Experience & Projects

재정정보 AI 검색 알고리즘 경진대회 2024.08

Description

· DACON 경진대회 참여

· PDF 파일 기반 질의응답 LLM 구축

What did I do

· 베이스라인 코드 기반 LLM 모델 구축 및 모델별 성능 테스트

What I learned

· RAG, Langchain 관련 개념(논문) 학습 및 실습

· GPU 부족으로 여러 모델과 finetuning을 시도하지 못해 아쉽지만, LLM을 학습하고 사용해볼 수 있었음

항공화물 수요 예측 프로젝트 2022.10 - 2022.11

Description

· 한국공항공사 일경험수련생 팀프로젝트(2人)

· 항공화물 실적 사내 데이터 기반 향후 1년치 김포공항 일일 화물량 예측

What did I do

· 항공화물 실적 데이터(한국공항공사), 공휴일 데이터(한국천문연구원) 전처리 및 파생변수 생성

· '주 단위 화물량'으로 Auto ARIMA를 수행 → ARMIA(0,1,1) 기반 '주 단위 예측값' 도출

· 이후 AutoML을 수행(CatBoost Regressor)하여 '최종 주 단위 예측값' 도출, 이를 변수로 하는 일 단위 데이터 완성

· 해당 데이터를 기반으로 AutoML(CatBoost Regressor) 수행 → 최종 '일 단위 화물량' 예측

· 향후 한 달치 예측하는 모델을 12번 반복하여 최종 1년치 예측값 도출 완성

What I learned

· 시계열 데이터 및 ARIMA 이론을 학습하고 실습함

· AutoARIMA, AutoML을 학습하고 적용함

· 통계학 전공 팀원과 협업하며, 데이터를 통계적 측면에서 바라보는 관점을 배울 수 있었음

· 일경험수련생으로서 활용가능한 데이터가 한정적이었기에 더 다양한 분석 프로젝트를 수행하지 못한 데 아쉬움을 느낌

어린이공원 실효성 분석 프로젝트 2022.08

Description

· 데이터분석 청년인재 양성사업 교육 팀프로젝트(6人)

· 서울시 어린이공원별 현황 분석 및 실효성 판단

What did I do

· 서울시 자치구별 공원 현황 데이터/공원시설물 현황 데이터(서울특별시) 등 수집 및 전처리

· QGIS 기반 위치별 어린이 관련 시설 확인 및 파생변수 생성 후 요인별 상관분석

· 공원별 이용 후기 데이터 텍스트마이닝 분석

· 전문가 대상 AHP 기반 주요 요인별 가중치 설정 후 최종 실효성 점수 도출

· 팀원별 다양한 역량을 발휘할 수 있도록 알맞은 역할을 분담, 전반적인 프로세스 관리

What I learned

· Data-Driven에서 벗어나 Biz-Driven 기반 분석 프로젝트를 여러 전공의 사람들과 수행해봄으로써, 모델링 중심이었던 데이터에 대한 편협한 사고를 한층 넓히게 됨

음악 장르 분류 및 생성 프로젝트 2022.03 - 2022.06

https://github.com/Doleeee/GTZAN-Data-analysis

Description

· 전공과목 [기계학습] 팀프로젝트(4人)
· 이미지 및 tabular 데이터 기반 음악 장르 분류 및 생성
· 음악 데이터(wav)로부터 추출된 feature 데이터와 이미지 데이터 기반 분류 모델과 생성 모델 구축

What did I do

· 분류모델 구축 부분을 주로 맡아 수행함
· 분류모델 학습 데이터의 변수가 음악의 feature를 과하게 세분화한 상태였음 → 차원축소를 시도하여 VAE(VAE_MLP, VAE_CNN)의 decoder를 통해 loss가 최소인 최적의 차원을 찾고자 함 → 기존 데이터를 학습시킬 때의 성능이 더 우수했음
· 기존 데이터 기반 분류 모델링 - LogisticRegression, Stochastic Gradient Descent, KNN, RandomForest, XGBoost, etc

What I learned

· 음성 데이터에 대한 지식 및 이론을 습득함
· 차원축소 모델과 생성 모델의 이론에 대해 학습하고, 차원축소의 경우 실제 적용해봄

반도체 시장 관련 정보 및 Trend를 제공하는 의사결정지원 도구 개발 2021.06 - 2022.06

https://github.com/spaceFreeze/newSpace

Description

· 전공과목 [산학연계SW프로젝트] 팀프로젝트(4人)
· 반도체 회사 및 시장 관련 인터넷 뉴스 기사를 요약 제공하는 사이트 개발

What did I do

· Selenium 기반 인터넷 댓글 웹크롤링 후 텍스트 전처리 수행
· 제목 및 내용이 동일한 기사를 제거한 이후에도 수많은 중복 기사 존재 → 내용상 중복되는 기사를 제거하는 알고리즘 필요
· 계층적 클러스터링으로 유사한 기사를 군집으로 묶고, 군집 내 코사인 유사도로 군집별 대표 기사를 선정 → 하나의 주제와 관련해서 하나의 기사만 최대한 남도록 설계
· 수집된 기사의 내용 요약 및 키워드(해시태그) 추출
· 웹서버(Flask, AWS Lightsail) 구축 및 데이터 수집·저장(SQLite3)하며 자동으로 작동하는 웹사이트를 완성
· 관련 내용을 정리하여 디지털콘텐츠학회 논문에 게재(다중문서 요약시스템(KCI)) 및 SW 저작권 등록(등록번호: C-2022-024242)

