웹크롤링을 활용한 데이터 수집 및 시장분석

(숙박 시설 예약 사이트 분석)

JH-lee95

Chapter1. 분석 개요

1. 분석의 목적 및 동기

데이터크롤링을 통해 시장을 분석 및 비교한다.

특정 시장에 진출하기 위해 가장 선행되어야 할 것은 시장조사 및 경쟁사 분석이다.

특정한 퀄리티의 제품을 어떤 가격대에 판매하여야 소비자들의 선택을 받을 수 있는지 판단하려면, 그 시장에 대한 분석이 필수적이다. 따라서 이번 프로젝트에서 시장 분석을 위한 크롤러를 구축하고, 크롤링한 데이터를 바탕으로 시장을 분석한다. 기업 입장에서는 시장 분석을 위해 활용할 수 있으며, 소비자 입장에서는 각자의 목적에 맞는 아이템을 선택하는데 도움이 될 수 있다.

2. 데이터 분석 대상 및 확보 방법

본 프로젝트에서는 모든 데이터 수집 과정을 웹크롤링을 통해 진행하였다. 데이터 크롤링 대상은 시중의 숙박 예약 사이트들이다.

- 대상 사이트 : 에이버앤비(Airbnb), 여기어때, 부킹닷컴(Booking.com)
- 대상 데이터 : 숙박 업소 이름, 가격, 리뷰 개수, 평점

3. 주 사용 라이브러리 및 도구

1) 크롤러 구축 : requests, BeautifulSoup, urllib

크롤러를 구축하기 위해 위의 라이브러리를 들을 이용하였다. Requests 라이브러리를 이용하여 수집 대상에 해당하는 url에 접근 요청을 보낼 수 있다. Urllib 라이브러리는 해당 url의 파라미터들을 손쉽게 확인할 수 있다. BeautifulSoup을 통해 찾고자 하는 정보가 있는 html 태그내의 정보들을 가져올 수 있다.

2) 텍스트데이터 전처리 : KoNLPy(Okt)

3) 시각화 및 데이터핸들링 : matplotlib, seaborn, pandas, wordcloud

Pandas를 통해 크롤링한 데이터를 데이터프레임화 하였다. 이후 matplotlib과 seaborn을 이용하여 "사이트별 가격 분포" 와 같은 시각화를 진행하였다. 또한 wordcloud 라이브러리를 이용하여 특정 숙소에 대한 리뷰를 키워드에 따라 시각화하였다.

4. 데이터 분석 과정

1) 크롤러 구축

각 사이트의 웹페이지 구조에 맞는 크롤러를 구축한다.

2) 크롤링

구축한 크롤러를 바탕으로 데이터를 크롤링하고, 데이터프레임으로 저장한다.

3) 기초 EDA

4) 데이터 시각화 및 분석

5. 주요 이슈

1) "야놀자"와 "호텔스닷컴" 제외

야놀자의 경우 쿼리를 위한 파라미터 설정을 하는데 에러를 겪어 크롤링 대상에서 제외하였다. 호텔스닷컴은 크롤링을 통해 수집되는 데이터의 개수가 17개를 초과하지 않아, 데이터의 개수가 다른 사이트들에 비해 현저히 적어, 분석 대상에서 제외하였다.

