H040-4-4139-01 텍스트 및 오피니언 마이닝

Team Project 최종 보고서

21세기 대중가요 가사 분석

멜론 (권순찬, 장석원, 김의진)

---

1. 서론

대중가요가 늘 현대인들에게 사랑 받는 이유는 무엇일까. 누군가는 듣는 음악 자체가 좋았을 수도 있지만, 지금까지 대중가요가 시대적 흐름과 유행에 따라서 계속되는 장르의 변화에도 꾸준히 사랑을 받고 있는 이유는 많은 사람들의 공감을 얻고 있는 가사, 글 때문이다. 현대인의 대다수가 매일 듣는 대중 가요만큼, 사람들이 자주 접하고, 사람들의 생각과 일상을 잘 나타내는 텍스트는 없다고 추론하였다. 어떤 가사들이 많은 사람들의 공감을 얻고 사랑을 받았는지, 사랑이나 이별과 같은 어떤 주제나 특정 단어였는지, 혹은 그 시대적 상황에 맞는 단어들이 자주 사용되면서 사랑을 받았는지, 그렇다면 어떤 단어들이 그 시대를 드러내는 역할을 했는지에 대한 궁금증이 생겨 이 주제를 선택하게 되었다.
처음에는 21세기의 대중가요 가사를 분석하려고 하였으나 시대의 범위가 다소 짧아 그 당시의 대표성은 드러내도 가사들의 변화를 유의미하게 나타내기에는 부족하다고 판단하여 1970년대 이후로 범위를 넓혀 가사분석을 진행하였다. 텍스트 분석을 진행하기 위해 멜론에서 제공하는 차트 중 선택한 년도에 흥행하였던 음원들을 1위부터 100위까지 소개하는 시대별 차트를 활용하였다. 1970년대, 1980년대와 1990년대는 시대별 차트에서 제공하는 1위부터 100위까지의 데이터를 다 활용하였고 21세기에 해당되는 기간의 대중가요들은 각 년도에서 20위까지의 데이터를 수집하여 활용하였다.

2. 내용

2.1. 데이터 전처리

1) 노이즈 캔슬링

- 맞춤법과 띄어쓰기는 가사 데이터 특성상 이미 완료가 되어있다고 판단하고 진행하지 않았다.
- 실제로 'hanspell' 을 통해 검증해본 결과 큰 차이가 없었다.
- 추가적으로 Lyrics 칼럼 노래가사 데이터에 있는 개행문자와 특수문자를 제거했다.

2) 토크나이징

- KoNLPy의 Okt를 활용하여 토크나이징 진행했다.

3) 형태소분석 / 품사태깅

- 어간 추출 작업과 KoNLPy의 Okt를 활용하여 pos 옵션을 통해 품사태깅 작업을 진행했다.

4) 불용어처리

- RANKS NL의 stopwords 말뭉치와 필요하다고 생각되는 불용어를 추가하여 불용어 제거했다.

5) 텍스트 벡터화

- LDA 모델 적용시 벡터화를 진행하였다..

2.2. 모델링

1) 빈도분석

(1) 워드 클라우드

시대별 노래가사를 워드클라우드를 통해 시각화 해보았다.


왼쪽 이미지는 1970년대, 오른쪽 이미지는 2020년대의 노래 가사를 시각화한 이미지이다. 두 시대뿐만 아니라 나머지 시대들도 워드클라우드를 통해 시각화한 결과 뚜렷한 차이는 발견할 수 없었다. 1970년대부터 2020년대까지 '사랑'이라는 단어는 항상 자주 등장해왔으며 그 외에 공통적으로 눈물, 마음, 사람 등의 단어가 자주 등장해오고 있음을 알 수 있었다.

(2) TF-IDF

텍스트에서 쓰인 단어가 자체적으로 고빈도어 이거나 혹은 그 텍스트의 고빈도어일 경우 단어의 중요도를 정확하게 판단하기 어렵다. 따라서 빈도에 적절한 가중치를 부여하는 것이 필요하다고 알려져 있다(강범모 2014:10) 여러 방법 중 TF-IDF값을 통해 분석한 결과는 다음과 같다.

