**□ 모듈프로젝트 산출물**

1. 클라우드 AI 융합 전문가 양성 과정 – 기반기술 인프라

| 1. **모듈프로젝트 1** |
| --- |

**프로젝트 명 : 리뷰를 기반으로 한 도쿄 맛집 분석**

2023-01-18

**6조**

이름: 윤예찬,박민이,이지원,고종현

**목차**

1. 개요 – 도쿄 음식점 분석

1.1 목적

1.1.1 기획의도

1.1.2 목표

1.2 방법

2. 데이터 분석

2.1 리뷰 추출

2.1.1 웹사이트 크롤링

2.1.2 리뷰 번역

2.1.3 csv 데이터 정제

2.1.4 키워드 단어 수집

2.1.5 데이터 정제 준비

3. 시각화 준비

3.1 키워드 별 카테고리 분류

3.2 식당 비교 분석

4. 시각화

4.1 음식점 별 그래프

4.2 도쿄 전체 음식점 그래프

5. 결론

**1.개요**

**1.1 목적**

**1.1.1 기획 의도**

코로나 19 확산으로 중단됐던 일본 무비자 관광이 재개되며 일본 여행객들이 급증하고 있습니다. 이에 따라 예비 관광객들이 참고하여 자신이 원하는 가치에 적합한 식당을 더 쉽고 빠르게 찾을 수 있도록 데이터를 재가공하려고 합니다.

**1.1.2 목표**

맛집을 찾아주는 사이트에서 데이터들을 수집해 분석하고, 리뷰 텍스트 안에서 반복적으로 나타나는 단어들을 특성 키워드 별로 재가공하여 관광객들에게 직관적인 자료로 제공하려고 합니다.

**2.1 방법**

활용한 데이터: [Tripadvisor](https://docs.google.com/document/d/1MN8ynlJpcNjytZPkptItKIF-O2nODYYN3Lb5kUleq_Q/edit#)(일본 지역별 맛집과 리뷰 데이터가 있는 홈페이지)

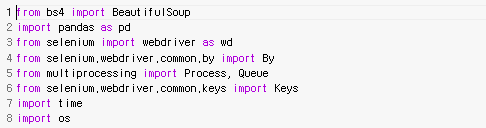
1. 인터넷 맛집 사이트에서Selenium과 BeautifulSoup를 이용하여 지역별로 각각의 음식점 정보와 리뷰 데이터를 크롤링합니다.
2. 크롤링하여 얻은 데이터들을 토큰화합니다. 그 후 반복적으로 나타나는 단어를 추출하고 가게의 특성을 보여줄 수 있는 키워드들을 추출합니다.
3. 추출한 키워드들은 비슷한 성격을 띄우는 것들끼리 묶어 하나의 해시태그를 만듭니다.
4. 각 음식점의 리뷰 코멘트에 키워드를 분석하고 적합한 해시 태그를 부여한다.
5. 부여된 해시 태그는 다시 음식점 별로 각각 합산한다.
6. 합산된 키워드를 가지고 임의의 계산식을 적용해 시각화하여 고객들에게 전달한다.