2) 텍스트데이터 전처리

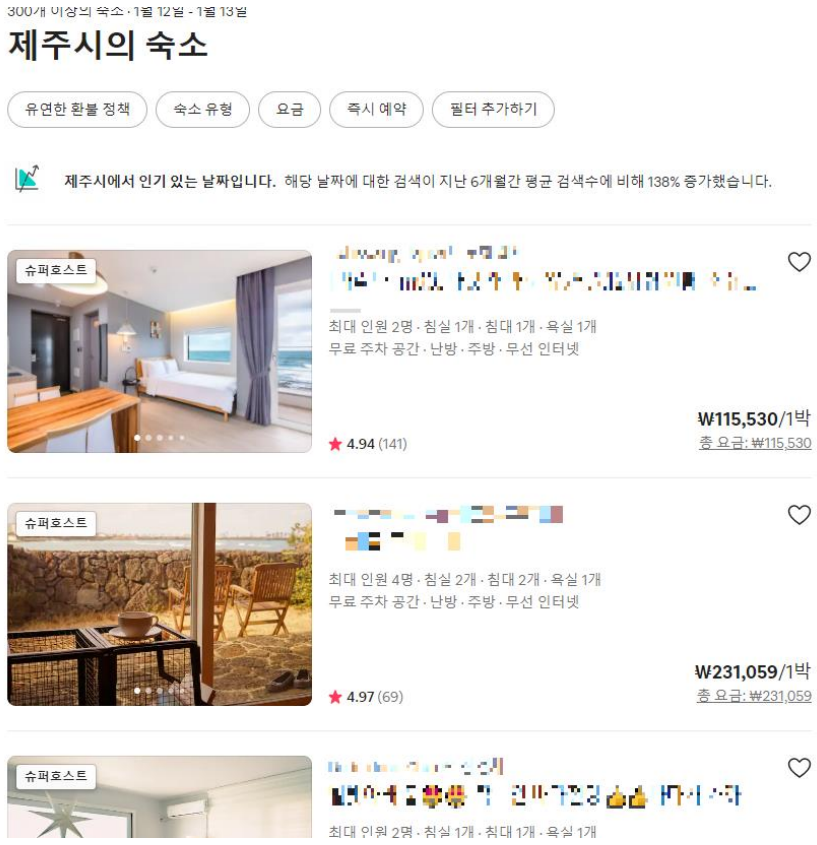
특정 숙박 시설에 대한 리뷰를 워드클라우드로 시각화를 시도하였다. 하지만 리뷰데이터는 텍스트로 이루어진 비정형데이터로써, 이를 분석에 활용될 수 있도록 일정한 형태로 정제할 필요가 있었다. 따라서 KoNLPy 라이브러리를 이용하여 형태소 분석을 통한 전처리를 시행하였다. 또한 빈도수를 기준으로 정렬하여 워드클라우드를 만들었다.

Chapter2. 단계별 분석 진행

1. 크롤러 구축

1) 사이트별 웹페이지 구조 확인

위 사진은 "에어비앤비(Airbnb)" 웹페이지에서 숙소를 검색하기 위해 만들어 놓은 검색기이다. 해당 검색기에서 위치, 체크인(아웃) 날짜 등의 키워드를 설정하고 검색을 시도하면 아래와 같은 화면이 나오게 된다.



각 숙박 시설 별 이름(모자이크 처리), 가격, 평점, 리뷰 개수 등이 나오는 것을 확인할 수 있다.

예를 들어 가장 위에 있는 숙박 시설의 평점은 4.94 이고 리뷰 개수는 141개, 가격은 1박당 115.530원 인 것을 확인할 수 있다. 따라서 이들 수집 대상의 html 구조를 확인하고, 정보를 수집하는 크롤링 코드를 만들어야 한다.

2) 쿼리를 위한 파라미터 설정

크롤링 코드를 작성하기 앞서, 키워드 설정에 따라 웹페이지가 어떻게 바뀌는지 확인할 필요가 있다.

예를 들어, 첫 번째 사진에서의 예시처럼, 이용자가 검색기를 이용하여 다음과 같은 키워드 ["제주시", "1월12일", "1월"13일]를 이용한다면 아래와 같은 주소의 웹페이지로 이동하게 된다.

https://www.airbnb.co.kr/s/%EC%A0%9C%EC%A3%BC%EC%8B%9C/homes?tab_id=home_tab&refinement_paths%5B%5D=%2Fhomes&checkin=2021-01-12&checkout=2021-01-13&source=structured_search_input_header&search_type=unknown&ne_lat=33.816138587050276&ne_lng=127.35836830810547&sw_lat=32.91580970298307&sw_lng=125.9534915991211&zoom=10&search_by_map=true

복잡해 보이지만, 입력한 키워드들이 그대로 나타나는 것을 볼 수 있다. 예를 들어

&checkin=2020-01-12

&checkout=2021-01-13

과 같이 나타난다.

