PORTFOLIO

김도연

안녕하세요.
Data 를 다루는 일에 열정과 자신이 있는 지원자 김도연 입니다!

CONTACT

rla04050@naver.com
010-7544-9816

ABOUT ME

안녕하세요.
정보통계학 기반의 지식과
실무형 프로젝트 경험을 갖추고,
Python / R / SAS / SQL을 활용한
데이터 정제 및 분석 및 모델링에 자신 있는
지원자 김도연입니다.

김도연 / Kim Do-yeon

1997.08.16 (만 24세)

Tel) 010-7544-9816
Email) rla04050@naver.com
Address) 서울특별시 성북구 장위3동
Github) https://github.com/dongnee

Graduation

16.03 - 21.02	동덕여자대학교	정보통계학과	졸업(3.22/4.5)
13.03 - 16.02	용화여자고등학교	인문계	졸업

Education

22.01 - 22.06	K-digital 프로젝트형 빅데이터 분석 서비스 개발(4회차)	멀티캠퍼스	수료

ABOUT ME

Certificate / Language

22.04	SQLD	한국데이터산업진흥원
22.03	ADsP	한국데이터산업진흥원
17.08	SAS Base	SAS
21.08	TOEIC	한국 Toeic 위원회

Hard Skills

구분	Skill
Programming Languages	Python, R, SQL, SAS, HTML, CSS
Server	MySQL, MongoDB
Tooling / DevOps	Github, bash
Environment	Windows, AWS

Awards

22.06	프로젝트형 빅데이터 분석 서비스 개발	개인 최우수상
22.05	빅데이터 분석 프로젝트	최우수상
22.02	인터페이스 개발 프로젝트	최우수상


multicampus
개인 최우수상
훈련과정 : 프로젝트형 빅데이터 분석 서비스 개발 (4회차)
훈련직종 : 빅데이터 분석(20010105)
훈련기간 : 2022.01.03 ~ 2022.06.14 (880시간)
훈련생명 : 김도연

multicampus
최우수상
<빅데이터 활용 프로젝트 >
훈련과정 : 프로젝트형 빅데이터 분석 서비스 개발
훈련직종 : 빅데이터 분석 (20010105)
훈련기간 : 2022.01.03 ~ 2022.06.14 (880시간)
주 제 : 뉴스가 주가에 미치는 영향분석 및 개선방안 도출

multicampus
최우수상
<인터페이스 개발 프로젝트 >
훈련과정 : 프로젝트형 빅데이터 분석 서비스개발
훈련직종 : 빅데이터 분석 (20010105)
훈련기간 : 2022.01.03 ~ 2022.06.14 (880시간)
주 제 : Angel in Star

