# [유튜브 유사 태그 추천 프로젝트 테크니컬 리포트]

게임소프트웨어학과 B977008 김재환

## 1. 서론

유튜브는 현대인들이 현재 가장 많이 소비하는 비디오 컨텐츠 플랫폼 중 하나이고 최근에 업데이트로 유튜브에서는 영상 검색 및 분류를 돕기 위해 태그를 사용합니다. 이번 프로젝트는 유튜브 영상 설명에 있는 태그 데이터 셋을 활용하여 단어 간의 연관성을 파악하고 시각적으로 보여주며 내가 원하는 단어와 관련된 다른 태그들을 추천해주는 프로젝트를 진행하였습니다.

데이터 셋을 만드는 과정에서 태그가 없는 영상도 있고 태그에 연애인 등 유명인의 이름이 들어가는 경우도 많아 사람의 이름이 많이 나올 수 있고 같은 의미에 영어와 한글을 분류를 하기 힘들 것 같았고 많은 데이터 셋을 모으지 못하였기 때문에 모델의 성능에 제한이 있을 것으로 보이지만 WordCloud와 Word2Vec를 활용하여 단어들의 관련성을 보기 쉽게 만들려 했습니다.

## 2. 본론

### 2.1 데이터 수집 및 전처리

데이터셋은 ScrapeStorm이란 프로그램을 활용하여 유튜브 api를 이용해 태그 데이터를 수집하였습니다. 유튜브와 유튜버의 이름, 링크, 조회수, 태그를 csv 파일로 저장했습니다. 추출한 데이터를 바로 사용하기에는 여러가지 문제가 있어 데이터를 정제해주는 과정을 가졌습니다.

우선 중복 검사를 통해 겹치는 제목을 가진 데이터들과 확인 후에 영상 내용에 태그를 가지고 있지 않는 영상들도 삭제처리를 하여 정리를 하였습니다. 그리고 태그 데이터들을 뽑아 보니 아래의 그림1과 같이 자꾸 반복되는 내용들도 보여 데이터들을 집합화 하여 반복되는 데이터들을 삭제하여 정리를 해주었습니다.

| | 제목 | name | view | tag |
|---|---|---|---|---|
| 0 | 야 니가먼저 사과해...😒찐친 완전체 \| 퇴근하고 한잔할레오? ep.5 | LeoJ Makeup | 81.24만 | 레오제이 레오제이 찐친 레오제이 찐친 브이로그 찐친 완전체 레오제이 찐친큐... |
| 1 | 조카들의 사랑을 갈구하는 어른들의 선물 배틀...카야&라니 생일파티 🎂 | 해쭈 [HAEJOO] | 36.35만 | 해쭈 가족 쁘큐보이 카야 새삼이 브이로그 호주 foodtrip foo... |
| 2 | ✨최초 공개✨ 은우에게 여자친구가 생겼어요😁♥ [슈돌 유튜브] KBS 231212 방송 | KBS 슈퍼맨이 돌아왔다 | 43.54만 | 제로베이스원 제베원 제배원 성한빈 장하오 리키 한빈 하오 ZEROBA... |
| 3 | [#습포이드] "그 결혼, 나랑 하지?" 송강의 손을 잡은 김유정, 그리고 시작된 ... | SBS Drama | 41.12만 | 마이데몬 김유정 송강 마이 데몬 마이대몬 김해숙 도도희 구원 도희구... |
| 4 | 정국 (Jung Kook) 'Hate You' Official Visualizer | HYBE LABELS | 774.14만 | HYBE HYBE LABELS 하이브 하이브레이블즈 정국 Jung Kook... |

