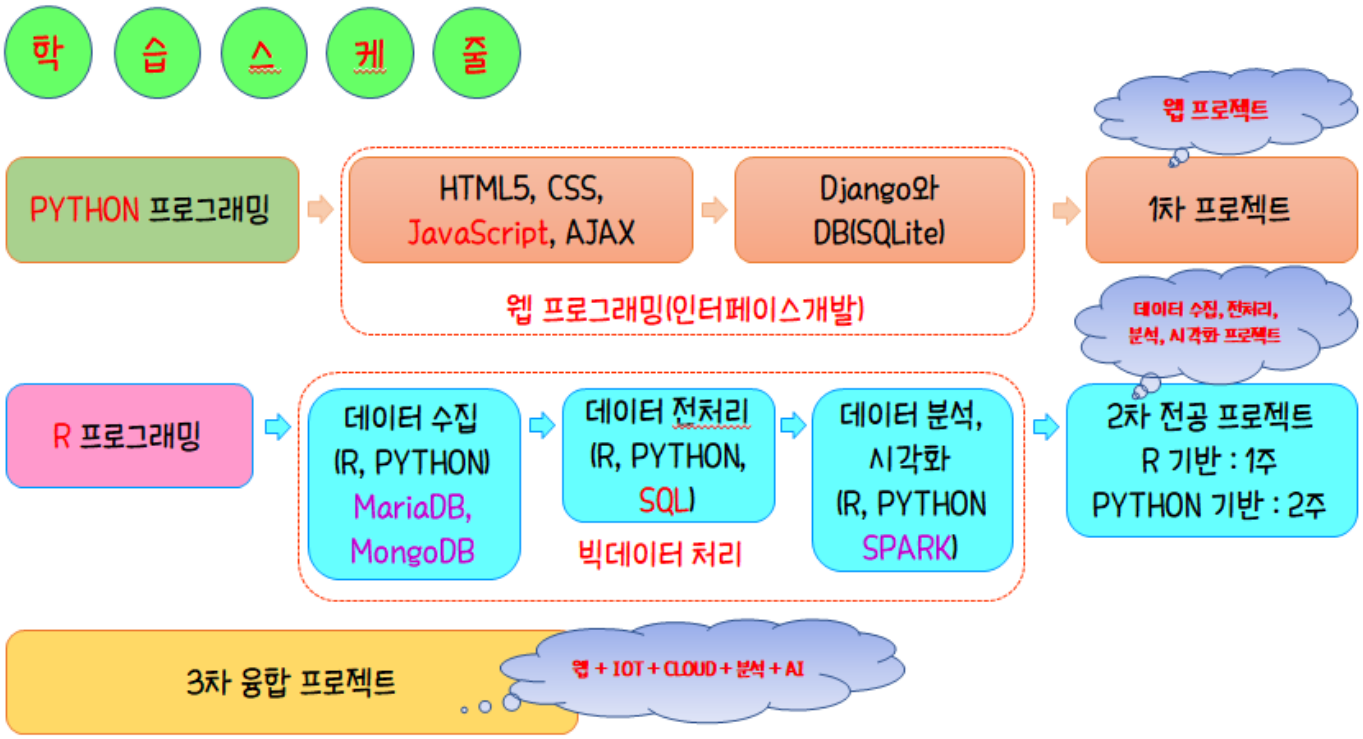
전공과목 순환교육 강의 자료 - 빅데이터 반

김 정현
(unicodaum@hanmail.net)

[빅데이터반의 전공과목 주요 내용]



R 프로그래밍

- R 구문
- 데이터 수집(정적 크롤링, 동적 크롤링, Open API)
- R에서의 MariaDB 연동
- 기본시각화
- dplyr 패키지를 활용한 데이터 전처리
- ggplot2 패키지와 ggmap 그리고 leaflet 을 활용한 고급시각화
- EDA, 통계분석 기초
- 상관분석, 회귀분석

파이썬 언어로 처리하는 데이터 분석

- 데이터 수집(정적 크롤링, 동적 크롤링, Open API)
- 파이썬에서의 AWS MariaDB 연동, AWS MongoDB 연동
- Pandas, Numpy, Matplotlib, Seaborn, Folium, Scikit-learn
- 상관분석, 연관분석, 분류분석, 군집분석, 회귀분석(사이킷런 API를 활용한 머신러닝 기초)
- AWS의 Spark를 활용한 데이터 전처리와 분석 기초

※ 빅데이터반의 전공 수업은 [R 기반의 빅데이터 처리], [Python 기반의 빅데이터 처리] 그리고 [빅데이터 활용 프로젝트] 입니다.

