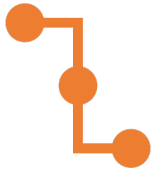
Data Science

Portfolio

of Suhoa-Yoo

프로젝트 목차



Markov-Chain 을 이용한 작곡

현대수학세미나 기말고사 프로젝트

2017.11 ~ 2017.12

베이커리 트렌드 변화,
6년 전과 비교

셀프 프로젝트

2020.07.16 ~ 2020.07.26

Nsshop+ 홈쇼핑 매출 예측

2020 빅콘테스트

2020.08.13 ~ 2020.09.28

Markov-Chain을 이용한 작곡

개요: 상세 내용

현대수학세미나 수업에서 진행한 프로젝트입니다. 기존의 곡 데이터를 기준으로 마르코프체인을 생성하고 이를 통해 자동으로 작곡을 하는 프로그램을 만들었습니다. 프로그램을 통해 기존의 곡과 비슷한 느낌의 새로운 곡을 만들어 낼 수 있습니다.

핵심

음악에는 tick(음이 연주되는 시간), tempo(음악의 박자), 음의 높낮이(pitch), 음의 길이(duration), 음의 세기(velocity) 등의 요소가 있습니다. 여러가지 요소 중 pitch, duration, velocity를 가장 중요한 요소로 고려했습니다. 어떤 음악과 다른 음악의 pitch, duration, velocity의 진행이 유사하다면 두 음악은 '비슷한 느낌'일 것이라고 가정합니다. 마르코프체인을 이용하여 **어떤 두 음이 연속해서 나올 경우**, **다음 음이 x일 확률**을 매트릭스를 구하고 이 확률에 따라 작곡을 진행했습니다.

프로젝트 과정

1) 음악의 요소를 추출하기 위해 Bruno Mars의 'Marry you', 'Just The Way You Are'을 선택
2) 각각의 midifile에서 pitch, duration, velocity를 추출
3) Markov-chain을 토대로 pitch, duration, velocity의 경우의 수 matrix 생성
4) 경우의 수 matrix의 원소를 rowSum으로 나누어 transition matrix 생성
5) Transition matrix를 이용하여 작곡
   - 주어진 초기 벡터로 시작하며 transition matrix를 따라 다음에 나올 pitch, duration, velocity를 결정

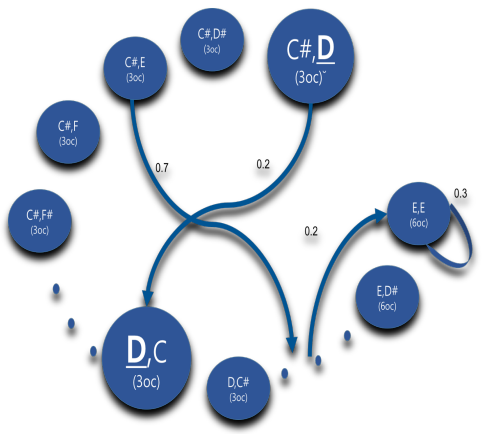
Markov-Chain을 이용한 작곡

사용한 데이터

- Bruno Mars의 ‘Marry you’, ‘Just The Way You Are’의 midi 데이터

사용한 기술

- midiutil package 사용
- midifile로 부터 pitch, duration, velocity 추출과 생성된 pitch, duration, velocity로부터 midifile 생성
- Order 2인 마르코프체인



정의

stochastic process (Zn)n∈N일 때, 시간 n에서 확률 분포 Zn+1이 Zn, … , Zn−2의 상태에만 의존할 때 Zn을 **메모리가 2인 마르코프체인** 혹은 **order 2인 마르코프체인**이라고 한다.

예) Order 2인 마르코프체인(pitch)

그림은 Bruno Mars의 두 곡에서 추출한 pitch(음의 높이)를 토대로 마르코프체인을 생성하여 시각화 한 것이다. 그래프에서 보면 C#, D가 연속으로 연주 되었을 때 그 다음 pitch가 C일 확률은 0.2이다.

