Neshat Jalali Heravi

Building a Content-based Movie Recommendation System

Goal: The objective of this project is to utilize Natural Language Processing on a movie meta-dataset and build an unsupervised learning model that creates different movie profiles using topic modeling, which can be used as a basis of recommendation system. The proposed recommendation system can be beneficial to new streaming services without sufficient user history or established streaming services' consumers who are looking for movies based on their current topics of interests rather than their user history!

Process: The dataset for this project contains information of 45,466 movies featured in the Full MovieLens dataset (movies that are released on or before July 2017) and is available on Kaggle. After the preliminary data cleaning and text preprocessing (including removing numbers and punctuations, text tokenizing, parts of speech labeling and lemmatizing etc.) the dataset is left with 3 feature columns (title, released_date, and overview_lemm) and 38,148 rows where each represents a movie. Next, TFIDF vectorizer followed by TruncatedSVD and NMF topic modelers were tested to extract topics from the preprocessed overview plots. I decided to continue with the TFIDF vectorizer and NMF topic modeler to create 15 topics. Sample of topics extracted using Truncated SVD and NMF can be found in the following pages.

From Here: The cosine similarity function will be used to find movies that shared similar topic profiles. Then the recommendation system will be designed to take in a movie title and output couple of most similar movies based on their topic profiles.

In [6]:
```
#Topic Modeling (15 topics, 15 words each)

my_stop_words = text.ENGLISH_STOP_WORDS.union(['this', 'when', 'each','film', 'by', 'as'])
corpus = data['overview_lemm']
preprocessor = None
vectorizer = TfidfVectorizer(stop_words=my_stop_words)
topic_modeler = NMF(15, random_state=10, max_iter=1000)
print_n_words= 15

topics_15 = extract_topics(corpus, preprocessor, vectorizer, topic_modeler,print_n_words);
topics_15
```

```
Topic 0:
 LIFE, CHANGE, LIVE, FOREVER, DAY, REAL, DEATH, TURN, GOOD, STRUGGLE, DREAM, PERSONAL, PAST, EXPERIENCE, PEOPLE

Topic 1:
 LOVE, FALL, MEET, MARRY, BEAUTIFUL, RELATIONSHIP, AFFAIR, ROMANTIC, DAUGHTER, MARRIAGE, TRIANGLE, HEART, ROMANCE, TRUE, LOVER

Topic 2:
 STORY, TELL, TRUE, BASED, BASE, SET, NOVEL, TALE, PEOPLE, FOLLOW, WAR, BOY, BOOK, SHORT, DIFFERENT

Topic 3:
 WOMAN, YOUNG, HUSBAND, MEN, BEAUTIFUL, LOVER, MYSTERIOUS, AFFAIR, MARRIED, COUPLE, SEXUAL, RELATIONSHIP, MARRIAGE, MEET, SEARC
H

Topic 4:
 SCHOOL, HIGH, STUDENT, TEACHER, NEW, COLLEGE, CLASS, GROUP, SENIOR, GRADUATE, TEAM, KID, BULLY, POPULAR, CLASSMATE

Topic 5:
 YEAR, OLD, LATER, BOY, RETURN, AGO, LIVE, AGE, PARENT, SUMMER, PAST, SPEND, SEVEN, PRISON, TIME

Topic 6:
 FAMILY, HOME, BROTHER, HOUSE, CHILD, RETURN, SISTER, MEMBER, PARENT, COME, SECRET, FORCE, STRUGGLE, NEW, YOUNG

Topic 7:
 GIRL, BOY, YOUNG, TEENAGE, LITTLE, PARENT, MEET, DREAM, COME, GUY, TEEN, SUMMER, AGE, RUN, NIGHT

Topic 8:
 MURDER, KILL, POLICE, KILLER, GANG, CRIME, FORCE, COP, DETECTIVE, CASE, LEAD, CRIMINAL, DEATH, AGENT, PRISON

Topic 9:
 TOWN, SMALL, LOCAL, SHERIFF, VILLAGE, COME, COMMUNITY, ARRIVE, BIG, PEOPLE, CITY, RESIDENT, CITIZEN, NEW, BOY

Topic 10:
 MAKE, WIFE, WORK, WANT, DAY, LEAVE, TIME, TRY, MEET, DECIDE, JOB, MONEY, START, COME, NEW

Topic 11:
 FRIEND, BEST, HELP, CHILDHOOD, GROUP, PARTY, FRIENDSHIP, TRIP, GIRLFRIEND, TURN, COLLEGE, CLOSE, ADVENTURE, WEEKEND, RELATIONS
HIP

Topic 12:
 MAN, YOUNG, WIFE, KILL, DIE, HIT, OLD, TRY, RICH, KNOW, PRISON, BLACK, DEAD, BODY, COMMIT

Topic 13:
 FATHER, MOTHER, SON, DAUGHTER, BROTHER, CHILD, DEATH, BOY, LIVE, SISTER, DIE, YOUNG, RETURN, MARRY, CARE

Topic 14:
 WORLD, STAR, MOVIE, DOCUMENTARY, PLAY, FEATURE, DIRECT, MUSIC, COMEDY, BAND, DIRECTOR, INCLUDE, FOLLOW, AMERICAN, MAKE
```