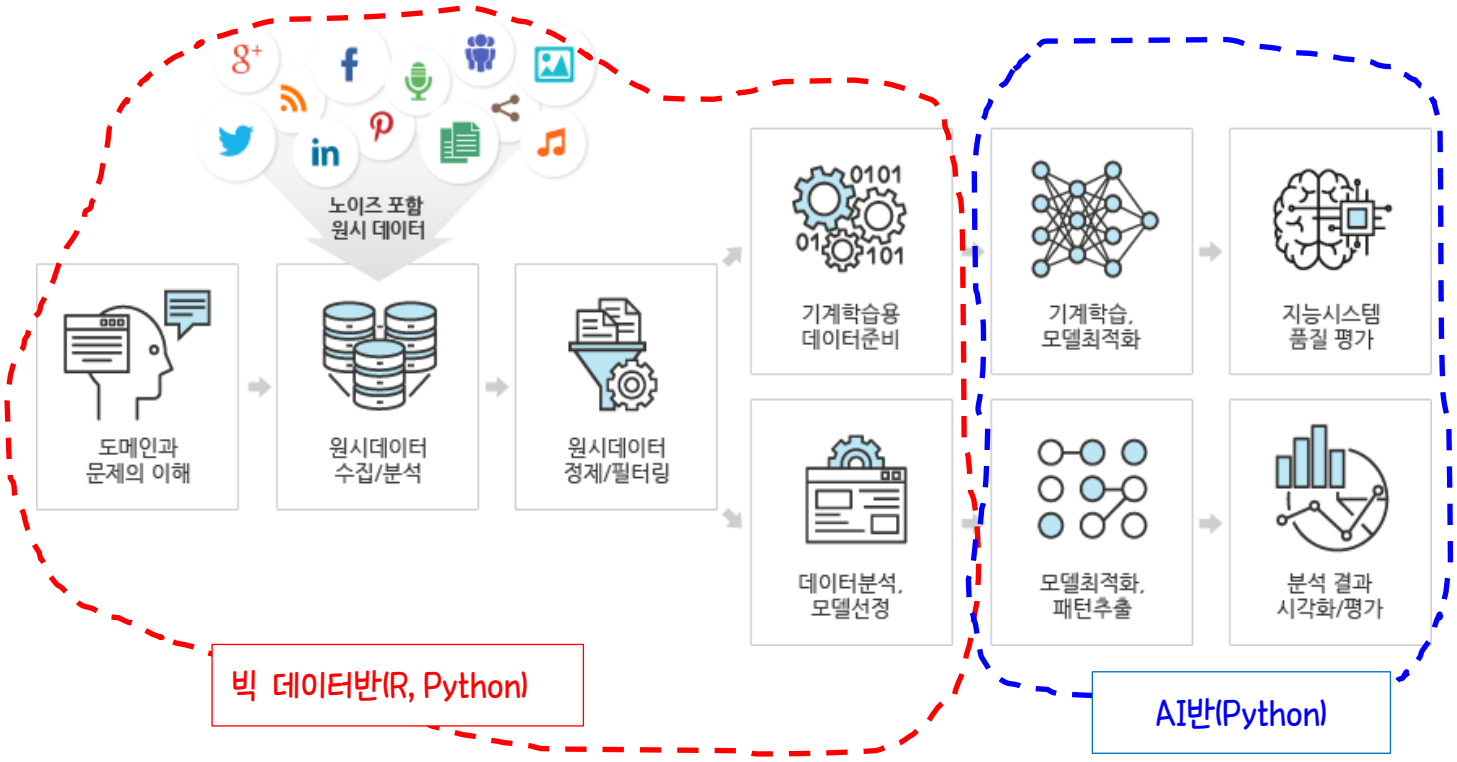
빅데이터 처리

데이터 수집과 저장 | EDA(데이터셋 파악) 전처리(정제&변환) | 데이터 분석과 검증 | 분석 결과 도출 시각화 표현

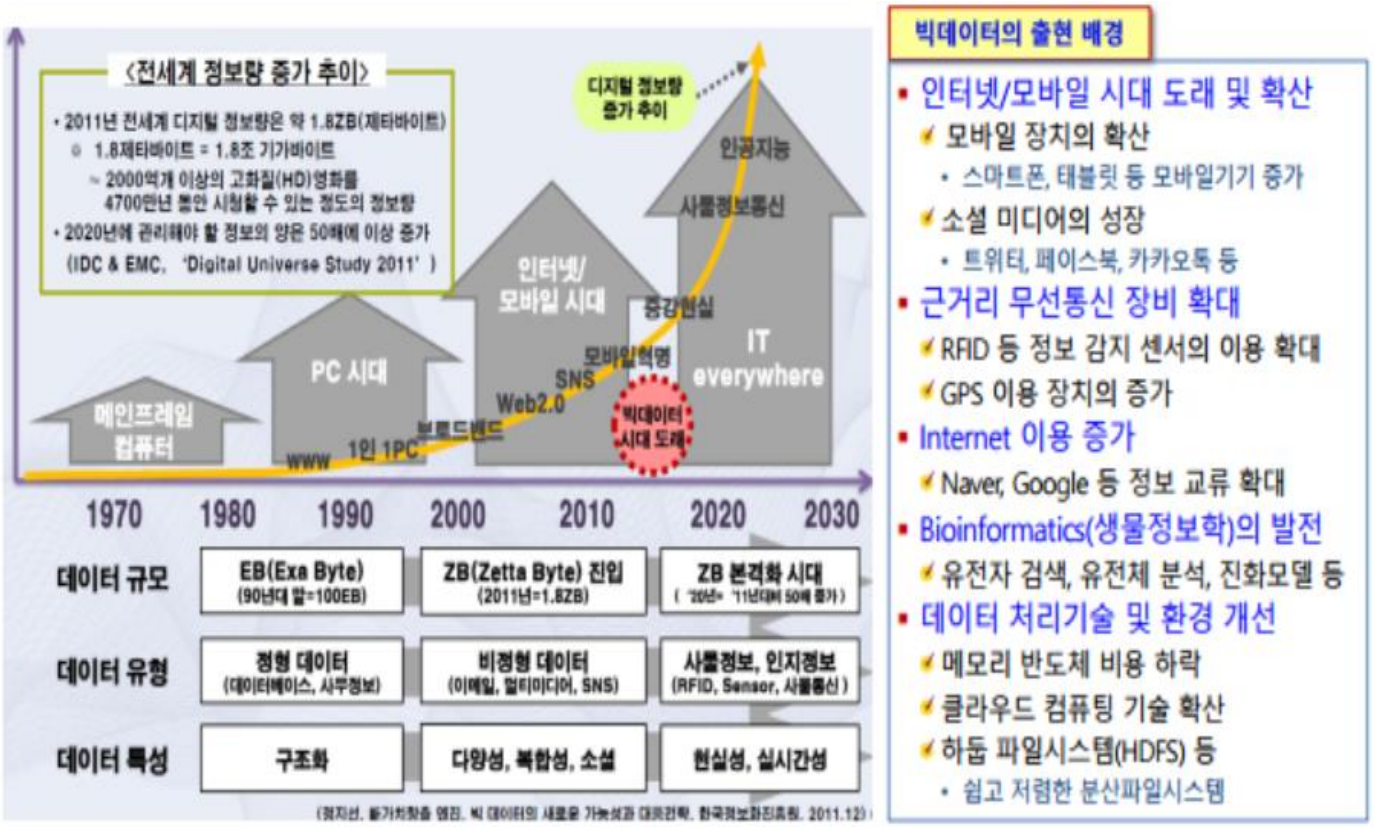
※학습 내용과 기간

2월 22일 ~ 3월 23일 – R 기반의 빅데이터 처리(15일)
– R 기반의 빅데이터 활용 프로젝트(5일)
R구문, 데이터 수집, 데이터 전처리&기본시각화, 텍스트분석,
빅데이터 분석을 위한 통계, 상관&회귀분석, 고급시각화(ggplot, 동적그래프, 지도시각화)