* oc: 옥타브

Markov-Chain을 이용한 작곡

본인의 역할

- midifile에서 필요한 데이터 추출 및 midifile 생성
- transition matrix 생성

베이커리 트렌드 변화, 6년 전과 비교 

개요: 상세 내용

개인적인 호기심과 관심으로 진행한 미니 프로젝트로 네이버의 유명 베이커리 카페 “먹은 빵 후기” 게시판의 게시글을 크롤링하여 이용했습니다. 2013.08~2014.07(이하 2013년)과 2019.08~2020.07(이하 2019년)의 데이터를 통해 6년 전과 최신의 베이커리 인기도나 메뉴 트렌드의 변화가 있는지 살펴보고자 했습니다.

사용한 데이터

- 베이커리 카페 “빵소담”의 2013.08~2014.07, 2019.08~2020.07의 서울지역 게시글 데이터
- 2013.08~2014.07 : 2788건
- 2019.08~2020.07 : 3970건

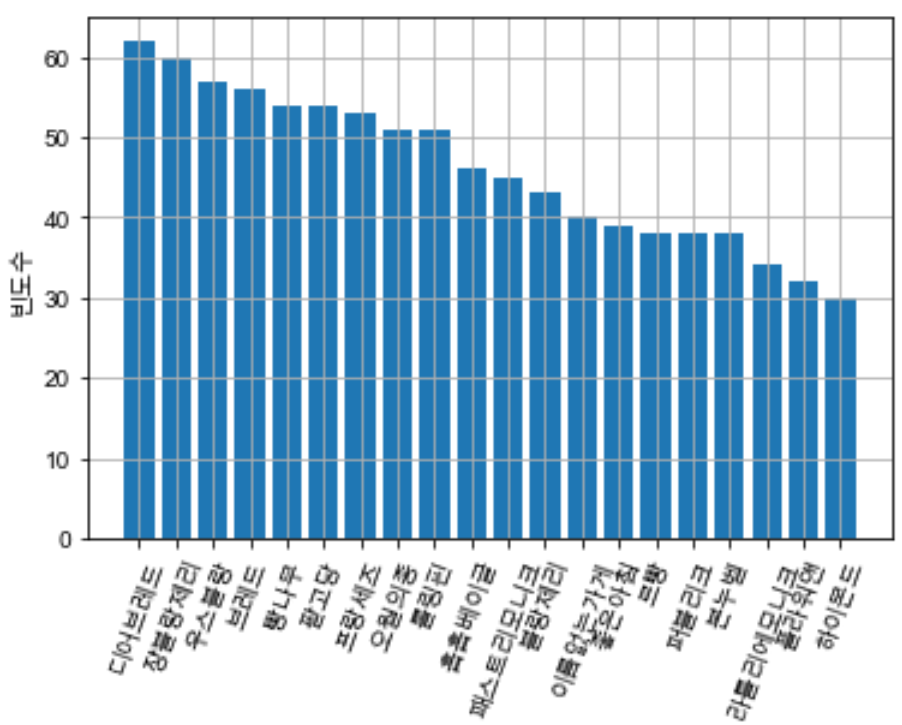
사용한 기술

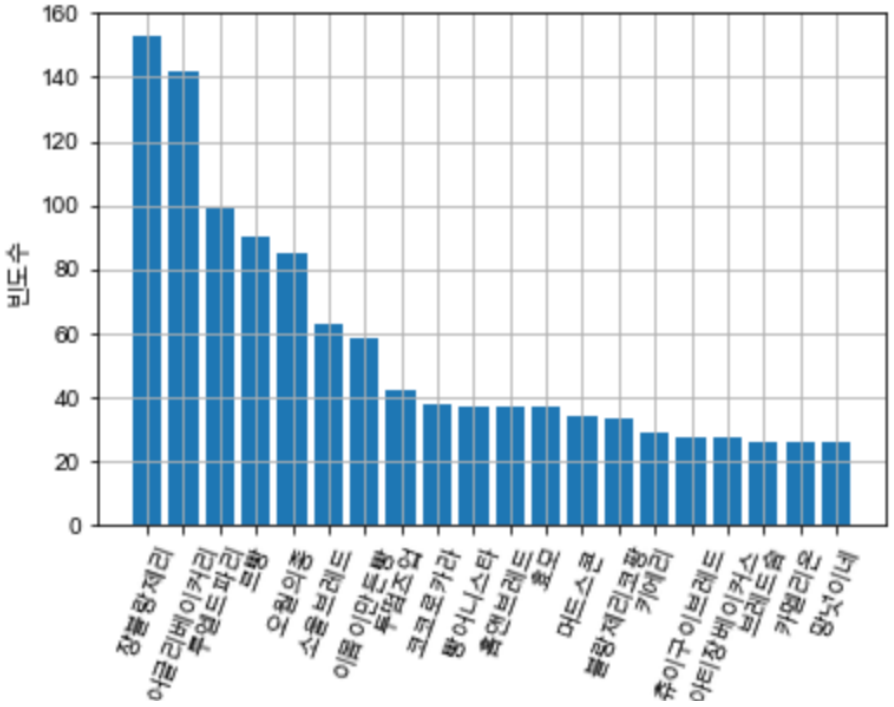
- Beautiful soup 등 python package를 이용한 웹 크롤링
- 텍스트 데이터 활용을 위한 전처리
- R, python의 wordcloud를 이용한 텍스트 마이닝
- Python konlpy 를 이용한 연관어 분석