PROJECT

1

뉴스와 주가의 연관성 분석

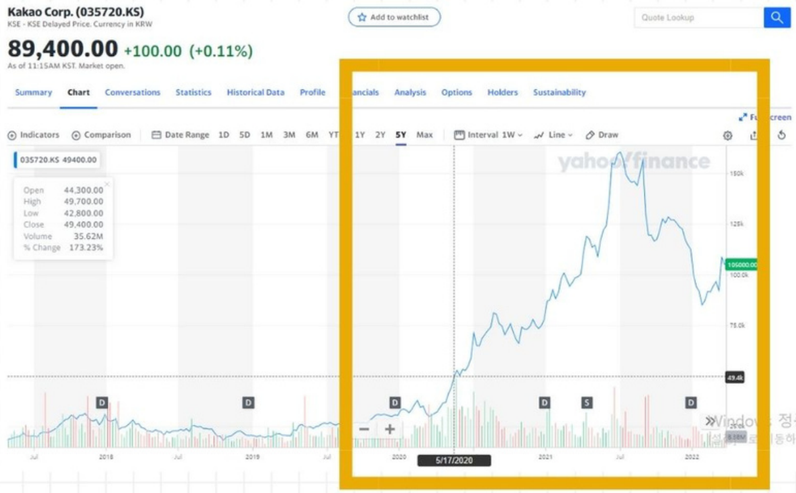
PROJECT 1

프로젝트 이름	뉴스와 주가의 연관성
기간	22.04.11 - 22.05.02 (한 달)
역할	팀장, 뉴스 크롤링 / 시각화 / 감성분석, 주가 예측 모델링
사용 언어	Python, R
목표	주가에 뉴스가 얼마나 영향을 미치는지 분석
결과	예상한 만큼의 연관성은 발견하지 못했지만, 유의미한 상관관계가 있음을 파악 주가 예측 모델로는 LSTM이 가장 설명력이 높음을 발견
개선점	향후 모든 과정을 자동화 시켜, 뉴스만 가지고도 주가를 예측하는 프로그램 생성 가능
수상	최우수상
링크	https://github.com/dongnee/BigOne

프로젝트 개요

주제 선정 배경	• 코로나 이후 커진 주식 시장 • 주식 거래 참고 사항에 뉴스가 큰 비율을 차지하는 것을 확인
타겟 기업 '카카오' 선정 이유	• 코로나 이후 주가 변동폭이 월등히 커진 기업 중 하나 • 대기업 중 하나이며, 크고 작은 이슈가 끊이지 않아 뉴스 감성분석에 용이하다고 판단
사용 데이터	1. 코로나 이후의 주가 데이터 (20.1.1 - 22.4.27) 2. 카카오 재무제표 데이터 3. 기간 동안의 카카오 관련 뉴스 크롤링 데이터 4. 기간 동안의 카카오 관련 SNS(네이버 종목토론실, 네이버 View, 트위터) 크롤링 데이터 5. knu 한국어 감성사전
사용 언어	Python, R
프로젝트 진행 순서	1. 관련 데이터 수집 / 크롤링 진행 2. 감성분석 진행 3. 감성분석 결과를 바탕으로 주가와의 상관관계 분석 (다중회귀분석) 4. 주가 예측 모델링 (LSTM / GRU / ARIMA)

• Yahoo Finance 페이지에서 다운로드



--> 코로나 이후 변동폭이 커진 카카오 주가 그래프

주식 주요 참고사항

- 많은 사람들이 뉴스를 통해 주식의 매도/매수를 결정하는 것으로 볼 수 있음

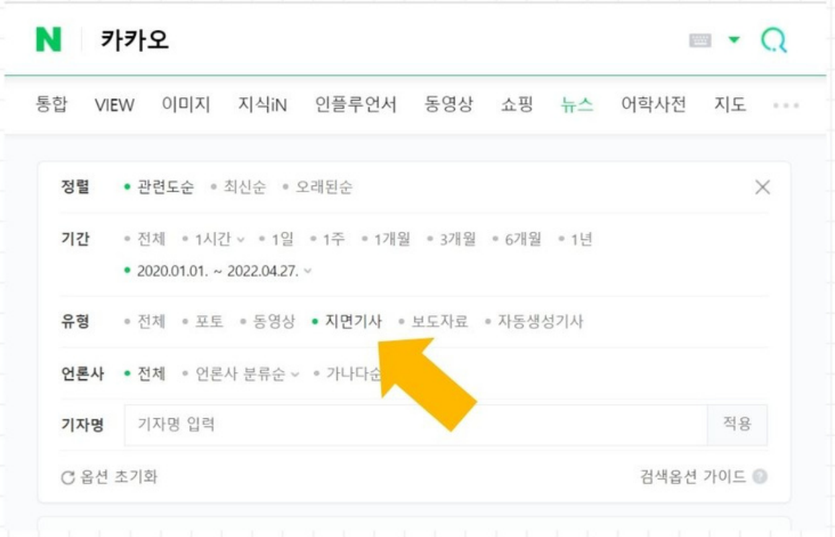
프로젝트 수행

1) 데이터 수집 (크롤링)

--> 프로젝트 기획 당시엔 뉴스 데이터 만을 활용하여 연관성을 분석할 예정이었지만,
sns 데이터도 합하여 분석하였을 때의 연관성도 추가하고자 sns 데이터도 크롤링 진행

