많이찾는의

Web Crawling

Team: Build Up

박경빈, 방울, 최유림, 한설아



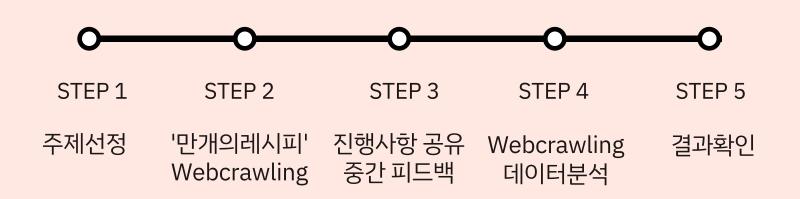






사람들이 많이 찾는 레시피

인간생활의 세 가지 기본 요소라고 불리는 의식주에서도 볼 수 있듯 우리의 삶에 있어서 빼놓을 수 없는 식(食)을 주제로 히여 사람들이 많이 찾는 레시피를 레시피 사이트에서 Crawling 하여 데이터를 분석해보고자 함.

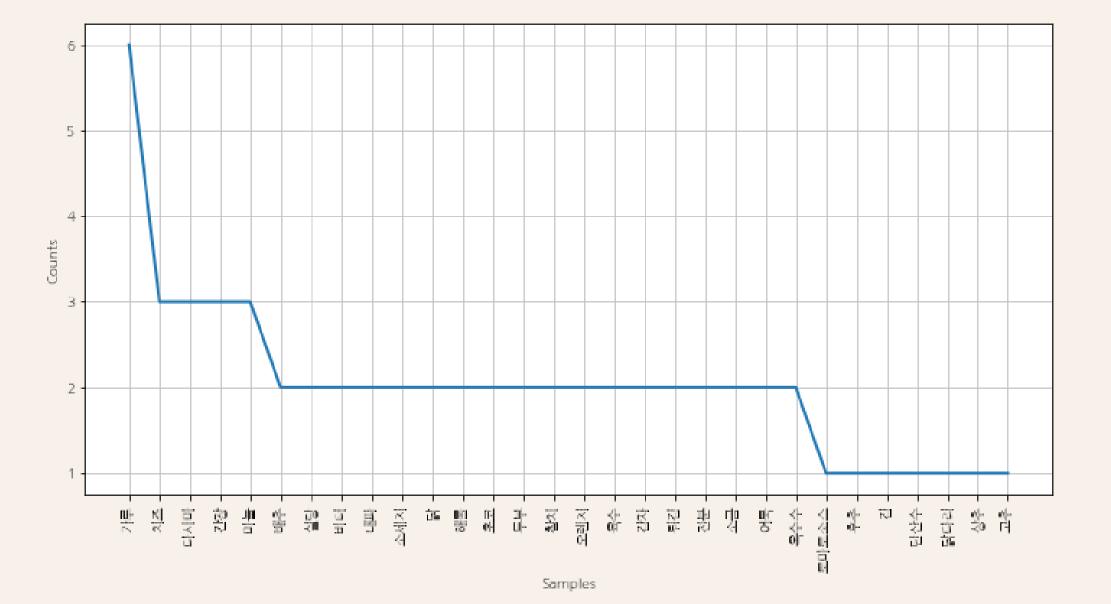


2

레시피 사이트 정보수집 (1)

1. 요리 레시피 사이트 웹 크롤링을 통한 1차 분석

- ❤ '만개의레피시' 사이트 웹 크롤링
- ❤ 주로 사용되는 '식재료' 도출
- ✅ 그래프와 워드클라우드로 시각화



```
import urllib.request
from bs4 import BeautifulSoup as bs
import re
import json, csv
def toJson(recipe_dict):
  with open('recipe.json', 'w', encoding='utf-8') as file :
    json.dump(recipe_dict, file, ensure_ascii=False, indent='\taut')
def toCSV(recipe_list):
  with open('ingredients.csv', 'w', encoding='utf-8', newline='') as file:
    csvfile = csv.writer(file)
    for row in recipe_list:
      csvfile.writerow(row)
def url_func(n, m):
 num_range = range(n, m)
 url = "https://www.10000recipe.com/profile/recipe.html?uid=gdubu33&page="
 url_list = []
  for num in num_range:
    reg = urllib.reguest.Reguest(url + str(num))
    code = urllib.request.urlopen(url).read()
    soup = bs(code, "html.parser")
    try:
      res = soup.find(class_ = 'cont_list st3')
      for i in res.find_all('a'):
       url_tmp = i.get('href')
       url_list.append(url_tmp)
    except(AttributeError):
        pass
  return url_list
```

