문화콘텐츠와자연어처리 기말프로젝트 결과보고서

#찐맛집_찾기

: 맛집 선별 해시태그 분석

목 차

그림 차례

Ⅰ. 서론

1. 연구 주제

#찐맛집_찾기: 맛집 선별 해시태그 분석

2. 연구 배경

인터넷 기술의 발달로 인해, 포털 사이트 및 SNS의 검색 기능을 활용하여 정보를 찾는 사람들이 증가했다. 더불어, 포털 사이트 및 SNS를 통한 마케팅도 급증했다. 이로 인해, 정보에 대한 접근성과 검색의 편리성이 향상되었다는 장점이 있지만, 허위광고 또는 과장광고로 인한 피해도 동시에 증가했다. 특히 SNS를 통한 음식 협찬과 광고가 많아졌고, 이로 인해 피해를 보는 경우가 증가했다. 따라서, 허위광고 또는 과장광고를 구분하고 맛집을 찾을 때 유의미하게 사용할 수 있는 해시태그를 선별하고자 한다.

3. 서비스 프로세스

맛집을 검색할 때 사용할 수 있는 해시태그 유형을 제시한다.

Ⅱ. 본론

1. 분석 데이터

1.1 맛집 크롤링

맛집에 대한 데이터는 '망고플레이트'라는 맛집 정보 공유 플랫폼의 2022 서울 맛집 TOP 100 페이지(https://www.mangoplate.com/top_lists/2960_seoul2022)를 크롤링하여 수집하였다. 크롤링에는 BeautifulSoup 모듈을 사용하였다. 먼저, [그림1]과 같이 'span'태그를 통해 맛집 이름에 하나씩 접근하여 크롤링하려고 했으나, 해당 페이지는 10개의 식당 이름만 표시하고 있어서 실패했다. 따라서, [그림2]와 같이 'meta'태그를 통해 맛집 리스트를 크롤링했다. 다만, '2022망고플레이트인기맛집', '2022인기맛집', '망고플레이트인기맛집' 등 식당 이름을 나타내지 않는 단어들은 제외하고 맛집 리스트를 구성하였다: '듁스커피', '아리아께', '허브족발', '아티장베이커스', '슬로우치즈', '작은피자집', '야키토리묵', '스시 시미즈', '가온', '패스트리부티크', '러시아케익', '오코노미야키식당하나', '이치류', '소이연남마오', '비엘티스테이크', '쿠이신보', '더파크뷰', '오향가', '카레', '서울케밥', '카와카츠', '미야자키상점', '부촌육회', '야상해', 'BISTROT de YOUNTVILLE', '세스크멘슬', '와려', '쉐시몽', '스시작', '왕스덕', '원조할아버지순두부', '메종조', '스시소라', '시라카와', '빠사삭', '티크닉', '우미노미', '노스티모', '모짜', '스시카나에', '인딕 슬로우', '맛짱조개', '회현식당', '연희미식', '서관면옥', '물랑', 'BAR CHAM', '버거파크', '소울브레드', '차만다', '유즈라멘', '타쿠미곤', '카밀로한남', '주옥', '우육면관', '바위파스타바', '따빠디또', '보트르메종', '신비갈비살', '도림', '시키카츠', '마루심', '있을재', '로바타탄요', '만가타', '껠끄쇼즈', '쎄쎄종', '쿠나', '쇼토', '북천', '우주옥', '밴건디 스테이크 하우스', '나인스케이트', '당도', '쥬에', '야스노야지로', '알라프리마', '미라이', '엘리스리틀이태리', '삐아프', '스시키', '호반', '농민백암왕순대', '호가양꼬치', '가디록', '산울림1992', '이누식당', '샐러드셀러', '이찌이스시', 'Osso파스타', '세이류', '치즈플로', '팔레드 신', '콴안다오', '오스테리아 오르조', '비플레이트 바이 브라운브레드',

