

유튜브 댓글 데이터를 활용한 새로운 음악 분류 방법 제안

- LDA(Latent Dirichlet Allocation) 토픽 모델링을 기반으로-

서울과학기술대학교 산업정보시스템전공 캡스톤디자인
지도 교수: 심재웅 교수님 20101672 한세희

Contents

01 연구 배경 및 소개

02 데이터셋 수집 및 생성

03 LDA 진행

04 연구 결과

05 결론 및 참고 문헌

연구 배경 및 소개

<연구 배경>

- 기존 음악 추천 시스템은 주로 장르를 기반으로 사용자에게 음악을 추천하고 있음. 그러나 음악에 대한 선호는 장르뿐만 아니라 다양한 특성에 의해 결정되기 때문에 단순 장르 기반의 추천은 청취자의 다양한 취향을 충족시키기에 한계가 있음
- 이를 해결하기 위해 장르 이외의 새로운 기준을 설정하여 접근한 기존 연구들이 있으나, 다음과 같은 한계 존재

(1) 가사 기반 음악 데이터 분석

- 사례: 음악청취 앱 'FLO'의 데이터 분석
- 가사만을 이용해 분류한 결과, 같이 묶인 곡들과 연관성이 떨어지는 곡 존재

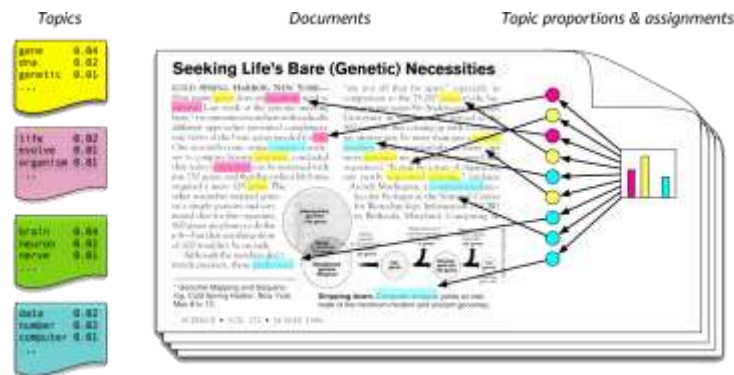
(2) 가사 해석 기반 음악 분류

- 사례: Music subject classification based on lyrics and user interpretations – Kahyun Choi 외 3인
- 사용자들이 올린 '가사해석' 데이터를 기반으로 음악 주제를 분류
- '가사' 기준 분류보다 '가사 해석' 기준 분류가 더 유용하다는 결론
- 그러나 가사 해석 데이터는 일부 노래를 대상으로만 존재 -> 분석의 한계로 작용