**2. 데이터 수집과 정제**

**2.1 리뷰 추출**

**2.1.1 웹사이트 크롤링**

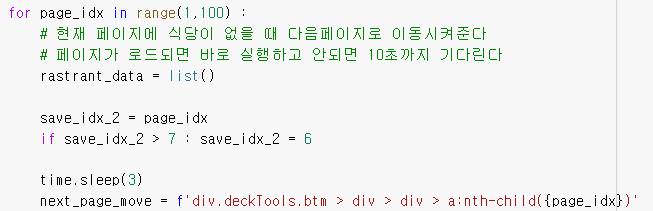
1. 필요한 라이브러리와 모듈을 불러온다



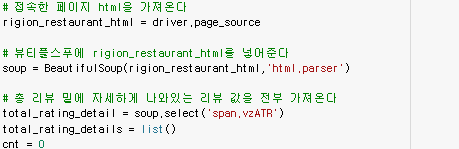
1. 드라이버 연결 및 URL 접속



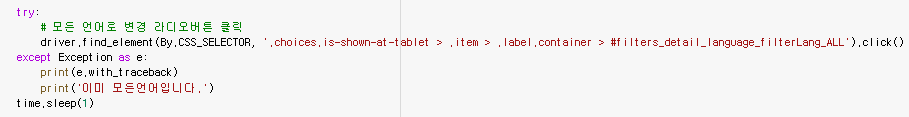
1. 다음페이지로 이동



1. BeautifulSoup으로 리뷰를 전부 가져온다



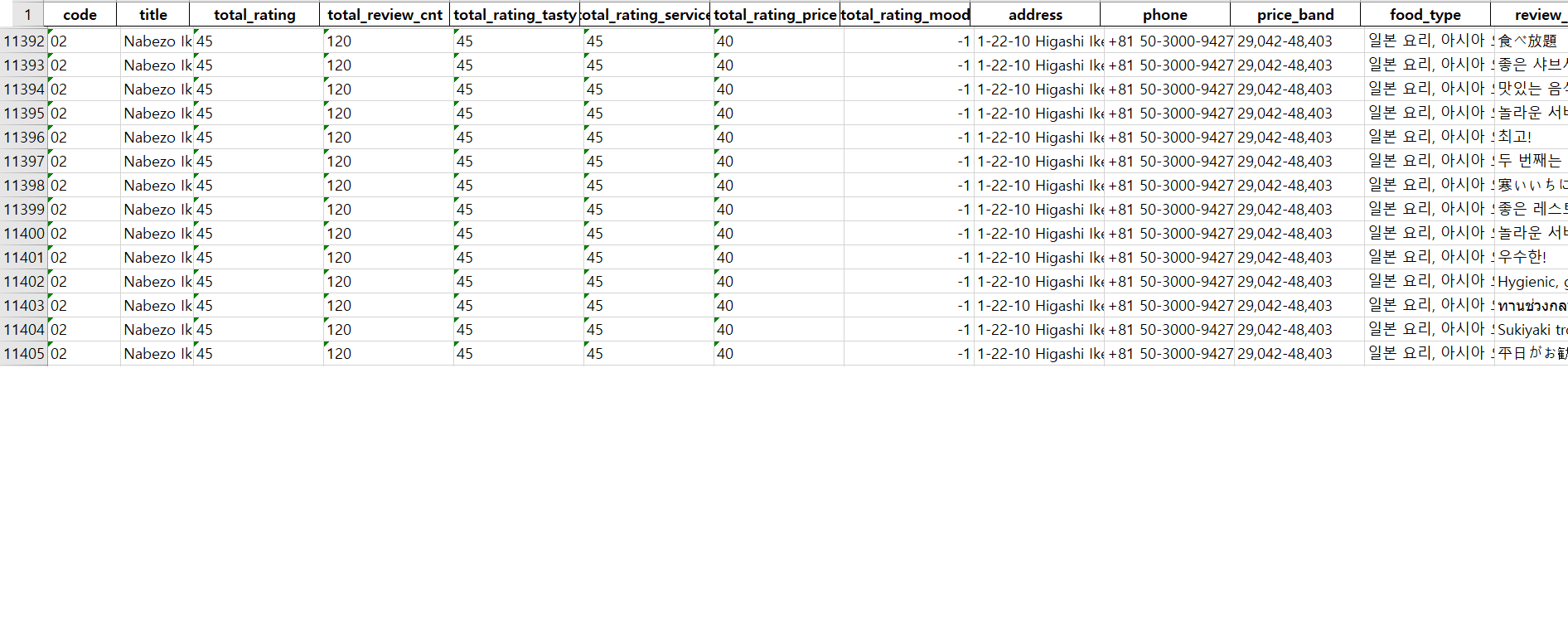
