1. 국회입법활동 빅데이터 시각화 공모전

**분석목적:**

입법이란 나라의 법률을 제정하는 것으로 국민 모두에게 영향을 미치는 중요한 과정이지만 일반인들에게 매우 생소한 분야이기도 함. 국회활동에 대해 일반인들의 관심을 제고할 수 있는 인사이트 발견 및 시각화가 주 목표 <https://dacon.io/competitions/official/235679/overview/description/>

**사용가능한 데이터:** 공모전 제공 데이터

* 역대 국회의원 현황(new\_people.csv) : 대수, 대별 및 소속정당, 이름, 생년월일, 본관, 출생지, 학력 및 경력, 종교 및 취미, 저서, 상훈, 사망일
* 국회의원 발의법의안(suggest.csv) : 의안ID, 의안번호, 법률안명, 소관위원회, 제안일, 처리상태, 대수, 상세페이지, 제안자, 제안자목록링크, 대표발의자, 공동발의자
* 본회의 처리 안건(process.csv) : 대수, 공포일, 예결위심사\_상정일,의결일, 회부일, 의안ID, 의안활동구분, 의안명, 의안번호, 기권, 소관위원회, 위원회심사\_상정일,의결일,회부일, 정부이송일, 법사위체계자구심사\_상정일,의결일,회부일, 반대, 의결결과, 제안일, 제안자, 제안자구분, 본회의심의\_상정일, 의결일, 총투표수, 찬성

**분석방향:**

각 대수 국회 입법활동 평가 : 1~21대별 국회의 발의건수와 가결율, 상임위원회별 법안처리

국회의원은 사회를 잘 대표하는 사람들일까? – 성별, 연령별, 출신지역별… 지역구의원과 비례대표 의원의 차이점이 있을까? 국회의원별 발의법안 등을 통해 클러스터링

시기별 법률안 키워드 : 시기에 따라 발의되는 법안의 트렌드가 존재할까? 상임위원회별, 정당별로도.. 토픽모델링 (LDA)

NLP 통한 국회의원 추천시스템 : 법안의 단어 벡터들을 임베딩하여 국회의원 벡터 구성 - 유사도 구하여 사용자에게 추천   
<https://github.com/hw79chopin/X-senators#-x-senators-%EA%B5%AD%ED%9A%8C%EC%9D%98%EC%9B%90-%EC%B6%94%EC%B2%9C-web>

**분석 장단점:**

* 우선 이미 끝난 공모전이고 사람들이 해놓은 결과물이 코드까지 모두 업로드되어 있음 – 남이 한 결과 따라하는 것 이상으로 더 끌어낼 수 있을까 하는 걱정..
* 텍스트데이터 생소 – 자연어처리 분야 + 거의 다 파이썬 활용해야 할 듯
* 도메인 지식 상당수 필요해보임 – 국회 입법 과정 등
* 공모전 목표 자체가 시각화이다보니 EDA 외에 할 수 있는 것이 있을까 하는 걱정… 국회의원 추천시스템을 한다면 너무 표절 같음…
* EDA 잘하려면 질문을 정말 잘 던져야 할 것 같다. (도메인 지식 부족)

1. 상권분석

**분석목적:** 코로나 19 발생 이후 지역경제의 위축으로 폐업하는 소상공인들이 많아진 만큼 이러한 시기적 특성을 반영한 소상공인 창업 시 폐업률과 생존율을 예측하고 싶었음 특히 온라인 소비가 급격히 증가하고, 신선상품, 당일배송, 새벽배송 등이 점점 활발해지면서 지역 상권에 구조적인 악영향 미치고 있음.

