Data Science Team Project

Final Report

**Team 1**

2012210080 이인엽

2012210077 박지현

2012210059 이주림

Contents

1. Introduction – Motivation and goal ………………………………… p. 2
2. Implementation ………………………………………………………………... p. 3
3. Procedure …………………………………………………………………………. p. 3
4. Result ………………………………………………………………………………... p. 9
5. Explanation ………………………………………………………………………. p. 10
6. Conclusion – Application area ………………………………………… p. 14
7. **Introduction - Motivation and goal**

These days many companies invest a lot of money for advertising and building marketing strategy. Comparing the cost, many companies can’t result in sales improvement. If we can configure if somebody likes specific brands also likes something else, this information can be used for target marketing. Also, in this case, enterprises want to buy this information for high price.

We thought that the decision making process would not be much different in the area of art and also there must be common factors feeling something is beautiful. We call it ‘taste’ and we believe the taste of person applies in the same way when they choose what to buy. In this project, we first focused on the relationship between the brand and art. In this case, art refers to movie, music, painting, pictures and so on, but we excluded the paintings and pictures because of less relation between them and brands comparing to movie and music.

Our goal is to find the preference of brand from one’s art works such as music, movie, and paintings. By making it to application, we want to extract the best brand related to the one’s favorites. With this application, we can suggest the company to place the pictures that customers most likely to prefer, play the music in the store or build any other marketing plans they can do with the art works. It would improve the effectiveness of advertisement, make people stay longer in the stores, and result in an increase of the possibility to spend money. I’ll explain in detail later.

프로젝트를 기획하면서 예술작품에 있어서의 취향이 브랜드를 선택하는데 영향을 미칠 것이라고 생각했다. 왜냐하면 어떤 것이 아름답다고 느끼고 끌리는지는 어떤 작품의 선호도를 통해 드러나고, 이러한 취향은 브랜드를 선택하는데 영향을 끼칠 것이라고 판단했기 때문이다.

따라서 작품의 선호도와 브랜드 선호도를 분석하여 어떤 브랜드를 선호하는 사람들이 어떤 작품을 좋아하는지 알 수 있다. 이 정보를 통해 회사들은 자신의 브랜드를 좋아하는 사람들 혹은 싫어하는 사람들에 맞춘 마케팅 전략을 세울 수 있다.

프로젝트 제안 때에는 음악과 영화에 더불어 예술 작품도 포함시키기로 했다. 하지만 실제 조사를 해보니 예술 작품과 브랜드의 연관성을 인스타그램에서 찾기 어려웠기 때문에 이번 프로젝트에서는 제외하기로 했다.

1. **Implementation**

* Language: Python
* Tool: ipython notebook, visual studio code, sublime
* Version control tool: github (https://github.com/inyeoplee77/DataScienceProject)
* Database: iMDB(6099 movies), Apparel search(342 brands)
* API: Instagram api, The Echo nest(3256 musics)
* Data machine learning model: SVM (<http://scikit-learn.org/stable/modules/svm.html>)

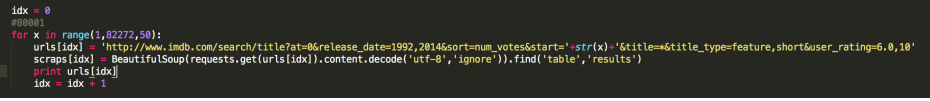
1. **Procedure**
   1. Construct Movie DB, Music DB, and Brand DB
      1. **iMDB :** iMDB has a great database of movies. It is an online database of information related to films, television programs, and video games, and so on. We could extract movie data such as title, ratings(0.0~10.0), votes, director, stars, genres.

Parsing movie date was an issue. We parsed all pages by increasing start from 1 to 82,271. (web page: [http://www.imdb.com/search/title?at=0&release\_date=1992,2014&sort=num\_votes&start=51&title=\*&title\_type=feature,short&user\_rating=6.0,10](http://www.imdb.com/search/title?at=0&release_date=1992,2014&sort=num_votes&start=51&title=*&title_type=feature,short&user_rating=6.0,10)) This web site only prints out rating 6.0 or higher movies in 1992 to 2014. This is because we though too old or to low-rated movies are unnecessary. We’ve got 7 features which are genre, director, rating, num of voted, title, year per actor, and if one of feature lack of information, we didn’t save it. To adjust the algorithm, we had to have all information. With this process, we got 68,316 movies.

