유럽 국가별 감정 분석을 통한 여행 추천 지수 서비스

4조. 단톡방

CONTENTS

Chapter 01

프로젝트 개요

기획 배경

- 위드 코로나 시대가 코 앞으로 다가온 지금, 규제가 완화되고 폭발할 국제 여행 수요를 대비하여 잠재 여행자들에게 유럽 국가로의 여행에 대한 정량적인 정보를 제공해주고자 함

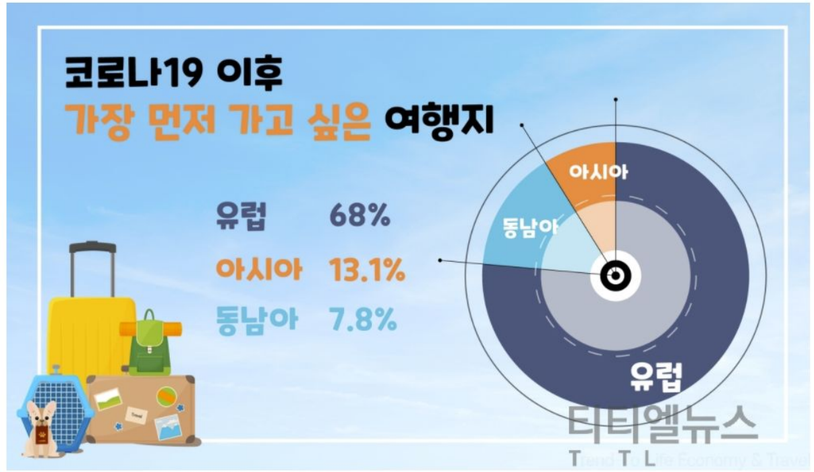
- 타 대륙보다도 유럽의 많은 국가들이 더욱 적극적으로 코로나19 봉쇄를 풀어 여행객에게 문을 열고 있는 현실을 고려하여 유럽으로 선정하게 됨

“한국인 33%, 1년내 해외여행 갈것… 하와이·호주·독일 가장 선호”

유럽 단체여행도 부활…출발일 확정상품에 고객 몰려

입력 2021-11-01 15:08:35 수정 2021.11.01 18:52:27 최성욱 기자

“코로나19 이후 가장 먼저 떠나고 싶은 해외여행지는 유럽, 응답자의 81%는 여행 제한 해제 후 1년 내로 해외여행 떠날 것”

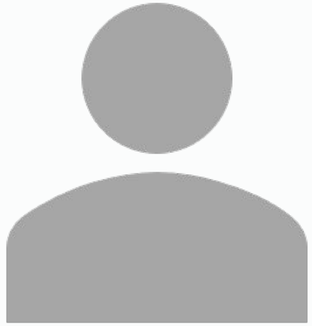


목표

- 여행자 리뷰에 대한 긍부정 점수 산출
- 유럽 각 국가에 해당하는 추천 지수를 시각화

  인터랙티브하게 해당 국가의 세부적인 여행 정보를 보여주는 웹 애플리케이션 개발

구성원 및 역할



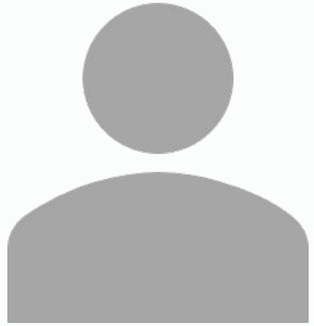
DS

신찬우

EDA, 모델링

검증 및 테스트

PPT, 발표



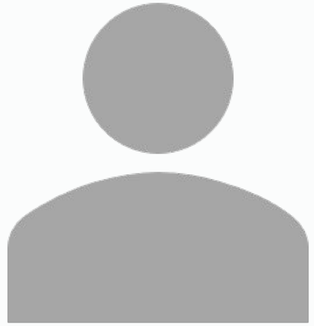
DE

김수민

데이터 수집

웹페이지 구현

PPT, 발표



DE

김정익

데이터 전처리, 수집

웹페이지 구현

기획안, WBS

WBS

WBS(Work Breakdown Structure)

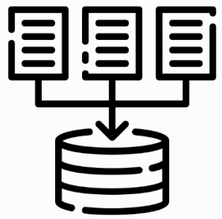
팀명	단톡방	Project Title				"유럽 국가별 감정 분석을 통한 여행 추천 지수 서비스"
구분	공정	완료율	기간	시작 날짜	완료 날짜	담당자
프로젝트 관리	프로젝트 수행계획 수립					
	프로젝트 착수 보고(Kick-Off)		1일	2021-10-21	2021-10-25	조장
	중간보고 및 확인		1일	2021-11-10	2021-11-11	조장
	종료보고 및 확인		1일	2021-12-03	2021-12-03	조장
프로젝트 수행	주제 선정					
	주제 탐색 리서치		2일	2021-10-21	2021-10-22	조원 전체
	기획안 , WBS 작성		4일	2021-10-25	2021-10-28	송광섭 (중도취업)
	주제 관련 도메인 리서치		4일	2021-10-26	2021-10-29	조원 전체
	데이터 수집 및 탐색적 분석		4일	2021-10-26	2021-10-29	DS
	데이터 수집 및 탐색적 분석		4일	2021-10-26	2021-10-29	DE
DE	웹 서비스 구현					김수민, 김정익, 김준길
	데이터 파이프라인 구축		7일	2021-11-01	2021-11-09	
	데이터 수집 및, 적재		7일	2021-11-10	2021-11-18	
	Web Site 구현 및 디자인		8일	2021-11-18	2021-11-29	
DS	데이터 모델링					김진목, 신찬우
	모델링 기법 선정		2일	2021-11-08	2021-11-09	
	데이터 전처리		6일	2021-11-10	2021-11-17	
	모델링 테스트		9일	2021-11-18	2021-11-30	
프로젝트 수행	프로젝트 발표					
	프로젝트 보고서 작성		2일	2021-11-30	2021-12-01	조원 전체
	프로젝트 PPT 제작		2일	2021-12-01	2021-12-02	신찬우,김수민
	프로젝트 발표		1일	2021-12-03	2021-12-03	신찬우,김수민

1일	2일	3일	4일	5일	6일	7일	8일	9일	10일	11일	12일	13일	14일	15일	16일	17일	18일	19일	20일	21일	22일	23일	24일	25일	26일	27일	28일	29일	30일	31일	32일	33일	34일	35일	36일	37일	38일	39일	40일	41일	42일	43일	44일
10.21	10.22	10.23	10.24	10.25	10.26	10.27	10.28	10.29	10.30	10.31	11.01	11.02	11.03	11.04	11.05	11.06	11.07	11.08	11.09	11.10	11.11	11.12	11.13	11.14	11.15	11.16	11.17	11.18	11.19	11.20	11.21	11.22	11.23	11.24	11.25	11.26	11.27	11.28	11.29	11.30	12.01	12.02	12.03
		주말							주말							주말							주말							주말							주말						