연구 배경 및 소개

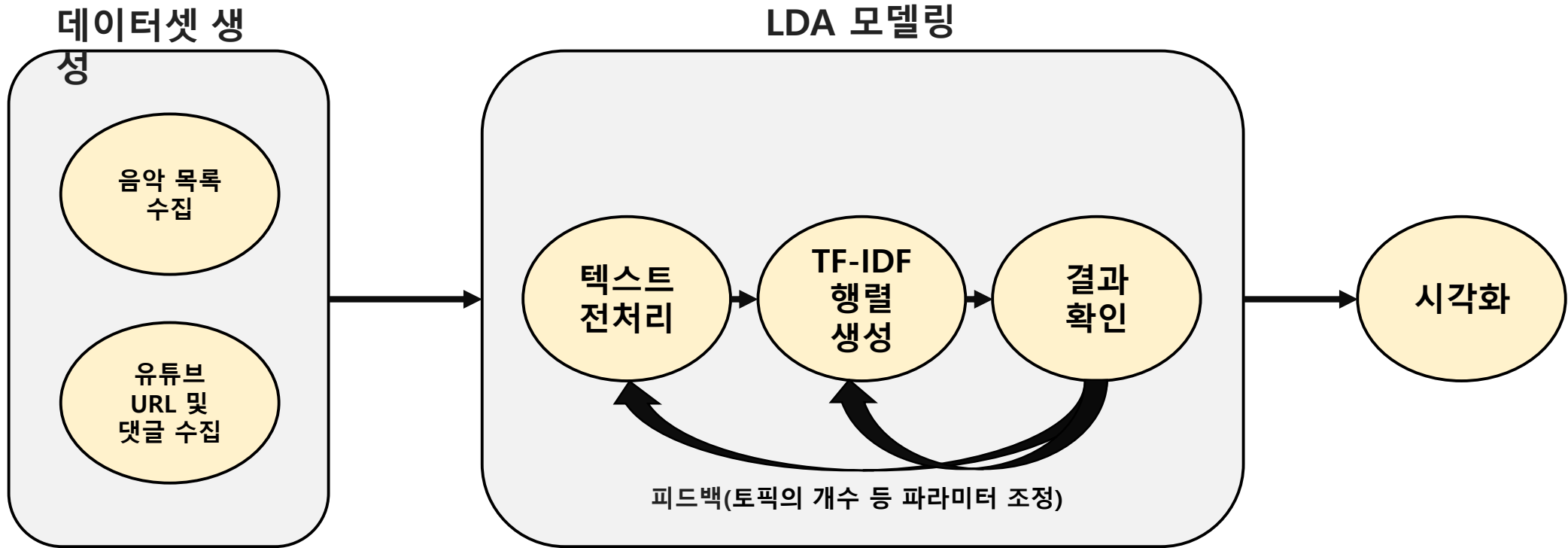
<연구 소개>

- 앞서 언급한 음악 분류 시스템의 한계를 해결하고자 '유튜브 댓글' 데이터 도입
- 유튜브 댓글의 특징
 - 가사해석보다 **방대한 양** 보유
 - **멜로디, 가사를 복합적으로 고려**해서 사용자들이 작성하는 감상평
=> 분류 시 토픽별 곡 '분위기'의 유사성이 높을 것으로 기대
- 본 프로젝트는 음악 분류에서 기존의 장르 중심 접근법을 벗어나, 음악의 분위기 및 특성을 보다 종합적으로 파악하여 더 나은 음악 분류 시스템을 구축하고자 함
- **이용 모델: 잠재 디리클레 할당(Latent Dirichlet Allocation, LDA) 기법**
 - 주어진 문서에 대하여 각 문서에 어떤 주제들이 존재하는지에 대한 확률모형
 - 토픽별 단어의 분포와 문서별 토픽의 분포를 모두 추정 가능
 - 댓글을 여러 토픽으로 분류하여 각각의 토픽이 어떤 음악적 분위기를 대표하는지 해석



연구 배경 및 소개

<진행 과정>



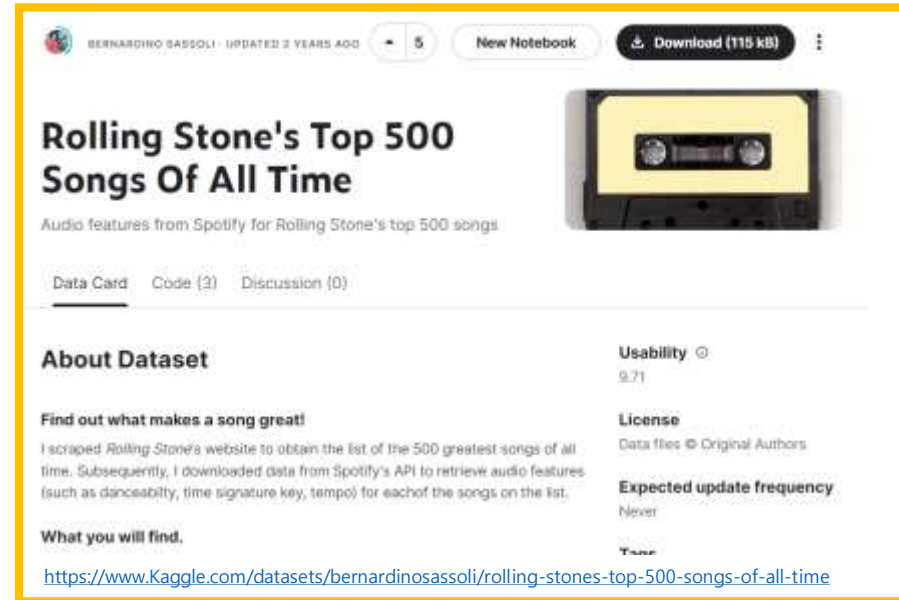
데이터셋 수집 및 생성

<음악 목록>

- Rolling Stone 's Top 500 Songs Of All Time
- 미국 음악 잡지 <롤링 스톤>이 선정한 역대 최고의 명곡 500선
- 1940년대~2000년대 발매된 곡을 다루고 있어 다양한 시대의 음악을 포함

