Data Science
- PORTFOLIO -

지원자: 송대훈


dhsong951030@gmail.com


https://github.com/dhsong95

https://github.com/coco-in-bluemoon


https://dhsong10.tistory.com/

SUMMARY

카카오 아레나 Music Playlist Continuation

추천 시스템의 Collaborative Filtering 방식과 Content Based 방식을 종합하여서 카카오 아레나 Music Playlist Continuation 참가 (공개 리더보드 13위)

추천 시스템에 대한 이해와 실제 서비스 환경에 대해 고민

데이터로 읽은 이야기: 호밀밭의 파수꾼

단어 빈도 분석을 통해 소설 호밀밭의 파수꾼이 욕설이 많은 소설인지 데이터를 통해 확인

word2vec 모델을 활용하여서 인물 임베딩을 만들어서 유사한 단어 확인

음악치료 연구 분야 동향 분석

한국음악치료학회에 개제된 논문을 수집하고 자연어 처리를 통해 음악치료 연구 분야의 동향을 파악한다.

분석 결과 “음악치료사”에 대한 연구가 2010년대에 크게 증가하였으며, 토픽 모델링을 통해 확인한 결과 음악치료사에 대한 연구가 주요 토픽으로서 위치한 것을 확인할 수 있다.

카카오 아레나 Music Playlist Continuation

INTRODUCTION

https://github.com/dhsong95/Hunchbrown-Melon-Playlist-Continuation

카카오 아레나에서 진행한 Music Playlist Continuation 경진대회 참가 (팀: 헌치브라운)

2020년 07월 한 달 동안 2인의 팀을 구성하여서 진행. 모델의 설계 및 구현 담당

Python 사용 / Git 버전 관리 / 주요 외부 라이브러리: implicit, numpy, scikit-learn, scipy

시드 아이템(노래 또는 태그)가 주어진 상황에서 숨겨진 아이템(노래 또는 태그)를 맞히는 과제.

추천 결과의 랭크(rank)까지 고려하기 위해 NDCG를 평가 지표로 사용.

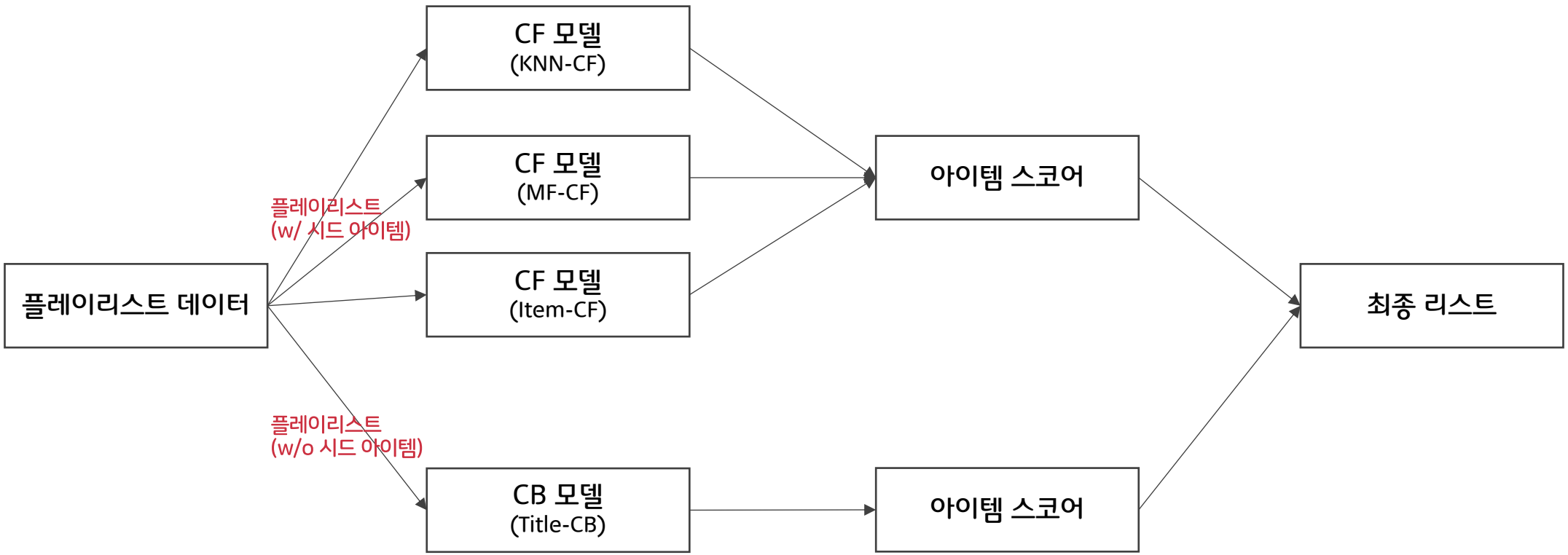
숨겨진 노래를 찾는 것에 높은 비중을 둔 평가 방식

추천 시스템으로 과제를 접근하여서 해결

추천 시스템에서 사용자-아이템의 관계가 플레이리스트-아이템의 관계에 대응

Collaborative Filtering 모델과 Content Based 모델 결합

OVERALL STRUCTURE



MODEL SPECIFICATION

CF 모델 (KNN-CF)