Chapter 02

프로세싱

Flow Chart



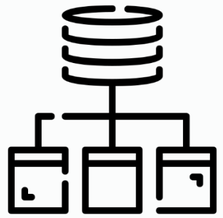
데이터 수집

여행지 리뷰

코로나 현황

전처리 및 EDA

자연어 처리, 분석



모델링

긍부정 점수 산출

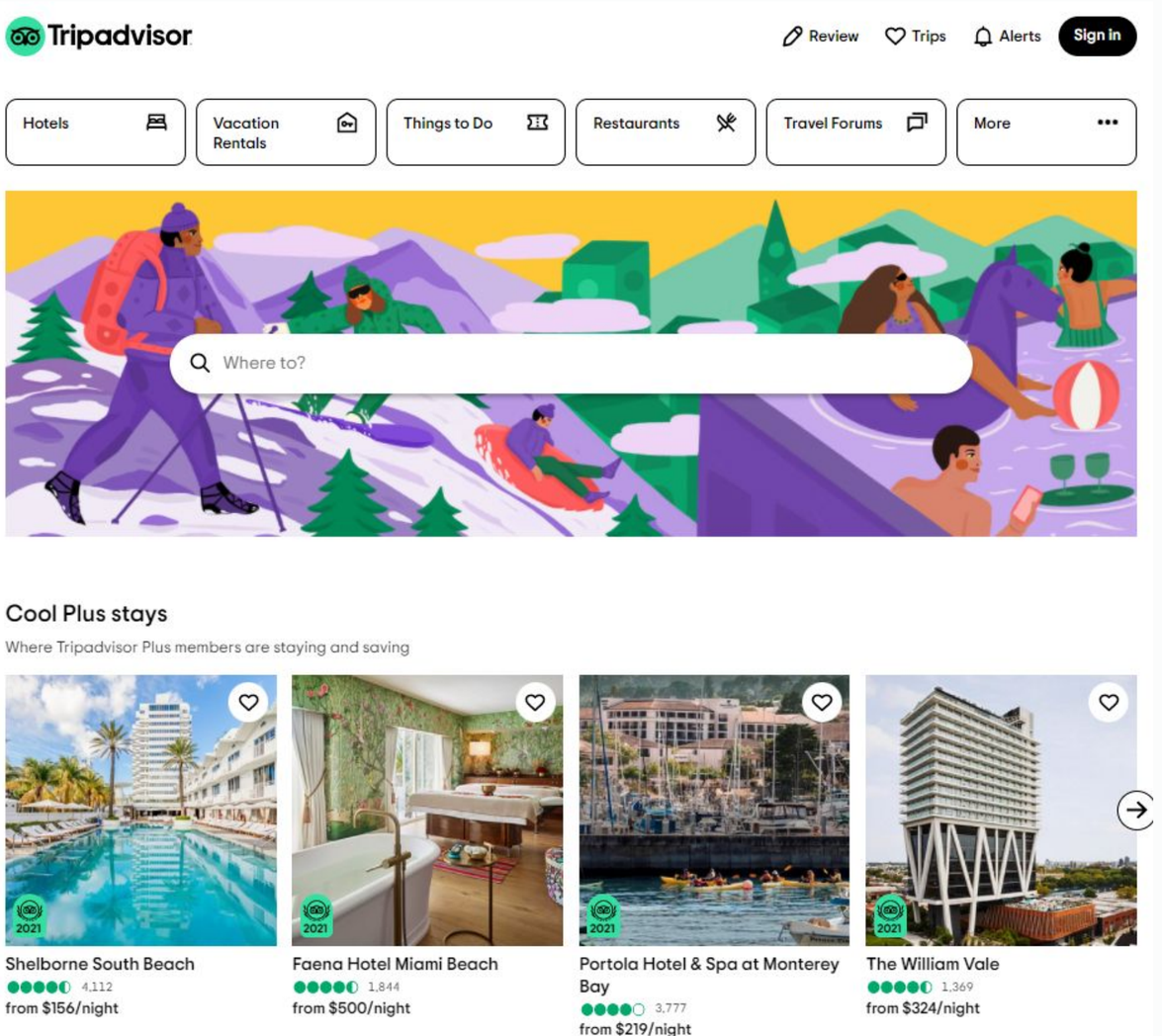
웹 페이지 구현

추천 지수 시각화

데이터 수집

1. 여행지 리뷰
2. 코로나 현황

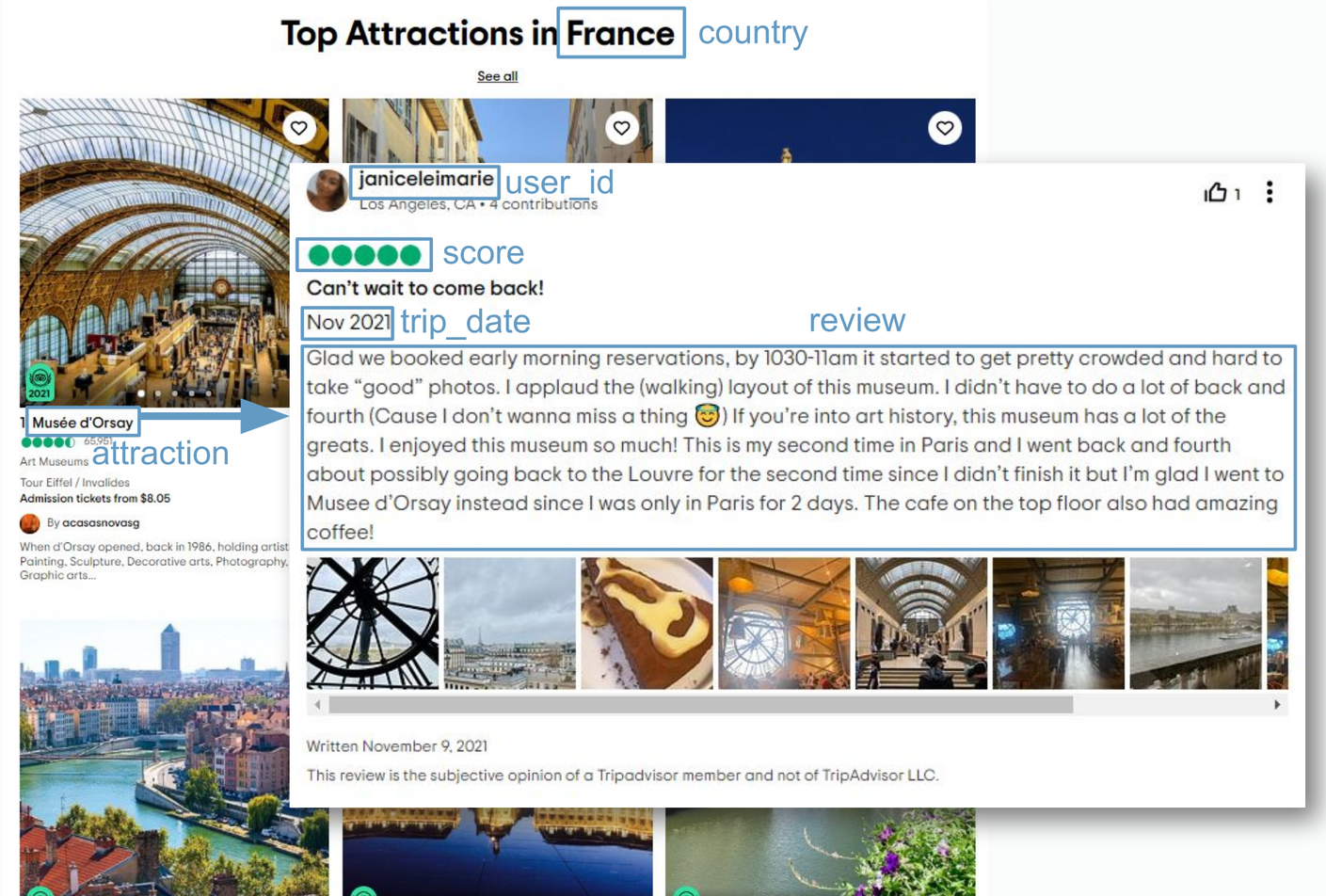
전 세계의 숙박시설, 맛집, 관광명소 등에 관한
여행자들의 리뷰와 의견을 확인할 수 있는
세계 최대 규모의 여행 플랫폼



