CS 360: ML Project

Non-Negative Matrix Factorisation

Ankit Jaiswal (180010004)

Devyani Maladkar (180020008)

Dewansh Chhatri (180010011)

Nidhish Sawant (180030017)

Siddharth Shah (180010027)

Siddharth Singh Solanki (180020025)

Introduction

Non negative Matrix factorization is used to factorize a given non-negative matrix into two matrices such that all the components of the factor matrices are non-negative

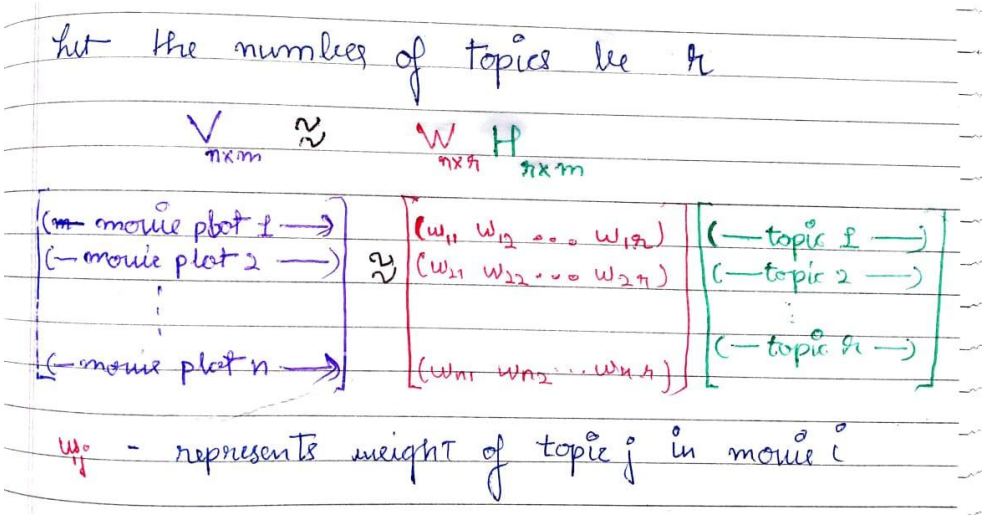
$$V \approx W*H$$

where V is an *nxm* matrix, W is an *nxr* matrix and H is an *rxm* matrix.

NMF gives us a part based representation of our data. We want that our data should not be lost as much as possible and thus a possible objective function is to minimize $||V - W*H||_F$. The problem is NP hard in general and we don't have any exact solutions, thus we go in with iterative optimization algorithms.

In the application of text mining let us say that we have a matrix of data V, where each row represents a movie plot. The factorized matrices can be interpreted as the following -

- Each row of H represents a topic.
- Each row of W represents the decomposition of the corresponding movie plot into the topics of H.



Implementation of NMF Algorithms

The NMF algorithms implemented as a part of the study are:

- Alternating Least Squares (ALS): This is a popular method which solves 2 non-negative least square problems to obtain H and W at each step.
- Multiplicative Update Rule (MU): The MU algorithm was proposed in the seminal paper by Lee and Seung.

We implement the algorithms in python and provide three applications of the implementation.

(Jupyter Notebook: NMF_Algorithm_implementations.ipynb, Data: Dataset-1 folder)

1] A model text based example

2] Image based decomposition of faces (Lee and Seung Paper)

3] Movie plot dataset (small subset = 35 plots)

Algorithm Details

1] Alternating Least Squares

The NMF problem is to find W and H such that W*H is as close as possible to A. The problem of determining W,H is such that

$$\min_{W,H >= 0} || A - W*H ||_F$$

where W, H >= 0 means all the elements of W and H are non-negative.

is a non convex problem in both the W, H variables taken together.
But taken in each variable separately, the problem is convex in nature. The algorithm for ALS uses this to solve alternately two optimization problems to find W and H.

```
Guess an initial value W(1).

for k = 1, 2,... until convergence
      i)Solve minH≥0 || A - W^(k) H ||_F , giving H^(k)
      ii)Solve minW≥0 || A - W H^(k) ||_F , giving W^(k+1).


Where,
|| ||_F Frobenius Norm
```

i) and ii) are the constrained optimization problems of non-negative least squares.