3월 24일 ~ 4월 27일 – Python 기반의 빅데이터 처리, SQL, Spark(16일)
– Python 기반의 빅데이터 활용 프로젝트(9일)
수집, Python의 데이터 분석 라이브러리 – Pandas, Numpy, Scipy, Matplotlib, Seaborn,Folium
텍스트분석, 피쳐엔지니어링, 머신러닝기초, SQL, Spark을 활용한 대용량 데이터 처리



[빅데이터의 등장 배경]



[정형화 정도에 따른 빅데이터의 종류]

정형 데이터	고정된 필드에 저장된 데이터 ex) RDBMS, Spread Sheet
반정형 데이터	메타데이터나 스키마 등을 포함하는 데이터 ex) XML, HTML
비정형 데이터	고정된 필드에 저장되지 않은 데이터 ex) Text, Image, Video, Audio 등

[빅데이터 처리 프로세스]

[빅데이터, AI, 클라우드 그리고 IOT를 활용한 서비스 개발]

[프로젝트 사례(1)]

1) 주제 선정 배경

현재 국내 청각 및 언어 장애인은 약 25만 3천명

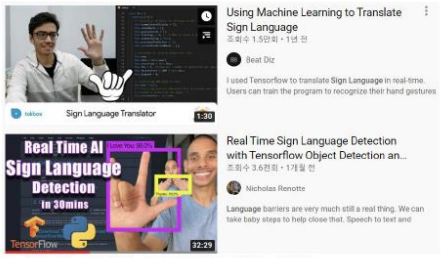
2020년 8월 AI 스피커가 독거노인의 "살려달라"는 요청을 듣고 노인을 구조할 수 있었다는 사례가 있음

그러나, 음성을 낼 수 없는 농인들에게는 기술을 누릴 선택지 조차 주어지지 않음

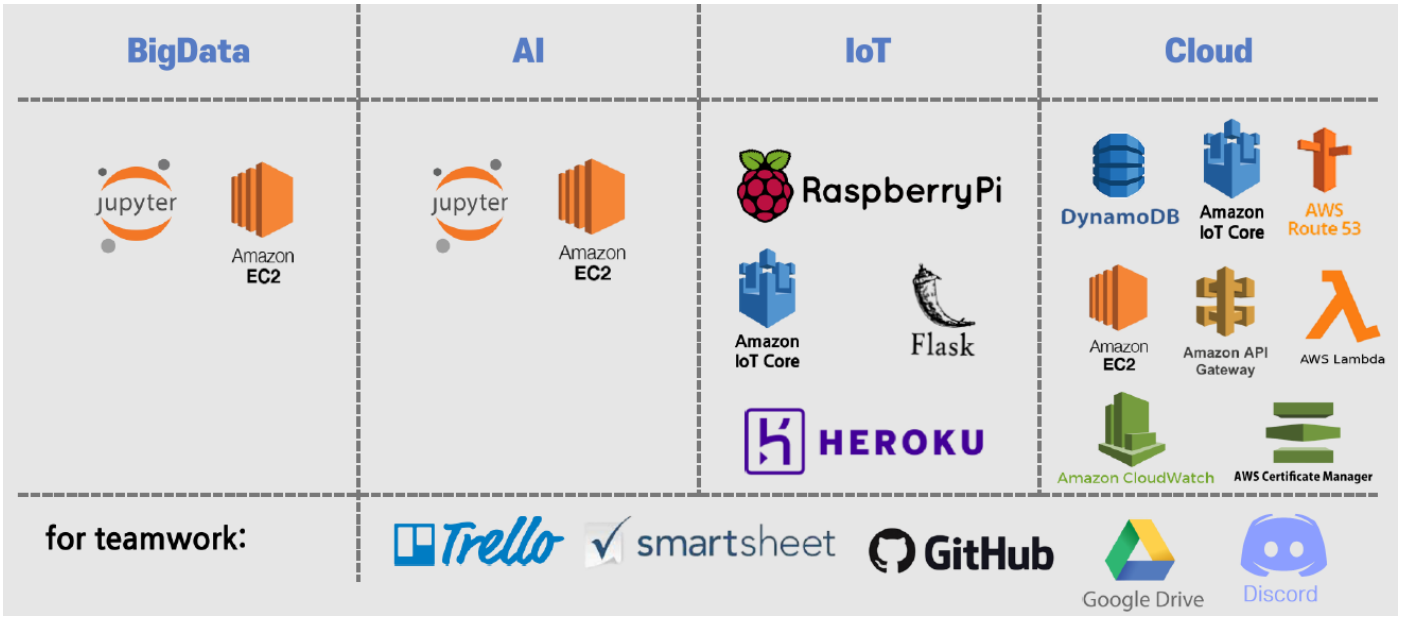
미국의 경우 마이크로 소프트, 아마존이 미국 수화 통역을 시도하였으나 국내는 실용적이지 못하다는 평가