데이터 수집

1. 여행지 리뷰
2. 코로나 현황

유럽 20개국 선정 후 각 나라별 관광지 10개 리뷰 크롤링 (관광지별 최대 500개)



MongoDB 적재

(유럽 20개국 총 93235개 데이터)


```
_id: ObjectId("61a599fa6233a5532799f3e7")
country: "France"
attraction: "Musée d'Orsay"
user_id: "janiceleimarie"
score: 5
review: "Glad we booked early morning reservations, by 1030-11am it started to ..."
trip_date: 2021-10-31T15:00:00.000+00:00
```

```
_id: ObjectId("61a599fa6233a5532799f3e8")
country: "France"
attraction: "Musée d'Orsay"
user_id: "Rexonaut"
score: 5
review: "Every time I am in Paris, I visit it. I like all their exhibitions wha..."
trip_date: 2021-10-31T15:00:00.000+00:00
```

```
_id: ObjectId("61a599fa6233a5532799f3e9")
country: "France"
attraction: "Musée d'Orsay"
user_id: "gacanuck"
score: 5
review: "I like the Orsay better than the Louvre, which to me is not worth all ..."
trip_date: 2021-10-31T15:00:00.000+00:00
```

```
_id: ObjectId("61a599fa6233a5532799f3ea")
country: "France"
attraction: "Musée d'Orsay"
```

데이터 수집

1. 여행지 리뷰
2. 코로나 현황

확진자 현황

disease-sh API를 호출해 Worldometer의 코로나 데이터 추출 후 MySQL 적재

	country	date	cases	today_cases	cases_per_million
		업데이트날짜	누적확진자	신규확진자	100만명당확진자
0	Austria	2021-11-21	1042571	15297	114851
1	Belgium	2021-11-21	1581500	21502	135640
2	Croatia	2021-11-21	566118	5614	139090
3	Czech	2021-11-21	1980889	22945	184500
4	Denmark	2021-11-21	446676	3795	76742
5	France	2021-11-21	7395222	22678	112950
6	Germany	2021-11-21	5341332	48245	63470
7	Greece	2021-11-21	874812	5944	84503
8	Iceland	2021-11-21	16435	0	47733
9	Ireland	2021-11-21	524783	5959	104660
10	Italy	2021-11-21	4915981	11555	81473
11	Netherlands	2021-11-21	2421643	21794	140896
12	Norway	2021-11-21	241441	1338	44059
13	Poland	2021-11-21	3326505	23455	88028
14	Portugal	2021-11-21	1119784	2333	110264
15	Spain	2021-11-21	5080663	0	108608
16	Sweden	2021-11-21	1188735	0	116697
17	Switzerland	2021-11-21	941216	0	107663
18	Turkey	2021-11-21	8550377	23347	99893
19	UK	2021-11-21	9806034	40941	143405

백신 접종 현황

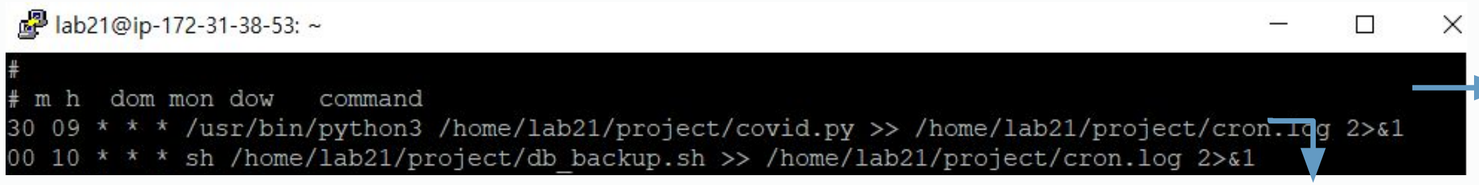
OWID GitHub에서 백신 접종률 데이터 추출 후 MySQL 적재

	country	date	vaccinated	fully_vaccinated	vaccination_rate	fully_vaccination_rate
		업데이트 날짜	1차 접종자수	접종 완료자수	1차 접종률	접종 완료율
0	Austria	2021-11-20	6222265	5822919	68.81	64.39
1	Belgium	2021-11-18	8791141	8647490	75.58	74.34
2	Croatia	2021-11-19	2111804	1889974	51.74	46.30
3	Czech	2021-11-20	6507226	6248339	60.68	58.26
4	Denmark	2021-11-18	4535368	4436345	78.02	76.31
5	France	2021-11-18	51644678	46587994	76.44	68.95
6	Germany	2021-11-19	58580138	56493326	69.82	67.33
7	Greece	2021-11-20	6903597	6512821	66.57	62.80
8	Iceland	2021-11-18	284528	280052	82.87	81.56
9	Ireland	2021-11-18	3841607	3773157	77.10	75.72
10	Italy	2021-11-20	46975851	44039915	77.82	72.95
11	Netherlands	2021-11-14	13171394	12618582	76.70	73.48
12	Norway	2021-11-17	4225203	3785677	77.30	69.26
13	Poland	2021-11-20	20619823	20264284	54.55	53.61
14	Portugal	2021-11-15	9053901	8925907	89.04	87.78
15	Spain	2021-11-18	38209702	37519860	81.74	80.26
16	Sweden	2021-11-19	7289691	6992273	71.75	68.82
17	Switzerland	2021-11-18	5806371	5649764	66.62	64.82
18	Turkey	2021-11-20	56026218	50012340	65.88	58.81
19	UK	2021-11-19	50734556	46129532	74.38	67.63