"제주시"라는 키워드가 들어간 파라미터가 인식되지 않는데, https://www.airbnb.co.kr/s/ 이후에 나타는 %EC%A0%9C%EC%A3%BC%EC%8B%9C 가 제주시를 의미한다.

```
base_url="https://www.airbnb.co.kr/s/%s/homes" %self.place
params={'query': [self.place],
 'federated_search_session_id': ['e296e38b-6ec1-4296-9a81-f762cfb2c92a'],
 'source': ['structured_search_input_header'],
 'search_type': ['pagination'],
 'tab_id': ['home_tab'],
 'checkin': [self.checkin],
 'refinement_paths[]': ['/homes'],
 'checkout': [self.checkout],
 'section_offset': ['2'],
 'items_offset': [offset]}
```

따라서 에어비앤비에서 키워드를 통해 검색하기 위한 파라미터를 설정해주기 위해서, 위와 같은 코드를 작성하였다.

3) 크롤링 코드 작성

사이트별 웹페이지 구조를 확인하고, 파라미터를 설정하였으므로, 이제는 크롤링을 위한 코드를 작성하면 된다.

수집하고자 하는 정보가 있는 html 소스들을 따라가면서 그 구조를 가져와야 한다. 예를 들어 각 숙소별 가격 정보를 가져오기 위해, 아래 그림과 같은 코드를 작성하였다.

```
resp=requests.get(base_url,params=params)
soup=bs(resp.text)
item_tags=soup.select("#FMP-target > div > div > div > div")
len(item_tags[0])


#해당 숙소에 대한 상세페이지 url을 가져옴
a_tags=item_tags[0].find_all("a")
url_list=[]
for i in a_tags:
    url_list.append("https://www.airbnb.co.kr%s" %i.get("href"))


#숙소들의 가격정보를 가져옴
price_tags=item_tags[0].find_all("span", class_="_1p7iugi")
price_list=[]
for i in price_tags:
    try:
        price=i.text.split(":")[2]
    except:
        price=i.text.split(":")[1]
    price=price.replace("₩","")
    price=price.replace(",","")
    price_list.append(float(price))
```

즉 requests를 통해 해당 url 접근하고 beautifulsoup을 사용하여 가격정보가 있는 html 태그에 접근하여 정보를 가져왔다.

4) 데이터 전처리

구축한 크롤러를 통해, 데이터를 수집할 시 모든 데이터가 "str" 타입이다. 가격, 평점, 리뷰 개수 등은 데이터분석을 위해 수치형으로 변환해 줄 필요가 있다. 따라서 데이터를 수집하는 과정에서 이를 바로 핸들링 하였다.

예를 들어 "여기어때"의 경우 가격 정보가 "58,000원" 과 같은 형태로 나온다. 따라서 콤마(",") 와 "원"을 제거할 필요가 있었다.

또한 각 사이트 마다 평점의 스케일이 다른 문제가 있었다. 에어비앤비의 경우 5.0만점인 반면에, 부킹닷컴의 경우 10.0만점이었다. 스케일을 동일하게 하기 위하여, 5.0만점에 맞추어 10.0 만점의 스케일을 갖는 사이트들은 나누기 2를 해주었다.

5) Class를 이용한 사이트별 크롤러 통합

위와 같은 과정을 거쳐, 각 사이트별 웹페이지 구조에 맞게 크롤러를 구축하였다. 이후 통합 검색할 수 있도록 Class를 이용한 통합 검색기를 구축하였다. 이를 통해 각 사이트에 등록된 숙박 업소들을 한번에 크롤링할 수 있고, 이를 자동으로 데이터프레임화 할 수 있게 하였다.

이 과정에서 각 사이트별 쿼리에 맞는 파라미터들을 조정할 필요가 있었다.