따라서 본 프로젝트는 수어를 인식하여 농인들도 사용할 수 있는 인공지능 비서를 개발하고자 함

[그림 1] 음성인식 AI 스피커

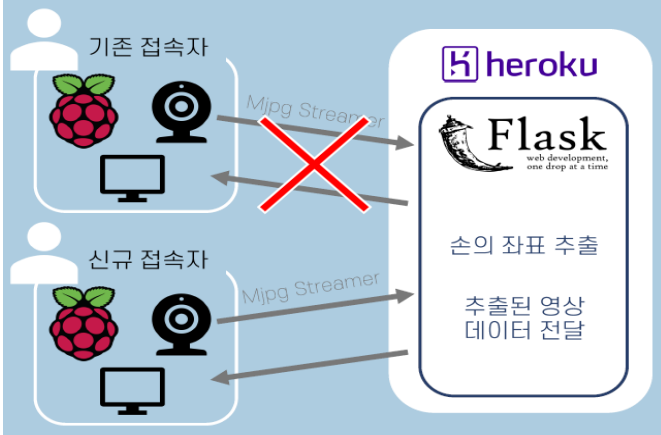


[그림 2] 딥러닝 수어 번역 개발 사례



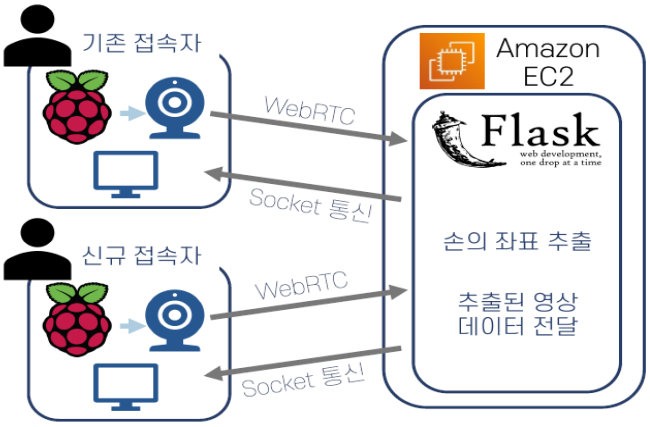
4) 통신 프로토콜

ISSUE



- 기존 접속자의 연결이 끊기는 현상
- Mjpg stream 사용시, 포트포워딩이 필수

SOLUTION



- 소켓통신으로 지속적인 접속유지
- WebRTC를 이용해 웹캠으로도 사용가능
- AWS EC2 환경에서 속도 개선

1) 모델 - 데이터

수어 데이터 확보
- AI Hub 수어 데이터 셋
- 국립국어원 한국 수어 사전
- 직접 촬영

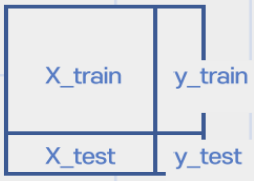
Text 파일로 저장
- x, y 순으로 텍스트 파일로 저장

좌표간 거리 값 추가
- 손 좌표(x, y) 21*2*2개 + 좌표간 유클리드 거리 21*20개(x)
- 숫자 라벨(y)