데이터 수집

1. 여행지 리뷰
2. 코로나 현황

Crontab을 사용해 데이터 자동 업데이트



매일 오전 9시 30분 코로나 확진자 현황, 백신 접종 현황 데이터 수집 후 DB 저장

매일 오전 10시 DB 백업

Crontab 로그 확인

```
Update Covid Info --- 2021-11-25 09:30:06.847969
DB Backup --- 2021:11:25 10:00:1637802001
Update Covid Info --- 2021-11-26 09:30:06.741637
DB Backup --- 2021:11:26 10:00:1637888401
Update Covid Info --- 2021-11-30 09:30:06.394303
DB Backup --- 2021:11:30 10:00:1638234001
```

전처리 및 EDA

VADER: 주로 소셜 미디어의 텍스트에 대한 감성 분석을 제공하기 위한 패키지입니다. 뛰어난 감성 분석 결과를 제공하며, 비교적 빠른 수행 시간을 보장해 대용량 텍스트 데이터에 잘 사용되는 패키지입니다.

Sentiment Analysis
using
VDER in Python

Vader 감정 분석을 통한
sentiment_type 열 생성

Vader_com이 0 초과이면 POSITIVE
Vader_com이 0 이하이면 NEGETIVE

vader_neg	vader_neu	vader_pos	vader_com	sentiment_type
0.000	0.777	0.223	0.9382	POSITIVE
0.000	0.515	0.485	0.8834	POSITIVE
0.051	0.693	0.256	0.8360	POSITIVE
0.000	0.805	0.195	0.8700	POSITIVE
0.000	0.706	0.294	0.9245	POSITIVE

전처리 및 EDA

```
df['sentiment_type'].value_counts()

POSITIVE    77059
NEGATIVE     3532
NEUTRAL      1247
Name: sentiment_type, dtype: int64
```

```
df['sentiment_type'] = df['sentiment_type'].replace(['POSITIVE','NEGATIVE','NEUTRAL'], [1,0,0])
```

문자를 숫자형으로 변경
POSITIVE = 1
NEGATIVE, NEUTRAL = 0

	country	review	sentiment_type
0	Austria	I suppose my friend and I went here as it the ...	1
1	Austria	Lovely place to visit. Good information provid...	1
2	Austria	Interesting former residence of the Habsburg k...	1
3	Austria	Fantastic experience! All was clean and sparkl...	1
4	Austria	Amazing. Absolutely beautiful and full of hist...	1

```
df['sentiment_type'].value_counts()

1    77059
0     4779
Name: sentiment_type, dtype: int64
```

나라와 리뷰, 감정
타입만 가지고 온다.

```
[ ] import nltk
    nltk.download('stopwords')

    [nltk_data] Downloading package stopwords to /root/nltk_data...
    [nltk_data]   Package stopwords is already up-to-date!
    True
```

```
[ ] from nltk.corpus import stopwords
    from nltk.tokenize import RegexpTokenizer

    def text_cleaning(text):
        example = text
        stop_words = set(stopwords.words('english'))

        tokenizer = RegexpTokenizer(r"\w+")
        word_tokens = tokenizer.tokenize(example)

        result = []
        for w in word_tokens:
            if w not in stop_words:
                result.append(w)

        return result
```

NLTK를 이용한
각각
리뷰들에대한
텍스트 크리닝

전처리 및 EDA

TF-IDF(Term Frequency - Inverse Document Frequency)

정보 검색과 텍스트 마이닝에서 이용하는 가중치로, 여러 문서로 이루어진 문서군이 있을 때 어떤 단어가 특정 문서 내에서 얼마나 중요한 것인지를 나타내는 통계적 수치이다.

```
[ ]  from sklearn.feature_extraction.text import CountVectorizer

    vect = CountVectorizer(tokenizer = lambda x : text_cleaning(x))
    bow_vect = vect.fit_transform(Austria['review'].tolist())
    word_list = vect.get_feature_names()
    count_list = bow_vect.toarray().sum(axis=0)
```

1.리뷰의 벡터화
2. Word List
3. Word의 개수

```
from sklearn.feature_extraction.text import TfidfTransformer

tfidf_vectorizer = TfidfTransformer()
tf_idf_vect = tfidf_vectorizer.fit_transform(bow_vect)
```

Bag of Words
벡터에 대해서
TF-IDF변환

모델링

감성분류 = Logistic Regression

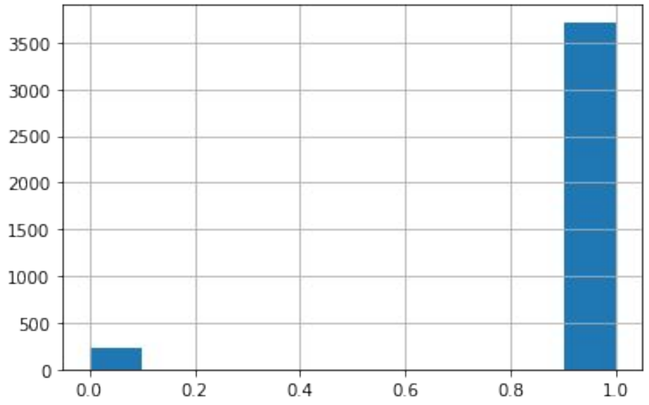
전처리된 리뷰 데이터를 활용하여 감성 분류 예측 모델을 만들어본다.

감성 분류 예측 모델이란, 이용자의 리뷰의 평가 내용을 통해 이리뷰가 긍정적인지, 부정적인지를 예측하여, 이용자의 감성을 파악하는 것이다.

따라서 모델의 X값은 Tripadviosr의 이용자의 리뷰가 되겠고 모델의 Y값은 Vader 감정분석 사전을 이용한 긍/부정 값이 된다.