1. 모든 리뷰를 불러오기 위해 ‘모든언어’ 버튼을 클릭한다



1. 리뷰가 길면 더보기 버튼을 클릭해서 한번 더 페이지를 추출한 후에 식당 정보, 리뷰 제목, 코멘트, 별점 등을 한번에 리스트에 추가한다

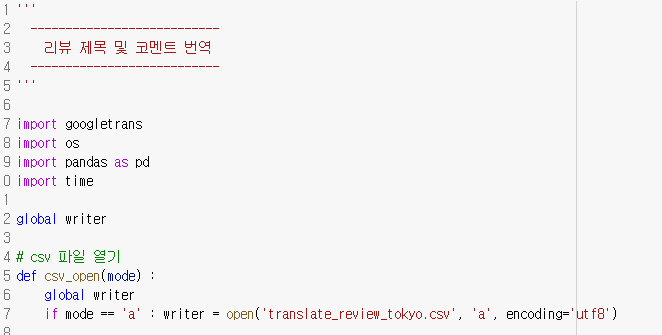


1. 엑셀로 저장된 데이터들

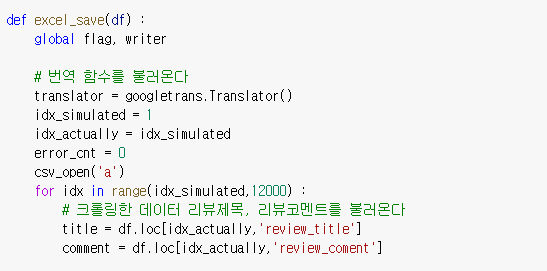
****

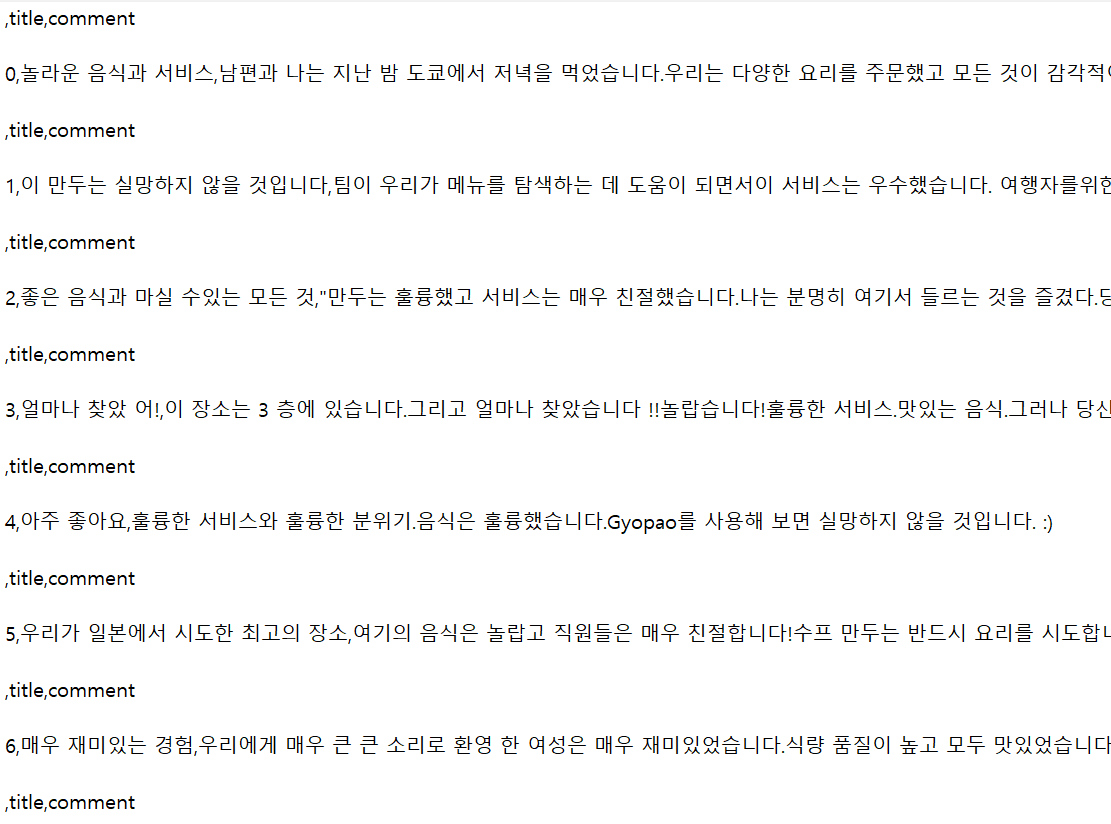
**2.1.2** **리뷰 번역**

1. 필요한 라이브러리와 모듈을 불러온다

****

1. 크롤링한 데이터 리뷰제목, 리뷰코멘트를 번역한다.