뉴스 크롤링

BeautifulSoup 패키지를 이용하여
20.1.1 ~ 22.4.27 기간의 네이버 뉴스 지면기사 크롤링
-> 각 기간을 월 단위로 나누어, 월마다 1000개 기사의 헤드라인과 날짜를 수집한 후
정제 과정을 거쳐 총 15000개 이상의 뉴스 데이터 수집




```
titles = []
dates = []

def news_crawler(news_url):
    for i in range(0,page):
        headers = {'User-Agent' : 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.
        res = requests.get(page_crawler(base_url)[i], headers=headers)
        html = res.text
        bs_obj = bs4.BeautifulSoup(html,'html.parser')
        for j in range(0,10):
            title = bs_obj.findAll('a',{'class':'news_tit'})[j]['title']
            date = bs_obj.findAll('span',{'class':'info'})[j].text
            titles.append(title)
            dates.append(date)

    news_df = pd.DataFrame({'제목':titles,'날짜':dates})
    return news_df
```

--> 크롤링 예시 : 뉴스 크롤링

프로젝트 수행

2) 감성분석 --> 수집한 데이터의 형태소 분석 후, 감성사전 구축 및 일자별 결합, 점수 부여

감성분석

자연어 처리(NLP), 텍스트 마이닝을 통해
문장(텍스트)에 표현된 감성이
긍정, 부정인지 혹은 중립인지 파악하는 것

수집한 자료 형태소 분석 ▶ 감성사전 구축 ▶ 긍정,부정,중립 파악 및 점수 합계 ▶ 긍정,부정,중립 분류

뉴스 데이터 형태소 분석

수집한 뉴스 데이터의 헤드라인을 감성사전 구축에 이용하고
그 후 감성사전을 이용하여 긍정/부정/중립을 분류하기 위하여 형태소 분석

1. 불용어 제거

--> 헤드라인에 포함되어 있는 특수문자들을 제거

```
In [77]: # 불용어 제거

         for i in range(len(title)):
             title[i] = re.sub(r'[\"#$%&\'()++,-./:·;<=>?@\[\]^_\`{|}~\\\]','', title[i])

         news_data['제목'] = title
         news_data
```

Out[77]:

	제목	날짜
0	네이버카카오 작년 연매출 사상최대	2020.01.01.
1	한진네이버카카오두산…국민연금 수익률 '효자'	2020.01.01.
2	백브리핑 돈보다 금배지	2020.01.01.
3	게임위 GO 애플 STOP19금 게임 이중검열 논란	2020.01.01.
4	달린 거리만큼 보험료 내는 '디지털 車보험' 나온다	2020.01.01.
...	...	...
11459	운수화학백화점 코로나 2년 반전스토리	2021.12.31.
11460	삼천피 못지킨 2021 증시…그래도 천스닥은 꿋꿋	2021.12.31.
11461	견제장치 없어 '표적수사' 폭주… 공수처 폐지론 확산	2021.12.31.
11462	연말 기업 신용등급 줄상향	2021.12.31.
11463	동학서학개미 올해 주식 100조원 어치 쓸어 담았다	2021.12.31.