데이터 셋

	country	review	sentiment_type
4311	Belgium	You can visit the place for free. We visited t...	1
4312	Belgium	It is a large square located in the center of ...	1
4313	Belgium	Simply beautiful and a must see both during th...	1
4314	Belgium	THIS IS A LOVELY SECTION OF BRUSSELS ,the stat...	1
4315	Belgium	Grand Place is a must to visit especially when...	1
...	...	...	...
8253	Belgium	The citadel is on top of a hill and gives a gr...	0
8254	Belgium	Amazing views and interesting tour. If you're ...	1
8255	Belgium	We stopped here on our way through Belgium. We...	1
8256	Belgium	We went to the citadel by car, free parking. W...	1
8257	Belgium	If entered on foot via the town of Dinant, the...	1




```
[16] Belgium['sentiment_type'].value_counts()

     1    3720
     0     227
     Name: sentiment_type, dtype: int64
```

모델링

Training Set / Test 나누기

```
from sklearn.model_selection import train_test_split

x = tf_idf_vect
y = Austria['sentiment_type']
x_train, x_test, y_train, y_test = train_test_split(x,y, test_size = 0.3, random_state = 1)
```

```
[ ] x_train.shape, x_test.shape

    ((3017, 10974), (1294, 10974))
```

모델 학습 및 평가

```
[ ] from sklearn.linear_model import LogisticRegression
    from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score

    lr = LogisticRegression(random_state = 0)
    lr.fit(x_train, y_train)

    y_pred = lr.predict(x_test)
```

```
[ ] print('accuracy: %.2f' % accuracy_score(y_test, y_pred))
    print('precision: %.2f' % precision_score(y_test, y_pred))
    print('recall: %.2f' % recall_score(y_test, y_pred))
    print('F1: %.2f' % f1_score(y_test, y_pred))

    accuracy: 0.94
    precision: 0.94
    recall: 1.00
    F1: 0.97
```

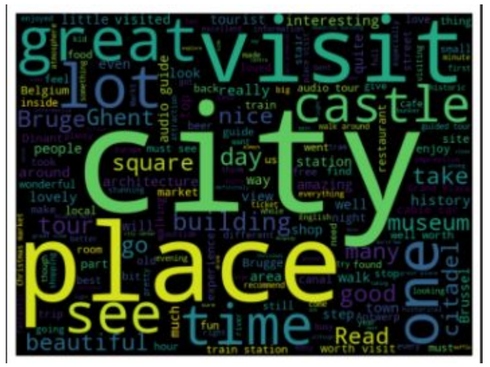
WordCloud

WordCloud

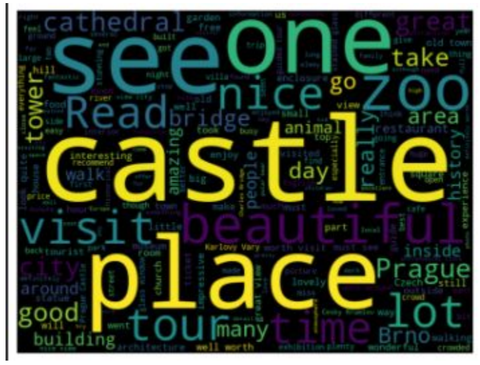
핵심단어를 시각화하는 기법으로 문서의 키워드, 개념 등을 직관적으로 파악할 수 있도록 핵심 단어를 시각적으로 돋보이게 하는 기법이다.

```
from wordcloud import WordCloud

cloud = WordCloud(width=800, height=600).generate(" ".join(Austria['review']))

plt.figure(figsize=(12,5))
plt.imshow(cloud)
plt.axis('off')
```

