안녕하세요 몇번째로 발표하게된 3팀 김가경 김수민 민채원입니다. 저희는 제 11회 문화 데이터 활용 경진대회에서 외국인 관광객을 위한 케이컬쳐 관광코스 추천 및 지도 시각화라는 주제로 공모전 출품을 목표로 한 것이 아닌 완성된 만족스러운 결과를 내는 것을 목표로 잡고 프로젝트를 시작하게 되었습니다. 한국 드라마 촬영지나 아이돌 촬영지와 그 주변의 식당과 숙박 시설을 추천해 주는 관광 코스를 목표로 해서 그와 관련된 데이터들을 모아 분석했습니다.

발표의 순서는 다음과 같습니다.

1
먼저 저희는 외국의 한류 키워드 중 언급이 높고 인기가 있는 것들을 알기 위해 키워드 빈도분석과 워드클라우드 시각화를 진행하였습니다.

호주의 한류 커뮤니티 키워드 데이터를 가져온 후에 나라 코드나, 모은 날짜 등 분석에 불필요하다고 여긴 열들은 제거했고 분석에 필요한 아이돌의 그룹명만 남길 수 있도록 드룹명과 관련된 단어이거나 그룹 멤버 개인의 이름이 키워드에 있다면 이를 그룹명으로 포함시켜서 분석할 수 있도록 전처리를 진행하였습니다. 그 후 내림차순으로 정리해 순위를 내본 결과 1위는 bts, 2위 youtube, 3위는 music 이 나왔습니다.

다른 데이터들의 전처리도 동일하게 하고 순위를 정리한 결과 미국데이터의 1위는 bts, 2위 youtube, 3위는 music 이,

영국데이터의 1위는 youtube, 2위 bts , 3위는 amp 가

베트남데이터의 1위는 bts, 2위 youtube , 3위는 scandal 이

인도네시아 데이터의 1위는 youtube, 2위 bts, 3위는 artist 이라는 결과가 나왔습니다.

이러한 빈도분석 자료를 가지고 워드클라우드를 이용해 한눈에 정보가 보이도록 시각화를 진행하였는데 빈도가 높은 자료들 중 불필요한 youtube데이터나 kpop은 시각화를 더 잘 보이도록 하기 위해 제거하고 진행하였습니다. 그렇게 다음과 같은 다섯개의 워드클라우드를 시각화했습니다.

그 후 다섯개 나라의 키워드 데이터들을 모두 하나로 합쳐 빈도수 최종을 낸 후에 전체 나라의 키워드 빈도를 가지고 빈도분석과 워드클라우드 시각화를 진행했습니다. 그 결과 키워드 언급량 1위는 4872라는 빈도수로 bts가 차지했습니다. 이 자료를 보고 분석에 필요한 아이돌 그룹명이나 배우명만 보고 빈도수를 순위로 내보자면 1위는bts, 2위got7, 3위는exo라는 사실을 알 수 있었습니다.

2. 크롤링 데이터(넘김)

최신의 데이터를 통해 시각화를 진행해보고자 카카오맵을 크롤링한 데이터를 수집해보았습니다. 동적 페이지를 크롤링하고 싶었기 때문에 셀레니움이라는 라이브러리를 사용하였는데요. 크롤링 코드가 다음과 같이 작동하도록 구현하였습니다. 먼저 셀레니움 웹 드라이버 라이브러리를 통해 크롬에 접속을 하고 뷰티풀숲 라이브러리를 통해 웹페이지의 데이터를 추출하고 파싱하고자 하였습니다. 본격적인 크롤링에 앞서 사용자로부터 검색어와 저장할 파일 명에 대한 입력값을 받은 뒤, 검색 결과의 가게명, 카테고리, 별점, 리뷰 개수, 음식점의 링크, 주소를 크롤링 하였습니다. 리뷰개수 경우 크롤링 과정에서 '54개'라는 데이터가 있다면 54라는 숫자 값만 반환하도록 전처리하였고, 쉼표가 포함된 숫자의 경우 쉼표는 제거하고 사용자로부터 초기에 입력받은 파일명으로 크롤링 데이터를 저장하도록 구성하였습니다. 왼쪽에 나오는 영상과 같은 과정을 거쳐 총 14개 시도 맛집 검색 결과를 크롤링하였습니다.

이렇게 크롤링된 데이터를 가지고 전처리를 진행해보았는데요. 먼저사전에 내려받은 csv파일과의 지역별 그룹화를 위해 주소 컬럼의 일부를 발췌하여 시도 단위의 컬럼 생성하였습니다.해서 사진과 같은 최종 데이터 컬럼의 형태를 완성하였습니다.