레시피 사이트 정보수집 (1)

2. 크롤링 1차 분석 결과

- ✓ 추출값에서 불용어가 많아 제거되는 값이 많았다.
- ✅ 해당하는 크롤링 결과값으로는 유의미한 결과를 얻기가 어려웠다.
- ✅ 페이지의 범위나 키워드에 따른 다른 목록값을 받아 결과를 도출해보자.
- ✓ 넓은 범위의 데이터를 수집 후 '다시마', '마늘'과 같은 양념 재료 외에 메인식재료에 대한 트렌드를 알아보자.

범위 랭킹, 키워드, 계절별 분류 등 크롤링 범위를 다양하게 레시피에서 많이 쓰이는 식재료 or 계절별로 많이 쓰이는 식재료



```
dmil or Loget data(n. m);
          # https://www.10000yec.ine.com/restime/flet.html?==1=11000051000=Filles#800100.terden=vendone=5
          It to 67/mov 10000 mm (net to 6/mm (net to 6/mm) for $2200 x $4200 x $4200 for $1200 for $120
          ■21Ue = [].
          s_authors = [[
          summer []
          for | In range(n.m+135
                   WY | https://www.it0000kecisia.com/yadking/Nime.html
                   page = trippen(wri)
                   soup = EvantifulSoup(page, htm: mareer)
                   print(("{|| 田田村 海陽 | {poup.title}")
                   racipes = soup find( ul : class = common ap listar( nod))
                   recipes_all = recipes.find_si(("| '_ithms_s'commonter.ii")
                   all_title = []
                   authors = 11
                   all_nums = [[
                   altauri * ii
                    for one in recipes_alt:
                             title = one find(s) v . chas_= componese_conting_tite(inc ) text
                              auth = one find( d) / stabs = common_cr_cast(an_FV_hume*) text
                             auth = auth_restace() ** . " )
                             nous = one find( spint, class = comman spiciot(on buyer ), text
                             one_srf = one_find( ) . cfass_= preson=== fine 31 hraf )
                             una_ull_url = "https://www.HECEsected.com" * upa_url
                             will_title:append(title)
                             authors.append(auth)
                             allimons appendinues
                             all_url.append(one_all_url)
                   smint(len(all_title)_ len(authors), len(all_tues), len(all_url)}
                   print(authors).
                   print(all_nums)
                   mrint(all_uel)
                   w_titin_estandiali_titie)
                   a_withors_extend(authors)
                   s_nums.extend(all_mass)
                   a_url materid[atl_url]
```

	349	宝石村	車前≑	W)
	AMERICAN PROPERTY.	##fp	2407 H1400	A
371		明明重	主動 物地	long to will be the long with the long to
40	19 10 10 10 10 10 10 10 10 10 10 10 10 10	Amount 1	手粉中型	https://www.p0000mile.com/millio-000millio-
100	サリヤキ ががり 東京市人中 中本キー	哲を発証す	251 1248B	https://www.communications
	新規模型 的复数 医乳管 医乳期医性炎	要は4.3.4円=		MINISTRA WATER STREET
W	QUEE BOFE FILE WAY BESINES TO HAN READ	11年20日	### ##################################	https://www.scooperace.com/wellers/file/file
*	医可含性的 计信息管 四十分的中央处理 网络沙克 有杂位	神智學學	#### 195 <u>9</u> #	https://www.moneupchehouse
91	### 4444 F18 E20	-tiampy	241 310	https://www.com/com/com/com/com/com/com/com/com/com/
100	意の業別 対色相相 美年外	長の同位	主帥 44.0 0	https://www.homouries.com/militare/1957.
	- 27 PP 10	8540	非职士的	https://www.2000.evaic.com/weige.chi/2472474
4	No. 1 A Sharey			

