중간보고 1차 계획서

4조 (남지원, 박준민, 유건욱)

1. Team Data 선정과정

팀원 개인별 주제 탐색 후 토론을 거쳐 하나의 주제를 선정하기로 했습니다.

1) 택배 관련 데이터

평소에 택배를 시키면 배송추적이 가능한데, 서울에서 출발한 상품이 옥천 등의 지방으로 갔다가 다시 도착지인 서울로 오는 바람에 택배가 오래 걸리는 경험이 있었습니다. 이 경험을 바탕으로 각 택배사 물류창고의 위치정보와 배송지, 발송지의 data를 통합해 어떠한 insight를 도출해보고자 했습니다.

하지만 각 택배사마다 여러 개의 물류창고가 있고 발송지의 정보 등의 data를 구하는데 어려움이 있을 것 같았습니다. 또한 개인의 택배 이동 정보가 필요할 것으로 보았는데 이 data 또한 개인정보이므로 수집이 힘들 것 같아 보류하였습니다.

2) 교통정보와 관광 및 코로나 데이터

교통사고 data, 지역별 관광객 data, 일별 코로나19 확진자 data, 날씨 data를 통합하여 데이터 분석을 진행해보려 했습니다. 위의 data들은 모두 시계열 데이터로 데이터 통합을 쉽게 할 수 있을 것으로 예상했습니다.

하지만 통합의 난이도가 너무 낮으며, 뻔한 결과가 나올 것으로 예상되어 보류하였습니다.

3) 공연장 및 공연 데이터

공공 데이터 포털을 탐색하다가 전국의 공연시설 데이터를 발견했습니다. 이 데이터를 어떻게 활용할 수 있을지 고민하다 공연시설의 주변 시설, 공연정보, 객석 수 등을 통합해 데이터 분석을 진행해보려 했습니다.

이 주제의 장점으로는 공연정보에 대해서는 사람들이 많이 이용하고 있지만, 공연장 데이터에 대한 수요와 활용이 적어서 유의미한 분석을 진행할 수 있을 것 같습니다.

하지만 직접 크롤링 해야 하는 작업이 많아 시간이 많이 걸릴 것으로 예상됩니다.

위의 세 가지 주제들의 장점과 단점을 비교해 본 결과, 적당한 난이도와 유

의미한 분석이 예상되는 공연장 및 공연 data를 Team Data로 선정하게 됐습니다.

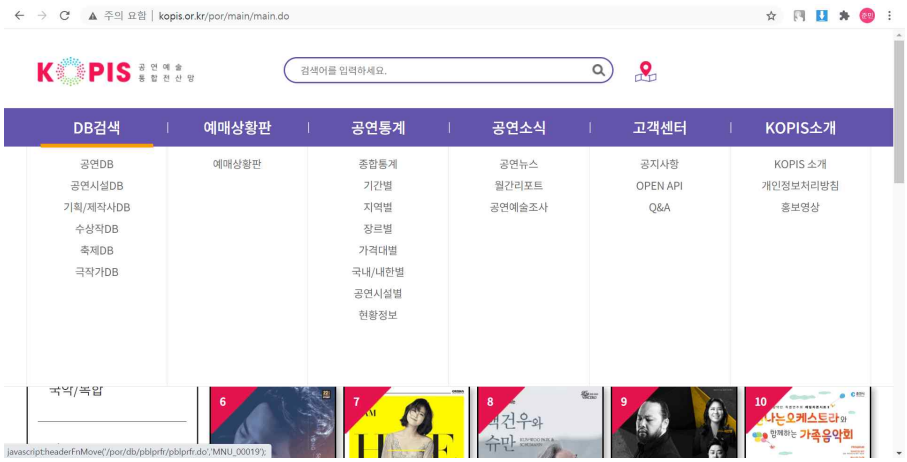
2. 데이터 통합 계획

공연장 데이터와 공연 데이터, 주변 시설 정보 등을 수집하기로 했습니다.

1) 공연장 및 공연 데이터

공연장 data와 공연 data를 어떻게 수집할 수 있을지 찾아보았습니다. 검색 결과 다음의 몇 가지 사이트를 찾을 수 있었습니다.