데이터 수집 → x, y 좌표 → Text 파일 → Train/Test → Input data

MediaPipe library
- 약 300개의 수어 데이터 영상에서 landmark x, y 좌표 값 추출
- EC2 GPU 사용
- 왼손 21*2, 오른손 21*2개 => 총 84개

Train, Test 분리
- Train(80%), Test(20%) 데이터 분리



1) 모델 – 성능 비교

오른손/양손 좌표 추출 정확도 비교

각 좌표 사이 거리를 변수로 추가

거리 변수 정규화 하여 손 크기에 따른 차이 최소화

모델 성능 보완 과정

	SimpleRNN	LSTM	CNN	SVM
정확도	Layer2개 - accuracy: 0.8167 - val_accuracy: 0.2167	Layer2개 - accuracy: 0.9833 - val_accuracy: 0.8333	accuracy: 0.8617 val_accuracy: 0.9709	0.9
	Layer3개 - accuracy: 0.9958 - val_accuracy: 0.2667	Layer3개 - accuracy: 0.9667 - val_accuracy: 0.8500		
etc		최종 선택	•숫자 수어만 학습 •실시간 수어 데이터는 예측 성능 하락	•숫자 수어만 학습

[프로젝트 사례(2)]

사회적 이슈가 되고 있는 배달원 안전문제

라이더/안전에 대한 뉴스기사 162건에 대한 내용 워드 클라우드 분석

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
0	(배달)	(라이더)	0.872727	0.172727	0.172727	0.197917	1.145833
1	(라이더)	(배달)	0.172727	0.872727	0.172727	1.000000	1.145833
2	(라이더)	(안전)	0.172727	0.554545	0.109091	0.631579	1.138913
3	(안전)	(라이더)	0.554545	0.172727	0.109091	0.196721	1.138913
4	(배달)	(민족)	0.872727	0.109091	0.109091	0.125000	1.145833
5	(민족)	(배달)	0.109091	0.872727	0.109091	1.000000	1.145833
6	(사고)	(배달)	0.136364	0.872727	0.136364	1.000000	1.145833
7	(배달)	(사고)	0.872727	0.136364	0.136364	0.156250	1.145833
8	(배달)	(안전)	0.872727	0.554545	0.472727	0.541667	0.976776
9	(안전)	(배달)	0.554545	0.872727	0.472727	0.852459	0.976776

라이더에 대한 뉴스기사 185건에 연관분석

기능1. 음성인식으로 안전한 배달업무 개선

아이템 기능

배달 업무 중 발생하는 **앱 조작을 음성인식을 통해 처리하고, 안전운전을 유도**하는 기능

1. 배달 중 앱 조작을 **터치에서 음성조작으로 변경**해 화면을 보지않아도 일할 수 있게
2. 센서에서 진동 감지, **과하게 흔들릴 때 라이더에게 알려줘** 안전운전과 온전한 음식배달을 유도함.