<유튜브 영상 URL 및 댓글>

- Selenium과 BeautifulSoup 라이브러리를 사용하여 유튜브 URL 및 댓글 크롤링 시행
- 500곡에 대하여 각 최대 100개의 댓글 수집



BERNARDINO SASSOLI · UPDATED 2 YEARS AGO · 5 · New Notebook · Download (115 kB)

Rolling Stone's Top 500 Songs Of All Time

Audio features from Spotify for Rolling Stone's top 500 songs

Data Card · Code (2) · Discussion (0)

About Dataset

Find out what makes a song great!

I scraped Rolling Stone's website to obtain the list of the 500 greatest songs of all time. Subsequently, I downloaded data from Spotify's API to retrieve audio features (such as danceability, time signature key, tempo) for each of the songs on the list.

What you will find,

<https://www.kaggle.com/datasets/bernardinosassoli/rolling-stones-top-500-songs-of-all-time>

Usability 9.71

License Data files © Original Authors

Expected update frequency Never

	artist	song	video_url
0	Smokey Robinson and the Miracles	Shop Around	youtube.com/watch?v=eE36-7_pwl0&pp=ygUsU2hwcC...
1	Weezer	Buddy Holly	youtube.com/watch?v=kemivUKb4t4&pp=ygUSQnVikZH...
2	The Rolling Stones	Miss You	youtube.com/watch?v=KuRxXRuAz-i&pp=ygUbTWlzc...
3	Bruce Springsteen	The Rising	youtube.com/watch?v=r5_8gpiSotI&pp=ygUcVGhlIF...
4	Jackson Browne	Running on Empty	youtube.com/watch?v=IKnnh8VDULs&pp=ygU8UnVubm...
...
495	Aretha Franklin	Respect	youtube.com/watch?v=A134trShx_gw&pp=ygUXUmVzcG...
496	Marvin Gaye	What's Going On	youtube.com/watch?v=H-kA3UtBj4tA&pp=ygUbV2hhdC...

LDA 진행 - 전처리

<데이터 전처리>

- 토픽 분류 단위: 한 곡에 대한 댓글 데이터를 하나로 병합하고 곡별 토픽 분류
- 곡별 활용된 댓글 개수: 10개 & 100개 → 분류 결과가 더 우수하게 나타난 '10개' 기준으로 최종 분류 시행

<텍스트 전처리>

- 이모지, 아랍어 등 ASCII 코드에 해당하지 않는 문자 삭제
- 불용어 제거: NLTK가 제공하는 영어 불용어 제거
- 표제어 추출(단수형, 인칭 및 시제 통일)
- 길이가 짧은 단어 제거(길이 3 이하 대상)
- 공통적으로 등장하는 단어 및 추가 불용어 제거
 - 분석을 진행하면서 직접 판단내린 stopwords 제거
 - 'song', 'best', 'great', 'favorite', 'like', 'music', 'good', ...

	artist	song	video_url	combined_comment
0	Smokey Robinson and the Miracles	Shop Around	youtube.com/watch?v=eE35-7_pwi0&pp=ygU2hvcC...	I never get tired of listening to them! I'm 28 a...
1	Weezer	Buddy Holly	youtube.com/watch?v=kemvUKb4f4&pp=ygU5QnVhZG...	Good to see Weezer is still going strong over
2	The Rolling Stones	Miss You	youtube.com/watch?v=KuRdXRuA2-l&pp=ygU6TWVzcy...	71 years old, still rolling my shoulders to th...
3	Bruce Springsteen	The Rising	youtube.com/watch?v=r5_8ppiSotI&pp=ygUcVGhtIF...	"Nobody wins unless everybody wins." — Bruce S...
4	Jackson Browne	Running on Empty	youtube.com/watch?v=iKnnh8VDULs&pp=ygUJfUwVubm...	Once a Jackson Browne fan, always a Jackson B...
...
495	Aretha Franklin	Respect	youtube.com/watch?v=A134h5lhx_gw&pp=ygUxUmVzcG...	Her voice could cut through solid rock. It's i...
496	Marvin Gaye	What's Going On	youtube.com/watch?v=H-lA3U8j4M&pp=ygU6V2hhdC...	This song is almost 50 years old and the messa...

```
0 [never, tire, listen, themI, Smokey, Robinson,...
1 [Good, Weezer, still, strong, year, Beethoven,...
2 [year, still, roll, shoulder, funky, help, mov...
3 [Nobody, unless, everybody, Bruce, Springsteen...
4 [Once, Jackson, Browne, always, Jackson, Brown...
Name: combined_comment, dtype: object
```