베이커리 트렌드 변화, 6년 전과 비교

주요 내용

1) 서울 지역 인기 베이커리 순위 : 2013년 vs 2019년



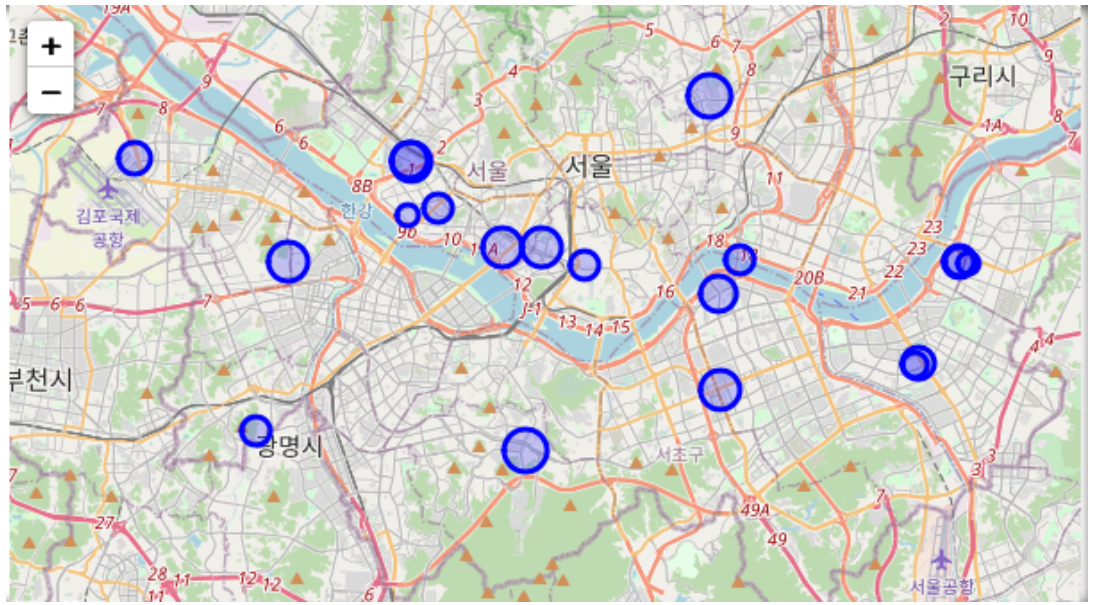


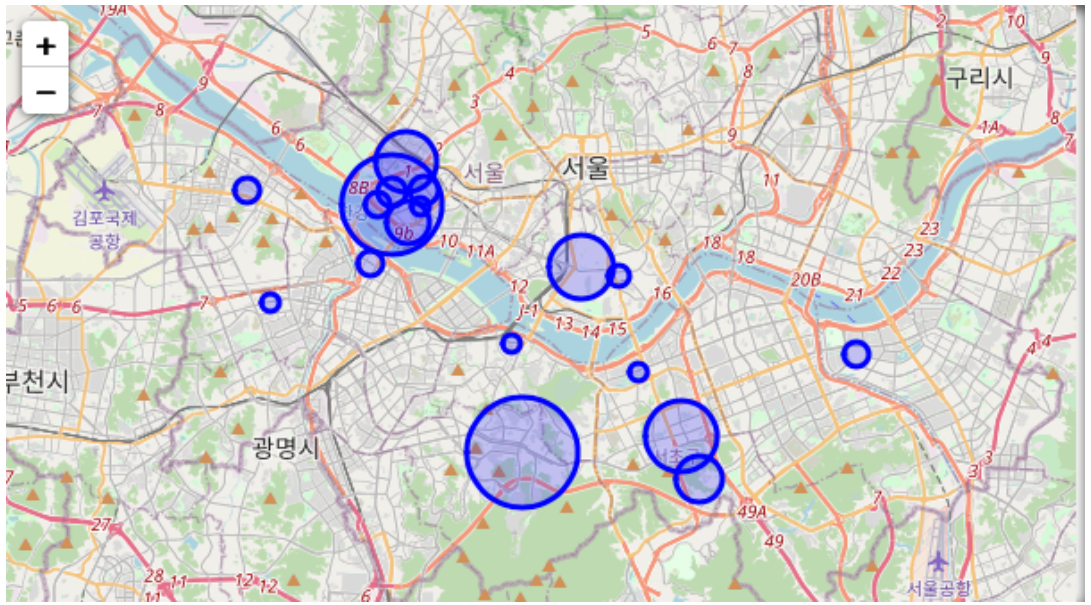
2013년에는 연간 인기 베이커리의 게시글의 수가 비슷합니다. 1위와 20위의 게시글 빈도수가 21건 차이로, 1위와 20위의 게시글 빈도수 차가 127건인 2019년과는 상당히 다른 양상입니다. 2013년과 다른 2019년에 나타난 특징을 통해 베이커리 트렌드의 변화를 추측해 보고자 합니다. 첫째, 2019년에 상위권을 차지한 베이커리는 대부분 “속재료가 풍부한” 베이커리로 빵 속에 들어가는 크림, 구황작물, 견과류 등이 많기로 유명한 베이커리들입니다. 둘째, “빵어니스타, 머드스콘, 키에리, 망넛이네” 등 비건을 내세운 베이커리들이 다수 상위권을 차지했습니다. 셋째. “머드스콘, 망넛이네” 과 같이 매장이 없이 택배로만 주문을 받는 곳이 등장했습니다. 이 세 가지 변화로 베이커리의 유행이 어떻게 변화하고 있는지 대략 알 수 있었습니다.