‘가담’, ‘파씨오네’, ‘와라야키 쿠이신보’, ‘운봉산장양고기전문점’.

```python
import requests
from bs4 import BeautifulSoup
from pprint import pprint

url = 'https://www.mangoplate.com/top_lists/2960_seoul2022'
r = requests.get(url, headers = header)

lists = soup.find_all('span', attrs = {'class':'title '})

restaurants = []
for i in lists:
  place = i.a.get_text()
  place = place.strip()
  num, name = place.split(". ")
  restaurants.append(name)

print(restaurants)
```

```
['미라이', '시라카와', '호반', '쉐시몽', '맛짱조개', '치즈플로', '연희미식', '마루심', '스시카나에', 'BISTROT de YOUNTVILLE']
```

[그림1] ‘span’태그를 활용한 크롤링

```python
import requests
from bs4 import BeautifulSoup
from pprint import pprint

url = 'https://www.mangoplate.com/top_lists/2960_seoul2022'
r = requests.get(url, headers = header)

lists = soup.find('meta', attrs = {'name':'keywords'})
```

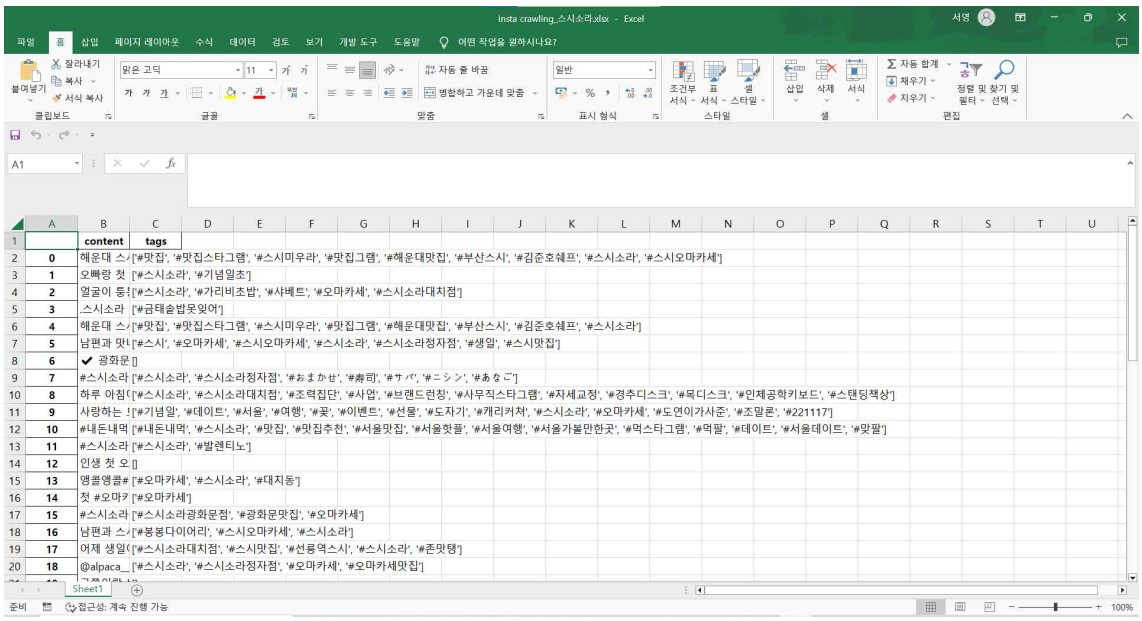
```
pprint(lists)
```

```
<meta content="2022망고플레이트인기맛집, 2022인기맛집, 망고플레이트인기맛집, 2022, 2022서울맛집, 2022서울, 2022인기, 2022망고플레이트, 망고플레이트, 서울맛집, 서울식당, 강남식당, 테헤란로식당, 압구정맛집, 서울식당,
```

[그림2] ‘meta’태그를 활용한 크롤링

1.2 인스타그램 본문 및 해시태그 크롤링

인스타그램에 각 맛집을 검색했을 때, 나오는 인기 게시물 100개에 대하여 각 게시물의 본문 내용과 해시태그를 크롤링하였다. 크롤링에는 Selenium 모듈을 사용하였고, 각 맛집에 대한 크롤링 결과는 [그림3]과 같이 엑셀 파일 형태로 저장하였다[1][2].