다음으로 맛집의 검증을 위해 score 즉 별점 값이 3.5 이상인 데이터들만 추출하였습니다. 카테고리

를 구분할 수 없는 null값과 카테고리 중 카지노, 등산로와 같이 맛집과는 무관한 데이터가 확인되어 이와 같은 값에 대해서는 drop 처리를 하였습니다. 원인을 파악하진 못했지만 맛집 크롤링 데이터의 카테고리 중 playground 카테고리 부합하는 동물원, 자연 휴양림등의 카테고리를 확인하여 그룹화를 진행하고자 cafe, restaurant, playground 총 3개의 카테고리로 분류 하였습니다.

-기존 맛집 데이터 프레임과 outer join을 통해 두 데이터 프레임을 merge 하였습니다. 두 데이터 간의 컬럼 내용에 차로 null 값이 다수 생기는 컬럼이 있었으나 value 값이 존재하는 데이터만이라도 활용할 가능성이 있어 데이터셋에 존재하는 모든 컬럼을 살려두는 outer 조인을 사용하였습니다. 최종 데이터 프레임을 그룹화 한 결과 cafe는 2057건, playground는 6158건, restaurant은 9699건이라는 결과가 나왔습니다.

P.16
앞에서 합쳐진 데이터프레임을 토대로 별점별 맛집을 지도에 시각화 해보았습니다. 먼저, 카카오맵 크롤링 데이터의 위도와 경도의 값이 전부 NaN이기 때문에 지오코딩을 진행하였습니다. 지오코딩은 주소와 장소 이름과 같은 텍스트 기반의 위치정보를 지리적 좌표로 변환해주는 기술로, 저희가 갖고 있는 데이터프레임의 주소 정보를 가진 'ADDR' 열을 통해 위도와 경도값을 알아냈습니다. 카카오맵에서 크롤링한 데이터를 활용했기에 카카오 개발자 콘솔에서 자바스크립트 API를 발급받아 지오코딩을 하였습니다.

p.17
알아낸 위도와 경도 값으로 지도 시각화 코드를 실행한 결과, 데이터가 약 17,000개로, 범위가 너무 겹쳐져 해석에 어려움이 생긴 문제가 발생하였습니다. 이를 해결하고자 별점 데이터가 있는 Score 값이 각각 3.5이상, 4.0이상, 4.5이상인 경우를 기준으로 데이터프레임을 분할하였고, 옆에 보이시는 사진과 같이 마커 색상도 다르게 지정하였습니다.

3
그 다음으로 저희는 전국의 숙박시설 데이터를 가지고 촬영지 근처 맛집과 그 지역에 있는 호텔도 함께 추천할 수 있게 분석을 진행하여 검색하고 찾아갈 수 있도록 시각화를 진행하였습니다.

먼저 2019년 지역별 선호 관광상품 데이터를 이용하였는데 이 데이터에 있는 지역의 숙박 시설 정보와 그 근처 관광 시설 정보들을 이용하여 그 전에 분석해 놓고 합쳐 만들어둔 맛집 데이터와 합쳐서 인터랙티브한 테이블로 시각화하고자 했습니다. 그래서 이 데이터에서 필요한 열들을 제외한 다른 열들은 제거하고 주소 정보도 시군구만 남겨두고 전부 삭제하였습니다.

2020년 데이터도 마찬가지로 열 제거, 주소 전처리를 해주었고

2021년 데이터도 마찬가지로 동일하게 전처리를 진행한 후에 주소의 열 이름이 달랐기 때문에 이를 동일하게 변경해주었습니다.

그리고나서 3개년도 데이터들을 한번에 합쳐주었는데 호텔의이름이 동일한 데이터가 있다면 가장 최근의 데이터가 남도록 하여 데이터를 합쳐주었습니다.

그 후 그 전에 크롤링한 데이터와 미디어 데이터를 합쳐 맛집정보 데이터를 만들었던 것을 가져와서 이 데이터의 주소도 시군구만 남기고 지웠고 불필요 열도 다 제거하는 전처리를 해주었습니다.

그리고 나서 3개년도 숙박 데이터와, 맛집데이터 두 데이터를 주소 열을 기준으로 인터랙티브 테이블로 합쳐서 그 주소의 촬영지 정보, 또 그 근처의 숙소 정보와 맛집 정보를 볼 수 있도록하고 키워드를 검색해 관광지를 찾을 수 있도록 하는 시각화를 진행하였습니다. 이 검색 테이블 시각화는 rstudio내에서는 잘 검색이 되는데 아직 저장해서 외부로 공유하는 법을 찾지 못하여 이 부분은 아직 더 진행하여야 합니다.