11464 rows × 2 columns

2. 토큰화

--> 불용어 제거된 헤드라인을 여러 형태소 분석 패키지를 제공하는 KoNLPy 라이브러리의 Okt 패키지를 이용하여 형태소 단위로 뽑아냄 (morphs 함수 이용)

```
In [78]: okt_list = []

         for i in range(len(title)):
             okt_list.append(okt.morphs(title[i]))

         print(okt_list)
```

```
[['네이버', '카카오', '작년', '연매출', '사상', '최대'], ['한진', '네이버', '카카오', '두산', '…', '국민', '연금', '수익률', ''', '효자', '''], ['백', '브리핑', '돈', '보다', '금', '배지'], ['게임', '위', 'GO', '애플', 'STOP', '19', '금', '게임', '이중', '검열', '논란'], ['달린', '거리', '만큼', '보험료', '내는', ''', '디지털', '車', '보험', ''', '나온다'], ['새', '로운', '보단', '익숙함', '…', '2020년', '주목', '해야', '할', '새해', '자리', '바꾸는', 'IT', '리더', '들', '혁신', '으로', '경제', '살릴까'], ['불변', '…', '명문', '대', '수시', '합격', '비결', '은', ''', '자율', '동아리', ''', '도', '믿음', '건', 'IT', '뿐'], ['민주당', '이낙연', '복귀', '초읽기', '…', '한국', '인'], ['네이버', '네이버', '페이', '투자', '유지', '로', '테크', '핀', '시장', '선', '굳히기', '헤이크', '핵', '도심', '속', '에서', '즐기는', '자연', '연출', '하다'], ['막막했던', '창업', '패키지', '로', '숨통'], ['씨줄', '날줄', '간편', '후불', '결제', '장', '사로', '이버', '경', '영진', '베트남', '대학', '캠퍼스', '찾아간', '까닭', '은'], ['작년', '은', '카톡', '국민', '10', '명', '중', '7', '명', '이', '사용'], ['배달', '의', '민', '족', '는', '與', '…', '업계', '진입', '장벽', '낮은데', '왜', '그러나'], ['대', '탈출', '…
```

점수 부여

구축한 감정사전 데이터를 이용하여,
뉴스, 종목토론실, View, 트위터 데이터에 각각 긍정/부정/중립에 대한 점수 부여

카카오 계열사 주가 나란히 약세…하루새 시가총액 46조 날아갔다

나란히 --> 긍정/부정에 모두 쓰일 수 있는 단어 -> 0점
시가총액 --> 단순한 경제 용어 -> 0점

약세, 날아갔다 --> 부정단어 -> -1점

→ -2 점

각 문장마다 부여된 긍정/부정/중립 점수

	date	text	score	senti
0	2020.01.01.	네이버카카오 작년 연매출 사상최대	1	1
1	2020.01.01.	한진네이버카카오두산…국민연금 수익률 '효자'	2	1
2	2020.01.01.	백브리핑 돈보다 금배지	1	1
3	2020.01.01.	게임위 GO 애플 STOP19금 게임 이중검열 논란	-2	-1
4	2020.01.01.	달린 거리만큼 보험료 내는 '디지털 車보험' 나온다	0	0
...	...	...	...	...
11459	2021.12.31.	운수화학백화점 코로나 2년 반전스토리	-1	-1
11460	2021.12.31.	삼천피 못지킨 2021 증시…그래도 천스닥은 꿋꿋	-1	-1
11461	2021.12.31.	견제장치 없어 '표적수사' 폭주… 공수처 폐지론 확산	-3	-1
11462	2021.12.31.	연말 기업 신용등급 줄상향	1	1
11463	2021.12.31.	동학서학개미 올해 주식 100조원 어치 쓸어 담았다	0	0

프로젝트 수행

3) 모델링 --> 앞서 구한 감정점수와 주가의 연관성을 다중회귀분석으로 분석 / LSTM, GRU, ARIMA 이용 주가 예측 모델링

모델링

다중회귀분석

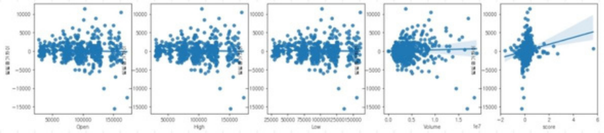
직접 구한 감정점수가
주가에 영향을 미치는지
파악하기 위한 모델링

ARIMA
LSTM
GRU

train data를 이용하여
test 기간의 주가를 예측하는
딥러닝 모델 세가지를 시행 후
비교 분석

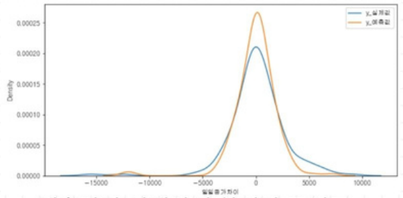
다중회귀분석

독립변수 : '20~'22년도의 시가, 고가, 저가, 거래량, 감성점수



MSE : 2129484.952 , RMSE : 1459.275
결정계수 Variance score : 0.706

실제값과 예측값의 차이는 여전히 크지만,
예측 정확도가 높아짐



예측값과 실제값의 차이가 줄어듦

결론 :
감성점수에 따라 일일종가차이에 유의미한 상관이
있을 것이라 예상했지만, 오히려 감성점수 이외의
다른 독립변수를 추가해야만 예측값이 실제값과 가까워짐을
알 수 있었다.
감성점수를 독립변수로 둘 경우에는 긍정,부정 사전을 보완해야
할 필요가 있다고 생각한다.