베이커리 트렌드 변화, 6년 전과 비교

주요 내용

2) 서울 Best20 베이커리의 위치 : 2013년 vs 2019년



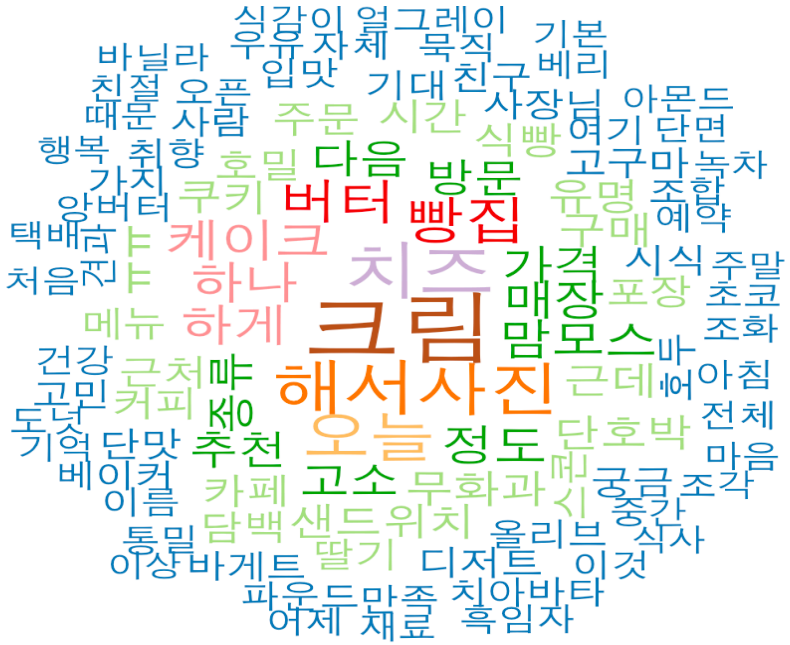


지도상 위치한 원은 인기 베이커리가 있는 자리입니다. 원의 크기는 게시글의 건수를 나타냅니다. 지도 시각화에서도 2013년에는 원의 크기가 비슷하고 2019년에는 원의 크기가 많이 차이 나는 것을 볼 수 있습니다. 또 비교적 인기 베이커리가 서울 중심부에 고르게 퍼져있던 2013년과는 다르게 2019년에는 마포구 쪽에 7개의 베이커리가 몰려있습니다. 이 중 “어글리베이커리, 투떰즈업, 이몸이만든빵, 코코로카라”는 2013년 이후 오픈한 베이커리입니다. 