The difficulty of parsing was that web site is not always same, so NoneType error has occurred many times. Also, exception made us took a lot of time, but it was not hard problem. Instead, execution time is too long, so if I want to modify the code, it took much time to get the result. Searching for document of BeautifulSoup also took a lot.

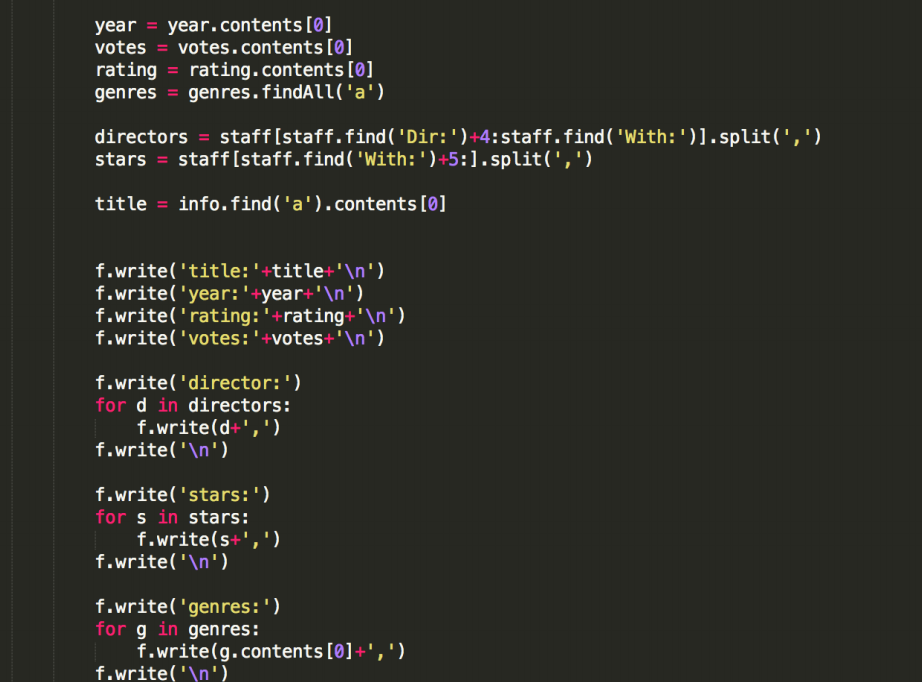
Source code is below:

First, send a request and get web pages.



Then, parse each page. Make exception and extract and write data which only passed through the exception.