[그림3] 해시태그 크롤링 결과

2. 연구 과정

2.1 인스타그램 본문 감성분석

인스타그램의 각 게시물의 해시태그가 '맛집'임을 나타내는 해시태그인지 판별하기 위해 게시물 본문 내용에 대한 긍정/부정 분류 감성분석을 진행하였다. 감성분석 모델은 Naive Bayes 모델과 Support Vector Machine 모델을 모두 학습시킨 후, 정확도가 높은 모델을 사용하였다. 학습 데이터로는 '긍정' 또는 '부정'으로 라벨링 된 네이버 영화 리뷰 데이터 10,000개를 전처리하여 사용하였다. 전처리는 형태소 분석 결과 명사/동사/형용사 추출 및 불용어 제거 과정을 거쳤다. 이후에는 tf-idf를 통해 자질을 추출했다. 학습 후 정확도 평가 결과, Naive Bayes 모델의 경우에는 78.6%의 정확도를 보였고, Support Vector Machine 모델의 경우에는 76.4%의 정확도를 보였다. 따라서, 본 연구에서는 Naive Bayes 모델을 사용하여 감성분석을 진행하였다. 5000개의 게시물 중에서 감성분석 결과 '긍정'으로 분류된 게시물은 2677개였다.

2.2 인스타그램 해시태그 키워드 추출

인스타그램 본문 감성분석 결과가 '긍정'으로 분류된 게시물들에 대해 해시태그를 분석하였다. 분석에 앞서, 해시태그 중 '광고' 또는 '협찬'을 포함하고 있는 경우, 광고 게시물이라고 판단하여 해당 게시물의 해시태그들은 제외하고 분석을 진행하였다. 따라서 최종적으로 2657개의 해시태그 세트들로 데이터셋을 구성하였다. 구성된 데이터셋은 'result_zip.xlsx'의 0번째 행과 같다. 키워드 추출은 빈도 기반 키워드 추출과 의미 유사도 기반 키워드 추출로 크게 두 가지 기법을 사용하여 진행하였다.

2.2.1 빈도 기반 키워드 추출

먼저, 빈도 기반 키워드 추출 기법 중, 해시태그 세트 내 단어 중에서 가장 큰 tf-idf 가중치를 갖는 단어를 추출하였다. 해당 기법을 사용한 결과, 중복을 제외하고 2065개의 단어가 출력되었다. 추출된 결과는 'result_zip.xlsx'의 1번째 행과 같다. 추출된 키워드들은 식당 이름 또는 메뉴 이름으로, 맛집을 선별할 수 있는 변별력이 상대적으로 낮다고 판단하였다. 또, 해시태그 세트 내에 tf-idf 가중치가 다른 해시태그 세트에서 추출된 단어의 tf-idf 가중치보다 높은 수준임에도 불구하고 하나의 단어만 추출되어 추출되지 못하는 한계점이 있다.

따라서, 빈도 기반 키워드 추출 기법 중, 해시태그 세트 내 단어 중에서 tf-idf 가중치가 임계치를 넘는 단어를 추출하였다. 임계값은 0.6으로 설정하였다. 해당 기법을 사용한 결과, 중복을 제외하고 651개의 단어가 출력되었다. 추출된 결과는 'result_zip.xlsx'의 2번째 행과 같다. 몇몇의 식당이름 또는 메뉴이름을 지칭하는 키워드들이 있었지만, 이전 분석에 비해 해시태그를 통한 맛집 선별의 일반성이 증가했다.

2.2.2 의미 유사도 기반 키워드 추출

빈도 기반 키워드 추출 기법 중, 해시태그 세트 내 단어 중에서 가장 큰 tf-idf 가중치를 갖는 단어를 출력했을 때 보다 키워드 변별력은 향상되었지만, 단순히 빈도만을 고려한 기법으로, 선별된 해시태그의 유의미성이 떨어지는 한계점이 있다. 따라서 의미 유사도 기반 키워드 추출 방식을 사용하여 분석을 진행하였다. 의미 유사도 기반 키워드 추출은 모두 KeyBERT를 사용했으며, 'distilbert-base-nli-mean-tokens' 모델을 활용하였다. 또한, 식당

이름은 불용어로 지정하였다.