따라서 지역상권 소상공인들을 위한 업종별 생존분석+ 생존요인 분석을 해보자! 가 목표

**사용가능한 데이터:**

LOCAL\_DATA : <https://www.localdata.kr/main.do> 지방행정 인허가데이터 개방 : 폐업/영업중인 상권 데이터 전부 받을 수 있음!!!! (대박)

골목상권분석정보 : <https://data.seoul.go.kr/dataList/3/literacyView.do> 서울상권 중 골목상권, 발달상권, 전통시장, 관광특구 및 각 상권의 배후지에 대한 성별, 연령별 인구정보와 지표 제공 – 상권 내 유동인구, 상주인구, 상권 기반이 되는 상권 배후지의 상주인구 및 아파트 단지, 점포 정보 확인 & 상권변화지표 및 상권별 추정매출 제공하며 현재 가게 운영 또는 창업에 관심있는 서울시민들에게 차목자ㅛ를 제공하여 가게 운영 및 창업에 도움이 되도록 함! – 데이터 절라많다.. (서울시만)

**분석방향:**

매출 데이터 같은 직접적인 정보는 개인적인 금융정보다 보니 구하기 어려움.. 대신 유동인구, 임대료 등 간접적이지만 역시 상권의 생존에 영향을 미칠만한 지역적 특징들을 중심으로 접근해보는건 어떨까? 라고 생각했는데 서울시 우리마을가게 상권분석 서비스에서 제공하는 데이터가 (추정매출 포함) 매우 다양…! 1496개의 상권(골목, 발달, 전통시장, 관광특구 총 4개 유형)별 데이터 제공 –

EDA : 지역별로 상권 특징 알아보기 – 클러스터링 & 시기별 추정매출 비교 (코로나 19전후 변화) & 서비스 업종별롤 추정매출에 영향을 주는 변수들 알아보기

생존분석: LOCAL DATA와 연결 지을 수 있다면? 상권별 폐업 정보도 알 수 있음! 폐업에 영향을 주는 변수들 알아보고 생존요인 분석

* 생존분석이란? 생존 시간자료를 분석하는 통계적 방법으로 치료방법, 예후인자 등이 생존에 미치는 영향을 추정하는 분석 ( 보통 바이오통계에서 활용)
* 이태원+청주시 예시에서는 **카플란-마이어 분석**으로 업종, 상권유형, 상권에 따른 생존율과 생존기간을 알아보고 위험함수 도출 -> **시간가변 콕스-비례위험모형**으로 폐업위험도에 영향을 미치는 요인 분석

**분석 장단점:**

* 우선 카드 매출 데이터 구하기 어려움…(끝난 공모전 데이터 or 유료 데이터-개비쌈…)
* 업종을 특정하는게 좋을까?
* 비슷한 주제(18년2학기 골목상권분석, 20년 1학기 핫플레이스 예측)
* 서울시 우리마을가게 상권분석 서비스 <https://golmok.seoul.go.kr/main.do>

이미 서울시에서 이렇게 구체적인 데이터로 소상공인/예비 창업자에게 정보 제공 중… 우리의 차별점?

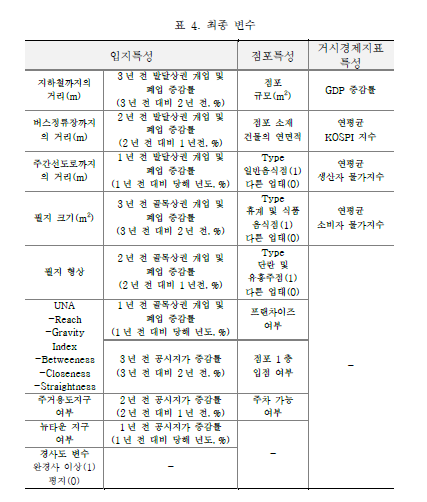
**참고자료:**

골목상권의 영역 설정 및 생존요인 분석 : 청주시 중심

<http://www.riss.kr/search/detail/DetailView.do?p_mat_type=be54d9b8bc7cdb09&control_no=9acd8d35c3d0f06dffe0bdc3ef48d419&outLink=K>

발달상권과 골목상권에 위치한 음식점 생존과 폐업 비교 분석 : 이태원 중심  
<https://s-space.snu.ac.kr/handle/10371/125505#export_btn>

변수 참고하기에 좋을 듯!!