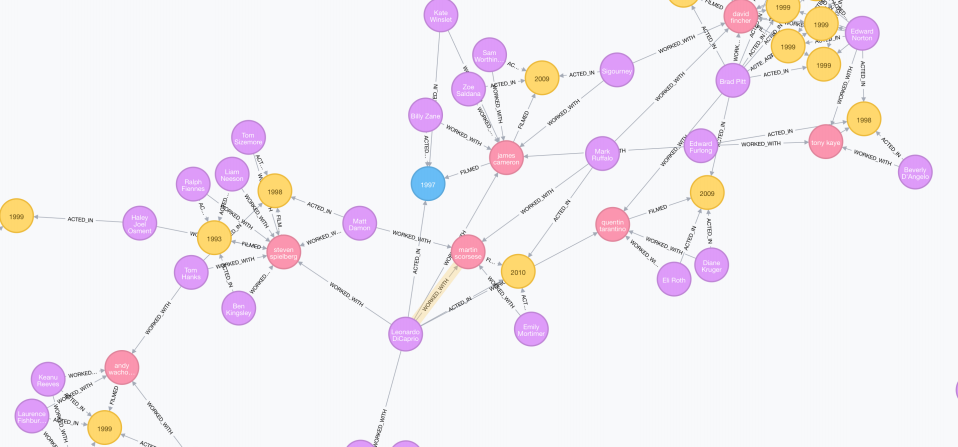


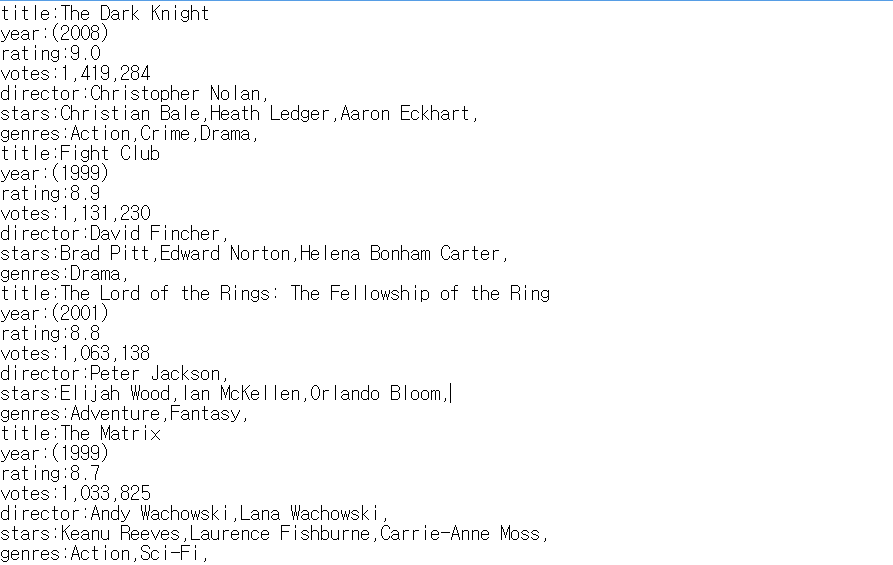


We constructed movie database with NEO4J. The procedure is below.

* + - 1. The way of representing special characters in HTML was different from utf8, so we added code to convert.
      2. In cypher, we couldn’t use ‘-‘, so converted it to ‘\_’
      3. It took time to make connection to neo4j
      4. It took time to make cypher query code.
      5. Read file and made code making query.
      6. Success.