LDA 진행 – TF-IDF 행렬 생성

- 사이킷런의 **TfidfVectorizer**를 이용하여 조건 설정
 - **max_df**: 단어의 최대 빈도값. → 30% 이상의 문서에 등장하는 단어 제거
 - **min_df**: 단어의 최소 빈도값. → 0.5% 이하의 문서에 등장하는 단어 제거
 - **max_features**: TF-IDF 행렬의 최대 컬럼 수(= 분석에 사용되는 단어의 종류 수). → 상위 2000개의 단어 보존
 - **gram_range**: 단어의 묶음 범위. → 1개짜리 단어 묶음과 2개짜리짜리 단어 묶음 사용

```
vectorizer = TfidfVectorizer(stop_words='english', max_df=0.3, min_df=0.005, max_features= 2000, ngram_range=(1, 2))  
X = vectorizer.fit_transform(text['combined_comment'])
```

```
# TF-IDF 행렬의 크기 확인  
print('TF-IDF 행렬의 크기 :', X.shape)
```

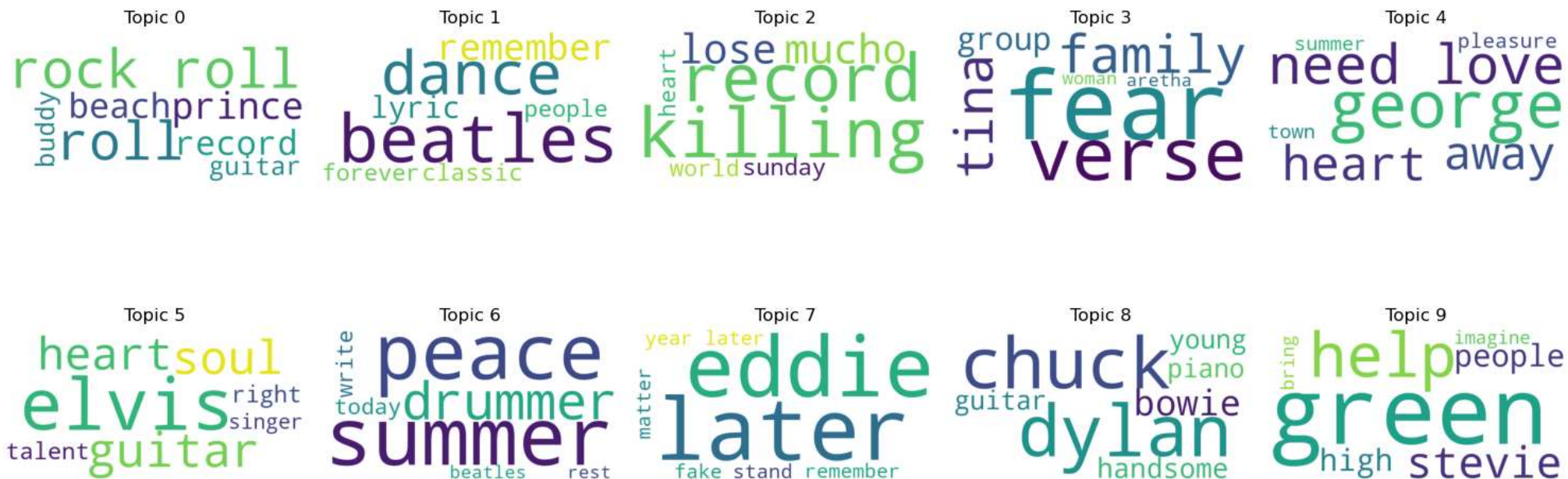
TF-IDF 행렬의 크기 : (500, 2000)