기능2. 배달원의 안전운전을 파악할 수 있는 데이터 수집 IOT

	id	userID	shopName	omAddre	destinatio	eliveryTi	alertCount	omLatitu	omLongitu	esLatitude	sLongitu	distance	assignDate	status	age	rank
0	1	user1	초록마을(서	서울 강남	서울시 강	20.5	6	37.48894	127.0682	37.50935	127.0389	3.445286	2020-11-09 18:00:00	0	27	100
1	2	user2	신현대상사	서울 강남	서울시 강	22.8	4	37.47916	127.0497	37.50935	127.0389	3.489975	2020-11-09 20:00:00	0	43	100
2	3	user3	포이현대부	서울 강남	서울시 강	18.7	4	37.47916	127.0497	37.50935	127.0389	3.489975	2020-11-09 22:00:00	0	25	100
3	4	user4	포이현대부	서울 강남	서울시 강	19.2	9	37.47916	127.0497	37.50935	127.0389	3.489975	2020-11-10 00:00:00	0	27	95
4	5	user5	주식회사명	서울 강남	서울시 강	20.5	5	37.51576	127.0326	37.50935	127.0389	0.899932	2020-11-10 02:00:00	0	58	85
5	6	user6	빙달(논)	서울 강남	서울시 강	20.2	6	37.51842	127.038	37.50935	127.0389	1.010807	2020-11-10 04:00:00	0	32	85
6	7	user7	세븐브릭	서울 강남	서울시 강	28.9	0	37.51635	127.0379	37.50935	127.0389	0.783285	2020-11-10 06:00:00	0	22	110
7	8	user8	계화기식당	서울 강남	서울시 강	26.4	1	37.5168	127.0381	37.50935	127.0389	0.830566	2020-11-10 08:00:00	0	46	100
8	9	user9	네네치킨	서울 강남	서울시 강	25.5	3	37.51012	127.0256	37.50935	127.0389	1.173108	2020-11-10 10:00:00	0	58	95

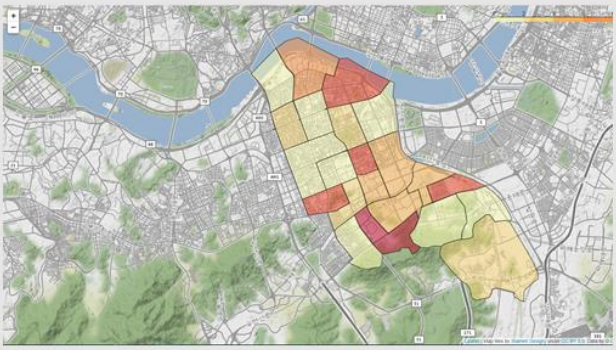
배달원 운행 데이터

보험회사가 사용할 수 있는 배달원의 운전데이터를 수집하고 제공

1. 주요 수집 데이터: **배달 출발지/ 도착지/ 배달시간/ 이상치 발견회수/ 운행거리** 등.
2. 배달원의 안전운행지수를 이용하여 **보험상품 설계에 도움** : 안전지수에 따른 차등 보험료 책정 등
3. 배달원이 안전하게 운전할 수 있게 하고 사고율을 줄여서, 보험금 지급 확률을 낮춤

모니터링 페이지

시간, 지역별 이상치 발생



- 시간, 지역별 위험신호 발생정도를 확인
- 라이더의 운전지역/ 시간에 따라 사고 위험도를 평가할 수 있음.

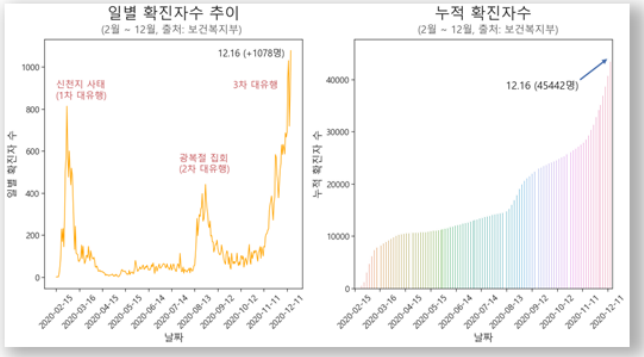
배달원 안전 운행 개선정도 평가

- 배달원들의 위험 운전 정도를 모니터링해 '헬라'가 배달원들의 안전운전에 얼마나 영향을 주는지 확인할 수 있습니다.