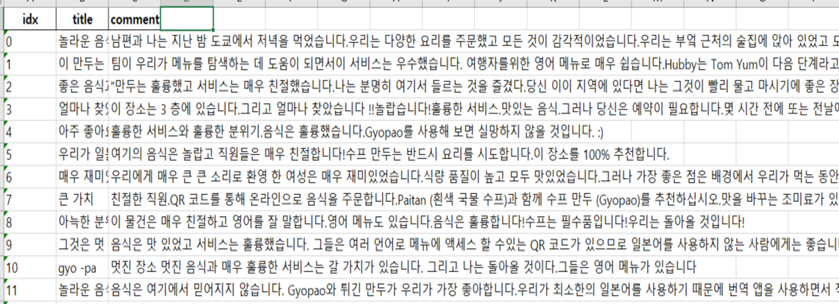
****

****

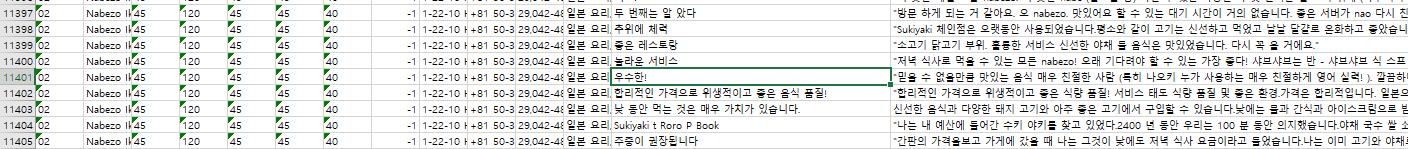
**2.1.3 csv 데이터 정제**

1. 번역한 리뷰에 띄어쓰기 및 필요없는 데이터 등을 정제하고 쪼개진 코멘트들을 join 함수로 묶어준다

****



1. 나온 값 excel에 매치시켜 붙여넣기한다

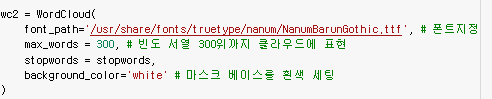
****

**2.1.4 키워드 단어 수집**

1. 모듈 불러오고 워드클라우드 생성



1. 워드 클라우드에 필요한 것들을 적용해준다



1. txt 파일로 정리한 불용어를 불러 리스트에 저장한다



1. 리뷰 코멘트를 합치고 저장된 불용어들을 적용한다

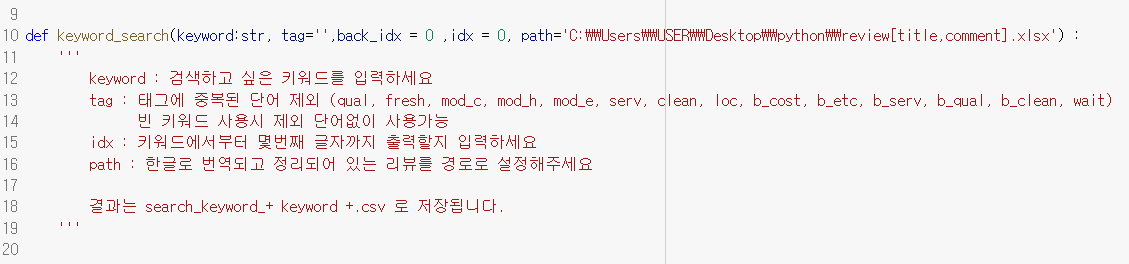




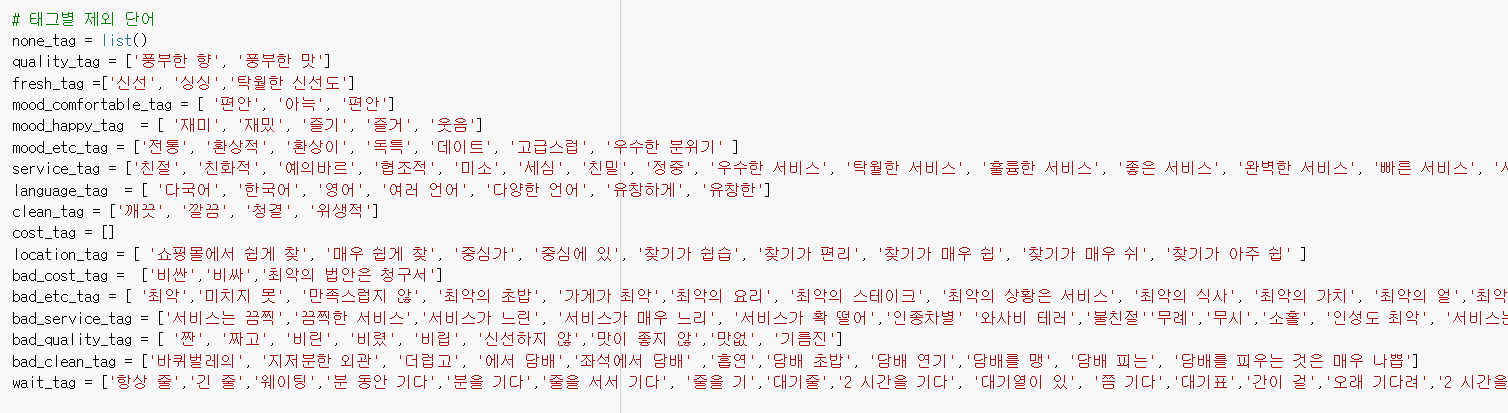
- 필요없는 단어를 제거하면서 쓸 수 있는 단어는 추출하여 키워드에 추가한다