LDA 진행 - 토픽 모델링

- 토픽 개수 **10**개로 설정
- Parameter 조정 및 모델링 과정에서 등장하는 **추가적인 불용어 제거**

```
lda_model = LatentDirichletAllocation(n_components=10, learning_method='online')  
  
lda_top = lda_model.fit_transform(X)  
  
print(lda_model.components_)  
print(lda_model.components_.shape)  
  
[[0.55552142 0.4847074 0.80020832 ... 0.74141284 0.56665087 0.98744765]  
 [0.49792507 0.68013455 0.56467881 ... 0.88720344 0.86648002 0.63624504]  
 [0.52029008 0.44411576 0.47332067 ... 0.48807291 0.45768686 0.60076026]  
 ...  
 [0.5871998 0.61948704 0.5947439 ... 0.54522605 0.50218346 0.49200966]  
 [0.5804983 0.47287653 0.54430651 ... 0.54017592 0.56244603 0.51033497]  
 [0.54478828 0.59789937 0.5820159 ... 0.73756693 0.47945356 0.93586827]]  
(10, 2000)
```

결과 – 토픽별 주요 키워드



토픽 0: 클래식한 락앤롤

토픽 1: 비틀즈와 추억의 명곡

토픽 2: 여유롭고 다양한 스타일

토픽 3: 감동적이고 감성적인 노래

토픽 4: 조지처럼 따뜻한 사랑 노래

토픽 5: 엘비스 프레슬리처럼 진한 소울과 기타

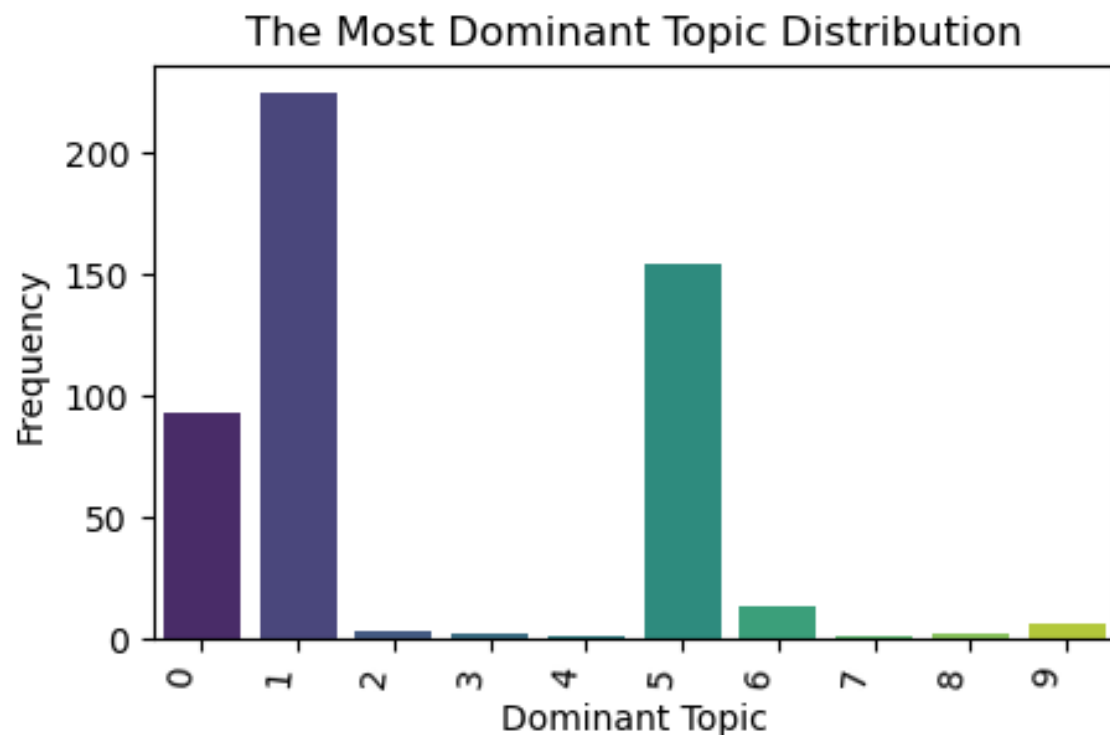
토픽 6: 여름 분위기와 긍정적 에너지

토픽 7: 지난 시간을 회상

토픽 8: 스타 솔로가수의 젊은 에너지

토픽 9: 건강하고 희망적인 스타일

결과 – 토픽별 문서 비율(1)