[프로젝트 사례(3)]

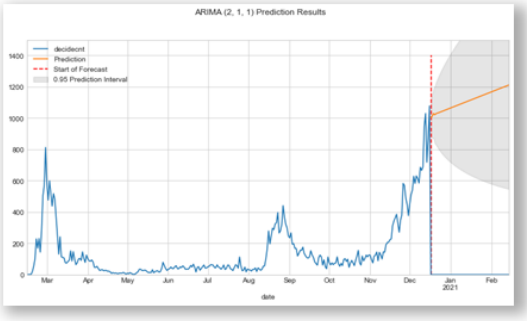
I. 프로젝트 배경

1. 코로나 확진자 현황 & 미래 예측



11월부터 급격히 증가하는 추세 (3차 대유행)

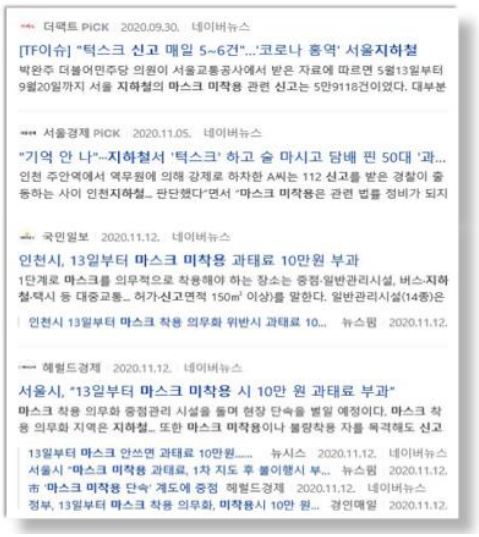
ARIMA 모델을 이용한 시계열 분석



앞으로도 계속 증가할 것으로 예측

I. 프로젝트 배경

2. 마스크 관련 이슈 & 마스크 미착용 신고건수



더팩트 PICK 2020.09.30. 네이버뉴스

[TF이슈] "턱스크 신고 매일 5~6건"...'코로나 흉역' 서울지하철

박완주 더불어민주당 의원이 서울교통공사에서 받은 자료에 따르면 5월13일부터 9월20일까지 서울 지하철의 마스크 미착용 관련 신고는 5만9118건이었다. 대부분

서울경제 PICK 2020.11.05. 네이버뉴스

"기억 안 나"...지하철서 '턱스크' 하고 술 마시고 담배 핀 50대 '과...

인천 주안역에서 역무원에 의해 강제로 하차한 A씨는 112 신고를 받은 경찰이 출동하는 사이 인천지하철... 판단했다"면서 "마스크 미착용은 관련 법률 정비가 되지

국민일보 2020.11.12. 네이버뉴스

인천시, 13일부터 마스크 미착용 과태료 10만원 부과

1단계로 마스크를 의무적으로 착용해야 하는 장소는 중점·일반관리시설, 버스·지하철·택시 등 대중교통... 허가·신고면적 150㎡ 이상)를 말한다. 일반관리시설(14종)은

인천시 13일부터 마스크 착용 의무화 위반시 과태료 10... 뉴스핌 2020.11.12.

헤럴드경제 2020.11.12. 네이버뉴스

서울시, "13일부터 마스크 미착용 시 10만 원 과태료 부과"

마스크 착용 의무화 중점관리 시설을 돌며 현장 단속을 벌일 예정이다. 마스크 착용 의무화 지역은 지하철... 또한 마스크 미착용이나 불량착용 자를 목격해도 신고

13일부터 마스크 안쓰면 과태료 10만원..... 뉴시스 2020.11.12. 네이버뉴스
서울시 "마스크 미착용 과태료, 1차 지도 후 불이행시 부... 뉴스핌 2020.11.12.
市 '마스크 미착용 단속' 계도에 중점 헤럴드경제 2020.11.12. 네이버뉴스
정부, 13일부터 마스크 착용 의무화, 미착용시 10만 원... 경인매일 2020.11.12.

계속되는 코로나 사태로 마스크 착용 의무화 되었지만,
여전히 마스크 미착용자가 많은 것으로 파악됨

I. 프로젝트 목표

프로젝트 목표

대중교통, 회사, 학원, 병원 등의 다중이용시설에 제공할 수 있는
자동으로 마스크 착용 여부와 체온을 검사하여 출입문을 통제하는 시스템을 개발

+

추가 확장안

실내 감염확산 방지를 위한 마스크 불량착용자 다중감시 시스템 개발

II. 개발환경과 수행도구