**2.1.5 데이터 정제 준비**

1. 함수 인자에 대한 설명



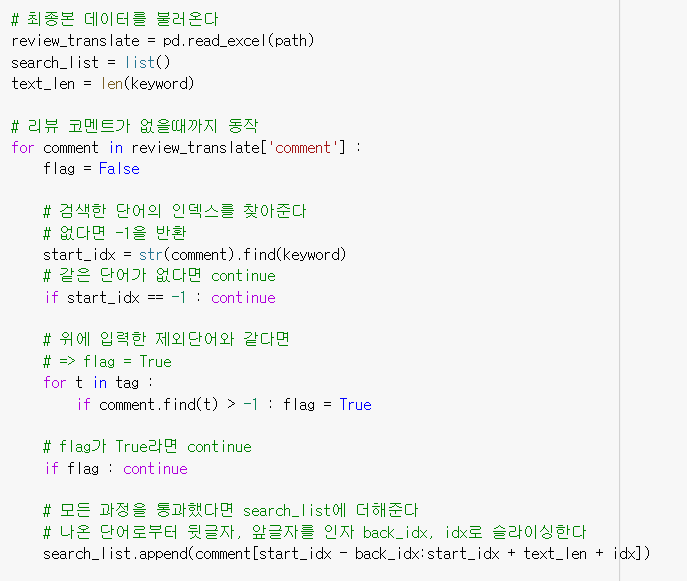
1. 태그 별 제외 단어 준비



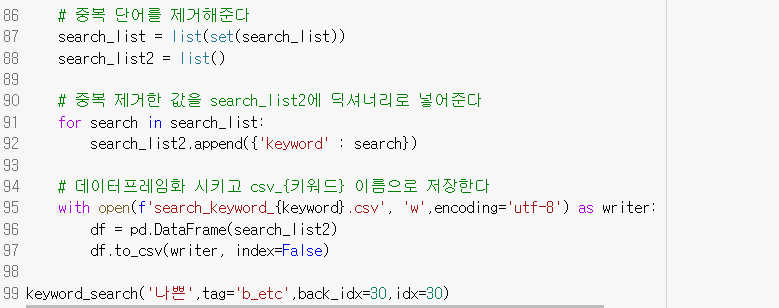
1. 인자에 입력한 태그 설정



1. 리뷰가 정리된 엑셀을 불러오고, 검색어가 포함된 리뷰를 추출



1. 정제 리뷰 중 중복 단어제거 후 csv로 저장



* 추출할 키워드와 그 키워드 중 제외할 키워드를 정확한 데이터로

정제할 수 있다

**3. 시각화 준비**

**3.1 키워드별 카테고리 분류**

- 키워드를 특성별로 묶어 리뷰 내 키워드가 존재한다면 해시태그를 부여한다.