LSTM 모델링 과정

```
# 실제주가, 예측주가 데이터프레임 확인
col_name = ['true', 'pred']
true, pred = pd.DataFrame(unscaled_y[start:]), pd.DataFrame(y_predicted[start:])
foo = pd.concat([true, pred], axis = 1)
foo.columns = col_name
foo.index= data1[50:].index
foo
```

Date	true	pred
2020-03-16	29500.0	36882.226562
2020-03-17	29500.0	35668.250000
2020-03-18	28200.0	34357.921875
2020-03-19	26800.0	33149.636719
2020-03-20	29900.0	33476.472656
...	...	...
2022-04-21	93600.0	96206.531250
2022-04-22	92000.0	96068.953125
2022-04-25	89700.0	95154.875000
2022-04-26	90200.0	93723.875000
2022-04-27	88400.0	92718.593750

523 rows × 2 columns

데이터의 처음 50일은 학습용도로 사용했기때문에
2020년 3월16일부터 예측이 시작되었음.
실제주가와 예측주가가 어느정도 오차는 보이지만
그래프상 추세는 따라가는것으로 보임

```
# 상관관계 확인
foo.corr()
```

	true	pred
true	1.00000	0.99562
pred	0.99562	1.00000

```
foo['true+1'] = foo['true'].shift(periods = 1)
foo[['true+1', 'pred']].corr()
```

	true+1	pred
true+1	1.000000	0.998238
pred	0.998238	1.000000

실제 주가와 예측 주가의 상관계수를 확인해보면
0.99562로 매우 유사하다고 볼 수 있지만,
예측주가가 이미 반영된 실제주가를 따라가는 추세는
실제 투자에 도움이 되지 않기때문에 실제주가의 날짜를
1일만 앞당긴다면 더욱 유사도가 높아진다.

PROJECT

2

헤어스타일 추천 스마트미러

PROJECT 2

프로젝트 이름	Our Mirror
팀원	총 8명 (빅데이터, AI, IoT, 클라우드 각 두 명씩)
기간	22.05.03 - 22.06.14 (6주)
목표	사용자의 얼굴형을 분석하여 어울리는 헤어 추천 및 유행 스타일 시각화 자료 띄우기
역할	얼굴형 분석 모델 구현 헤어스타일 추천 시스템 구현 최신 유행 헤어스타일 크롤링 및 시각화
사용 언어	Python(Jupyter Notebook, Colab), R, AWS
결과	시간상 기획했던 모든 것을 구현해내진 못하였지만, 처음 겪어보는 융합 프로젝트를 완성 시켰다는 성취감을 얻었음
개선점	추후 헤어스타일을 합성시켜 3D로 구현해낸다면, 더 수준 높은 프로젝트가 될 것이라 예상
링크	https://github.com/dongnee/OurMirro