What I learned

· 중복 기사 제거 알고리즘을 구축하면서 여러 임베딩 기법과 클러스터링, 유사도 검사 기법을 테스트하고 텍스트 요약을 위해서도 모델을 비교분석함. 이를 통해 데이터와 프로젝트의 목적에 따라 최적의 기법과 방법이 다양하게 고안될 수 있다는 것을 느낌
· 당시 연결된 회사로부터 데이터나 가이드라인을 받지 못한 상태에서 프로그램을 만들어야 했기에, 팀원들 간에 소통하는 시간이 굉장히 중요했음. 서로의 의견을 존중하고 경청함으로써 방향을 잡아가고 프로젝트를 무사히 완성할 수 있었음

자연어처리 및 AI를 활용한 유튜브 악성 댓글 블라인드 자동 처리 시스템 2021.03 - 2022.02

https://github.com/commentcover/COmmentCOver

Description

· 학생주도형 설계학기 [참빛설계] 팀프로젝트(3人)
· 유튜브에서 악성 댓글로 인한 1차적 피해 예방을 목적으로 악성 댓글 자동 블라인드 처리하는 크롬 확장 프로그램 개발

What did I do

· Github, 유튜브 댓글 웹크롤링 등 데이터 수집 후 텍스트 전처리 수행
· 악성 댓글 여부 판별 모델링을 위한 데이터 라벨링(일반 댓글: 0, 악성 댓글: 1) 작업
· 악성 댓글의 판단 기준 문제 → 관련 논문과 서비스(네이버 클린봇 등)의 기준을 참고하여 욕설 및 혐오 표현, 그리고 문맥상 긍정적인 비속어 등을 기준으로 설정하여 라벨링을 완료함 → 총 23,534개 댓글 수집
· 모델(LSTM, GRU, BERT) 설계 및 학습 - 정확도 87%

· Flask 기반 웹서버 구축 및 크롬 확장 프로그램 개발 → 실시간 자동 블라인드 시스템 완성
· HCI 학술대회에 참가하여 프로젝트 관련 논문 작성 및 발표(자연어처리 기술을 활용한 유튜브 악성 댓글 자동 블라인드 시스템 - 한국HCI학회 학술대회)

What I learned

· 첫 장기 프로젝트인 만큼 수많은 오류와 시행착오를 겪으며 스스로 지식적 한계와 부족함을 확인하였고, 이를 통해 실패하거나 다시 도전하는 것을 두려워하지 않게 되어 더욱 발전하고자 하는 열의를 얻게 됨
· 잘 모르던 웹 서버(Flask)와 클라이언트(크롬 확장 프로그램), 그리고 통신 방법을 학습하고 직접 구현하며 관련 이론과 경험을 터득함
· 기존 유튜브의 경우 유튜버가 직접 악성 댓글 키워드를 설정하거나 특정인의 댓글을 차단하는 방식이었기에 최소 한 번 이상 악성 댓글에 누군가 노출될 수밖에 없었음. 이러한 기존 시스템의 한계를 보완하기 위해 해당 프로젝트를 구상하게 되었고, 목표했던 대로 구현한 것에 보람과 자부심을 느낌
· 팀원 모두 각자의 자리에서 열심히 임하고 성실히 수행하며 좋은 팀워크를 발휘한 프로젝트였고, 서로 부족한 부분을 채워주고 어려운 점도 같이 해결해나가며 협력하는 공동체의 시너지를 깊이 경험할 수 있었음

항공사 승객 만족도에 영향을 미치는 주요인 분석 2021.04 - 2021.06

https://github.com/jaieun-l/DataMining_proj

Description

· 전공과목 [데이터마이닝] 팀프로젝트(4人)
· 항공사 승객의 만족도 요인 분석 및 저가 항공사별 텍스트 리뷰 분석

What did I do

· 항공사 승객 만족도 데이터(kaggle) 기반 EDA, 전처리, 모델링 수행 후 모델별 변수 중요도 확인
· 데이터 모델링 - LogisticRegression, DecisionTree, RandomForest, AdaBoost, GBM, etc
· 웹크롤링 기반 항공사 텍스트 리뷰 데이터(TripAdvisor) 수집 후 전처리 수행
· 항공사별 1/5점/전체 리뷰 텍스트 분석(워드 클라우드)

What I learned

· 데이터 마이닝 이론 학습 후 처음 수행한 분석 프로젝트로, 학습한 기법과 이론들을 직접 실습해보며 데이터 사이언스에 입문함
· Chrome driver, Selenium 기반 자동 웹 크롤링 구현하는 방법 터득함
· 국내 항공사 승객 만족도 데이터를 얻을 수 없어 외국 데이터를 활용한 것이 아쉬웠고, 좋은 분석을 위해선 먼저 좋은 데이터가 필요하단 것을 느꼈음

유튜브 영상 자막 및 댓글 분석 2021.04 - 2021.06

https://github.com/jaieun-l/TextMining_proj

Description

· 전공과목 [텍스트및오피니언마이닝] 팀프로젝트(4人)
· 특정 유튜버의 영상 댓글, 자막 데이터를 토대로 자연어처리 분석을 수행

What did I do

· 유튜브 API 기반 영상 자막 및 댓글 데이터 크롤링
· 텍스트 데이터 전처리 - EDA, 맞춤법 검사, 불용어 처리, 토크나이징 등
· 토픽모델링(LDA), 감성분석(LSTM, CNN, BERT, KoBERT, GPT2 등), 텍스트 요약(textRank, lexrank) 수행

What I learned

· 자연어처리 기본 개념 학습 및 유튜브에서의 실습을 통한 다양한 모델 기법을 적용해봄
· 텍스트 데이터에 대한 흥미를 느끼기 시작함