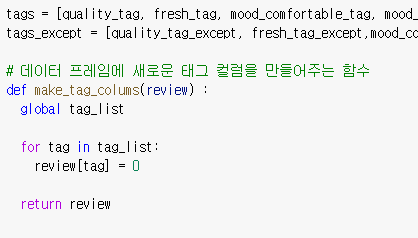
1. 필요한 자료들을 리스트로 정리한다

****

1-1. 필요한 자료들을 리스트로 정리한다

****

2. 최종 엑셀에 여러개의 태그 컬럼을 추가시켜준다

****

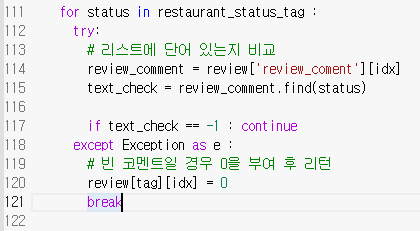
3. 각 태그에 값을 부여해준다

* 키워드와 동일한 리뷰가 있다면 +1

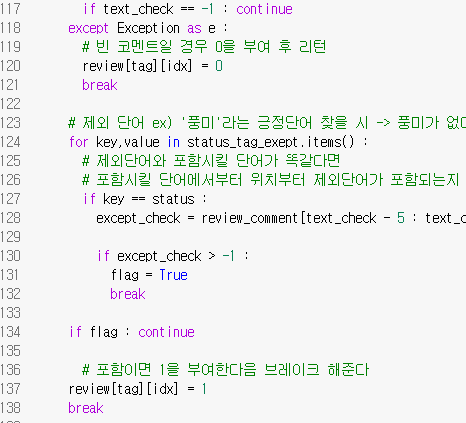


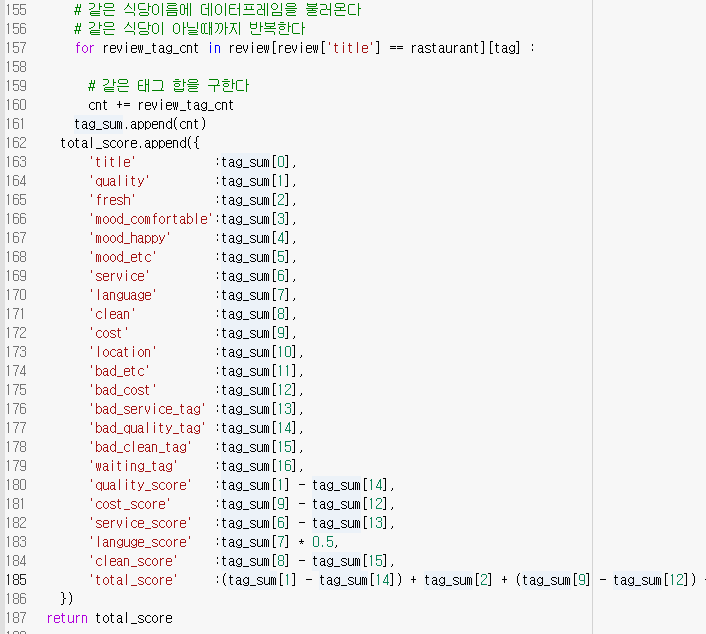
4. 리스트에 위에 정리한 식당이름, 키워드 단어, 제외단어 등을 리스트를 넣어주고,

