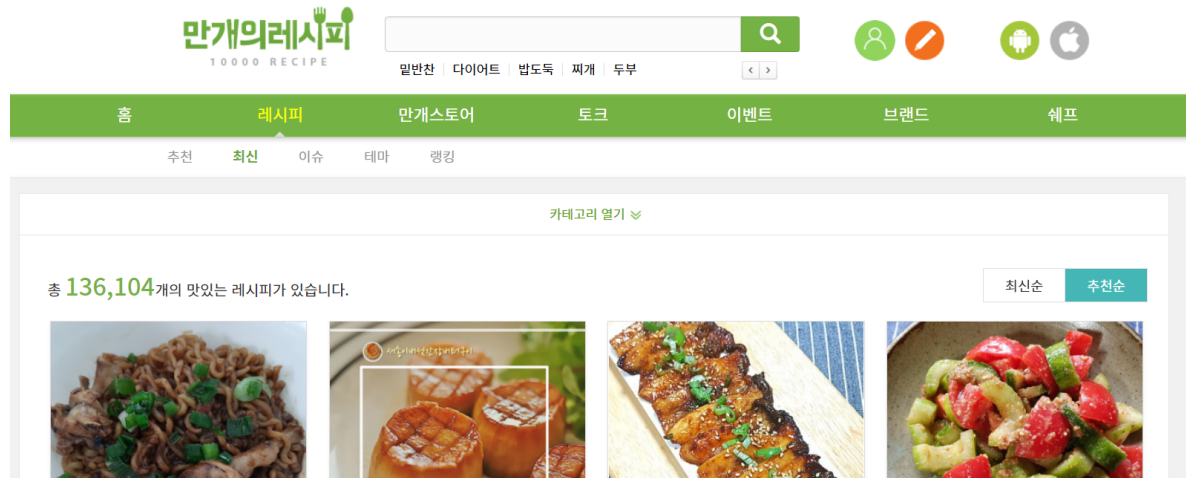


1. 크롤링

레시피 데이터는 '만개의 레시피' 사이트를 크롤링해 수집한다.



레시피 카테고리의 1페이지부터 2400페이지까지, 약 95000여 개의 레시피 데이터를 수집한다.

수집할 내용은 다음과 같다.

```
id = [] // 고유 아이디
title = [] // 제목
writer = [] // 작성자
viewers = [] // 조회수
size = [] // 찜인분
time = [] // 조리시간
level = [] // 조리난이도
ingredients = [] // 모든 재료
description = [] // 설명

main = [] // 주재료
minor = [] // 부재료
source = [] // 양념

complete = [] // 완성 사진

total_quantity = [] // 모든 재료 용량
main_quantity = [] // 주재료 용량
minor_quantity = [] // 부재료 용량
source_quantity = [] // 양념 용량
for i in range(25):
    globals()['step{}'.format(i)] = []
    globals()['img{}'.format(i)] = [] // 조리순서
```

2. 데이터 정렬

이미지 디텍션을 위한 학습에 사용할 재료 이미지가 필요하다. 레시피에 많이 포함되는 재료들을 우선적으로 학습시키기로 한다. 어떤 재료가 많이 사용되는지 알아보기 위해 크롤링한 데이터를 정렬한다.

```
ingredients = dataset.loc[:,['ingredients']]
ingredients
```

	ingredients
0	['오리고기 3조각, 짜장라면 1개, 물 300ml']
1	['새송이버섯', '버터', '어린잎채소', '간장', '맛술', '올리고당']
2	['새송이 버섯', '청양고추', '대파', '버터', '간장', '올리고당', '...
3	['콩나물', '비엔나소시지', '대파', '양파', '물', '고추장', '고추가...
4	['자른박이', '깻잎', '팽이버섯', '숙주나물', '미소된장', '마요네즈',...
...	...
95698	['오리 새고기', '새송이 버섯', '제철 잎채소', '청양고추', '대파', '...

ingredients 칼럼의 데이터들을 적절히 가공한다.

```
ingredients = []

for i in range(0,95698):
    for j in dataset.loc[i]
[8].replace(' ','').replace('[','').replace(']','').replace('"','').split(', '):
        ingredients.append(i)
ingredients
```

```
['오리고기 3조각',
'짜장라면 1개',
'물 300ml',
'새송이버섯',
'버터',
'어린잎채소',
'간장',
'맛술',
'올리고당',
'새송이 버섯',
'청양고추',
'대파',
'버터',
'간장',
'올리고당',
'다진마늘',
```

그리고 자주 사용된 순서대로 정렬해 출력한다.

```
from collections import Counter
Counter(ingredients).most_common(200)
```

```
[('소금', 30437),
('설탕', 28594),
('양파', 28465),
('참기름', 23251),
('다진마늘', 22241),
('물', 21513),
('간장', 21356),
('대파', 18034),
('후추', 17020),
('고춧가루', 15037),
('마늘', 11158),
('달걀', 11074),
('통깨', 10621),
('고추장', 10599),
('청양고추', 10299),
('페라' 9146)
```

양념과 부재료들은 제외하고 선별하여 '감자, 양파, 대파, 계란, 돼지고기, 고추' 이미지를 우선 학습시켜
보기로 결정했다.