예를 들어 에어비앤비의 경우 체크인 날짜를 "2021-01-05"와 같이 한번에 입력하는 반면에 부킹닷컴은 "년", "월", "일"을 각각 따로 입력해야 한다. 따라서 이를 조정해주었다.

2. 크롤링

본 단계에서는 구축한 크롤러를 통해, 실제로 크롤링을 진행하고 데이터프레임을 형성한다.

- 장소 : 제주도
- Checkin : 2021-01-07
- Checkout: 2021-01-10

위와 같은 키워드를 이용하여 데이터를 수집한 결과, 아래와 같은 데이터프레임을 얻을 수 있었다.

	이름	평점	리뷰 개수	가격	사이트
0	제주안뜰_안채	4.88	34	146337	AirBnb
1	제주동쪽 정갈한 돌집과 가드닝창고_스테이보보	4.96	46	188544	AirBnb
2	스테이 일면식	4.94	91	269569	AirBnb
3	"우리동네하가리"상업적공간이 아닌\n낭이가 갑인 공간입니다\n공지사항 필독후예약...	4.72	62	23106	AirBnb
4	여성전용 별채숙소 노리민박	4.95	39	71206	AirBnb
...	...	...	...	...	...
342	J2 패밀리 호텔	4.00	374	142500	부킹 닷컴
343	취다선 리조트	4.65	47	478500	부킹 닷컴
344	굿데이펜션	4.55	220	210000	부킹 닷컴
345	휴 안 스테이	4.30	293	121500	부킹 닷컴
346	제주 모슬포 호텔	3.65	133	138240	부킹 닷컴

이후에는 위의 데이터를 이용하여, 시각화 및 데이터 분석을 진행한다.