2] Multiplicative Update Rule

The multiplicative update rule is obtained from implementing a gradient descent method on an objective function based on KL divergence. The gradient descent is performed with a fixed step size so that at each update the positive nature of the entries of H and W is maintained.

```
W, H random initial value

while (not converged)
        update(W)
        update(H)

Where,

update(W) : W=W.*(V*H')./(W*(H*H')+epsilon)
update(H) : H=H.*(W'*V)./((W'*W)*H+epsilon)

.* is element wise multiplication
./ is element wise division
```

Reference:
http://www.ru.ac.bd/stat/wp-content/uploads/sites/25/2019/03/106_02_Elden-Matrix-Methods-in-Data-Mining-and-Pattern-Recognition_2007.pdf

Real World Application

The NMF algorithm is popularly used for text based applications. Be it determining the theme of documents or tags for abstracts and stack overflow questions, NMF can be seen as a tool with a varied number of applications.

For the project the application we chose consists of determining the topics from the movie's dataset. The details of the dataset are as follows:

Dataset Source : https://www.kaggle.com/jrobischon/wikipedia-movie-plots

The dataset contains descriptions of 34,886 movies from around the world.

Column descriptions are listed below:

- Release Year - Year in which the movie was released
- Title - Movie title
- Origin/Ethnicity - Origin of the movie (i.e. American, Bollywood, Tamil, etc.)
- Director - Director(s)
- Cast - Main actresses and actors
- Genre - Movie Genre(s)
- Wiki Page - URL of the Wikipedia page from which the plot description was scraped
- Plot - Long form description of the movie plot.

For the application we used movie titles and plots.

Example -

Movie: Master Of Thunder

Plot: For 1400 years, Spiritual Guardians have watched over the mountains of Japan and defeated the evil spirits there. The nearby Kikyo Temple is rumored to have been the home of these legendary Guardians known as the "Blue Seven Dragons." Only two survivors of the long battle between good and evil remain, the martial monks Santoku (Yasuaki Kurata) & Genryu (Sonny Chiba). Now the fate of the world must be decided once and for all in a final ferocious battle between the force of good and evil.

Overview of Methodology

The objective of the implementation was to use NMF to determine the themes/topics for the movie plots. 24420 of the movies were used for training. The movie plots were used without any other metadata. The dataset was preprocessed and vocabulary generated for vectorising the data from text to a matrix. The matrix was then used for decomposition and the topics were obtained.

Preprocessing

The following preprocessing steps were used:

- The text was converted to lowercase.
- Words containing numbers were removed.
- Brackets and other punctuations were removed.
- The data was lemmatised (it can be replaced by stemming as an alternate)
    - Lemmatising is a way of reducing the word length in a smart way.
      For instance :

      Sentence -
      "Our meeting today was worse than yesterday,
      I'm scared of meeting the clients tomorrow."

      Lemmatization:
      ['our', 'meeting', 'today', 'be', 'bad', 'than', 'yesterday', ',', 'i', 'be', 'scar', 'of', 'meet', 'the', 'client', 'tomorrow', '.']
- Stop words and Names are dropped from the data since they do not add to the theme of the data. The names dataset used for collecting names to drop is taken from here.

The preprocessed data was then converted to numerical data using tf-idf implementation of sklearn. (Dataset folder : Moviedata folder, Code file : Data_Preprocessing.ipynb)

The term frequency (tf) and inverse document frequency (idf) for the words in the plot text are used to generate a score for each word. For the words in the text and in the vocabulary the tf-idf weight is used. For words not in the vocab zero is considered. This then vectorizes

the data from words to numbers. The result is that each text/document is a vector of numbers and the entire dataset then is a matrix.