반복문으로 돌려준다

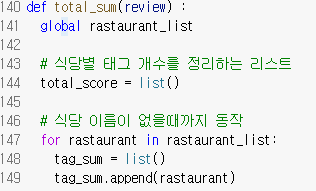
****

4. 키워드가 포함되고 제외키워드에 포함되지 않으면 tag에 값을 1 증가시킨다.



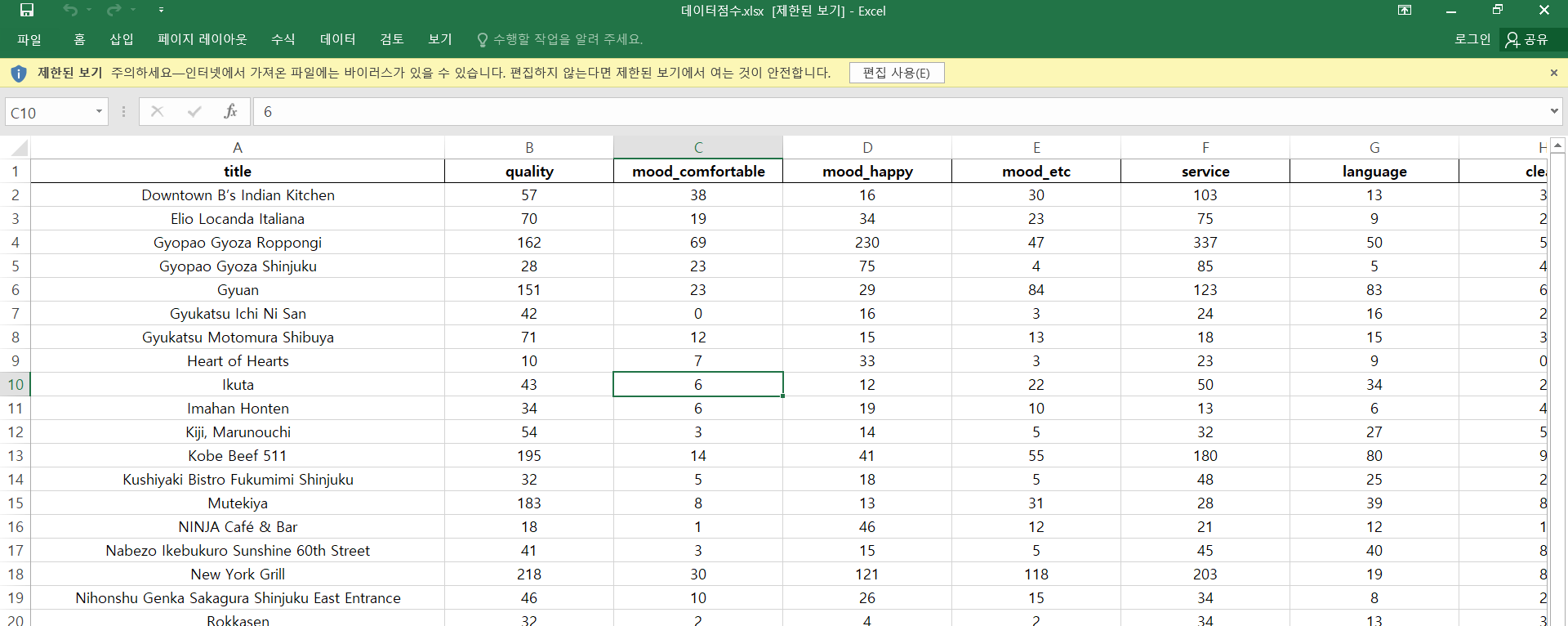
5. 가게별로 각각의 태그 값들을 합산해주고 리스트에 추가하여 반환한다

6. 각각의 해시태그를 더해 합산을 해주는 함수이다



7. 합산한 결과를 리스트에 추가하여 엑셀로 저장한다



8. 결과

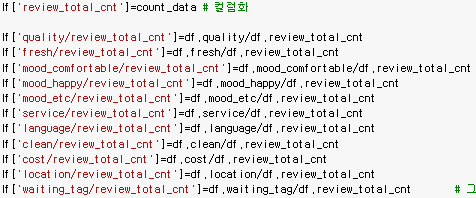
**4.시각화**

1. 모듈 불러오기



1. 아래 엑셀표의 해쉬테크 컬럼들을 기반으로 그래프를 만들기 위해 비율 용 컬럼들 생성한다

**비율 = 가게의 해당 태그의수 / 가게당 리뷰의 수**

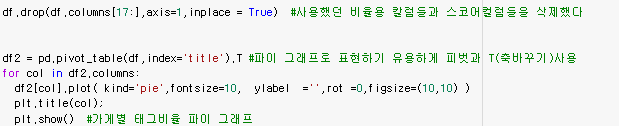


**4.1 키워드 별 그래프**

****

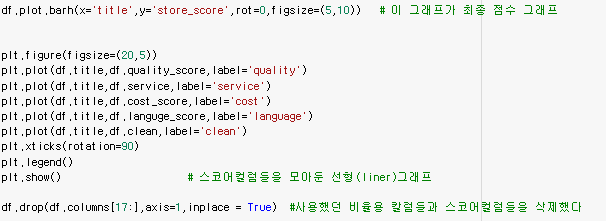
|  |  |
| --- | --- |
| **# 퀄리티 좋은** | **# 신선한** |
|  |  |
| **# 편안한 분위기** | **#즐거운 분위기** |
|  |  |
| **# 특별한 분위기** | **# 서비스가 좋은** |
|  |  |
| **# 영어를 사용하는** | **# 청결한** |
|  |  |
| **# 가성비 좋은** | **# 위치가 좋은** |
|  |  |
