자연어 처리를 통한 노래 추천 시스템

PORTFOLIO

CONTACT
Cv00214@gmail.com
010-9924-9801

성장하는 개발자 박정호입니다

박정호 / Jeongho Park

1997.02.14 / 경기도

Tel. 010-9924-9801
Email. cv00214@gmail.com
경기도 시흥시 신천동

GRADUATION

2016 은행고등학교 졸업
2020 학점은행제 멀티미디어학과 전문학사 (수료)

SKILL

Python 95
Linux 05

ABOUT

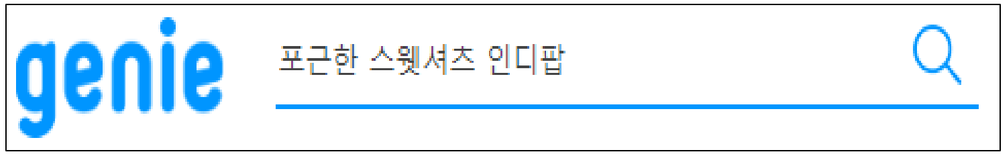
메디치 교육센터
비즈니스 인사이트 도출을 위한 데이터 분석 과정 (수료)

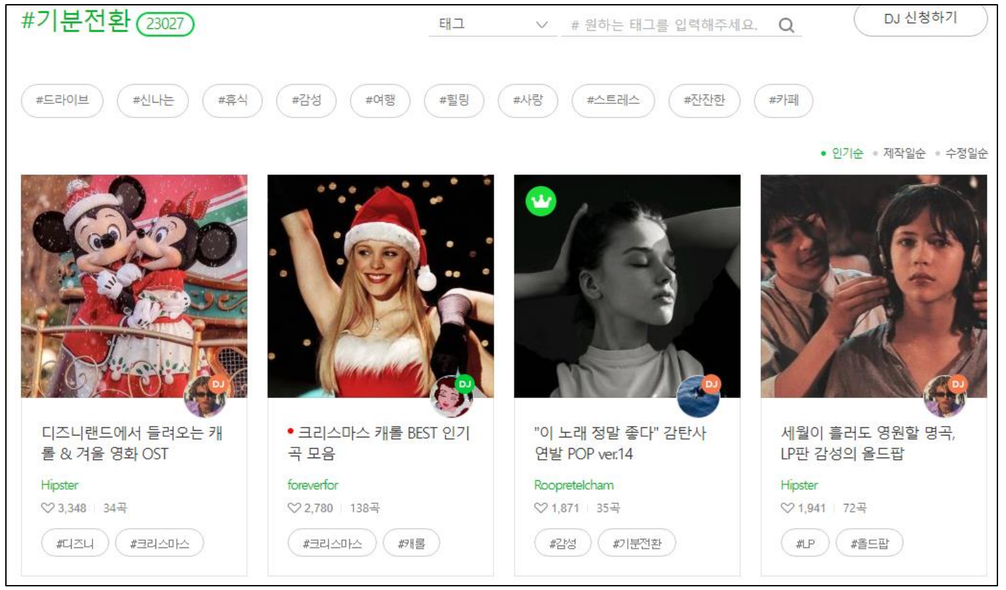
구름
AI기술 자연어 처리 과정 2기(진행 중)

PROJECT

자연어 처리를 통한 노래 추천 시스템

1. 문제 정의





기존의 음악 검색의 시스템은 입력 값에 대한 정확한 정보의 전달이 목적입니다.

하지만 플레이리스트의 태그와 타이틀을 이용해
기존의 검색이 아닌 사용자의 의도, 의사를 반영한 노래 추천을 하게 된다면
소비자가 원하는 노래를 빠르게 찾을 수 있을 것입니다.

또한 소비자는 새로운 플레이리스트를 만드는 번거로움 없이
노래를 추천받을 수 있게 됨으로 cold start 문제가 해결될 것이라는 기대가 있습니다

2. 데이터 명세

Kakao Arena Melon data (2020.9.23)