Library Implementation of NMF

The sklearn package was used for the following functionality :

- tf-idf
- NMF algorithm

The inhouse implementation lacks in code optimisation and uses a less complex objective function. The sklearn package allows easy utilisation on a larger dataset.
(Code file : NMF_library_implementation.ipnyb)

The objective function for the sklearn NMF is :

$$
\begin{aligned}
0.5 * ||X - WH||^2_{loss} + alpha * l1_{ratio} * ||vec(W)||_1 \\
+alpha * l1_{ratio} * ||vec(H)||_1 \\
+0.5 * alpha * (1 - l1_{ratio}) * ||W||^2_{Fro} \\
+0.5 * alpha * (1 - l1_{ratio}) * ||H||^2_{Fro}
\end{aligned}
$$

Where,
X is the data matrix
W, H is the NMF decomposition for X
alpha controls regularisation

Results, Findings, Observations

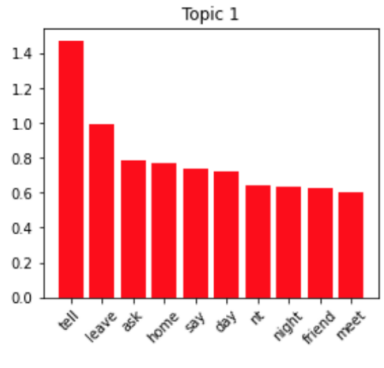
The NMF algorithm output consists of the H and W matrices for the movie plots dataset. Using the vocab and values in the H and W matrix inference and analysis is done in this section.

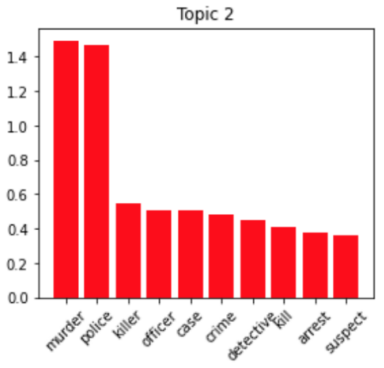
The matrix X consists of Movie plots as rows and columns as the words of the vocabulary.

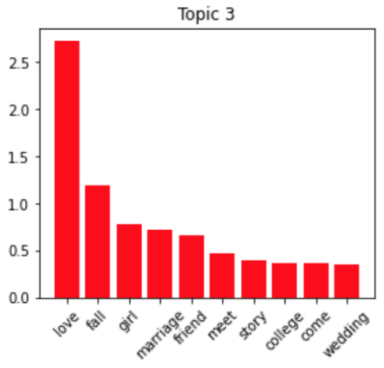
Interpreting H

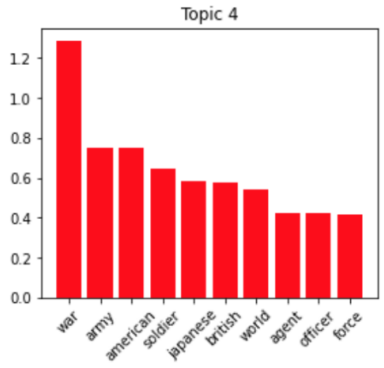
H enables us to interpret the various words for each topic. The output for the movie plot implementation was performed with k = 18. The 18 topics obtained were the following, for each topic the words corresponding to the top 10 words are as below.