| **# 대기가 있는** | **음식점별 최종 점수** |
|  | |

**4.2 음식점 별 그래프**

****

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

**4.3 도쿄 전체 음식점의 특성별 점수 비교 선 그래프**

****

**5.결론**

일본 여행객이 급증하고 있는 시점에 의미있는 주제를 선정하여 진행하게 된 데이터 분석입니다.

분석을 진행하며 찾을 수 있었던 여러 결과로는

1. 대부분의 리뷰에서 서비스 평가가 높은 비율인것으로 볼때 일본의 음식점들을 대체로 손님들을 친절하게 맞아주는 경향일 것으로 예측됩니다.
2. 수많은 리뷰들을 전부 읽지 않고도 각 음식점마다 어떤 강점이 있는 곳인지 직관적으로 알 수 있습니다
3. 사이트에서 별점으로는 표현되지 않았던 장단점들을 키워드들로 추출해 다시 점수를 매겨 사이트의 맛집 등수보다 정확한 평가를 보여줄 수 있습니다.

다양한 음식점 랭킹 사이트들이 존재하지만 기존의 음식점 평가 서비스들과는 다르게 좀 더 구체적이고 다양한 해쉬테그를 제작하였고 부정 리뷰 또한 추출하여 하나의 해쉬태그로 만들며 차별화할 수 있었습니다.

외국 음식점을 분석하여 각 가게만의 개성적인 특징을 유의미한 결과로 얻을 수 있었습니다. 도쿄 말고 다른지역 데이터를 분석 못한게 아쉽지만, 코드 인덱스만 변경하면 다른지역 데이터도 추출하고 분석할 수 있기에 큰 아쉬움이 남진 않았습니다. 여러가지로 시도도 많이 해보고 상당히 유익한 프로젝트 였습니다.