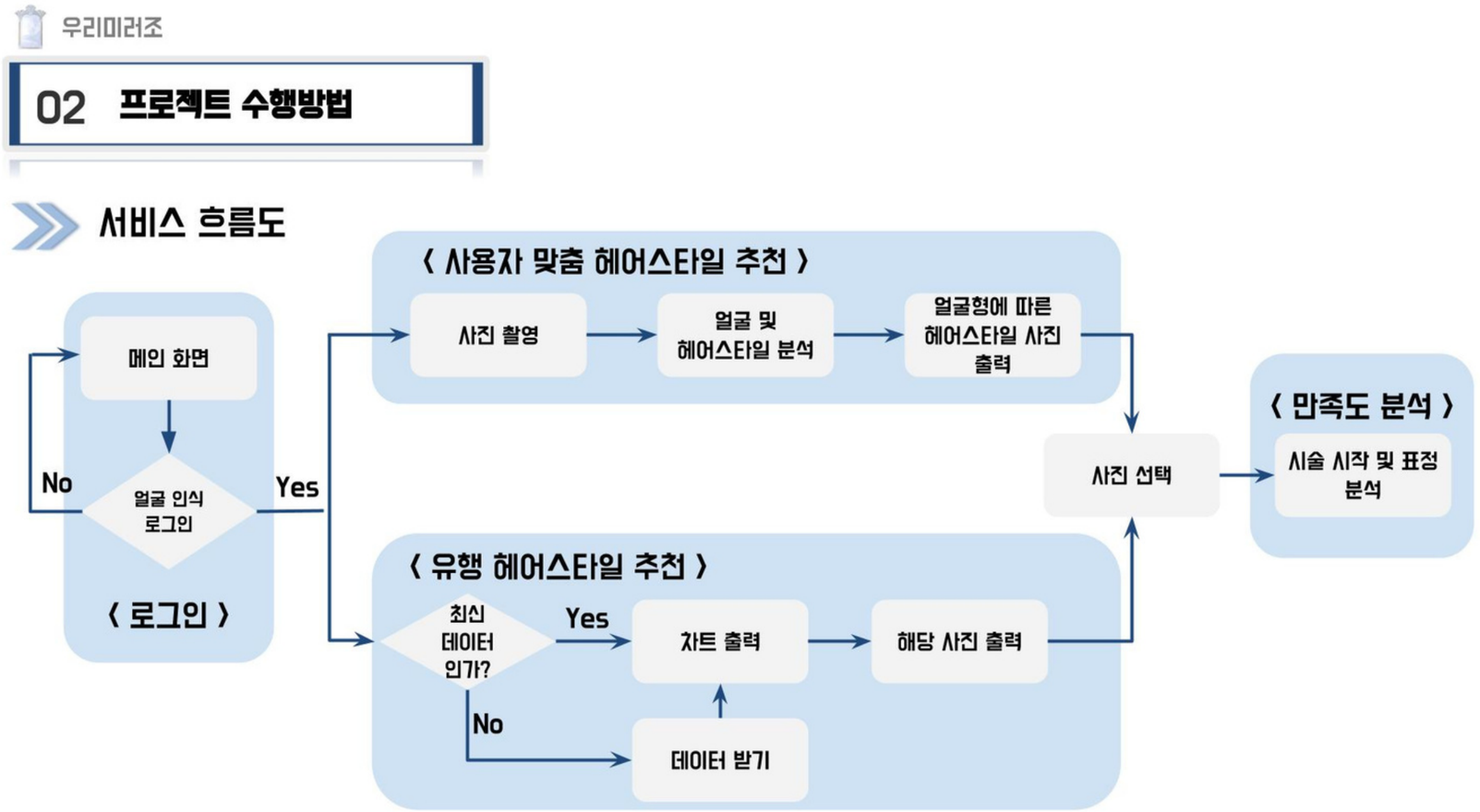
프로젝트 개요

주제 선정 배경	• 국내외 스마트미러 시장의 성장 • 헤어 합성 어플의 누적 다운로드 수를 통해, 시술 전 헤어 합성에 대한 고객들의 니즈 확인 • 스마트미러 도입 헤어샵의 성공 사례
타겟	헤어샵
기대 효과	• 고객은 시술 전 헤어 가상 체험 가능 --> 만족도 증가 • 헤어샵은 신규 고객 유치 및 기존 고객 관리 --> 매출 증가 • 추후 K-beauty 관련 해외 수출 가능
주기능	• 얼굴형 분석 후, 얼굴형에 어울리는 헤어 추천 • 추천하는 헤어 합성 기능 • 유행 헤어스타일 시각화 자료 제공 • 음성인식 기능 제공 • 고객의 표정을 통해 실시간 만족도 확인
사용 데이터 (빅데이터)	1. kaggle : 얼굴형 데이터 5000장 2. 인스타그램 해시태그를 통한 크롤링 데이터
사용 언어 (빅데이터)	Python(tensorflow, keras, pytorch 등), R, AWS