준비

플레이리스트(테스트용)와 플레이리스트(학습용)의 코사인 유사도 계산

추론

플레이리스트(테스트용)와 유사한 유사한 K개의 플레이리스트(학습용)의 아이템 벡터를 유사도와 곱한 합으로 아이템 스코어 산출

CF 모델 (MF-CF)

준비

플레이리스트(학습용 + 테스트용) 행렬을 implicit 라이브러리로 분해하여 플레이리스트 벡터와 아이템 벡터 산출

추론

플레이리스트(테스트용) 벡터와 아이템 벡터를 내적 연산하여서 아이템 스코어 산출

CF 모델 (Item-CF)

준비

플레이리스트(학습용)에서 아이템(태그 또는 노래)간의 코사인 유사도 계산

추론

플레이리스트(테스트용)가 가지고 있는 아이템에 대해서 유사도 벡터를 합하여서 아이템 스코어 산출

CB 모델 (Title-CB)

준비

플레이리스트(학습용) 타이틀을 형태소 분석하여서 주요 형태소-아이템의 빈도 행렬 구축. TF-IDF 변환을 통한 전처리 과정 수행

추론

플레이리스트(테스트용) 타이틀이 가진 주요 형태소의 아이템 벡터를 형태소-아이템 행렬에서 찾아서 연산

CONCLUSION & IMPROVEMENTS

추천 시스템을 이해하고 관련 모델을 구현함으로서 모델의 장단점 파악

- 사용자(플레이리스트)와 아이템의 관계에 대한 정보가 있다면, 다양한 방법론의 Collaborative Filtering 모델을 적용할 수 있으며 성능도 준수
- 사용자(플레이리스트)와 아이템의 관계에 대한 정보가 없다면, CF 모델 보다는 사용자 또는 아이템의 특징(feature)를 기반으로 하는 Content Based 모델 필요

과제의 결과를 생성하는데 많은 시간이 소요되었으며 실제 서비스에서 이러한 시간 지연은 서비스의 질을 떨어뜨릴 수 있다.

- 행렬 기반의 모델은 행렬 연산에 많은 시간과 자원 소요 → numba 라이브러리를 통해 numpy 연산 개선
- 빅 데이터 환경에서의 서비스: 알고리즘은 딥러닝 기반의 모델 사용 / 아키텍처는 분산 처리 기반의 시스템 사용

https://github.com/dhsong95/Movielens-Personal-Recommender-System

딥러닝 기반의 모델 성능 확인: movielens 데이터로 딥러닝 기반의 모델(AutoEncoder, Word2Vec)기반 추천 시스템 동작 구현

- 새로운 데이터를 처리할 때) 행렬 기반의 모델(특히 MF-CF)은 행렬의 재구축 과정 필요하지만, 딥러닝 기반의 모델은 학습된 파라미터로 결과 생성이 가능
- AutoEncoder는 Matrix Factorization과 다르게 학습 데이터만을 가지고 학습 가능.
- Word2Vec을 기반으로 아이템 임베딩을 사용하여 Item-CF 개선 가능

데이터로 읽은 이야기: 호밀밭의 파수꾼

INTRODUCTION

https://github.com/dhsong95/the-catcher-in-the-rye

본 프로젝트는 데이터를 통해서 소설과 관련된 궁금증을 해결하는 개인 프로젝트로 **호밀밭의 파수꾼 영문판**을 데이터로 사용

호밀밭의 파수꾼이 정말로 **욕설**이 많은지 확인하고자 데이터를 통해서 확인

word2vec 모델을 구현하고 이를 통해서 주인공와 유사한 단어는 무엇인지 확인

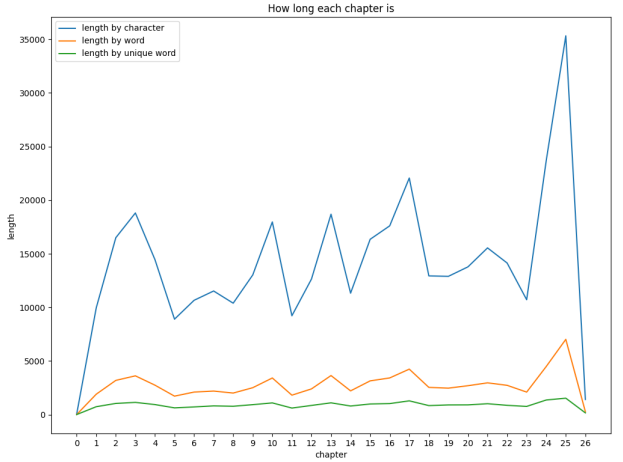
DATA ANALYSIS

소설의 길이?