순위	단어	TF-IDF
1	사랑	8.469884
2	마음	5.650371
3	사람	5.646274
4	그대	4.956854
5	눈물	2.880280
6	추억	2.753804
7	세월	2.640072
8	생각	2.590047
9	노래	2.440626
10	슬픔	2.425240

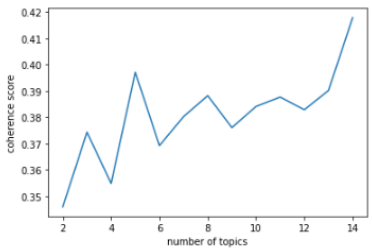
그림 1 1970년대

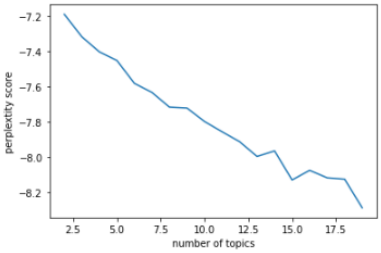
순위	단어	TF-IDF
1	사랑	16.377951
2	그대	11.150462
3	사람	6.021220
4	눈물	5.599424
5	마음	3.889788
6	추억	3.767096
7	가슴	3.487240
8	세상	3.377184
9	생각	3.233919
10	기억	3.217356

그림 2 2000~2006년

공통적으로 사랑, 마음, 사람, 그대, 눈물 등의 단어들이 높은 TF-IDF 값을 가졌다. 또한 1970, 1980년대는 추억, 세월, 슬픔 등의 단어가 주류를 이루었고 2000년대에 들어서 '세상'이라는 단어가 높은 TF-IDF값을 가지기 시작했다. 최근으로 올수록 슬픈 감성의 단어 보다는 '순간', '사이' 등 시간적 개념이나 공간적 개념의 단어가 높은 TF-IDF값을 가지는 특성을 보인다.

2) 토픽 모델링





가사 데이터를 리스트 형식으로 불러온 후 Coherence score와 Perplexity score를 참고하여 토픽의 개수를 지정해주었다. 2010년대 노래의 경우 위의 그래프를 참고하여 토픽의 개수를 14개로 지정해 주었다. Coherence score는 토픽의 일관성을 의미하며 높을수록 좋다. Perplexity score는 모델의 평가 지표로 낮을수록 좋은 모델임을 의미한다.

LDA 모델을 돌리기 전 노래가사 데이터에서 형용사, 동사, 명사만 추출하여 토크나이징을 진행했다. LDA는 Count기반의 Vetorizer만 적용되기 때문에 CountVectorizer를 사용하였고 2개의 문서 미만으로 등장하는 단어는 제외시켰으며 전체 10%이상으로 자주 등장하는 단어 또한 제외시켰다.

	Doc_Num	Topic	Percentage	year	rank	name	genre	lyricist	lyrics
0	0	11	0.977352	2010	1	Bad Girl Good Girl	댄스	박진영	You don't know me You don't know me You don't ...
1	1	2	0.986345	2010	2	잔소리	발라드	김이나	늦게 다니지좀 마 술은 멀리좀 해봐 열살짜리 애처럼 말을 안듣니 정말 웃음만 나와 ...
2	2	10	0.983990	2010	3	죽을 만큼 아파서	랩/힙합	박장근	I found the way to let you leave I never reall...
3	3	12	0.977352	2010	4	못해	발라드	민연재	이제는 밥을 먹어도 눈물없이는 삼키지 못해 억지로 먹고 먹어도 속이 늘 허전해 노랠...
4	4	2	0.974206	2010	5	죽어도 못 보내	발라드	방시혁	어려도 아픈건 똑같아 세상을 잘 모른다고 아픈걸 모르진 않아 괜찮아 질거라고 왜 거...