곡 정보 * 약 70만 곡

장르	발매 일자	곡 명	가수 명

플레이 리스트 정보 * 약 11만 개

업데이트 날짜	장르	타이틀	곡 명

개별 유저 데이터가 존재하지 X

제한된 데이터로 유저 의도를 파악할 것

유저가 임의 지정한
플레이리스트의 태그와 제목을 적극적으로 활용

개별 유저를 특정하는 기준을 마련할 것

개별 **플레이리스트** → 개별 **유저**로 규정

3. 해결 방식 고민 : 전처리

특정 단어의 경우 같은 단어이나 형태가 다르게 쓰인 경우가 보입니다.
타이틀 같은 경우 '음악', '노래'와 같이 지나치게 많이 사용되어 오히려 편향을 가져올 수 있는 단어가 존재합니다.
따라서 자연어 전처리를 통해 단어의 제한과 통일을 두게 되면 추천의 질을 향상시킬 수 있을 것입니다.

	tags	id	plylst_title
0	[락]	61281	여행같은 음악
1	[추억, 회상]	10532	요즘 너 말야
2	[까페, 잔잔한]	76951	편하게, 잔잔하게 들을 수 있는 곡.-
3	[연말, 눈오는날, 캐럴, 분위기, 따듯한, 크리스마스 캐럴, 겨울노래, 크리스마스,...	147456	크리스마스 분위기에 흠뻑 취하고 싶을때
4	[댄스]	27616	추억의 노래 ㅋ

Konlpy의 Mecab을 사용

국어사전에 등록되어 있지 않는 가수의 이름, 신조어 등을 추가해 최대한 다양한 단어를 사용할 수 있도록 노력했습니다.
사용자 사전에 등록이 되어 있지 않거나 단어 비용이 지나치게 높은 경우
아래의 예시처럼 의도하지 않은 결과를 보여줄 가능성이 높습니다.

예시

"센치한" → "센" "치한" "센치" "한"

또한 후에 모델 학습에 용이하게 하기 위해
고유명사, 일반명사, 형용사, 동사, 외국어, 숫자, 어근에 해당하는 품사만 추출한 후
단어들을 모두 합쳐 놓은 new_tag라는 파생 변수를 생성했습니다.

new_tags
[락, 여행]
[회상, 추억]
[편하, 잔잔, 까페]
[취하, 분위기, 따듯, 왕국, 캐럴, 눈, 연말, 겨울, 크리스마스]
[댄스, 추억]

3. 해결 방식 고민 : 모델 선택

Word2Vec VS Sent2Vec

Word2Vec

Word2Vec는 단어 단위에 대해서는 매우 유용한 수준의 의미론적 표현이 가능합니다.
하지만 문장, 단락을 넘어 문서 단위의 긴 텍스트에 적용하는 일은 쉽지 않습니다.
또한 비지도 학습으로 유용한 방식을 찾기는 매우 어려운 일입니다.

Sent2Vec

Word2Vec에서 사용한 서브 샘플링과 다이나믹 컨제스트 윈도우를 비활성화하고
일반적인 n-그램의 의미가 아닌 비아그램의 최대 거리를 구하는 n-그램을 사용합니다.
이것으로 인해 Sent2Vec을 통해 문장에서 유용한 표현이 가능합니다.

KNN

Sent2Vec을 통해 형성된 임베딩 값을 Cosine 유사도를 통해 유사도를 구하게 됩니다.
이때 사용하는 KNN은 HNSW 기반의 비지도 학습의 KNN입니다.

Word2Vec – Sent2Vec 비교

입력 단어

"잔잔한 노래"

입력 단어

"밤이나 새벽에 듣기 좋은 잔잔한 노래"

위의워드클라우드는 입력 값의 변화를 주며 Word2Vec와 Sent2Vec을 비교

입력된 단어의 길이가 늘어날수록 Word2Vec는 문장과 관련된 단어들이 줄어들고 입력 단어에 집중합니다.
하지만 Sent2Vec 문맥을 파악하고 문장과 관련된 많은 태그 단어들을 출력합니다.
추천의 다양성을 고려해 Sent2Vec을 선택하는 것이 옳다고 판단했습니다.

4. 최종 결과

웹사이트를 구성한 모습입니다.

사용자는 ~~g~~워너비, 발라드, 한국 힙합과 같이 단어 단위로 입력을 할 수 있고
카페에서 공부하면서 듣기 좋은 노래, 여름에 드라이브하면서 듣기 좋은 노래와 같이 문장 단위로 입력을 할 수 있습니다.

따라서 사용자는 사용자가 의도한대로 추천을 받을 수 있습니다.

감사합니다!
잘 부탁드립니다!

PORTFOLIO

CONTACT
Cv00214@gmail.com
010 9924 9801