<문서별 가장 dominant한 토픽>

- 토픽 1(비틀즈와 추억의 명곡), 토픽 0(클래식한 락앤롤), 토픽 5(엘비스 프레슬리처럼 진한 소울과 기타)이 높은 문서 비율을 가짐
- 나머지 토픽은 the most dominant topic으로 선정된 비율이 비교적 낮음

토픽 0에 대한 노래:

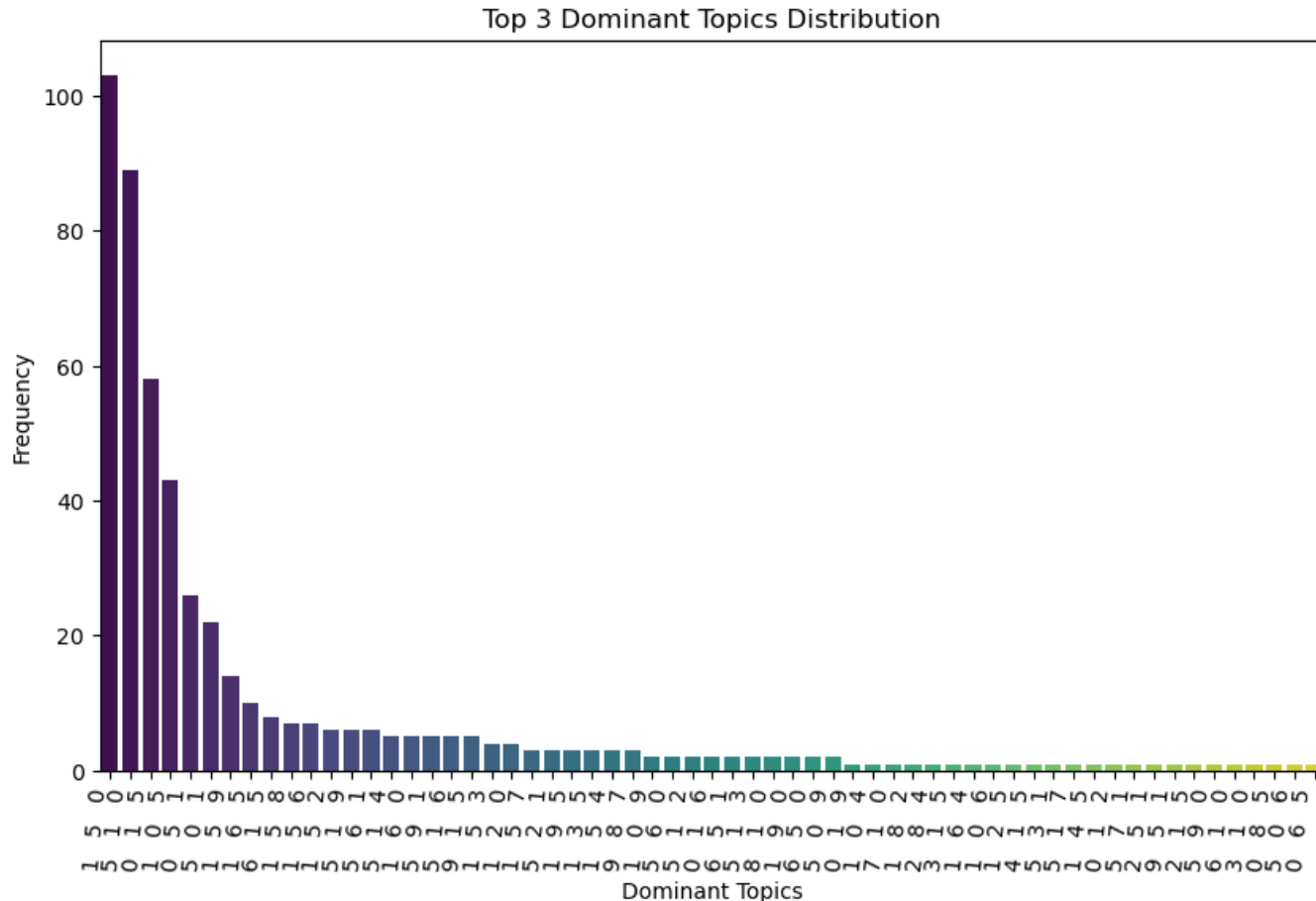
	artist	song
2	The Rolling Stones	Miss You
5	The Rolling Stones	Brown Sugar
13	Alice Cooper	I'm Eighteen
14	David Bowie	Young Americans
19	Rick James	Super Freak