문서별로 가장 확률이 높은 토픽으로 할당시켜준 후 기존의 데이터셋과 합쳐보았다. 그 후에 토픽별 개수를 확인하고 토픽별 가장 연관도가 높은 단어 10개를 추출한뒤 토픽의 주제를 라벨링하는 작업을 진행했다.

	1970년대	1980년대	1990년대	2000-06	2007-12	2013-16	2017~
1	애절한 사랑	시련/아픔	시련/아픔	사랑에 대한 아픔	추억과 그리움	사랑 고백	사랑에 대한 아픔
2	다짐/극복/짝사랑	애절한 사랑	이별에 대한 걱정	사랑에 빠짐	이별에 대한 슬픔	사랑에 대한 아픔	사랑 고백
3	희망/꿈/미래	자유	지침	이별에 대한 슬픔	사랑 고백	사랑에 대한 설렘	이별에 대한 슬픔

라벨링을 진행한 후 시대별로 상위 3개의 주제를 정리해본 결과 위의 표와 같았다. 20세기에는 애절한 사랑 노래와 시련과 아픔, 자유와 희망, 미래에 관한 노래가 주를 이루었던 반면에 21세기에는 비교적 사랑과 이별에 대한 노래가 많음을 알 수 있다.

3) 기타 분석

(1) 영어 사용 비율 변화 분석

우리나라 대중가요의 가사 속에는 영어가 얼마나 많이 포함되어 있는지, 그리고 그 비율이 시간 변화에 따라 어떻게 달라지는지 알아보았다. 먼저 70년대에는 전체 음절 기준 알파벳 비중이 26%를 차지했다. (단어 비율이 아닌 알파벳 비중이라 대체적으로 높게 형성된다.) 80년대는 27%, 90년대는 29%로 점점 증가하다가 00~06년대에는 37% 그리고 07~12년에 52%로 한국 대중가요 가사에 영어의 사용이 대폭 증가했다는 것을 알 수 있다. 13년부터 17년까지는 우리말 사용 여파로 다시 40%대로 줄었지만 최근 대중 가수들의 글로벌화로 인해 현재는 전체가사의 약 62%정도가 알파벳으로 이루어져 있음을 알 수 있다.

(2) 장소 어휘 분석

앞선 분석들이 해당 시대상과 가치관을 담아내기에 충분하지 않다고 생각해 추가적으로 장소분석을 실행하였다. 장소관련 분석은 해당 년도에 장소 관련 어휘의 종류와 빈도를 보여주는 분석이며 구글 검색을 통해 얻은 장소 단어에 일정 부분을 더 추가하여 장소 말뭉치를 구성해 활용했다.

70년은 '밭', '고향', '휴전선' 등의 단어를 통해 산업화의 시작과 농촌 생활, 전쟁의 여파가 남아있는 모습을 알 수 있다. 80년대 또한 '고향'의 빈도가 가장 높았다. 아직 고향을 그리워하는 것을 알 수 있고 도시화의 시작이라고 볼 수 있는 '지하철', '도시' 등의 단어도 볼 수 있다. 90년대는 대표적인 도시 단어들이 많이 보인다. 도시화가 어느 정도 정착되어 있음을 의미한다. 2000년대는 도시를 넘어 여행, 세계 관련 단어들을 볼 수 있다. 2010년대부터 현재까지는 크게 3가지로 나뉘는데 '파티', '클럽' 등 유흥 관련 단어들을 통해 삶의 걱정과 고향에 대한 그리움 등이 대체되었고 그 연장선으로 '우주'와 같은 경험해보지 못한 세계에 대한 열망도 드러난다. 마지막으로 지역 명이 많이 등장하는데 미디어와 통신의 발달에 따라 SNS와 대중매체에 노출된 지명의 등장이 증가한 것으로 보인다.