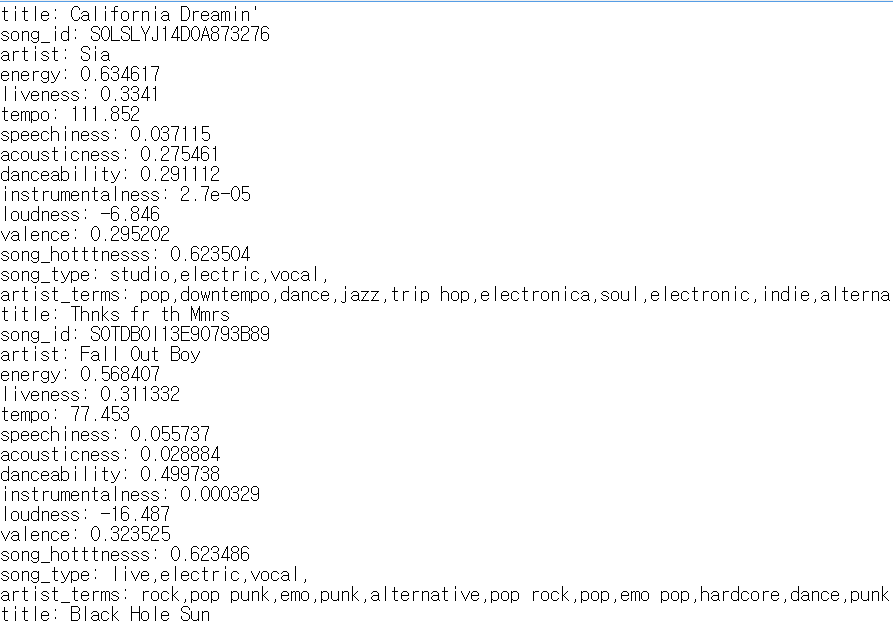






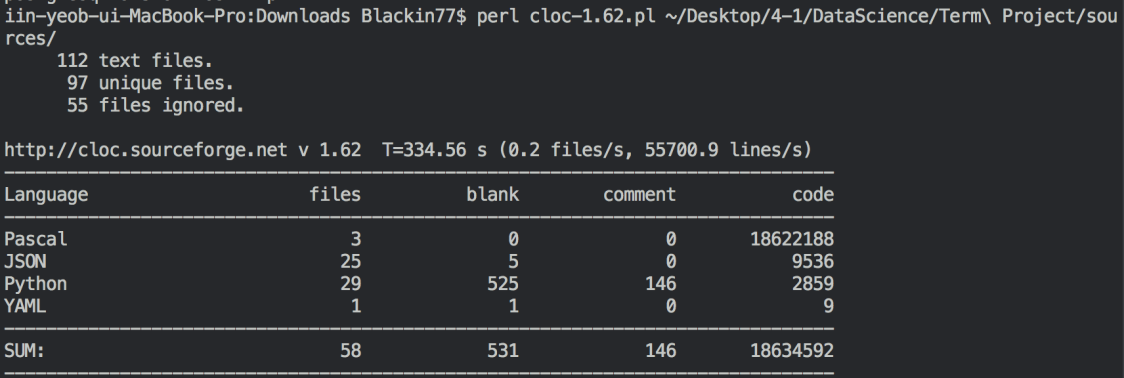
하지만 결국 neo4j는 사용하지 않기로 결정되었다. 원래 neo4j를 이용하여 영화간의 유사도 관계도 고려하려고 했었다. 하지만 계산량이 너무 많고 프로젝트가 너무 커지는 것 같아 포기하게 되었다. 대신 txt파일을 쓰기로 했다.

* + 1. **The Echo Nest:** When we first tried to get music data from Discog api, allmusic website, there were many problems. Using Discog api, there was protocol message problem which we couldn’t solve. Therefore, we tried to discover new method to get music database, using allmusic.com website. However, we couldn’t get valuable data because of the shortage of data of old musics. It took over 1 weeks to get music data, but because of the lack of organized music data, we deleted a lot of codes. Finally, we got to know the echo nest api, which has quite big database with a lot of features that we were looking for. Features we used in this project are Title, Song\_id, Artist, Energy, Liveness, Tempo, Speechiness, Acousticness, Danceability, Instrumentalness, Loudness, Valeance, Song\_hotttnesss, Song type, Artist\_terms