3. 기초 EDA

(1) 기초 정보

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 329 entries, 0 to 328
Data columns (total 5 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   이름      329 non-null    object
 1   평점      328 non-null    float64
 2   리뷰 개수   329 non-null    int64
 3   가격      329 non-null    int64
 4   사이트     329 non-null    object
dtypes: float64(1), int64(2), object(2)
memory usage: 13.0+ KB
```

크롤링의 결과로 총 329개의 데이터를 확보하였다.

(2) 결측치 확인

	0
이름	0
평점	1
리뷰 개수	0
가격	0
사이트	0

해당 데이터의 "평점" 컬럼에서 결측치가 1개 있는 것으로 확인된다.

	이름	평점	리뷰 개수	가격	사이트
221	그랜드 하얏트 제주	NaN	0	766260	부킹 닷컴

해당 결측치는 위의 그림에 해당하는 관측치에서 나왔다. 하지만 각 사이트별 가격 분포 확인을 위해 해당 데이터를 삭제하지 않았다.

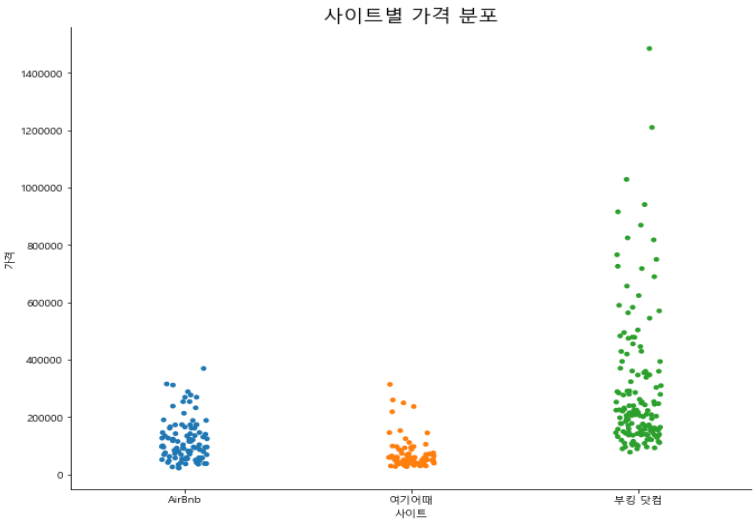
(3) 기술통계량

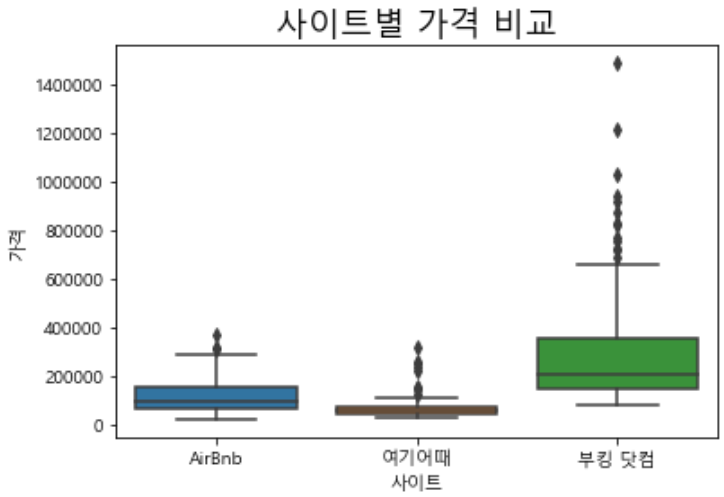
	평점	리뷰 개수	가격
count	328.000000	329.000000	3.290000e+02
mean	4.397927	433.422492	1.889060e+05
std	0.407167	532.324774	1.932830e+05
min	2.700000	0.000000	2.310000e+04
25%	4.100000	74.000000	6.931800e+04
50%	4.400000	228.000000	1.377000e+05
75%	4.750000	586.000000	2.243800e+05
max	5.000000	3414.000000	1.485000e+06

	가격
count	329
mean	188905
std	193283
min	23100
25%	69318
50%	137700
75%	224380
max	1485000