(3) 장르별 핵심어 분석

우리나라 대중 가요는 다양한 장르로 이루어져 있고 해당 장르별로 사용되는 단어가 어떻게 다른지 알아보기 위해 장르별 가사 분석을 행했다. 분석은 TF-IDF를 이용하였고 성인가요/트로트, 록/메탈, 발라드, 댄스, 랩/힙합 이렇게 5가지 장르를 대상으로 하였다. 트로트, 록, 발라드 같은 경우 한글 가사들이 많이 포함되어 있어 영어 단어를 제외하고 TF-IDF 분석을 했을 때, 충분한 관측치가 존재했다. 공통적으로 사랑, 그대, 마음과 같은 사랑과 슬픔관련 단어가 높은 TF-IDF 점수를 가졌고 그 중 트로트의 경우 여인, 바람 등 특유의 장르 단어들이 높은 점수를 얻은 것을 확인할 수 있다. 댄스와 랩/힙합 장르의 경우 한글 가사만으로는 분석을 행할 수 없어 영 단어를 포함하여 분석을 실행하였고 그 결과 love, you, baby와 같은 영단어가 높은 점수를 얻은 것을 알 수 있다.

(4) 작사가 별 핵심어 분석

21세기 노래로 한정하여 작사가 순위를 추출한 후 우리가 알만한 상위 5명의 작사가를 골라냈다. 그 결과 박진영, G-DRAGON, 김이나, 아이유, TEDDY 총 5명의 작사가가 선정되었다. 우리는 각 작사

가별 단어빈도와 핵심어를 관찰하기 위해 명사를 추출하여 분석해보았다.
첫 번째로 데이터셋에서 가장 많은 곡을 작사한 TEDDY 작사가 노래의 경우 <사랑, 지금, 남자, 세상, 시간> 의 단어들이 많이 쓰인 것을 확인할 수 있었다. 총 18개의 곡 중 11개가 댄스곡이었고, 랩/힙합이 4곡, 일렉트로니카, 발라드, R&B/Soul이 각각 1곡씩 있었다. 두 번째로 박진영 작사가 노래의 경우, <생각, 사랑, 여자, 남자, 행복> 의 단어들이 많이 쓰인 것을 알 수 있었다. 박진영 노래의 경우 총 15개의 노래 중 발라드가 5곡, 댄스곡이 10곡이었다. 세 번째, GD의 경우 <사랑, 눈물, 하나, 생각, 마음>의 단어가 많이 쓰였으며 총 13개의 곡 중 장르가 댄스인 곡이 5곡, 랩/힙합이 7곡, 록/메탈이 1곡이었다. 네 번째, 아이유의 곡을 살펴보면 데이터셋에 총 10개의 곡이 존재하였고 <사랑, 꽃잎, 우리, 지금, 사람> 의 단어들이 많이 쓰였다. 아이유의 경우 발라드, 댄스, R&B/Soul, 록/메탈 등 장르가 다양한 것을 확인할 수 있었다. 마지막으로 김이나 작사가의 노래는 총 8곡이었다. <사랑, 내가, 지금, 남자, 나의> 단어들이 김이나 작사가의 노래에서 자주 등장하였고, '내가', '나의' 단어들이 많이 등장하는 것으로 미루어볼 때, 노래의 내용이 대체적으로 주체적임을 알 수 있다. 김이나 작사가의 경우 댄스 장르가 5곡으로 가장 많았으며, 발라드, 성인가요/트로트가 각각 2곡, 1곡으로 뒤를 이었다.
전체적으로 사랑이라는 주제를 많이 쓰는 것을 확인할 수 있었으며 아이유의 꽃잎을 제외하면 작사가별 뚜렷한 개인의 단어는 발견하기 힘들었다.

3. 기대 효과 및 결론

이번 프로젝트에서 최종 목적이었던 Attention기반 모델을 통해 대중가요의 가사 가이드라인을 만드는 활동으로 이어지지는 못했다. 프로젝트 진행 중, 멜론과 같은 스트리밍 위주의 플랫폼이 대표적 대중가요를 선택하는 기준이 명확하지 않다는 점과 대중가요가 가지고 있는 반복적인 가사와 그 시대에 반영된 유행어들을 처리를 시도하여 분석을 진행하였지만 애매한 경향이 있다는 아쉬움이 있었지만 다음과 같은 결론을 얻을 수 있었다.