그림.1 전처리 전 데이터 셋

## 2.2 전처리 진행 후 데이터 셋

```
...
[list(['현석', '모음', '김지오', '친구', '퀀주호', '찐친', '레오제이', '브이로그', '지오', '주호다', '연말', '큐앤에이', '찐친연말', '친구들', '서기채널', '완전체', '주호', '한잔할
 list(['새삼이', '해쭈', 'foodtrip', '해쭈먹방', '카야', '호주', 'australia', '먹방', '브이로그', 'mukbang', '먹방브이로그', '요리', 'foodie', '고매요리', '쁘큐보이', 'cooking', '
 list(['싱글맘', '육아', '슈퍼맨이', '슈돌레전드', '차줘홀', '성한빈', '현실육아', '하오', '도장TV', '준', 'ZEROBASEONE', '난이도', '삼둥이', '지웅', '육아로그', '똥별', '제로베이
 list(['김유정', '명장면', '마이대몬', '마이데몬다시보기', '악마', '마이데몬송강', '도도희', '구원도희', '마이데몬1회', '몇부작', '마이', '도희구원', '구원', '도희', '송강', '김해
 list(['하이브', 'LABELS', '방탄', '방탄소년단', 'Kook', 'BTS', 'Jung', '정국', 'HYBE', '하이브레이블즈'])
 list(['소개', '인테리어', '별장', '제주', '단독주택', '꾸미기', '만들기', '집', '건축', '주택'])
 list(['j-hope', 'Jung', 'k-pop', 'BANGTAN', '슈가', '방탄', '방탄소년단', 'Kook', 'BTS', '지민', 'SUGA', 'Jimin', 'V', '정국', '알엠', 'RM', '제이홉', 'Jin', 'jhope'])
 list(['a', 'n'])
 list(['korean', '김장김치', 'kimchi', '김장', '절임배추20kg김장', '김장육수만들기', '김장배추절이는법'])
 list(['해외', '외국인', '귀여운', '네덜란드', '여행', '강아지', '먹방', '반응', '국제', '진돗개', '아기', '커플', '이민', '유럽', '혼혈', '가족'])
 list(['free', 'kit', 'mico', 'talk', 'pass', 'battle', 'stars', 'dani', 'brawler', 'royale', 'larry', 'lawrie', 'plus', 'brawl', 'skins', 'mobile', 'game', 'cartoon', 'supercell',
 list(['internationalcouple', 'vlog', '국제커플'])
 list(['love', 'drama', '사랑', '여사친', '드라마', '로멘스', '연애', 'web', '남사친'])
 list(['Actor', 'STANDARDFRIENDS', '고르며', '배우', 'K-pop', 'Zion.T', 'Ant', 'YGEntertainment', '겨울', '최민식배우', 'YG', 'Teaser', 'A', '알앤비', 'Stranger', '이수현', '최민식
 list(['짬뽕', '타노스', '서장훈', '주엽TV', '입짧은햇님', '백종원', 'PD', '오므라이스', '쭈엽', '투뿔', '짜장면', '요리', '채널', '최피디', '찍먹', '볶먹', '빡침', '맛집', '쭈엽E
 list(['별명', '내이름은카더가든', '아바타소개팅', '소개팅', 'SNL', '부럽지가', '헌팅', '카더가든', '아바타', '미팅', '장기하', '연애', '차정원', '알어', 'carthegarden', '아필소'],
 list(['주현미', '박나래', '남승민', '심수봉', '김다현', 'KCM', '공훈', '최윤하', '볼타는장미단', '전종혁', '양세형', '볼타는트롯맨', '한강', '이수호', '신성', '박현호', '박민수',
 list(['yéu', 'क्या', '소녀의행성', 'Cəgi', 'вельш-корги', 'planet', '리트리버', 'वस्त्र', 'पामेरेनियन', 'कोगी', 'girls', '웰시코기', 'ретривер', 'puppy', 'कुत्ते', '포메
 list(['배우', '김혜수', 'PD', '박진영', '제주도', '댄스', '방랑식객', '셀러브리티', '안무연습', '인간극장', 'When', 'We', '더먹고가', '밥상', '식사하셨어요', '송윤아', 'Coffee',
 list(['korean', '미국', 'vlog', '일상', '미국반응', '한미국제커플', '미국사는', 'husband', '한국', '외국인와이프', '국제부부', 'daddy', '국제커플', '한국남편', '외국인', 'isabell
 list(['김민선', '스피드스케이팅월드컵4차_토마슈프마조비에츠키', '500m'])
```

그림.2 전처리 진행 후 태그의 데이터 셋

## 2.3 데이터 셋 분석

전처리 전 태그 데이터의 단어 개수는 총 35134개로 나오고 데이터를 출력해보면 겹치는 단어들이 역시 너무 많이 나왔습니다. 하지만 호기심에 이를 워드클라우드를 이용해 뽑아보도록 하겠습니다.

```
temp_data = ' '.join(df["tag"].astype(str))

print(f"총 단어 개수: {len(temp_data)}")
print(temp_data)
```

```
총 단어 개수: 35134
레오제이  레오제이 찐친  레오제이 찐친 브이로그  찐친 완전체  레오제이 찐친 류앤에이  레오제이 찐친 모음  한잔할레오  레오제이 지오  주호  현석  퀸주호  서기채널  지오  김지오  주호다
```


역시 특정 인물과 관련된 영상들이 있으면 그 인물의 이름과 관련된 태그가 많기 때문에 이러한 일이 일어나는 것 같습니다.