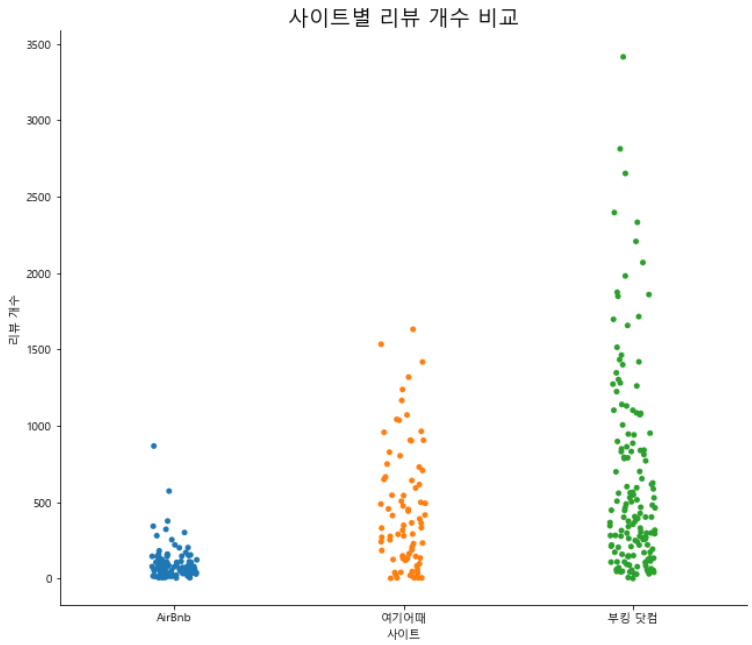
4. 시각화를 통한 데이터분석

1) 사이트별 가격 분포





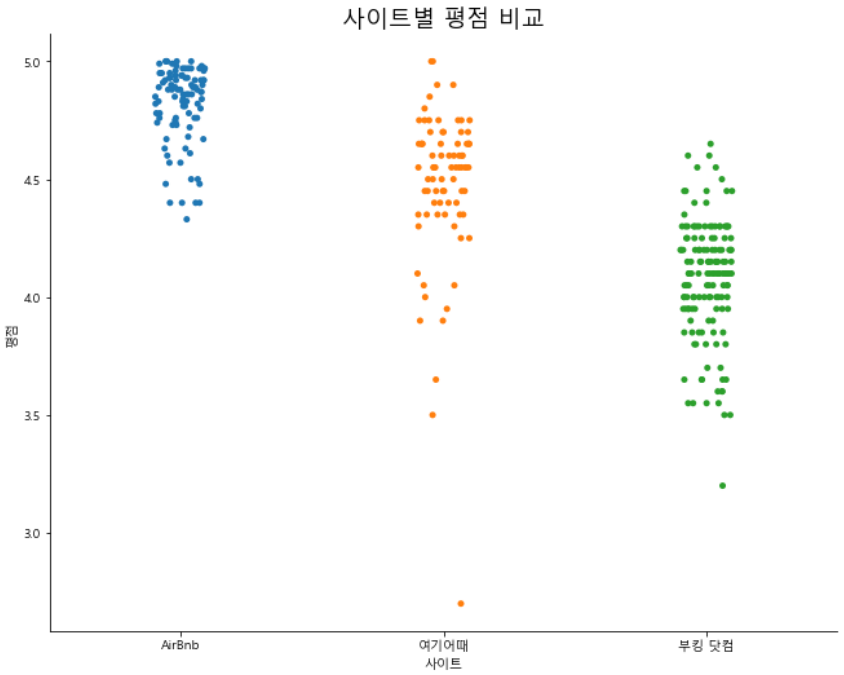
사이트별 가격대 분포를 살펴보자면, 에어비앤비와 여기어때에는 저가격대의 숙박시설들이 많이 모여 있는 것을 알 수 있다. 이는 에어비앤비에 공급되는 숙박 시설들은 자신의 집을 쉐어링 하는 형태이기 때문에 상대적으로 저렴한 가격에 숙박 시설을 공급할 수 있기 때문인 것으로 판단된다. 그리고 여기어때는 주로 고급가격대의 호텔보다는 중저가의 모텔이 많이 모여 있기 때문인 것으로 판단된다. 반면에 부킹닷컴은 가격의 분포가 10만원대 중후반부터 시작하여, 100만원이 넘는 가격의 숙박 시설도 보유하고 있는 것을 알 수 있다.



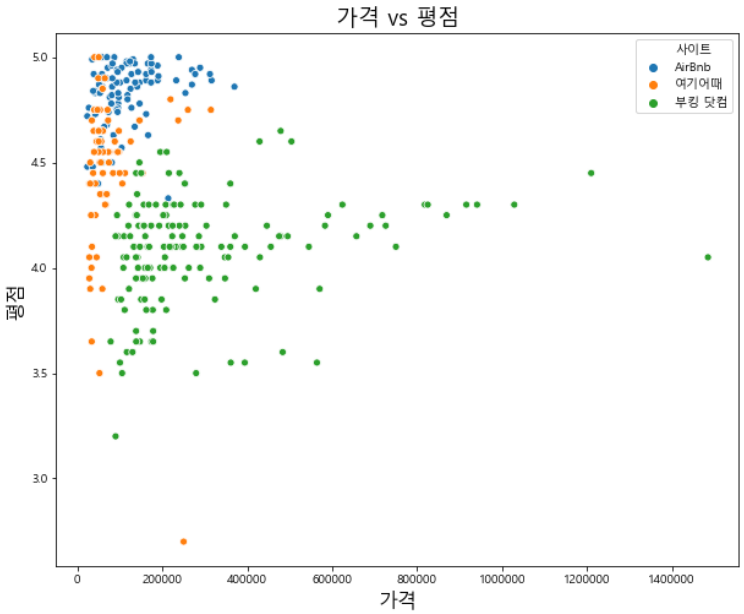
이를 사업자(기업)의 입장에서 활용한다면, 자신의 숙박 시설이 가성비를 위주로 하는 시설이거나 백팩커와 같은 저렴한 가격의 숙박 시설을 찾는 여행자를 타겟으로 한다면, 에어비앤비와 여기어때에 공급하는 것이 좋은 선택으로 보인다. 또한 바로 위의 사이트별 리뷰 개수 그래프를 참고했을 때 에어비앤비 보다는 여기어때에 리뷰개수가 많은 것을 확인할 수 있다. 이는 여기어때에 더 많은 사용자가 있다는 것을 유추할 수 있다. 따라서 공급하는 숙박 시설이 충분한 방 개수를 확보하고 있고, 더 많은 사람을 끌어들이기를 원한다면 여기어때에 공급하는 것이 좋은 선택으로 보인다. 반면에 호화 여행자 또는, 특별한 기호가 없는 여행자를 타깃으로 한다면, (리뷰 개수로 판단했을 때) 사용자 수가 가장 많아 보이는 부킹닷컴에 숙박 시설을 공급하는 것이 좋은 선택으로 보인다.

사용자 입장에서는, 자신의 목적에 맞는 숙박 시설을 찾을 때 위의 그래프를 활용할 수 있다. 예를 들어 전반적으로 저렴한 숙박 시설을 찾는다면, 부킹닷컴보다는 여기어때에서 숙박 시설을 찾는 것이 좋아보인다. 반면에 "호캉스"와 같은 호화 여행을 원한다면 부킹닷컴에서 숙박 시설을 찾는 것이 좋은 판단이다.

2) 사이트별 평점 분포



사이트별 평점의 분포를 살펴보면, 에어비앤비는 다른 두 사이트에 비해 평점이 상대적으로 높은 것이 보인다. 반면에 부킹닷컴은 최대치가 4.7 정도이고, 더 넓은 평점의 분포를 가지는 것을 확인할 수 있다.