6년간 여러 미디어매체에서 “빵투어”, ”빵지순례”라는 이름으로 가까이에 있는 유명 베이커리들을 하루에 들르는 문화가 소개된 점, “핫플레이스”라는 단어의 부상으로 카페를 이용하는 다수의 청년층이 “연남동”, “망원동”에 몰리게 된 점 등을 여러 베이커리가 마포구에 생기고 성공하게 된 이유로 예상해봅니다.

베이커리 트렌드 변화, 6년 전과 비교

주요 내용

3) 게시글 내용에 많이 언급된 단어 : 2013년 vs 2019년



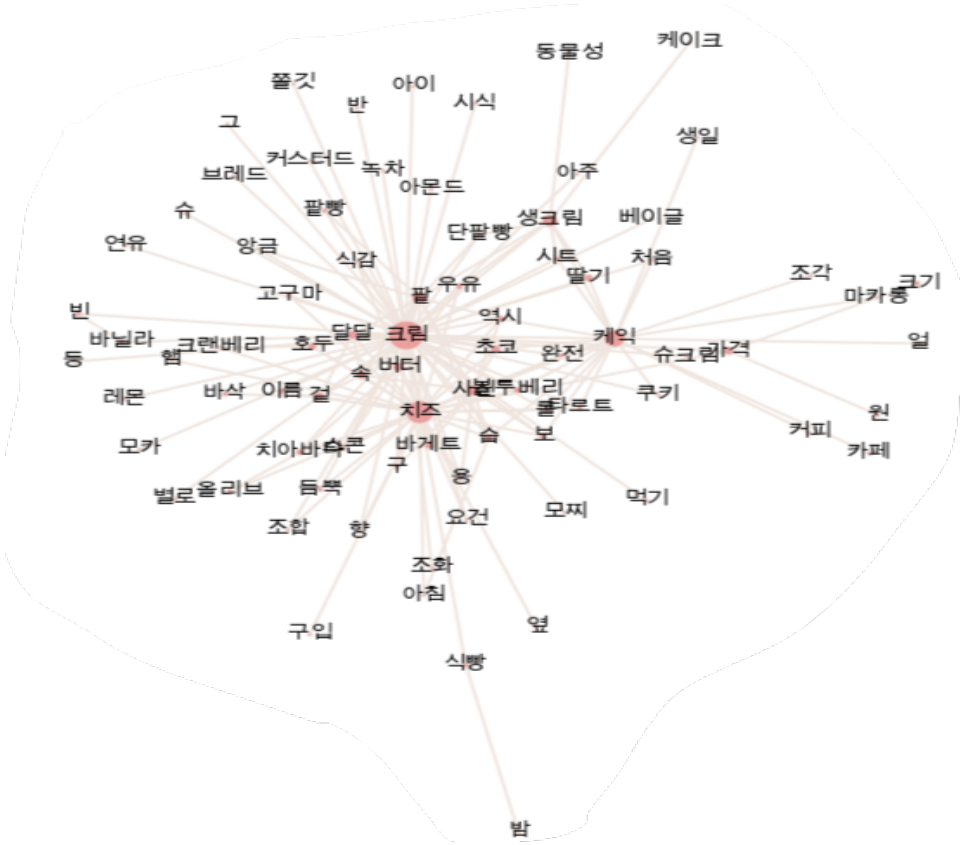
신기하게 게시글 내용에서 자주 언급된 단어들은 2013년과 2019년에 아주 비슷합니다.