--> 실제 사용 화면

프로젝트 개요

1) 서비스 흐름도



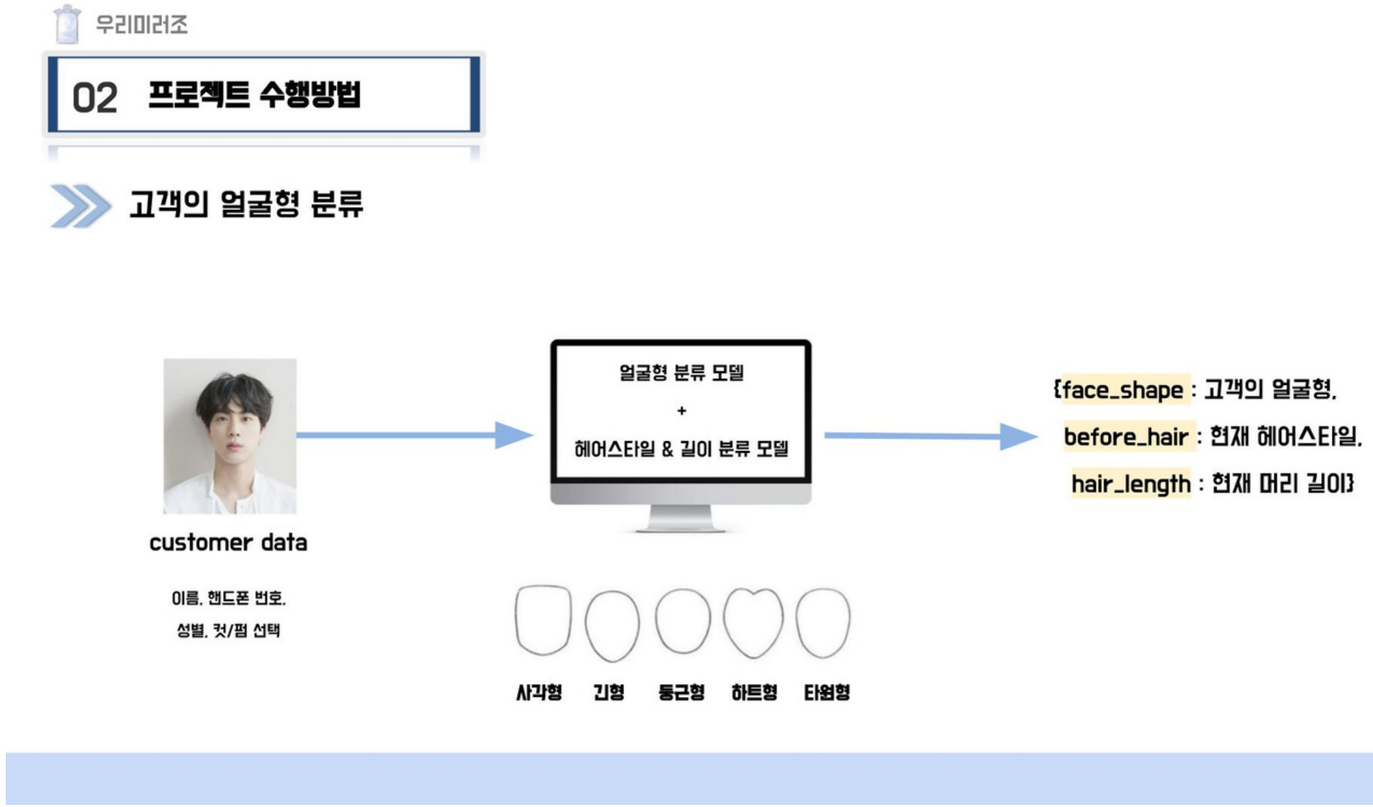
--> 빅데이터 분야에서 얼굴형 분석,
헤어 추천, 유행 헤어스타일 추천 부분 담당