가격과 평점의 관계를 보여주는 위의 그래프를 통해 살펴보았을 때, 전반적으로 저가격대의 숙박 시설들이 높은 평점을 획득하는 것을 볼 수 있다. 이는 시설 및 서비스의 퀄리티와 상관없이 "가격적 메리트"가 소비자에게 긍정적인 어필을 하는 것으로 여겨진다. 반면에 부킹닷컴의 호텔들은 3점대 중반에서 4점대 중반까지의 분포를 보이며 상대적으로 에어비앤비와 여기어때에 비해 저조한 평점을 보여준다. 높은 가격대의 호텔들도 고평점을 획득하는데 한계가 있는 것으로 보이는데, 이는 가격이 높아짐에 따라 그에 상응하는 서비스의 퀄리티를 기대하기 때문인 것으로 판단할 수

있다.

한편, 각 사이트별 평균 평점은 다음과 같이 형성 되어있다.

- 여기어때 평균 평점 : 4.48
- 에어비앤비 평균 평점 : 4.82
- 부킹닷컴 평균 평점 : 4.08

이는 에어비앤비가 전반적으로 가성비가 좋은 숙소들을 많이 보유하고, 부킹닷컴이 상대적으로 그렇지 못하다고 판단할 수도 있지만, 에어비앤비의 평점에 인플레이션이 있는 것으로 판단할 수도 있다. 따라서 이용자가 사이트간 비교를 통해 숙박시설을 결정한다고 했을 때, 단순히 높은 평점의 숙소를 정하는 것이 아니라, 사이트별로 평점의 분포를 고려해야 할 것이다.

(4) 워드클라우드를 통한 리뷰 시각화

위의 그림은 제주도의 "호텔 난타"에 대한 호텔스닷컴에 남긴 실사용자들의 리뷰이다. 해당 그림을 통해 숙소가 깨끗한 편이라는 것을 알 수 있으며, 한라산이 잘 조망되는 곳에 숙소가 위치해 있다는 것을 유추할 수 있다. 반면에 "수건","냄새"와 같은 키워드 역시 눈에 띄는 것을 확인할 수 있다.

실 사용자의 리뷰를 확인해보면 다음과 같다.

'늘 만족하는 호텔 난타에요. 자주 와서 묵는데 이번에는 수건에서 약품 냄새가 너무 진해서 씻고 사용하는데 머리가 아플 정도였어요..... 수건 냄새 빼고는 늘 만족하는 숙박입니다. 청결하고 친절해요. 얼른 코로나 괜찮아져서 룸서비스도 저녁까지 해주셨으면 좋겠어요'

즉 "수건"과 "냄새" 라는 키워드는 해당 호텔에 대한 부정적인 의견을 남긴 것으로 판단된다. 이는 해당 호텔이 개선해야할 항목으로 여겨진다.

이처럼 워드클라우드를 통해 사용자가 특정 기업 및 상품에 대해 남긴 리뷰를 분석할 수 있다. 기업 입장에서는 향후 개선 방안 및 마케팅 전략 등을 구성하는데 활용할 수 있으며, 사용자 입장에서는 구매 계획중인 상품의 전반적인 평가를 확인하여, 합리적인 소비에 도움이 될 수 있다.