소비자 감성기반 상권분석 시스템 개발”   
<https://www.koreascience.or.kr/article/JAKO201810852361476.pdf>

이건 약간 다른 건데 소비자들이 상권에 대하여 수행하는 웹 탐색 활동과 감성평가를 반영하는 데이터인 지역구 연관 감성어휘를 기반으로 서울시 내 대형 상업 공간으로 정의할 수 있는 지역구 간의 연관 감성 네트워크에 대한 소셜 네트워크 분석! 근데 만약 골목상권 위주로 간다면 관계 없을 듯…

1. 애니 추천 시스템

<https://www.kaggle.com/CooperUnion/anime-recommendations-database>

Kaggle 애니 추천 시스템 데이터베이스!

<https://laftel.net/> 라프텔 : 개인화 추천 기반 애니매이션 스트리밍 서비스

**분석목적:** 요즘 예전 투니버스 애니 보는 팀장..^.^ 게임추천 or 애니추천 시스템 알아봤는데 둘 다 캐글에 데이터 존재! 게임추천의 경우 스팀 게임 데이터! 근데 사실 본인이 잘 몰라서 대신 애니 추천을 가져옴..

분석목적은 최근 넷플릭스나 왓챠에도 애니 컨텐츠 많아지고 향유층들은 꾸준히 있음. 애니는 장르도 다양하고 종류도 많은데 어떤 작품이 있는지 모르고 하다 보니 뭐 하나 시작하기 쉽지 않음… 내 취향에 맞는 애니 추천해주는 시스템을 만들어보자..는 목적??

**사용가능한 데이터:**

우선 캐글에서 제공하는 데이터 (myanimelist.net API 여기서 제공)

* 캐글에서 제공하는 anime / rating 데이터 + 깃헙 업데이트된 애니 데이터 제공(myAnimeListData)

라프텔(애니ootd서비스)을 크롤링해볼수있을까?  
<https://velog.io/@zetbouaka/%EB%9D%BC%ED%94%84%ED%85%94-API-%EB%9C%AF%EC%96%B4%EB%B3%B4%EA%B8%B0-2> - 코드는 있는데 되는지는 모르겠음… 된다면 하고싶다…

라프텔에서 크롤링할 수 있다면? 할 수 있는 데이터 : 평점, 에피소드 수, 리뷰, 제작정보, 태그 (태그 작품에 대한 주요 정보이지 않을까?,,,??)

**분석방향:**

<https://eda-ai-lab.tistory.com/524> : 추천 시스템 설명 참고

어떤 추천시스템으로 가면 좋을까? 컨텐츠 기반 or 협업 필터링 기반

유저데이터를 사용한다면 **협업필터링 기반 추천** 가능하지만… 유저데이터의 신빙성 문제…(무시해도됨)

**컨텐츠 기반** 추천시스템 하려면 근데 작품에 대한 데이터가 더 많이 필요하지 않을까? 영화의 경우 감독이나 배우들 정보 네이버 영화에서 바로 크롤링 가능한데 그런 정보 제공하는 곳이 있을까? 식별 가능한 정보가 태그밖에 없긴 함 – 유사도 행렬 어떻게 구할지?

**분석 장단점:**

* 캐글 대회 4년전 대회라서 데이터가 얼마나 최신 컨텐츠를 갖고 있는지 모르겠음… 최신 데이터를 위해 직접 크롤링해야 하는가? 근데 어떤 대단한 사람이 깃헙에 올려줌… 언제 업데이트된건진 모르겠음.. <https://github.com/Dibakarroy1997/myanimelist-data-set-creator>
* Myannimelist 유저 데이터 제공하는 곳이 미국 기준이라 한국 정서와 안 맞을 수 있음… 재미없는 작품도 작화만 좋으면 선호하고 오래된 작품은 비선호하는 경향이 있다고 함…
* <https://anidb.net/> 애니DB 해외사이트.. 데이터 어떻게 끌어올지 모르겠음