Resume

About ME

- 김채형 (Chae Hyeong Kim)
- E-mail : cheris8@naver.com OR cheris8h@gmail.com
- GitHub : https://github.com/cheris8
- Blog : https://cheris8.github.io

Education

- 연세대학교 문헌정보학과 (응용통계학과 부전공) 졸업 예정 (2017.03 ~ 2021.08)
- 교내 수강 목록
  - 미분적분학, 선형대수
  - 통계학입문, 통계방법론
  - 수리통계학(1)
  - 시계열분석
  - 컴퓨팅적사고와SW프로그래밍, 프로그래밍기초, R과파이썬프로그래밍, 정보기술론, 정보처리연습
  - 텍스트정보처리론, 텍스트마이닝, 컬처로믹스
  - 데이터사이언스입문
  - 딥러닝
- 교외 수강 목록
  - Stanford CS231n & CS224n
  - datacamp (Click here for more info.)
  - 개별적으로 5주 간 R을 활용한 데이터사이언스 교육 이수
  - 머신러닝 스터디 (Click here for Git Repo.)
  - 파이썬 알고리즘과 자료구조 스터디 (Click here for Git Repo.)
  - NLP 기초 스터디 (Click here for Git Repo.)

Activities

- 연세대학교 통계&데이터사이언스 학회 ESC (2020.03 ~ 2021.02)
- 연세대학교 빅데이터 학회 YBIGTA (2021.01 ~)
- 데이콘에서 주관한 제주 신용카드 빅데이터 경진대회 참가
- 빅콘테스트에서 주관한 퓨쳐스리그 데이터 분석 경진대회 참가

Skills

- Python, R
- SQL
- Tableau
- 데이터분석준전문가(ADSP)

Project

인스타그램 해시태그 추천 시스템

- 입력 이미지에 대하여 자동으로 해시태그를 생성해주는 해시태그 추천 시스템
- 이미지를 입력하면 입력 이미지와 유사한 K개의 이미지의 해시태그를 바탕으로 태그 전환 및 확장을 반복하여 추천 태그를 생성하여 출력
- `Python` `Keras` `Deep Learning` `CV` `NLP`
- Click here for Git Repo.

1. Data Collection
    - 인기 해시태그 Top 100을 기준으로 해시태그 1개 당 포스트 100개씩 수집
    - 인스타그램 포스트로부터 이미지 url, 이미지 파일, 장소 태그, 본문 해시태그, 댓글 해시태그 수집
2. Data Preprocessing & EDA
    - 전체에서 출현 빈도가 1인 해시태그 제거
    - `customized konlpy`를 활용하여 사용자 사전을 적용하여 해시태그 토큰화
    - 토큰화 결과 중 동일한 토큰이 연속하여 나오는 경우 토큰 확장에서 성능 저하로 이어져 제거
3. Modeling
    - 토큰 간 유사도를 계산하기 위해 워드 임베딩 방법론 중 `Word2Vec` 사용
    - 이미지 간 유사도를 계산하기 위해 ImageNet이 학습된 `VGG19`를 완전분류기를 제거하여 사용
    - 단어 벡터 간 유사도를 활용하여 **토큰 전환** 알고리즘 개발
    - 출현 빈도에 기반한 조건부 확률을 활용하여 **토큰 확장** 알고리즘 개발

트랜스포머 구현

- Attention is all you need 논문을 리뷰하고 이를 바탕으로 트랜스포머 구현
- `Python` `Pytorch` `Deep Learning` `NLP`
- Click here for Git Repo.

텍스트마이닝을 통한 드라마 가치 요인 개발

- 2018년 방영 드라마를 대상으로 네이버 뉴스기사, 네이버 블로그, 유튜브 댓글을 크롤링하여 토픽 모델링, 공기어 분석, 감성 분석을 통해 드라마 가치 요인을 규명
- `Python` `Gephi` `Machine Learning`
- Click here for Doc.

1. Background
    - 오늘날 방송 플랫폼이 다원화되고 VOD, 넷플릭스 등을 통한 후속 시청이 많아짐에 따라 방송콘텐츠의 가치를 고전적인 가치 평가 기준인 시청률로 측정하는 데에 한계가 발생했다.
    - 그리하여 시청자의 다양한 시청 반응을 포괄하여 방송콘텐츠의 가치를 평가하고자 하는 노력이 계속되어 왔으나, 질적인 요소를 고려하지 않은 채