Chapter3. 결론

1. 요약

본 프로젝트에서, 시장 분석을 위한 웹크롤링과 시각화를 통한 데이터 분석을 진행하였다. 기업 입장에서는 데이터 크롤링 및 분석을 통해, 자사 제품의 적절한 공급 가격을 설정하고 시장 진출 전략을 산출해낼 수 있다. 또한 사용자들의 리뷰를 간편하게 확인하여 개선 방안을 도모할 수 있다. 사용자 및 소비자 입장에서는, 구매하고자 하는 아이템의 적절한 가격을 측정할 수 있으며, 이에 따라 합리적인 소비가 가능해진다.

2. 한계점 및 구현하지 못한 내용

1) 크롤링 대상 사이트의 한계

본 프로젝트의 크롤링 대상 사이트는 부킹닷컴,에어비앤비,여기어때 3개의 사이트로, 이는 현재 운영되고 있는 예약 사이트들의 개수에 비해 현저히 적은 숫자이다. 따라서 더욱 합리적인 시장 분석을 위해서는, 분석 대상 사이트들을 늘려 데이터의 절대적인 숫자를 늘릴 필요가 있어 보인다.

2) 워드클라우드와 감성분석

워드클라우드를 시행함에 있어, 300여개가 넘는 숙박 시설들을 모두 분석하는데 한계가 있어, 1개의 호텔을 분석하는데 그쳤다. 이를 해결하기 위해서는 장고와 플라스크와 같은 웹개발 프레임워크를 활용하여, 분석 결과를 웹페이지 상에서 시각화 하는 방법이 있다. 즉 리뷰를 확인하길 원하는 숙박 시설을 선택하여, 그 시설에 대한 리뷰만을 보여주는 형태로 기능을 구현할 수 있다.

또한 기획서(중간과제)에 기재하였던 내용중 감성분석을 구현하지 못하였다. 감성분석을 통해 이용자들이 어떤 점에 만족하고 어떤 점에 불만족하는지 간단하게 확인할 수 있다. 하지만 크롤링 코드를 구현하는데 많은 시간을 소모하여, 감성분석기 구축에 실패하였다.

3. 활용 및 발전 방향

본 프로젝트에서 사용된 크롤링 매커니즘은, 시장 조사 및 분석이 필요한 모든 비즈니스 분야에서 사용 가능하다. 본 프로젝트에서는 "숙박 예약 사이트" 들을 대상으로 하여 가격, 리뷰개수, 업체이름, 평점을 크롤링하였다. 이는 오픈마켓에서도 그대로 적용 가능하며, 기업이 자사의 제품을 출시하려고 할 때, 가격대 설정을 위해 본 프로젝트에서 사용된 방식의 크롤링이 사용가능하다. 그 외에도 중고차 시세와 같이, 기업과 소비자 간의 정보의 불균형이 있는 경우, 적절한 매매가격 확인을 통해 합리적인 소비를 할 수 있도록 도와줄 수 있다.