먼저, 2657개의 해시태그 세트들로 구성된 데이터셋에 대해 키워드 후보의 ngram 길이를 최대 3으로 설정하여 의미 유사도 기반 키워드 추출을 진행하였다. 각 해시태그 세트들에 대해 키워드는 1개씩 추출하였다. 그 결과, 2745개의 키워드가 추출되었다. 추출된 결과는 'result_zip.xlsx'의 3번째 행과 같다. 추출된 키워드를 분석한 결과, 식당 이름 또는 해당 메뉴에 대한 설명이 대부분이었고 실질적으로 맛집을 검색할 때 사용할 수 있는 해시태그 키워드를 찾기 어려웠다. 따라서 ngram 길이를 줄여서 다시 분석을 진행하였다.

두 번째로, 2657개의 해시태그 세트들로 구성된 데이터셋에 대해 키워드 후보의 ngram 길이를 최대 2로 설정하여 의미 유사도 기반 키워드 추출을 진행하였다. 각 해시태그 세트들에 대해 키워드는 1개씩 추출하였다. 그 결과, 2183개의 키워드들이 추출되었다. 추출된 결과는 'result_zip.xlsx'의 4번째 행과 같다. 추출된 키워드들 역시, '식당 이름 + 메뉴 이름' 조합의 키워드들이 대다수였다. 따라서 ngram 길이를 1로 설정하여 다시 분석을 진행하였다.

세 번째로, 2657개의 해시태그 세트들로 구성된 데이터셋에 대해 키워드 후보의 ngram 길이를 최대 1로 설정하여 의미 유사도 기반 키워드 추출을 진행하였다. 각 해시태그 세트들에 대해 키워드는 1개씩 추출하였다. 그 결과, 2183개의 키워드들이 추출되었다. 추출된 결과는 'result_zip.xlsx'의 5번째 행과 같다. 추출된 키워드들은 '식당 이름' 또는 '메뉴이 름'을 나타내는 경우들이 대다수였다. 또, 추출된 키워드 개수가 많아 맛집 선별에 활용하기 어려움이 있다고 판단했다. 따라서, 전체 해시태그 세트를 대상으로 의미분석을 진행하는 것이 아닌, tf-idf 빈도 기반 키워드 추출 기법으로 추출된, 즉, 등장 빈도가 높은 단어들에 대해서 의미 유사도 기반 키워드 추출을 진행하였다.

마지막으로, 빈도 기반 키워드 추출 기법 중, 해시태그 세트 내 단어 중에서 tf-idf 가중치가 0.6이상인 단어로 출력된 651개의 단어를 사용하여 의미 유사도 기반 키워드 추출을 진행하였다. 키워드 후보의 ngram 길이는 최대 1로 설정하였으며, 각 해시태그 세트들에 대해 키워드는 1개씩 추출하였다. 그 결과, 562개의 단어가 출력되었으며 출력된 단어는 'result_zip.xlsx'의 6번째 행과 같다.

Ⅲ. 결론

1. 연구 결과

빈도 기반 키워드 추출 기법을 사용하여 tf-idf 가중치가 0.6 이상인 단어들에 대해 의미 유사도 기반 키워드 추출을 진행했을 때 가장 유의미한 결과를 얻을 수 있었다. 분석 결과로 출력된 단어들 중, 식당 이름 또는 지역명을 포함한 경우를 제외하고 다음의 유의미한 해시태그를 선별할 수 있었다: '언제나그랬듯', '점심뭐먹지', '클메이거먹어봐', '맛있는점심', '엄마랑데이트', '잘먹었습니다', '손맛이란이런거', '먹스타그램쉐프님께서', '오늘도행복한시간', '또가자', '내서타일이야', '먹스타그램가끔은', '인스타먹성', '진솔한맛이야기', '결혼기념일', '옴뇸_종로', '메바쥬맛집', '뚜언니_종로', '뚜언니_용산', '명희맛집_', '민정맛집', '수민푸디_서울', '강푸파_당산', '구미얌_서울', '미식가김용준', '돈방구_서울', '돈방구_합정역', '푸드트래블러 용산', '빵또기in강남', '엣티제인_소울브레드', '썬푸_신사'.