전처리 후 데이터 셋은 이제 한 영상에서 반복되는 태그들을 리스트에서 집합화를 시켜 하나만 남고 나머지는 삭제하고 다른 종류의 태그들만 남도록 하였습니다. 그 결과 총 태그의 수는 5002개로 많이 줄어들었지만 카운트를 하자 이제 사람의 이름보단 먹방, 브이로그 등 영상과 관련된 주제와 맞는 태그들이 나오기 시작합니다.

```
result = []
for item in tag_list:
    result.extend(item)

print(result)
print(f"총 단어 개수: {len(result)}")
print(Counter(result))

temp_data = ' '.join(result)
```

```
['현석', '모음', '김지오', '친구', '퀸주호', '찐친', '레오제이', '브이로그', '지오', '주호다', '연말', '류앤에이', '찐친연말', '친구들', '서기채널', '완전체', '주호', '한잔할레오', '
총 단어 개수: 5002
Counter({'먹방': 33, '브이로그': 23, 'mukbang': 17, '예능': 16, '여행': 15, 'korean': 13, 'a': 11, '맛집': 11, '유튜브': 11, '노래': 11, '요리': 10, 'n': 10, 'vlog': 10, '토크쇼': 9,
```


워드 클라우드에 올라온 태그들을 보면 먹방, mukbang과 같이 같은 의미를 가지고 있지만 언어가 달라 다르게 표시되는 태그들을 정리를 해주어야 하는데 이는 제가 수작업으로 해야 할 것 같아 너무 시간이 오래 걸릴 것 같아 하지는 못하였고 바로 모델 학습을 진행했습니다.

## 2.4 모델

### 2.4.1 Word2Vec 모델

Word2Vec 모델은 Gensim 라이브러리를 활용하여 학습시켰습니다. 만들어질 워드 벡터의 차원(size)은 100이고 컨텍스트 윈도우(window)의 크기는 5, 학습에 이용하는 프로세스(workers)의 크기는 4이고 흔히 사용하는 학습 방식(sg)인 CBOW를 사용했습니다. 기존에는 단어 최소 빈도 수 제한(min_count)을 5로 두었는데 데이터 셋의 양이 그리 많지 않아 사용할 수 있는 단어의 수가 적어 제한을 4로 바꾸었습니다.

```
[71] # 모델 생성
     model = Word2Vec(sentences = tag_list, vector_size = 100, window = 5, min_count = 4, workers = 4, sg = 0)
     print(model)

     Word2Vec<vocab=137, vector_size=100, alpha=0.025>
```

```
[72] model.wv.vectors.shape

     (137, 100)
```

모델을 학습시킨 후에 모양을 확인해보니 단어는 137개와 설정한 것과 같이 100차원으로 학습했다고 나옵니다. 이제 단어들의 유사도를 확인해 보겠습니다.

```
print(model.wv.most_similar('먹방'))
print(model.wv.most_similar('케이팝'))
```

```
[('약뮤', 0.3052310347557068), ('KPOP', 0.24758237600326538), ('kpop', 0.2405254095792704), ('유튜브', 0.23257288336753845), ('나혼자산다', 0.2115024
[('Korean', 0.21265195310115814), ('뮤직비디오', 0.20730389654636383), ('코믹쏫무비', 0.1967858970165252), ('토트넘', 0.1942834109067917), ('정국', 0
```

직접 입력한 태그와 관련된 태그들이 나오기는 하지만 유사도가 너무 낮습니다. 이를 해결하기 위해 위에 단어 최소 빈도 수를 늘리는 등 여러가지를 변경을 해보았습니다만 데이터 셋의 크기가 작아 아직 모델이 충분히 학습을 못하는 것 같습니다.