4

p.25
다음은 촬영지 데이터를 동적 지도로 구현해보았는데요.

p.26
활용한 데이터는 한국문화정보원에서 제공한 미디어콘텐츠 영상 내 유명지 데이터로, 한국 드라마, 영화, 티비 프로그램, 아티스트 촬영지 정보를 제공합니다. 뒤에 나올 전처리 과정에서 자주 쓰이는 컬럼들을 짧게 설명드리겠습니다. MEDIA_TY는 촬영 유형을 나타낸 것으로, 값은 drama, movie, show, artist가 존재하며, TITLE_NM고 PLACE_NM은 각각 촬영한 미디어 제목과 장소를 의미합니다. PLACE_NM 옆에 있는 PLACE_TY는 촬영한 장소의 유형을 의미하며, cafe(카페), restaurant(식당), stay(숙박) 등의 값을 가집니다. 이 데이터를 통해 지역별, 아티스트명별 지도 시각화를 진행하고자 다음과 같은 전처리 과정을 거쳤습니다.

p. 27
먼저 지역별 시각화를 위해 주소 열 값 중 앞에서 두 텍스트만을 추출해 지역명 컬럼을 새로 생성하였고, MEDIA_TY 열을 기준으로 군집화하여 밑에 보이시는 4개의 데이터프레임을 생성하였습니다. 마지막으로 NaN값이 많거나 불필요한 컬럼을 제거하였습니다.

P. 28
지역명 컬럼의 unique값을 뽑은 결과, 총 14개의 지역이 출력되는데, 대한민국의 행정구역 구분을 토대로, 1개의 특별시와 특별자치도, 특별자치시와 6개의 광역시, 8개의 도로 나누어 진행하였습니다. 8개의 도의 경우, 경상북도와 경상남도는 경상도로 통일하여 진행하여 도는 총 5개로 분류하여 진행하였다는 점 참고해주시면 감사하겠습니다. 기본적으로 folium 라이브러리를 활용해 지도를 구현하였고, 지도 검색의 편리성을 더하기 위해 minimap을 사용하여 지도 하단에 작은 맵을 추가하였고, 여러 지역이 겹치는 문제 해결을 위해 markercluster로 여러 마커를 클러스터로 묶어 표시하였습니다. 다음은 경기 지역을 시각화한 코드 예시이며, 위도와 경도 값이 잘못표시된 값들은 구글맵스를 통해 직접 위도와 경도 값을 찾아내어 수정하였습니다.

p.29
다음은 서울 특별시를 시각화한 결과이며, 클러스터를 반복적으로 클릭하여 해당 장소 마커를 누르면 우측 사진과 같이 장소명, 장소유형, 촬영 내용, 주소 정보를 사용자에게 보여줍니다. 서울 지역 외에도 13곳 시각화한 결과를 웹페이지에서 확인해볼 수 있습니다.

p.30
다음은 아티스트명별로 지도 시각화하는 과정을 설명드리겠습니다. 앞에서 보여드린 한류(K-POP) 키워드 데이터에서 상위 10개의 아티스트명을 다음과 같이 추출하여 촬영지 데이터에 해당 아티스트명이 존재하는지 여부를 파악하였습니다.

p.31
파악한 결과, 각각 44위와 45위에 해당하는 jyp와 bigbang 값이 존재하지 않아 51위에 해당하는 세븐틴으로 대체한 결과 다음과 같이 상위 10명의 아티스트가 확정되었습니다. 이를 토대로 검색한 아티스트명과 일치하는 장소를 필터링하는 함수를 우측과 같이 구현하였으며, 해당 장소의 개수 정보를 제공하여 전국과 지역별로 지도에 시각화하였습니다.

p.32
다음은 TITLE_NM의 값을 방탄소년단일 경우의 시각화 결과이며, 다른 아티스트명을 입력할 시, 해당 함수가 실행되어 다른 지도도 볼 수 있습니다.

5. 웹페이지 구현(넘김)

마지막으로 이렇게 만들어진 결과물의 시각화를 위해 웹페이지를 구현해 보았습니다. 왼쪽의 제목을 클릭하시면 구현된 웹사이트를 방문하실 수 있습니다. 현재 검색기능의 구현은 완료하지 못한 상태라 깃허브의 페이지 기능을 활용하여 임의로 페이지를 공개해둔 상태 입니다. 추후 플라스크라는 프레임워크를 활용하여 그룹명을 통한 검색 기능을 구현해 볼 예정입니다.

이상으로 발표를 마치겠습니다. 감사합니다.