In [12]:
```
my_stop_words = text.ENGLISH_STOP_WORDS.union(['this', 'when', 'each','film', 'by', 'as'])
corpus = data['overview_lemm']
preprocessor = None
vectorizer = TfidfVectorizer(stop_words=my_stop_words)
topic_modeler = TruncatedSVD(15, random_state=10)
print_n_words= 15

extract_topics(corpus, preprocessor, vectorizer, topic_modeler, print_n_words);
```

```
Topic 0:
 LIFE, YOUNG, LOVE, YEAR, FAMILY, MAN, WOMAN, FRIEND, OLD, STORY, FATHER, MAKE, GIRL, LIVE, TIME

Topic 1:
 LOVE, FALL, YOUNG, WOMAN, FAMILY, FATHER, MOTHER, MEET, GIRL, OLD, MARRY, SON, DAUGHTER, MAN, LIVE

Topic 2:
 STORY, LOVE, LIFE, TELL, WORLD, DOCUMENTARY, FALL, STAR, TRUE, MOVIE, BASE, WOMAN, FEATURE, FOLLOW, DIRECTOR

Topic 3:
 WOMAN, YOUNG, MAN, MURDER, LOVE, KILL, FALL, KILLER, POLICE, HUSBAND, WIFE, MEN, DETECTIVE, BEAUTIFUL, TRY

Topic 4:
 SCHOOL, FRIEND, GIRL, LOVE, HIGH, STUDENT, FALL, BEST, MEET, TEACHER, MAKE, GROUP, NEW, COLLEGE, HELP

Topic 5:
 YEAR, STORY, OLD, GIRL, YOUNG, TELL, MURDER, BOY, SCHOOL, WOMAN, TRUE, MAN, BASED, STUDENT, HIGH

Topic 6:
 LIFE, WOMAN, YEAR, MAN, OLD, YOUNG, LIVE, CHANGE, FRIEND, DAY, WORLD, START, HUSBAND, JOB, BEST

Topic 7:
 YOUNG, GIRL, WOMAN, SCHOOL, FAMILY, BOY, TOWN, HIGH, SMALL, STUDENT, LIFE, MOTHER, GROUP, NEW, LIVE

Topic 8:
 MURDER, LIFE, STORY, SCHOOL, KILLER, HIGH, POLICE, STUDENT, GIRL, WIFE, TELL, CASE, KILL, DETECTIVE, CRIME

Topic 9:
 TOWN, SMALL, MAN, LOVE, LIFE, BOY, YEAR, LOCAL, PEOPLE, OLD, MEET, GANG, CITY, FALL, SHERIFF

Topic 10:
 LIFE, WORLD, LOVE, FORCE, WAR, YOUNG, FIGHT, FALL, GROUP, BROTHER, SAVE, FAMILY, BATTLE, EVIL, KILL

Topic 11:
 FAMILY, FRIEND, WOMAN, BEST, STORY, MAN, HOUSE, OLD, HOME, NIGHT, YEAR, MURDER, TELL, COME, PAST

Topic 12:
 MAN, FAMILY, SCHOOL, STAR, HIGH, STUDENT, NEW, DOCUMENTARY, DIRECT, DAUGHTER, MOVIE, FEATURE, TEACHER, MAKE, OLD

Topic 13:
 WOMAN, SCHOOL, HIGH, STUDENT, HUSBAND, WIFE, CHILD, TEACHER, SON, YEAR, WAR, WORLD, NEW, BEGIN, TOWN

Topic 14:
 MURDER, WOMAN, WORK, POLICE, KILLER, DOCUMENTARY, LOVE, FALL, FAMILY, YEAR, CRIME, DETECTIVE, CASE, GANG, NEW
```