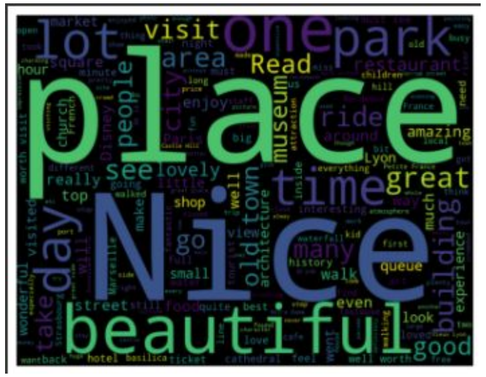
```
(-0.5, 799.5, 599.5, -0.5)
```



벨기에



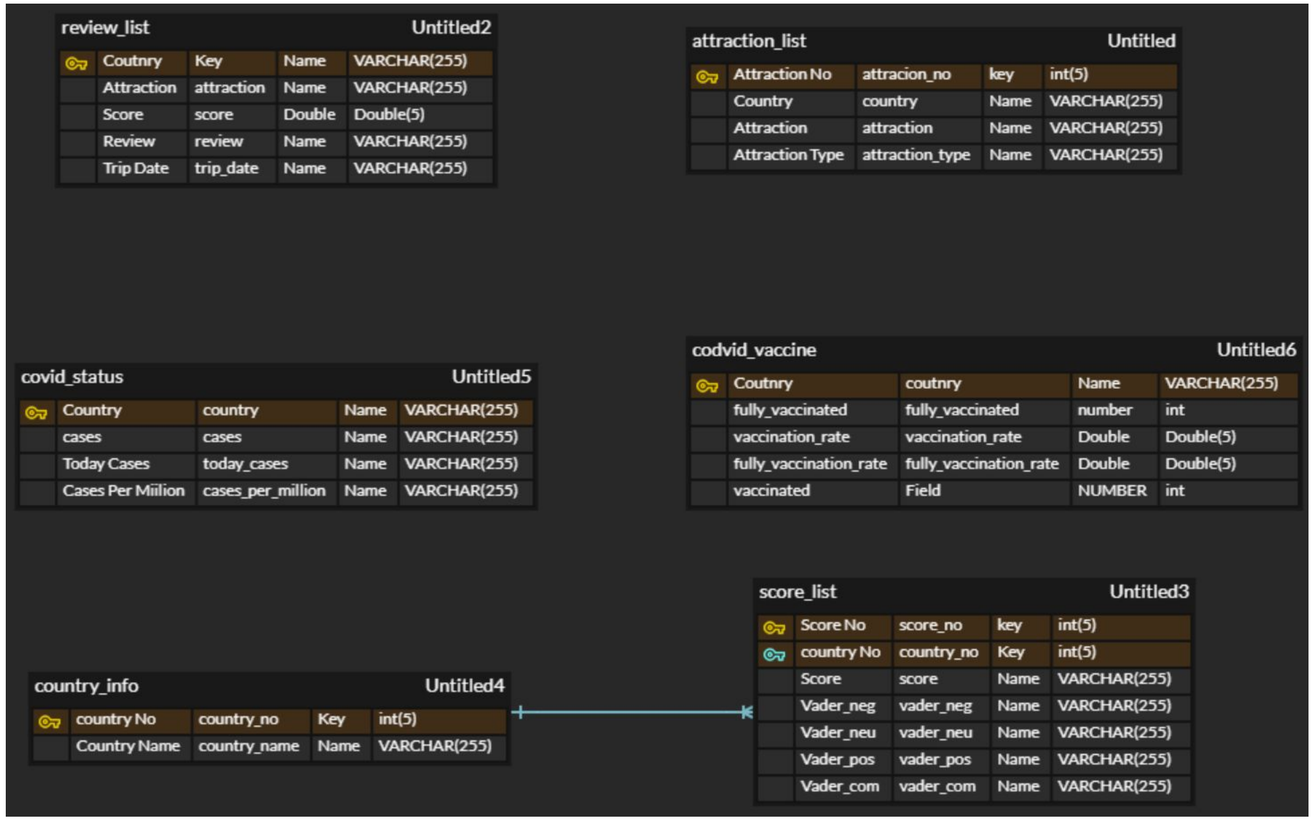
체코



덴마크

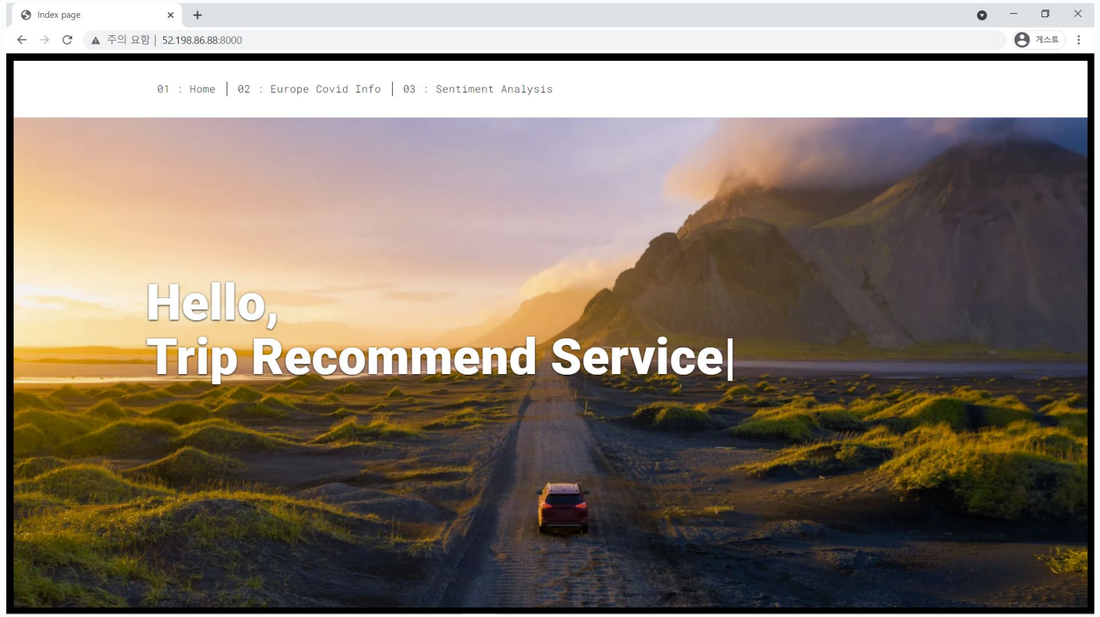
ERD

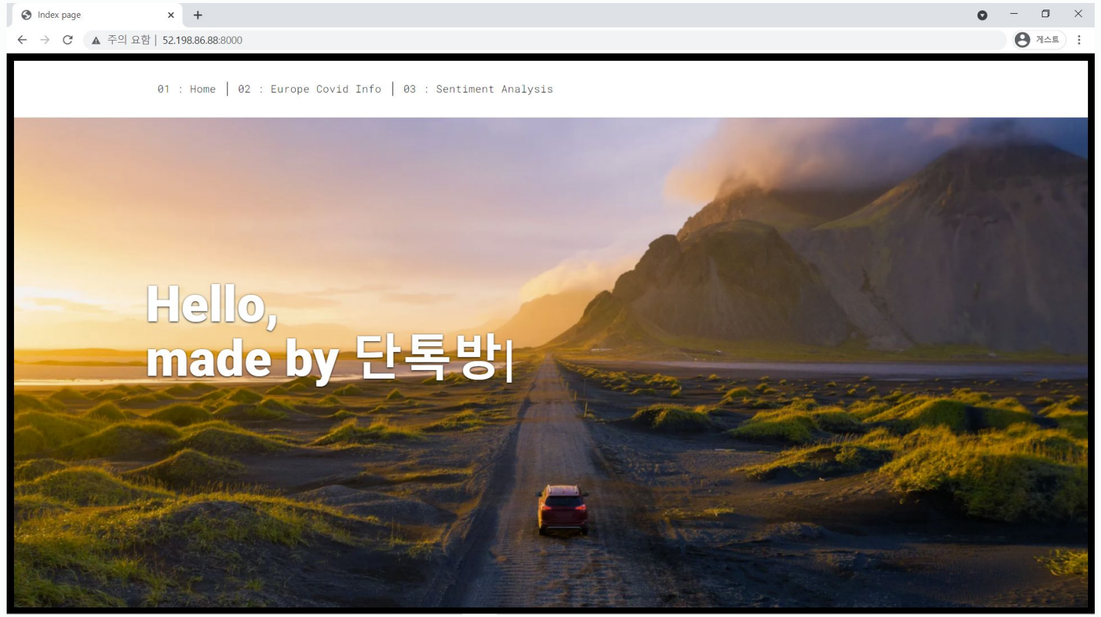
Table 생성 전에 ERD를 활용하여 테이블 설계



웹 페이지 구현

Index

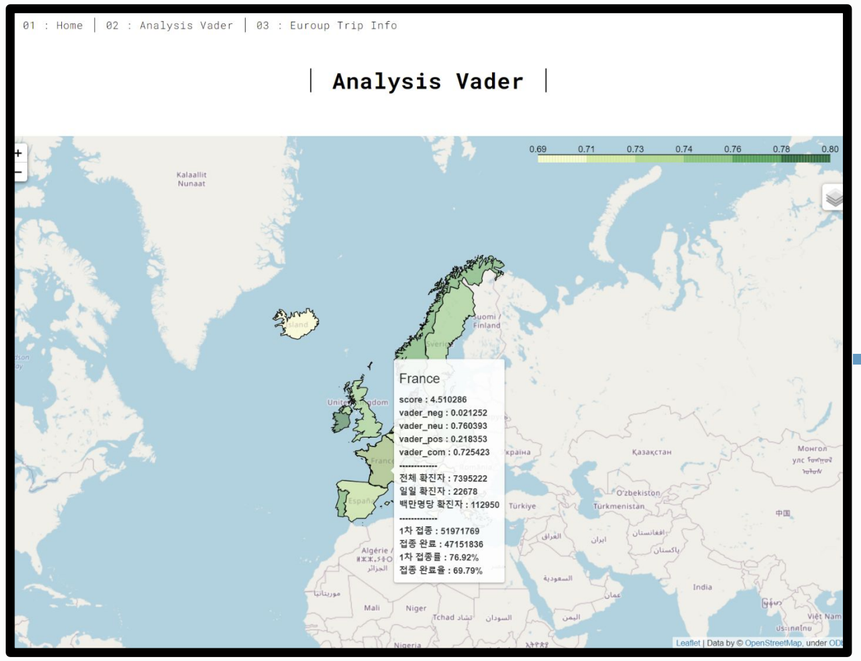


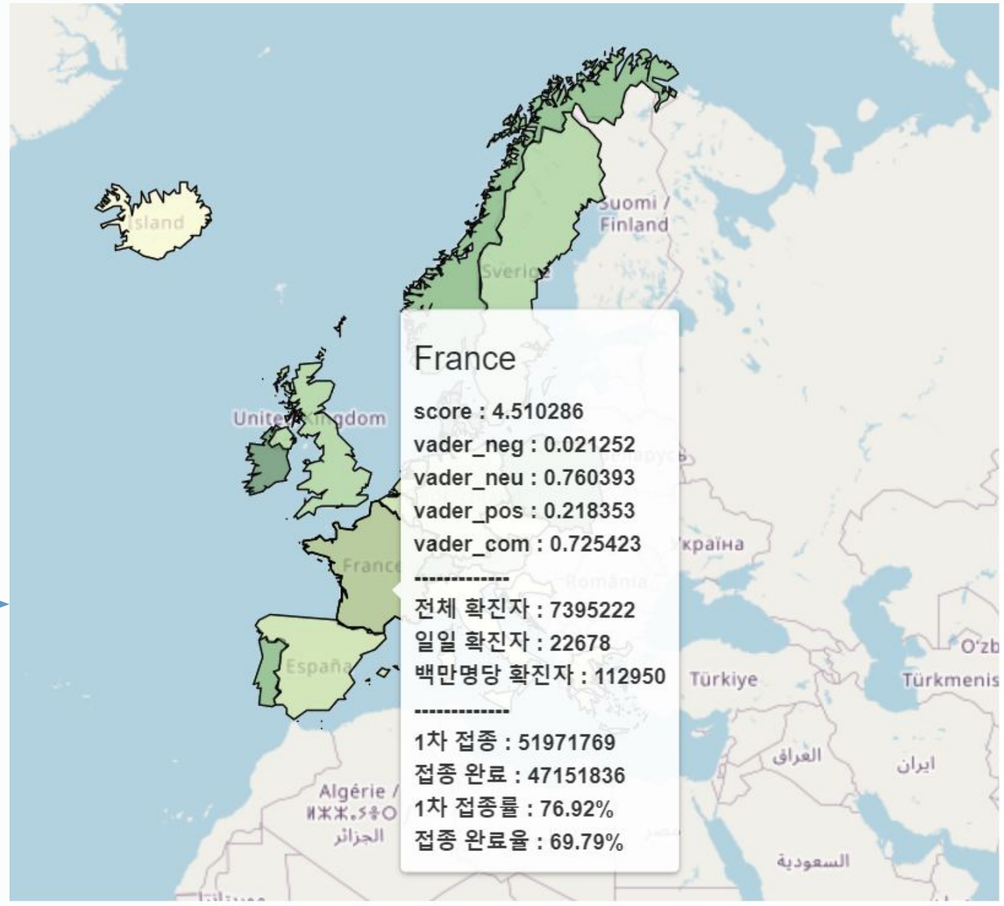


웹 페이지 구현

Analysis Vader

Vader 감정 사전을 이용해 산출된 compound 값을 기준으로 지도 시각화 (진할수록 높은 점수)





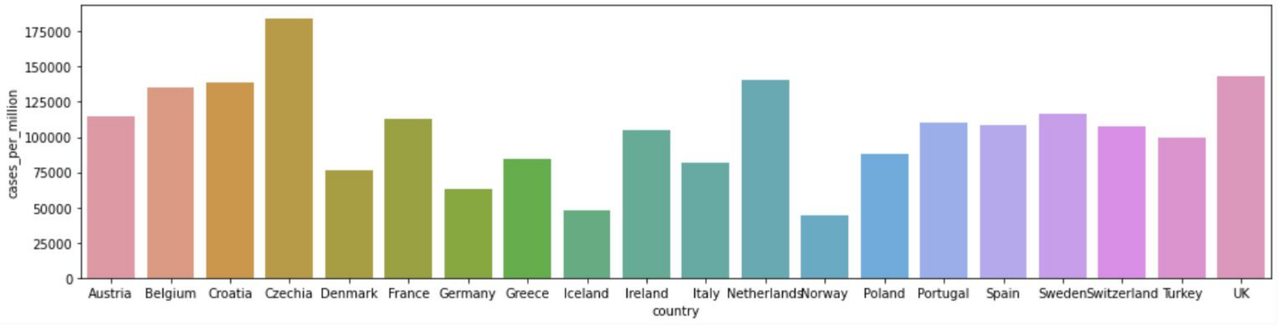
웹 페이지 구현

코로나 확진정보

최근 날짜로 업데이트된 코로나 확진 정보를 table과 bar chart로 확인

코로나 확진 정보 table bar chart

country	date	cases	today_cases	cases_per_million
Austria	2021-12-02	1170362	10367	128907
Belgium	2021-12-02	1766035	16566	151448
Croatia	2021-12-02	613914	5709	150861
Czech	2021-12-02	2172084	21973	202297
Denmark	2021-12-02	492521	5120	84610
France	2021-12-02	7725114	49610	117980
Germany	2021-12-02	5953310	71887	70735
Greece	2021-12-02	945095	6192	91305
Iceland	2021-12-02	18055	161	52428
Ireland	2021-12-02	573905	3790	114419
Italy	2021-12-02	5043620	15085	83592
Netherlands	2021-12-02	2661691	18515	154853
Norway	2021-12-02	271623	3780	49555
Poland	2021-12-02	3569137	29064	94452
Portugal	2021-12-02	1151919	4670	113438
Spain	2021-12-02	5174720	10536	110617
Sweden	2021-12-02	1207498	0	118517
Switzerland	2021-12-02	1020311	4902	116685
Turkey	2021-12-02	8818144	22556	102988
UK	2021-12-02	10276007	48374	150254