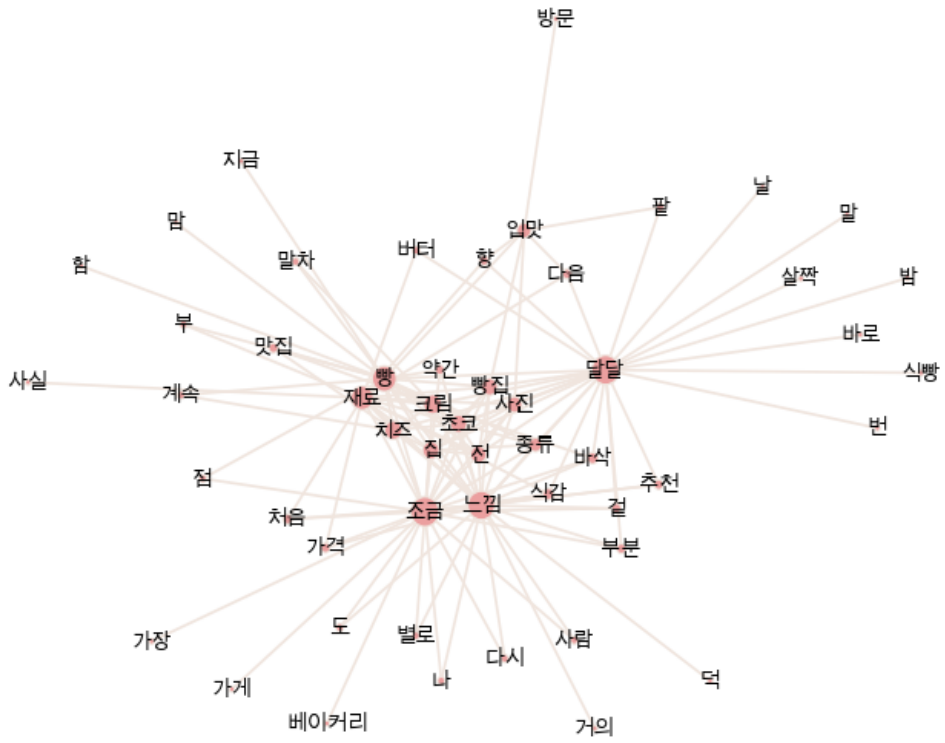
두 해에 공통되지 않은 단어 중 2019년에 눈에 띄는 단어는 “취향”, “만족” / “친절”, ”사장님” / “택배” / “흑임자”, “맘모스”, “앙버터” 입니다. 빵 후기를 남기며 자신의 취향이었는지, 만족했는지 주관적인 의사 표현이 많아지고 빵 자체의 평가 뿐 아니라 매장의 친절도에 대한 평가도 늘어났다는 것을 알 수 있습니다. 1)번의 베이커리 순위에서도 소개했지만 택배만으로 운영되는 베이커리가 많아졌고, 많은 매장에서 택배 서비스를 병행하여 제공하기 시작했기에 ”택배”라는 키워드가 등장했습니다. 마지막으로 최근에 유행하여 여러 프랜차이즈 베이커리에까지 등장한 메뉴인 맘모스, 흑임자, 앙버터도 볼 수 있습니다.