1970년대에 사용한 어휘들을 분석한 결과, 전쟁으로 인해 '휴전선', '대동강' 등의 어휘들이 대두되었다. 1980년대에는 이의 여파로 인해 '고향', '우리땅' 등의 어휘들이 자주 등장했다. 1990년대에는 국제 대회 및 박람회들의 유치로 인해 '전시회장'이라는 단어가 키워드로 사용되었다. 2000년대에는 '비행기', '바닷가'와 같은 여행과 관련된 어휘들이 자주 등장했다. 2010년대에는 '여수', '양화대교'와 같은 구체적인 지역의 명칭이 직접적으로 언급된 경우가 많아졌다. 이와 같이 각각의 시대상을 반영하는 어휘들은 간접적으로 보여졌다. 2010년대 전후를 비교하면 전에는 '세상', '우리', '하나'와 같은 어휘들이 자주 쓰였다면 그 이후에는 이러한 표현들 보다 '내가'라는 표현이 강조되었다. 이는 사회 전체에 비해 개인의 자아에 더 집중하는 시간을 가질 수 있게 된 사회적 변화로 보인다. 추가적으로 진행한 장르별 가사분석과 작사가 별 핵심어 분석을 통해서도 가사적 특징을 확인 할 수 있었다. 장르별로 가사를 분석한 결과, 트로트에서는 '여인', '바람'등의 단어가 자주 사용되었다는 것과 댄스와 랩/힙합의 경우는 영단어, 특히 'love', 'you'등의 어휘들이 많이 사용되었다는 것을 확인할 수 있었다. 작사가 별 핵심어까지 분석을 진행하면서 이를 통해 한국의 시대적 배경을 알아보고 사회 전반이 개개인에게 미치는 영향을 예측하는 연구에 도움을 줄 것으로 기대된다. 또한 대중가요의 가사 분석에 대한 연구가 아직 많이 이루어지지 않은 상태이기 때문에 이를 계기로 연구가 더 활발하게 이루어질 것이라고 기대

된다. 또한 이를 활용한 가사 예측 모델과 같은 여러 프로그램으로 개발되어 실용화할 수 있을 것으로 기대된다.

4. 참고 문헌

[1] 장유정 (Eu Jeong Zhang). "기획논문 : 1970-80년대 한국 대중가요 가사의 특징 -공중파 방송 인기곡을 중심으로-." 공연문화연구 0.24 (2012): 79-113.

[2] 문숙희.(2012).「1970-80년대 서울 관련 대중가요의 두 모습」 에 대한 토론문.한국문학과 예술,10(),206-208.

[3] Noh Young-Hae. "Main Themes of Korean Popular Songs in the Last 30 Years." 음악과 문화 5.- (2001): 149-183.

[4] Noh Young-Hae. "Main Themes of Korean Popular Songs between 1940 and 1970." 음악과 문화 7.- (2002): 113-142.

[5] 김용학. "한국 대중가요의 의미 연결망－1960년대부터 2000년대까지의 변화를 중심으로." 대중서사연구 21.1 (2015): 145-171.

[6] 장소원 (So Won Chang). "한국 대중가요 가사의 문체 분석." 텍스트언어학 39.- (2015): 283-311.

[7] 장유정 (Zheng Eu-jeong). "1990년대와 2000년대 '서울 노래'의 두 모습." 人文科學 112.- (2018): 7-29.

[8] 홍성규,이주원,and 서재혁. "1980년대 초기 한·일 헤비메탈음악의 발전 양상과 그 의미에 관한 고찰 － 대표적인 밴드들의 작품 활동을 중심으로 －." 일본학연구 55.- (2018): 95-120.