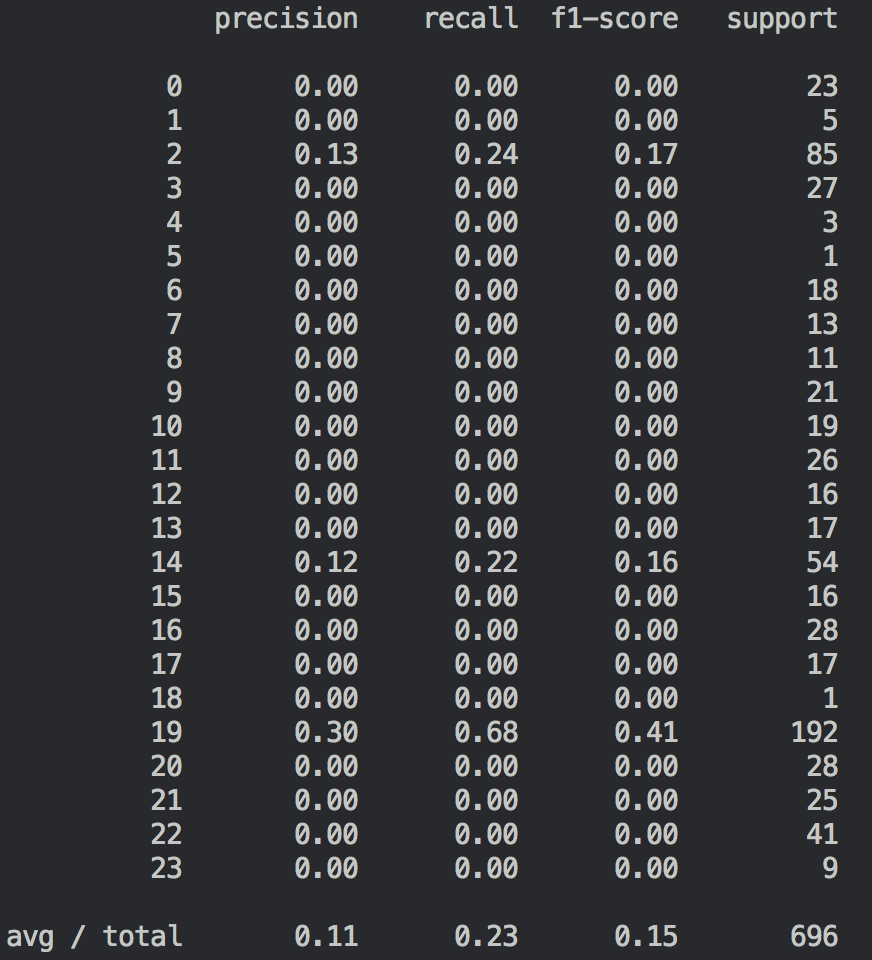
* + 1. **Apparel search:** In Apparel search website, we got the list of apparel brands.
  1. Search Instagram user id after searching movie or music tags
  2. Search brand name from each user id
  3. Train the data with SVM model
* While making data vector for training set,
* The reason why we chose SVM model instead of other models such as neural network model is that we’ve found out using neural network model takes too much time for training and needs to decide number of layers and nodes. Also, SVM model shows better result than others.
  1. Make application
* If you enter Instagram id, the result brand which the algorithm considered most preferable comes out.
  1. Test

1. **Result**
   1. The number of codes we’ve made effort was this:



저희는 많은 코드를 작성했습니다. 하지만 이 보고서에 코드를 많이 담지 않은 까닭은 코드에 자신 없어서가 아니라 보고서가 코드로 지저분해 지는 것이 싫어서 입니다. 첨부한 압축파일을 풀어 코드를 보시고 저희가 얼마나 열심히 했는지 확인해주시면 감사하겠습니다.

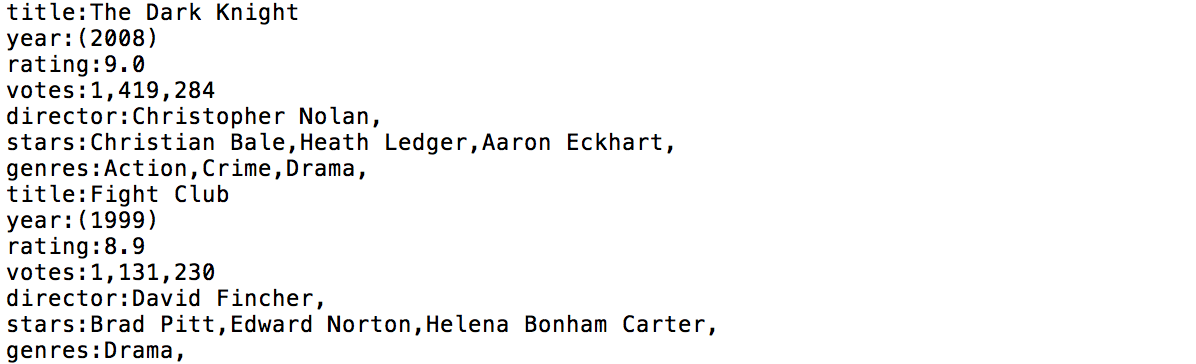
* 1. Result: The left side is answer and the right side is prediction. 이때 제출하는 어플리케이션은 Instagram의 아이디를 넣으면 해당 사람이 좋아할 만한 브랜드를 출력한다. 또한 이 어플리케이션은 전체 시스템의 데모이며, 실제로 기업에서 사용을 하는 경우에는 기업이 여기에서 도출해 낸 결과를 비교 분석해서 자신의 상품 홍보 및 고객 유치 등에 이용을 할 수 있다. 결과가 그렇게 좋지 않은데, 이는 training set이 적기 때문이라고 생각된다. 시간이 너무 오래 걸려서 많은 데이터를 뽑아내지 못했다. 하지만 데이터가 많아지면, 충분히 좋은 성능을 낼 것이라 생각한다. 그리고 데이터를 뽑는 알고리즘은 모두 구현되어있으니 구하는 것은 단지 시간문제이다.