베이커리 트렌드 변화, 6년 전과 비교

주요 내용

4) 본문 내용에서 언급 단어 간의 연관어 분석 : 2013년 vs 2019년


동물성
케이크
쫄깃
아이
반
시식
그
생일
커스터드
녹차
브레드
아주
아몬드
슈
팥빵
생크림
베이글
연유
앙금
단팥빵
식감
시트
처음
딸기
조각
크기
고구마
팥
우유
마카롱
빈
역시
바닐라
크랜베리
호두
달달
크림
케익
얼
초코
완전
슈크림
속
버터
레몬
바삭
이름
걸
사본투베리
쿠키
치즈
원
모카
바게트
커피
카페
구
용
별로
올리브
모찌
먹기
조합
향
요건
조화
아침
구입
옆
식빵
밤



Nsshop+ 홈쇼핑 매출 예측 

개요: 상세 내용

NIA 한국정보화진흥원, KBO 빅데이터포럼에서 주최한 공모전 '2020 빅콘테스트'에 참가했습니다.
Nsshop+ 홈쇼핑에서 제공한 2019년의 실제 판매 데이터를 토대로 2020년 6월 홈쇼핑 상품 판매실적을 예측했습니다.

핵심

본 공모전의 홈쇼핑 판매실적 예측은 6월 한 달 동안 20분 간격으로 편성 되어있는 다양한 상품의 20분간의 판매 실적을 각각 예측해야하는 문제였습니다. 상품이 모든 시간에 고르게 편성되지 않으므로 각 상품의 판매 기록을 시간 별로 집계할 경우 **데이터의 편중이 심각**하였고, 특정 상품-시간 조합에 판매 기록이 없어 예측 성능이 저하되는 것을 발견했습니다. 이에 **SVD(특이값 분해)를 통해 상품-시간 별 매출 예측 값을 학습데이터에 추가해주어** 예측 성능을 높였습니다. 또한 상품의 세부 품목별로 판매실적의 추이가 비슷한 것을 발견하여 각 상품의 '중분류', '세분류' 파생변수를 추가해주었습니다. 마지막으로 홈쇼핑 상품은 소위 "대박"이라고 말하는 순간이 존재합니다. 다른 날과 비슷한 시간에 비슷한 상품을 팔았음에도 20분간의 판매실적이 아주 높게 나타나는 것인데, 이런 특이한 데이터는 제거하여 트리 기반의 회귀 모델을 사용할 시 더 보편적 예측을 가능하게 만들었습니다.

사용한 데이터

- 방송일시별 상품명과 상품의 5가지 정보, 매출액이 feature인 데이터
- 일자별 날씨 데이터

Nsshop+ 홈쇼핑 매출 예측 

사용한 기술

- 데이터 희박성 해소를 위한 SVD(특이값 분해)
- Random forest, XGBoost 등 트리기반 회귀 모델
- RNN 기반의 딥러닝 예측

프로젝트 과정

1) 예측의 보편성을 위해 매출액이 지나치게 높은 데이터(홈쇼핑 대박)과 지나치게 낮은 데이터(홈쇼핑 쪽박)를 제거
2) 연구결과 매출에 영향을 미치는 날씨, 요일, 중분류, 세분류 등의 파생변수 추가
3) 특정 시간대에 편성 기록이 없는 상품은, 데이터 희박성 해소를 위해 SVD(특이값 분해)를 이용하여 데이터를 추가
4) Random forest, XGBoost 등 트리기반 회귀 모델과 RNN 모델로 예측

본인의 역할

- 시각화를 비롯한 EDA와 이상치 제거 및 데이터 전처리 (상세내용 링크 : EDA1, EDA2, EDA3)
- 파생변수 아이디어 제공
- 데이터 희박성 해소를 위한 SVD(특이값 분해)
- Random forest, XGBoost 등 트리기반 회귀 모델