인터넷을 통한 프로그램 직접 검색자 수, 소셜미디어 버즈 양 등의 수치에만 의존하는 경향이 존재한다.
    - 따라서 텍스트마이닝을 통해 방송콘텐츠의 가치에 영향을 미치는 여러 요인들을 규명함으로써 방송콘텐츠가치평가지표 개선에 도움이 되고자 한다.
2. Data Collection
    - 분석 대상 : 2018년 방영 드라마 중 평균/최고/최저 시청률 상위 5위 진입 횟수, 네이버 뉴스 기사/네이버 블로그 포스팅/유튜브 동영상 댓글 수를 종합적으로 고려하여 총 15개의 드라마 선정
    - 텍스트 데이터 수집 대상 : 각 드라마에 대한 네이버 뉴스 기사, 네이버 블로그 포스팅, 유튜브 동영상 댓글
    - 텍스트 데이터 수집 기간 : 각 드라마 별로 방영시작일 1개월 전 ~ 방영종료일 12개월 후
3. Data Preprocessing
    - 한글 형태소 분석기로 Komoran/Mecab 사용
    - 각 드라마 별로 불용어 사전 생성 및 적용, 빈도 수에 기반한 불용어 처리
    - uni-gram 및 bi-gram 활용
    - emoji 모듈 사용
4. Model & Algorithms
    - Topic Modeling (네이버 뉴스 기사, 네이버 블로그)
    - Coward Analysis (유튜브 동영상 댓글)
    - Sentiment Analysis (유튜브 동영상 댓글)
5. Conclusion
    - 드라마 촬영지 : 드라마가 특정 시대를 배경으로 할 경우 이를 얼마나 잘 재현해 냈는지가 중요한 요소 중 하나이다. 또한 아직 시청자들에게 드라마 촬영지로 소개되지 않은 이색적인 국가 및 도시를 배경으로 할 경우 화제성이 높다.
    - 드라마 OST : 해당 드라마의 각종 OST의 음원차트 진입 횟수 등을 통해 드라마 OST에 대한 가치 평가를 진행한 후 이를 드라마 가치 평가에 반영해야 한다.
    - 드라마 원작 : 드라마 화제성을 파악하고자 할 때 해당 드라마의 원작이 존재한다면 원작에 대한 가치 평가 또한 함께 수행되어야 한다. 원작이 웹 기반 콘텐츠인 경우 원작의 화제성이 드라마 화제성에 더 큰 영향을 미치는 것으로 보여진다.
    - 드라마 관련 이슈 : 드라마 방영 당시 해당 드라마에 있어 어떠한 이슈가 등장했고, 그 이슈에 대해 대중들이 얼마나 긍정 혹은 부정적으로 반응했는지에 대한 부분을 살펴볼 필요가 있다.
    - OTT 플랫폼 활용 : 해당 드라마가 얼마나 다양한 OTT 플랫폼에서 제공되는지, 방영 종료 후 얼마나 즉각적으로 제공되는지(동시 제공 등)에 관한 부분을 고려해야 한다.

서울시 아파트 가격 예측 대시보드 구축

- 네이버에서 서울시 아파트 정보를 크롤링하여 DB에 저장하고 이를 불러와 가격을 예측하여 아파트 이름을 입력하면 가격을 출력하는 대시보드 구현

- `R` `Python` `SQL` `Machine Learning`
- Click here for Git Repo.

1. Data Collection
    - 아파트 : 네이버 부동산 웹사이트로부터 서울시 아파트의 단지 정보, 시세/실거래가, 학군 정보를 수집하여 데이터베이스에 저장
    - 지하철 : 카카오 API를 통해 아파트에서 가장 가까운 지하철 역 정보를 수집하여 데이터베이스에 저장
2. Data Preprocessing & EDA
    - 각 변수 별 결측치 대체, 이상치 제거, 분포 확인
    - 아파트 별 거래 분포 확인
    - 클러스터링에 사용할 변수를 선택하기 위해 각 변수와 amount의 상관관계 파악에 중점
    - Feature Selection : EDA 과정에서 불필요하다고 판단한 변수를 제거
    - Scaling : 수치형 변수들에 대하여 이상치에 robust한 스케일링 적용 (추후 보다 간편한 인버스 스케일링을 위해 스케일러를 model 경로에 저장)
    - Encoding : 범주형 변수들에 대하여 `LabelEncoder` 적용 (추후 보다 간편한 디코딩을 위해 인코더를 model 경로에 저장)
3. Modeling
    - 클러스터링 : 아파트 가격에서의 결측치를 대체하기 위한 것으로 K-Prototype Clustering 사용
    - 시계열 예측 : 아파트 가격을 예측하기 위한 것으로 ARIMA 모형 사용
4. Conclusion
    - 쉘 스크립트를 작성하여 터미널에서 사용할 수 있도록 지원

KBO 정규시즌 팀별 승률, 타율 및 방어율(평균자책점) 예측

- 주최 : BIGCONTEST
- 기간 : 2020년 7월 ~ 2020년 10월
- `R` `Python` `Machine Learning`
- Click here for Git Repo.

제주 신용카드 빅데이터 경진대회

- 주최 : DACON
- 기간 : 2020년 6월 ~ 2020년 8월
- `R` `Python` `Machine Learning`
- Click here for Git Repo.

서울시 행복도 예측

- 2014 서울서베이 데이터를 바탕으로 행복도에 영향을 미치는 요인 파악 및 행복도 예측
- `R` `Python` `Machine Learning`
- Click here for Doc.

데이터사이언스 분야 분석 및 시각화

- 2019 캐글 서베이 데이터를 바탕으로 데이터사이언스 분야 업무 종사자 현황, 시장 동향, 기술 동향 등을 분석 및 대시보드로 시각화
- R Tableau Visualization

서울시 아파트 가격 예측

- 다방 데이터를 크롤링 하여 서울 소재 아파트 가격 예측
- R Machine Learning