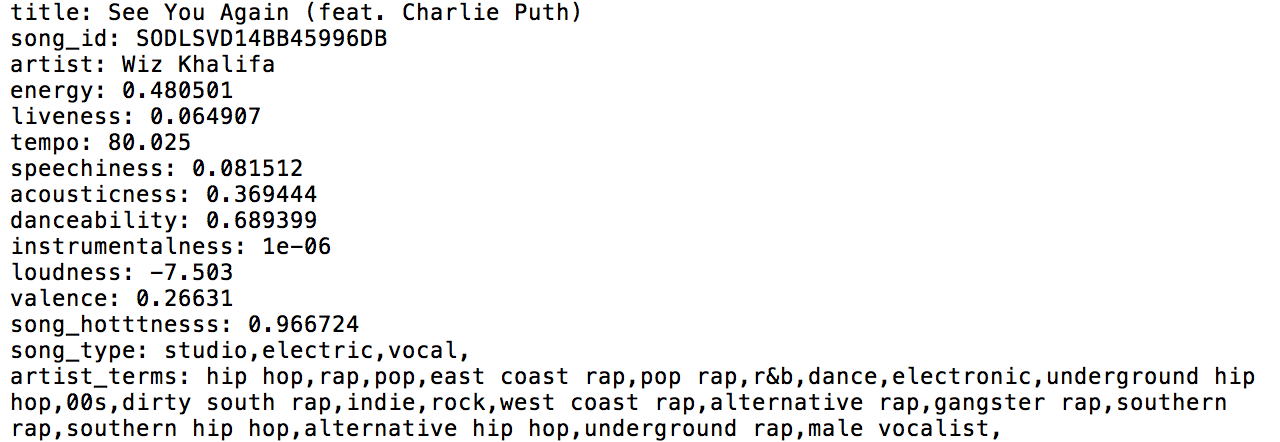
힘들었던 점은 여러가지가 있다. 다양한 api를 이용해서 각자 다른 형태의 데이터를 다루어서 하나로 합쳐야 한다는 것도 힘들었고, 그 데이터를 어떤 모델에 어떻게 적용해야 하는지도 힘들었다. 한 예로, 사용자마다 각각 다른 길이의 데이터가 나오기 때문에 이를 학습 모델에 바로 넣기는 어려웠다. 그래서 우리는 모든 feature를 가진 큰 벡터를 만들고 사용자의 데이터에 포함되는 feature만 1로, 나머지는 0으로 만들어서 인풋으로 넣었다. 또, 처음에는 만족할 만한 결과가 나오지 않아서 학습모델의 파라미터를 gridsearch 방법으로 최적화했다.

1. **Explanation**

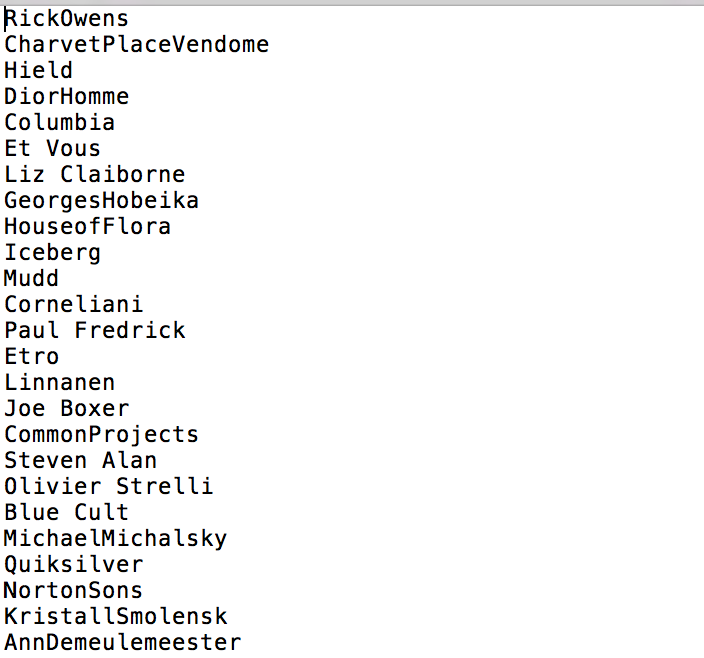
소스코드는 Movie, Music, Train, Instagram, Brand, Application 폴더에 분류되어 생성되어있다. 먼저 Movie 폴더에는 parser.py라는 코드가 있는데 이 코드는 imdb 사이트를 파싱하여 Movie\_DB.txt를 만들어낸다.