레시피 사이트 정보수집(2)

3. 요리 레시피 사이트 웹 크롤링을 통한 2차 분석

- ✓ 개별 질문과 피드백을 통한 코드 수정
- 생 행킹(일간 TOP100)의 요리 리스트의 '요리명' '요리자' '조회수' 'URL' 추출

```
歯を 三型
dates = []
for one in all_url:
 print(one)
 uril = one
 page = urlopen(url1)
 soup = BeautifulSoup(page, 'html parser')
 print( soup.title )
 print( soup find( p . class_='viev_notice_date') find( b ) text )
titus //www.10000recipe_com/secipe/5832651
<title>[참처 감원장] 밥도록 참처 감된장으로 밥함꼳기 뚝딱</title>
등폭일: 2015-08-24
 thes://www.lubburecips.com/rectps/693688F
<title>스템훈두뿌찌개 원하고 완전 밥도둑!!</title>
들복일 : 2020-07-17
http://www.10000rective.com/nactio=/6902902
<title>진정한 밥도둑~잘의은 배추감치로 들갈비감처집 만드는법**^*</title>
등통일: 2018-12-26
```

프로젝트를 통해 배운 점

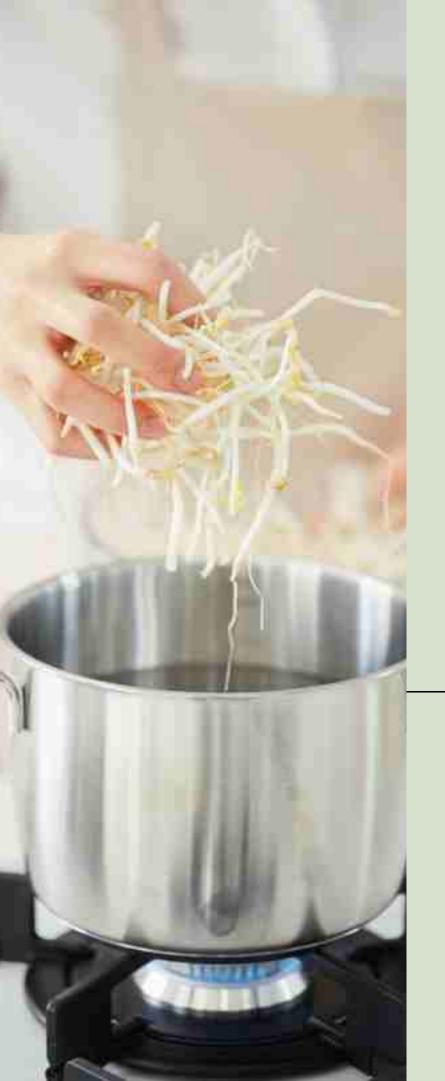


프로젝트를 통해 배운 점

- 웹 크롤링과 데이터 수집, 분석 등을 위해서는 Base(기초학습)이 중요함.
- 직접 코드를 입력, 수정해보면서 BeautifulSoup이나 urllib, pandas 등 적용이 익숙해짐.
- 소스코드를 활용(수정)하거나 새로 구상하여 결과 도출하는 것이 어려웠음.
- 데이터수집이 실제로 다양하게 활용되고 있으며, 데이터에 대한 효율성 극대화함.

더 분석하게 된다면

- (미완성한 주제) 계절별로 많이 쓰이는 식재료에 대한 데이터 분석
- 우크라이나침공, 상하이봉쇄 등 경제 상황으로 인한 식량 가격 상승의 상관관계 분석
- 우리나라 메인 식재료로 육류가 곡물 소비를 추월했다는 결과에 대한 식재료 소비량 분석



Thank you!

감사합니다.

 Team **Build Up**

 박경빈, 방울, 최유림, 한설아