작가의 인사말을 제외하면 총 26장으로 구성

25장이 가장 길며, 이는 작가가 방황을 마친 홀든과 피비의 모습에 집중한 것을 의미

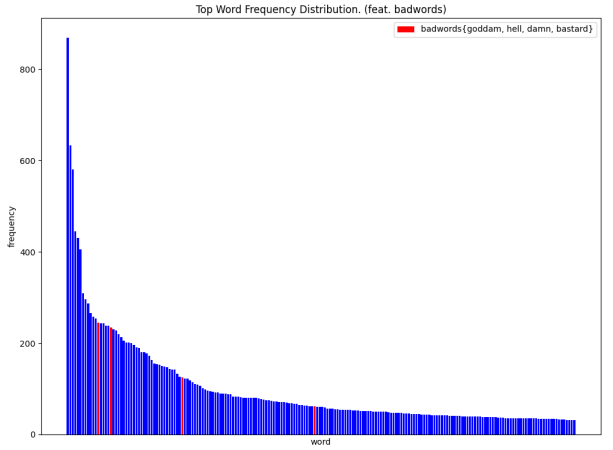
1장과 26장은 홀든의 인사말로 이야기를 시작하고 끝맺으므로 가장 짧음



얼마나 많은 욕설 등장?

불용어를 제외한 상위 빈도 100개 내에 구어체 욕설 “goddam”, “hell”, “damn”, “bastard”가 포함됨

일반적인 단어(right, even)들보다 goddam이 더 많이 사용됨 = 실제로 소설에 욕설이 많이 등장



주인공과 유사한 단어?

Keras 기반으로 word2vec 구현하고 이를 통해 소설 내의 단어 임베딩 학습

주인공 Holden과 가장 유사한 단어는 Holden의 성인 Caulfield

주인공 Holden이 사랑하는 동생 4b도 유사한 단어로 선택됨

CONCLUSION & IMPROVEMENTS

데이터 분석을 통해 알게 된 호밀밭의 파수꾼

- 호밀밭의 파수꾼은 후반부에 많은 이야기를 담고 있다.
- 호밀밭의 파수꾼에서는 실제로 욕설이 많이 등장한다.
- 호밀밭의 파수꾼의 주인공 홀든은 동생 포비와 유사한 것으로 학습된다.

소설에 대해서 분석한 내용은 블로그 포스팅(https://dhsong10.tistory.com/50)을 통해서 공유한다.

음악치료 연구 분야 동향 분석

INTRODUCTION

https://github.com/dhsong95/Music-Therapy-Article-NLP

본 프로젝트는 2019년 **음악치료의 연구 분야**에 대해 질문하고 그에 대한 답은 데이터를 통해서 찾기 위해 시작하였다.

본 프로젝트는 한국음악치료학회에서 발간한 **논문**을 RISS에서 수집하고 전처리하였으며, **자연어 처리** 기법을 활용하여서 분석했다.

본 프로젝트를 통해 음악치료 연구 분야로서 **"음악치료사"**가 2010년대부터 증가하며, 하나의 토픽으로 자리매김한 것을 확인할 수 있었다.

CONCLUSION & IMPROVEMENTS

본 프로젝트를 통해 문제 해결을 위한 데이터 분석의 효용을 확인할 수 있었으며 다음과 같은 사항을 보완할 필요가 있다.

Specific Topics

- 연구 분야에 대해 질문을 던졌을 때 기대한 것은 "노인음악치료", "청소년음악치료"와 같은 구체적인 주제
- 토픽 모델링의 결과는 연구 방법 또는 넓은(broad) 의미로서의 음악치료 연구 주제
- 일반적인 연구와 관련된 단어를 불용어 처리하여서 구체적인 주제가 나오도록 제어 필요

More Data

- 한국음악치료학회가 음악치료분야 전체를 대표할 수는 없음
- 불용어를 통해 일반적인 연구 관련 단어가 제외된다면 다른 학회의 논문 데이터를 추가하여서 구체적인 주제가 유의미하게 나오도록 조정할 필요가 있음
- 한국어 이외의 영어 논문을 사용하면 자연어 처리에서 정교한 형태소 분석 가능