'언제나그랬듯', '점심뭐먹지', '클메이거먹어봐', '맛있는점심', '엄마랑데이트', '잘먹었습니다', '손맛이란이런거', '먹스타그램쉐프님께서', '오늘도행복한시간', '또가자', '내서타일이야',

‘먹스타그램가끔은’, ‘인스타먹성’, ‘진솔한맛이야기’, ‘결혼기념일’과 같은 해시태그들은 해당 해시태그 자체만으로도 맛집 선별에 유의미한 기능을 한다. 즉, 해당 해시태그만 검색해도 맛집에 대한 정보를 얻을 수 있다.

‘옴뇸_종로’, ‘뚜언니_종로’, ‘뚜언니_용산’, ‘명희맛집_’, ‘민정맛집’, ‘수민푸디_서울’, ‘강푸파_당산’, ‘구미얌_서울’, ‘미식가김용준’, ‘돈방구_서울’, ‘돈방구_합정역’, ‘푸드트래블러용산’, ‘빵또기in강남’, ‘엣티제인_소울브레드’, ‘썬푸_신사’과 같은 해시태그들의 경우에는 맛집 인플루언서들이 개인적으로 사용하는 해시태그들이다. ‘옴뇸_종로’의 해시태그를 검색할 경우, 해당 인플루언서가 작성한 종로 맛집에 관한 게시물들이 나타난다. 개개인의 입맛에 따라 편차는 있겠지만, 자신과 입맛이 비슷하다고 판단되는 인플루언서를 알고 있는 경우에는 해당 방법을 사용하여 맛집을 검색할 수 있다.

2. 의의 및 한계

광고 맛집이 아닌, 정말 맛있는 맛집을 선별하기 위한 해시태그를 선별하였다. 이를 위해 ‘망고플레이트’라는 맛집 정보 제공 플랫폼의 공식적인 맛집 리스트를 활용하여 사람들이 해당 맛집에 대한 게시물을 업로드할 때 사용하는 해시태그들을 수집하고 분석하였다.

하지만, 본 연구는 여러 가지 측면에서 한계점이 있다. 먼저, 데이터 수집 과정에서 최대한 객관적인 데이터를 수집하고자 ‘망고플레이트’에서 제공하는 맛집 리스트를 크롤링하여 사용했지만, 이 역시 사람들의 주관적인 견해가 반영된 데이터이다. 인스타그램 게시물을 크롤링하는 과정에서 식당 이름으로 검색하여 크롤링을 진행하였는데, 식당에 대한 리뷰 게시물과 관련 없는 게시물들이 크롤링 된 경우들이 있다. 예를 들어, [그림4]와 같이 ‘가온’이라는 식당을 검색한 경우, 아이돌, 강아지, 피어싱 가게 등과 관련된 게시물들이 크롤링 되었다. 또 다른 예로는, 인스타그램을 통해 마케팅하는 사람들이 팔로워수를 늘리기 위해 인기 해시태그를 포함하여 게시물을 작성하는데, 이때 해당 식당이 포함된 경우들이 있다. 식당에서 인스타그램 계정을 운영하여 마케팅하는 경우들도 있었는데, 이런 경우에는 맛집을 나타내는 해시태그들이 사람들의 보편적인 견해를 대표하지 못했다. 이렇게 식당 리뷰와 관련없는 게시물들이 크롤링되어 데이터 구축에 한계점이 되었다. 분석 과정에서의 한계점으로는 인스타그램 게시물의 본문 내용에 ‘존마탱’, ‘유죄’ 등의 신조어 및 밈이 포함되어 있어서 감성분석을 진행할 때 긍정/부정 분류가 정확하지 않은 부분이 있었다.

[그림4] 인스타그램 크롤링의 한계점

참고문헌

[1] [파이썬] 인스타그램 해쉬태그(#) 검색결과 크롤링하기_최신ver, 지표덕후, 지덕智德 블로그, 2022년 8월 13일 수정, 2022년 11월 30일 접속, https://mokeya.tistory.com/166
[2] [python/selenium] 파이썬으로 인스타그램 크롤링하기 1편. 로그인하기, 코딩유치원 블로그, 2022년 9월 23일 수정, 2022년 11월 30일 접속, https://coding-kindergarten.tistory.com/224