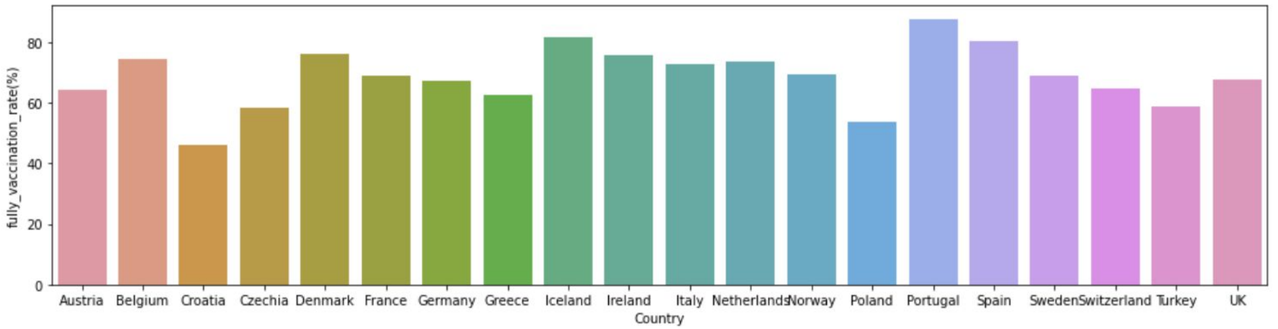
웹 페이지 구현

백신 접종정보

최근 날짜로 업데이트된 백신 접종정보를 table과 bar chart로 확인

코로나 백신 정보 table bar chart

country	date	vaccinated	fully_vaccinated	vaccination_rate	fully_vaccination_rate
Austria	2021-11-30	6347150	5958554	70.19%	65.89%
Belgium	2021-11-29	8832155	8689751	75.93%	74.7%
Croatia	2021-11-29	2180860	1937416	53.43%	47.47%
Czech	2021-11-30	6654320	6366599	62.05%	59.36%
Denmark	2021-11-29	4565137	4454053	78.53%	76.62%
France	2021-11-30	51971769	47151836	76.92%	69.79%
Germany	2021-11-30	59407188	57024545	70.81%	67.97%
Greece	2021-11-30	7047414	6606015	67.95%	63.7%
Iceland	2021-11-30	285721	281192	83.21%	81.89%
Ireland	2021-11-29	3858325	3793477	77.43%	76.13%
Italy	2021-11-30	47226119	44061830	78.23%	72.99%
Norway	2021-11-29	4239290	3826882	77.56%	70.02%
Poland	2021-11-29	20829997	20418316	55.11%	54.02%
Spain	2021-11-29	38339403	37615143	82.02%	80.47%
Sweden	2021-11-30	7580661	7134359	74.61%	70.22%
Switzerland	2021-11-29	5843124	5694544	67.04%	65.34%
Turkey	2021-11-30	56261687	50458355	66.16%	59.33%
UK	2021-11-29	50963718	46367149	74.72%	67.98%



웹 페이지 구현

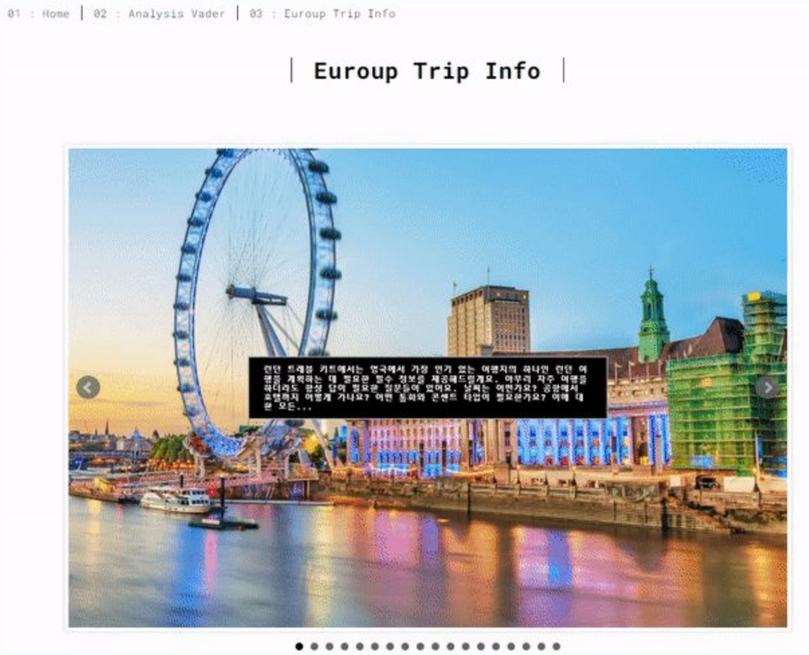
Wordcloud

리뷰데이터의 각 단어별 빈도수를 wordcloud로 변환하여 관심단어 확인

웹 페이지 구현

Euroup Trip Info

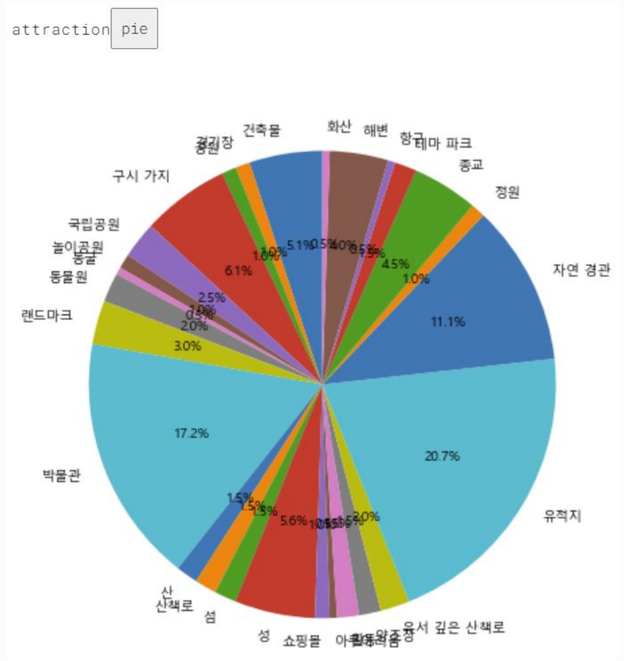
유럽 여행 추천 정보를 슬라이더 형식으로
보여주는 페이지



웹 페이지 구현

attraction type

유럽 전체 관광지의 유형 비율을 pie chart로 확인



Chapter 03

기대효과 및 개선 사항

기대 효과

- 유럽 국가별로 리뷰 분석을 통해 산출한 긍부정 점수와 여행에 필요한 필수 정보인 코로나 현황을 함께 제공해 여행자들의 편의성을 높임

개선 사항

- 리뷰를 분석하는데 있어 전세계 여행지가 아닌 유럽 국가로 한정하고 감정분석을 해보니 대부분의 리뷰가 긍정적인 리뷰였다. 그래서 모델을 만드는데 있어 정확도가 높게만 나오는 상황에 직면하게되었다. 수집한 데이터에 의해 이미 높은 정확도를 보여주는 모델이 나와 추가적인 모델을 구현하는데는 무의미한 상황이었다. 현재 감정분석에 가장 많이 사용하는 BERT를 사용하여 모델을 구현하고 싶었지만 모델을 만드는데 까지는 구현하였지만 활용하는 것에 대해 시간이 부족하여 구현하지 못하였다. 제대로 된 감정 분석을 위해서는 긍 부정 비율이 일정해야하지만 이번 프로젝트에서는 부정 리뷰를 많이 가져오지 못했다. 부정 리뷰를 긍정 리뷰만큼 가져온다면 훨씬 좋은 모델이 나올 것으로 예상된다. 또한 유럽국가로 한정하지않고 전세계 여행지에대한 리뷰에 대한 분석을 진행하였다면 더 좋은 비지니스 모델이 나올 것으로 예상된다.

- 코로나 관련 정보를 실시간으로 수집해 다양하고 구체적인 정보를 사용자에게 제공