Music 폴더에는 get\_music\_data.py가 있고, 이 코드는 echo nest API를 이용해 Music\_DB.txt 를 만들어낸다.

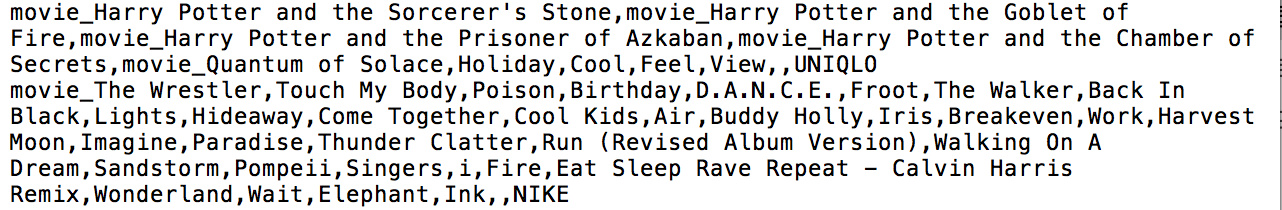


Brand 폴더에 brandList.py가 있고, 이 코드는 brand 리스트를 특정 사이트에서 받아온다. 결과 brand\_DB.txt를 만든다.



이제 인스타 그램에서 유저의 데이터를 받아와 검색하기 위한 기초 자료들이 완성되었다. Instagram 폴더에 위 과정의 결과들이 포함되어 있다.

먼저 우리가 해야하는 일을 순서대로 설명하겠다. 먼저, Moive\_DB에 있는 영화 제목으로 인스타그램 태그를 검색하여 가장 많은 게시글이 달린 태그를 받아와 그 태그가 달린 게시글을 쓴 유저들을 받아온다. 이때 영화 제목으로 가능한 검색 조합을 생성하여 검색한다. 예를 들면 영화 제목이 Harry Potter and the Deathly Hallows: Part 2 라면, 생성가능한 검색어는 harry potter, harry potter and the deathly hallows, the deathly hallows, Harry Potter and the Deathly Hallows: Part 2 가 될 것이다. 영화영화-유저 정보가 담긴 자료가 user\_movie\_DB.txt 이다.