토픽 1에 대한 노래:

	artist	song
1	Weezer	Buddy Holly
4	Jackson Browne	Running on Empty
6	R. Kelly	Ignition (Remix)
7	MGMT	Time to Pretend
9	Joan Jett and the Blackhearts	I Love Rock 'N Roll

(토픽0, 토픽1로 분류된 문서 중 각 5곡에 대한 정보)

결과 - 토픽별 문서 비율(2)



<문서별 distribution이 높은 토픽 Top 3>

- 문서별로 비중이 높은 토픽을 순서대로 세가지씩 추림 -> 한 음악에 여러 내용을 가진 댓글의 특성에 적합
- 토픽 1, 토픽 0, 토픽 5 이외의 토픽은 앞서 보인 the most dominant topic으로 선정된 비율은 적었으나, 왼쪽 그림과 같이 두세 번째로 높은 비중으로 distribution을 제공하는 경우를 볼 수 있음

연구 결과 및 의의

<결과>

- 20세기 음악이 많이 포함된 데이터셋을 이용한 결과, 분류된 토픽에서 클래식한 음악을 암시하는 키워드 등장 多
- 또한 해당 셋리스트 500곡에 힙합 음악이 많이 포함되어 있지 않아, 분류 결과에서도 힙합과 직접적으로 관련된 토픽은 나타나지 않음
- 동일한 곡 내에서도 댓글별로 주제가 다르기 때문에 여러 개의 토픽이 비슷한 확률분포로 기여하는 경우가 다수 존재. 따라서 dominant topic 범위를 top1이 아닌 top3로 설정하였을 때 더욱 광범위한 분류가 가능해짐

<한계 및 의의>

- **비지도 학습**의 특성상 분류 성능을 객관적으로 평가하기에 어려움이 있음
- 음악의 인지도가 낮거나 최근에 발매된 곡인 경우 양질의 **댓글이 부족**하여 분류 정확도가 저하될 수 있음
- 댓글은 사용자의 의견에 의존하기 때문에 특정 음악에 관한 댓글이 **특정 사용자 집단에 집중**되었을 경우 해당 집단의 주관이 편향되어 개입될 수 있음
- 단, 향후 연구에서는 유튜브 이외 플랫폼에서의 리뷰나 댓글을 종합적으로 수집함으로써 데이터 보완이 가능할 것임
- 기존의 장르나 가사, 가사해석만으로는 분석이 어려웠던 음악 분류 분야와 달리 '**사용자 의견**'을 통해 음악의 분위기를 종합적으로 이해하는 새로운 시각을 제공함
- 사람들의 주관적인 감상을 반영함으로써 실제 '**음악 추천**'에 더욱 적합한 음악 분류의 가능성을 제시함

참고 문헌

- 음악청취 앱 'FLO'의 데이터 분석 (<https://maily.so/musicdata/posts/7251b9>)
- Music subject classification based on lyrics and user interpretations – Kahyun Choi 외 3인 (<https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/pra2.2016.14505301041>)
- 이현수, 홍성은, 방준일 and 김화중. (2020). 데이터 임베딩을 활용한 사용자 플레이리스트 기반 음악 추천에 관한 연구. 한국정보기술학회논문지, 18(9), 27-34.
- Fei Li. (2020). 한국 대중가요 가사 유형화에 토픽 모델링(LDA 및 STM) 적용 실용성 시론 - 70년대 이후 "사랑" 주제어 관련 가사를 중심으로 -. 텍스트언어학, 49, 275-328.
- SAMUEL CORTINHAS. NLP6 - Topic Modelling with LDA (<https://www.kaggle.com/code/samuelcortinhas/nlp6-topic-modelling-with-lda>)
- Topic Modeling, LDA (<https://ratsgo.github.io/from%20frequency%20to%20semantics/2017/06/01/LDA/>)