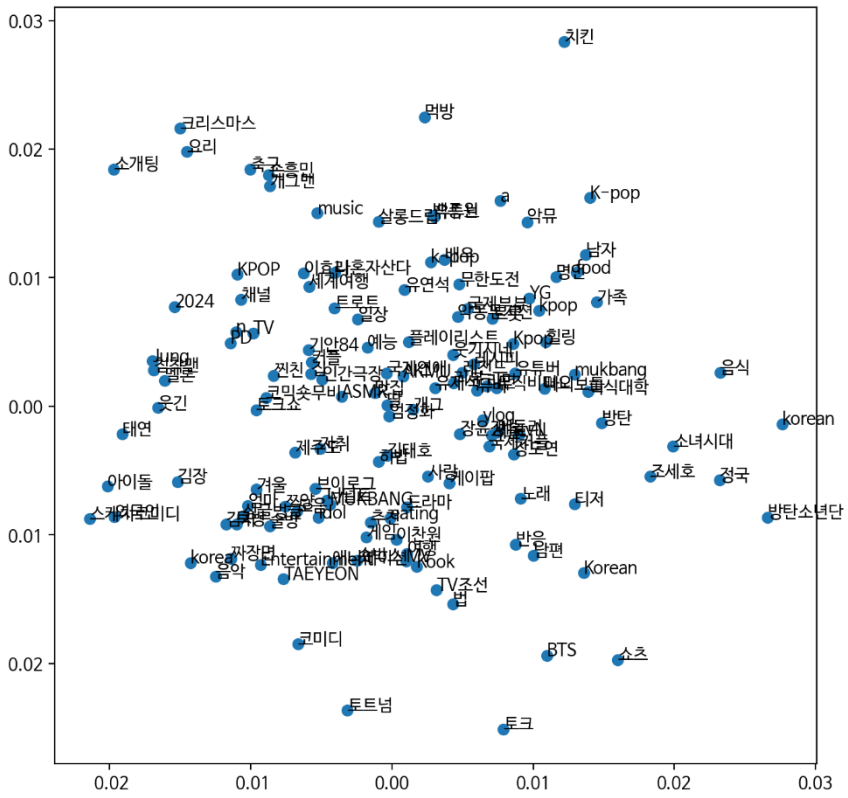
### 2.4.2 모델의 임베딩 결과 시각화

학습시킨 Word2Vec 모델의 임베딩 결과를 시각화 하는 방법을 찾아보니 학습으로 뽑은 여러 차원의 벡터를 시각화하기 위해서는 2차원 또는 3차원으로 벡터를 축소해야 하는데 이를 사이킷런의 PCA 모듈이 지원합니다.

n_components에 몇 차원으로 축소시킬 것인지 입력하고 가지고 있는 워드 벡터를 형태에 맞게 변환시킨 후 2차원 벡터의 x값과 y값을 각각 xs, ys에 저장을 하고 2차원 그래프로 출력해보겠습니다.

```
word_vectors = model.wv
vocabs = model.wv.index_to_key
word_vectors_list = [word_vectors[v] for v in vocabs]

pca = PCA(n_components=2)
xys = pca.fit_transform(word_vectors_list)
xs = xys[:,0]
ys = xys[:,1]
```




벡터들이 가운데에 많이 모여 있지만 벡터의 양이 많지 않고 100차원의 벡터를 2차원으로 나타내다 보니 서로 어떻게 연결되어 있는지 보기가 좀 아쉬운 것 같습니다.

## 3. 결론

수업에서 받은 데이터들로 모델들을 학습시켜서 잘 몰랐지만 일단 데이터 셋을 처음에 크롤링하여 모으는 것도 생각보다 힘들고 왜 시간과 비용이 많이 드는지 알게 되었습니다. 그렇게 모은 데이터 셋의 크기도 많이 작아 모델에서 나온 단어들의 유사도가 낮았습니다. 계속 작업을 하게 된다면 데이터 셋의 크기를 지금보다 몇 배는 더 키워야 할 것 같겠다는 생각이 들었습니다. 그리고 처음에 목표했던 것처럼 조회수를 점수로 이용하여 더 인기가 있는 태그를 분류하고 학습하여 한 주제에 관련된 인기있는 태그를 추천하게 만들어 보고 싶습니다.

4. 참조 사이트 및 부록

데이터 크롤링 타겟 사이트: https://kr.noxinfluencer.com/youtube-video-rank/top-kr-all-video-day

유튜브 인기 동영상 데이터 분석: https://95pbj.tistory.com/29

Word2Vec 표현과 학습 방식: https://wikidocs.net/22660

기말 프로젝트 실습 진행 collab 링크:

https://colab.research.google.com/drive/1rt2njb0D0ar-ivkf8IMKZYiOCuYA-6HI#scrollTo=J5CcJ-8rLlxT

실습 데이터 셋:

https://drive.google.com/file/d/1glcO3m3HdpFend1I8MV3wYFTYmLeVS2H/view?usp=sharing