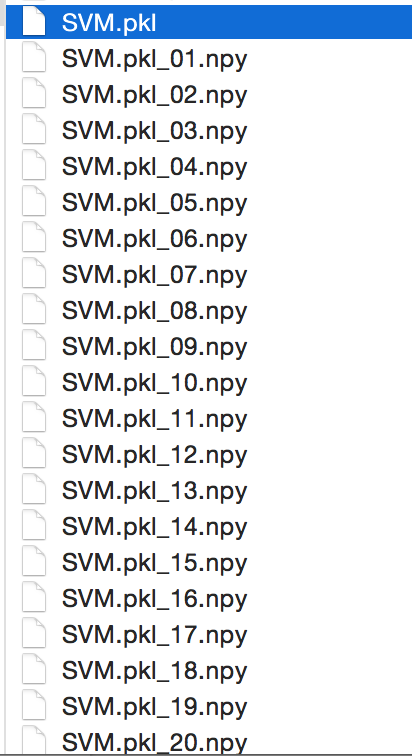


이제 insta\_training\_data.py에서 user\_movie\_DB.txt를 읽어온다. 이젠 그 유저가 올린 모든 게시글을 탐색하면서 태그된 음악과 브랜드를 모두 찾아낸다. 이때, 태그를 검색해야 하는데, 매번 검색하기 번거로워서 검색 가능한 태그를 모두 뽑아놓는데, 그게 tag\_search.py이고 여기서 만들어지는 태그들은 music\_tag.txt, brand\_tag.txt 파일로 만들어진다. 이 과정이 모두 끝나면, training.txt가 만들어지는데, 여기 들어간 자료는 어떤 사용자가 태그를 건 영화들과 음악들, 그리고 브랜드가 들어간다. 즉, 트레이닝 데이터에서 x 벡터가 영화와 음악이 되고, y가 브랜드가 된다.

하지만, 사람마다 태그가 달린 영화와 음악 갯수가 다르므로 x벡터가 가변길이가 된다. 따라서, 모든 feature들을 포함한 약 9000이상의 dimension을 가지는 벡터를 만들었다. 이 feature란 영화의 감독, 장르, 배우, rating, vote와 음악의 여러 정보들을 말한다. 예를 들면, 사용자가 harry potter를 봤다고 하면, 간단히 말해서 벡터에서 david yates = 1, daniel radcliffe = 1, Adventure = 1 나머지는 모두 0 이 된다.



그리고 y벡터는 모든 브랜드가 있고 그 중 하나만 1이고 나머지는 모두 0인 벡터가 된다. 이런 벡터를 만드는 과정에 data\_vector.py이다. 이 과정의 결과는 x.p, y.p 파일이다. 그럼 한 사용자마다 약 9000개 중에 수십개가 1이 되는 벡터를 가지게 된다. 그리고 y는 한개만 1이 되는 벡터이다. 이제 러닝 모델에 넣으면 된다.그 과정이 train.py 이다. 우린 SVM을 사용했고 파라미터는 grid search를 통해 결정했다. 이 결과 만들어지는 모델을 SVM.pkl로 저장했고 이는 이제 Application에서 쓰이게 된다. Application.py에서는 사용자의 인스타그램 username을 치면, 그 유저의 게시글을 분석해 vector로 만들고 svm에 넣어 결과를 보여주는 프로그램이다.



이 어플리케이션은 우리가 만든 시스템이 실제로 작동한다는 것을 보여주기 위한 하나의 예일 뿐이고, 우리가 실제로 하고 싶었던 것은 이러한 시스템을 구축하는 것이었다. 이제 이 시스템을 브랜드 회사에서 이용하여 자신의 브랜드를 사용하는 사람들이 어떤 영화나 음악을 좋아하는지 알게 되면, 그에 맞춰 마케팅 전략을 더욱 효율적으로 짤 수 있을 것이다.

1. **Conclusion - Application area**

With this application, many apparel companies and also non-apparel companies can make advertisement and make many people get to know about their products. By analyzing preference of people by inserting titles of movie and music, there are many things to do. I’ll introduce few specific examples.

First, imagine you are window shopping in the shopping mall with many clothes stores. When you walk across the Abercrombie & Fitch store, you heard your favorite song, which make you stop by the store. It would make people buy their products and consequently, it will connect to increase of sales.

Second, these days we can see a lot of product placement advertisements, as known as PPL. Sometimes, PPL seems weird because the product seems not relevant and to the original soap drama, movie, and so on. Also, sometimes PPL products can’t interest people because people actually don’t care about products. Therefore, if advertisement agency can analyze the relation between their brand and preference of the program, they can easily advertise their products.

Third, collaboration of art and movie will be amazing. To be specific, assume the result shows that some people who like movie ‘The Avengers’ series like the apparel brand ‘Adidas’. Actually, there are already products of ‘Adidas’ collaborated with ‘The Avengers’. However, if more brands which people couldn’t even imagine collaboration of two brands got collaborated and many awesome products been made, it couldn’t be better than anything else for their fans.

With these specific examples, the importance of finding out the relation between art and brands is evident. Considering the increasing scale of the advertisement, fusion of different element with brands will pioneer new area of advertisement of brands.