프로젝트 수행

1) 얼굴형 분석 모델 구현

--> 얼굴형 분류 모델

사용 데이터	kaggle - 얼굴형 이미지 5000장
label	사각형, 긴형, 둥근형, 하트형, 타원형
적용 모델	CNN - EfficientNet
Accuracy	0.79
반환값	json 형태로 서버에 전달



프로젝트 수행

2) 헤어스타일 추천

우리미러조

02 프로젝트 수행방법

헤어스타일 추천

- 사용자의 선호도 데이터 수집의 어려움으로 각 얼굴형에 맞는 헤어스타일 4가지를 임의로 정하여 제안하는 방식으로 구현
- 논문이나 sns 상의 정보를 활용해 고객 정보(성별, 얼굴형, 현재 헤어스타일/길이)에 맞게 제안
- 향후 사용자가 선호하는 헤어스타일 데이터를 수집하여 사용자 기반 추천 시스템으로 향상시키고자 함

〈예시 ① - 여성〉

〈예시 ② - 남성〉

→ 개인화 시키기 위한 데이터 부족으로,
임의로 정하여 제안하는 방식으로 진행

→ 추후 사용자의 만족도 데이터가 쌓이면,
사용자 기반 추천 시스템으로 향상시키고자 함

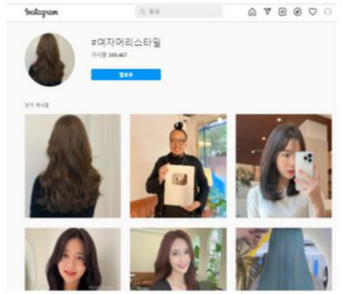
프로젝트 수행

3) 유행 헤어스타일 시각화

우리미러조

02 프로젝트 수행방법

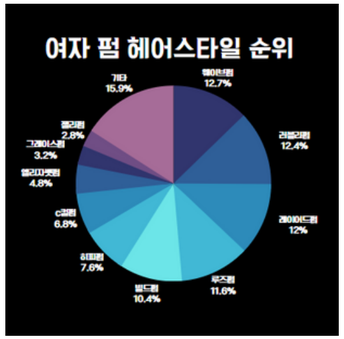
최신 유행 헤어스타일 분석 및 시각화



새 게시물 업로드가 빠른 인스타그램을
해시태그검색으로 데이터 크롤링

text	hashtag
입니다! 오늘 소개해드릴 디자인은 스타일 입니다 미디움기장에서 많은 사랑을 받고 부...	[#자홍롱홍석준, #보더펌, #자홍, #자홍롱, #자홍롱마곡점, #자홍롱마곡점홍석준...
하온헤어 수빈입니다🌽@beauteous_subin .. 꾸준...	[#엘리자벳펌, #사이드뱅, #빌드펌, #그레이스펌, #페미닌펌, #엘리자벳펌, #...
시그니처 시그니처 레이어드컷에 잘어울리는 가벼운 층 많은 레이어...	[#레이어드컷, #레이어드cs컬펌, #피치브라운❤️, #강남미용실, #강남역미용실,...
머리가 다 끊긴 숏컷에서 볼임머리와 볼임머리민트염색으로 여신이 되셨어요💛 실물이 더...	[#볼임머리, #숏컷볼임머리, #볼임머리염색, #선릉볼임머리, #강남볼임머리, #역...
히피펌 -가격- 베이직 트위스트펌(롤드펌 방식의 히피펌) 150,000원 개성있고 ...	[#역곡, #역곡미용실, #부천, #부천미용...

셀레니움을 이용한 게시글의
내용 및 해시태그 수집



헤어스타일별 태그 횟수 시각화
(성별, 컷/펌)

→ 셀레니움 및 BeautifulSoup 패키지를
이용하여 해시태그 수집

→ R을 통해 수집 데이터 정제 후,
원하는 내용만 뽑아내어 파이차트로 시각화

감사합니다!