- 공연예술 통합전산망 (http://www.kopis.or.kr/por/main/main.do)



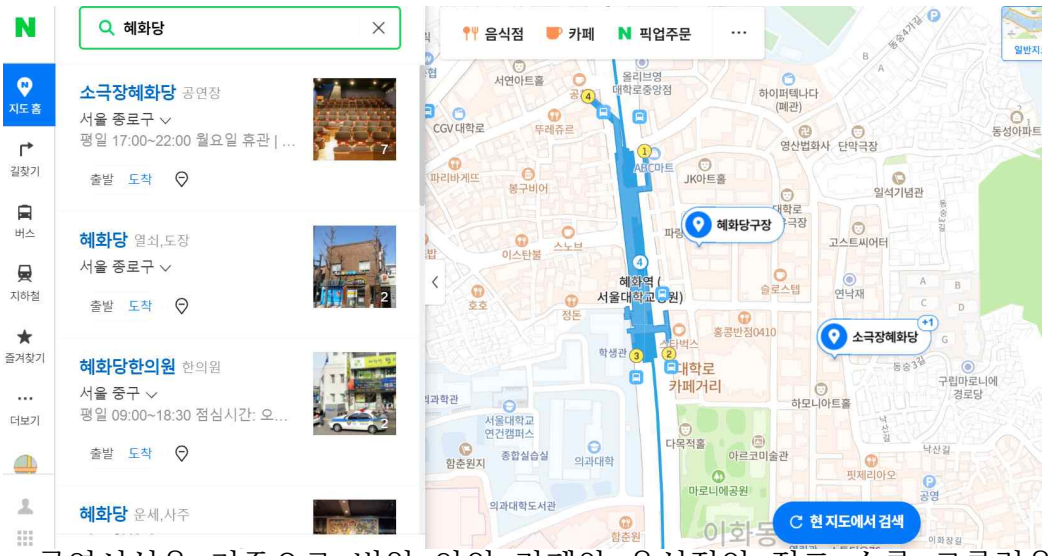
- 플레이db (http://www.playdb.co.kr/Index.asp)

위와 같은 사이트에서 공연장의 주소와 공연장의 편의시설, 장애인 시설, 주차시설 등의 주요시설, 편성되었던 공연의 관객 수, 예매율 등의 정보를 수집

할 계획입니다. 공연장의 주소와 특성 객석 수 등의 자료는 위의 사이트에서 csv, 또는 xlsx 파일로 다운 받을 수 있으며, 편의시설, 주차시설, 장애인 시설 등의 주요시설과 편성되었던 공연에 대한 정보는 크롤링을 통해 얻을 수 있습니다.

2) 주변 시설 data

네이버 지도에서 공연장을 검색한 후 주변의 카페와 음식점 등의 주변 시설 정보를 수집할 계획입니다.



공연시설을 기준으로 범위 안의 카페와 음식점의 점포 수를 크롤링을 통해 데이터를 통합할 계획입니다. 기준 범위와 크롤링 방법은 팀원 간 논의를 통해 결정할 것입니다.

3. 최종 통합된 데이터 예상

공연장 정보와 공연정보, 주변 시설 정보를 통합해 공연장 이름, 지역, 시설 특성, 객석 수, 주요시설, 개관연도, 주변 시설, 해당 공연장에서 공연한 공연 이름, 장르, 공연 기간, 관람객 수, 예매율 등을 열로 가질 것입니다.

ID	공연장 이름	세부 공연장	주소	지역(시,구)	시설특성	객석 수	주요시설	개관연도	주변시설 카페	주변시설 음식점	공연 이름	장르	기간	가격	평균관람객	예매율
1	영화의 전당	하늘 연극장	부산광역시 해운대구 수영강변대로 120 (우동)	부산	공공	841	주차장	2011	10개	8개	마티네 콘서트	클래식	2020.8	35000	400	40%
2		비프씨어터	부산광역시 해운대구 수영강변대로 121 (우동)	부산	공공	4000	주차장									
3																
4																
5																

1) 변수 상세 설명
- 공연장 이름 : 공연장의 이름 ex) 영화의 전당
- 세부 공연장 : 공연장 안의 세부 소극장 등 ex) 하늘연극장, 비프씨어터
- 주소 : 공연장의 도로명 주소
- 지역 : 광역시, 구 등의 세부 주소 (몇 개로 나눌지 추후 논의 예정)
- 시설특성 : 공공 시설인지 민간시설인지
- 객석 수
- 주요시설 : 주차장, 편의시설 등의 유무를 categorical data로 2~3개 예정
- 개관연도
- 주변시설 : 네이버 지도를 통해 크롤링한 음식점, 편의점 개수
- 공연 이름 : 해당 공연장에서 공연한 공연 이름
- 장르 : 클래식, 무용, 뮤지컬 등
- 기간 : 공연기간
- 가격 : 해당 공연의 가격 (좌석 등급이 나누어져 있는 경우 논의 예정)
- 평균 관람객
- 예매율

데이터 통합과정을 거치면서 추가적으로 추가할 수 있는 data가 있으면 추가할 계획입니다.