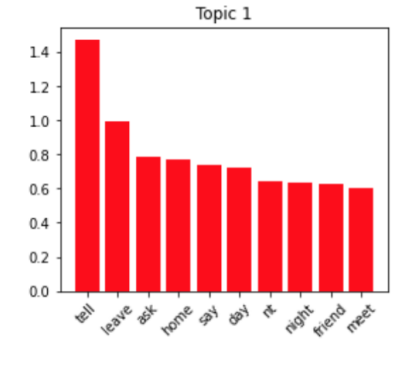
```
'tell leave ask home say day nt night friend meet',

'murder police killer officer case crime detective kill arrest suspect',

'love fall girl marriage friend meet story college come wedding',

'war army american soldier japanese british world agent officer force',

'film story life movie character set revolve star director end',

'school student teacher girl high college boy class friend parent',

'child wife husband baby son daughter affair couple pregnant life',

'gang money bank town steal robbery brother plan horse gangster',

'kill house shoot attack car escape body death dog group',

'man woman young old town husband lady life beautiful lover',

'father mother son daughter boy live die home parent year',

'team game player win coach football race play lose match',

'family brother sister son house home young live daughter parent',

'dr doctor hospital patient nurse medical psychiatrist scientist experiment surgeon',

'village villager ram people singh come city son story land',

'ship island crew alien pirate earth sea boat aboard planet',

'play band music song role singer star stage perform sing',

'bug rabbit bunny hole cartoon carrot tree fudd duck chase'
```

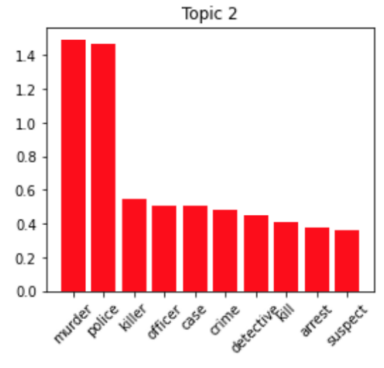

Topic 1
tell
leave
ask
home
say
day
nt
night
friend
meet

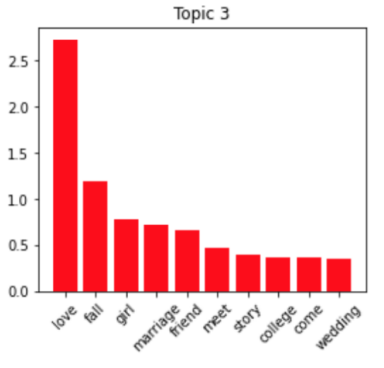

Topic 2
murder
police
killer
officer
case
crime
detective
kill
arrest
suspect

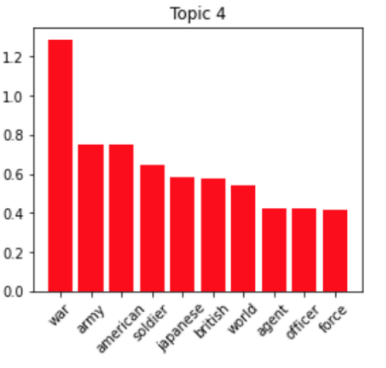

Topic 3
love
fall
girl
marriage
friend
meet
story
college
come
wedding

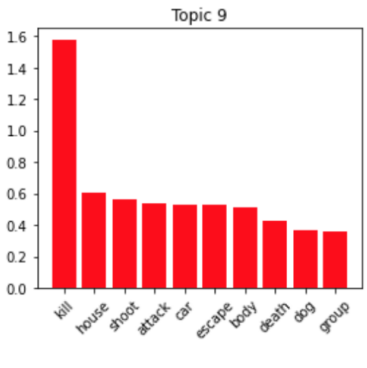

Topic 4
war
army
american
soldier
japanese
british
world
agent
officer
force

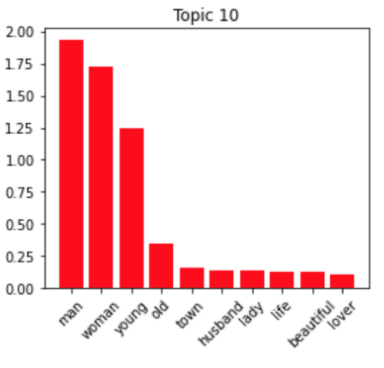

Topic 1
tell
leave
ask
home
say
day
nt
night
friend
meet

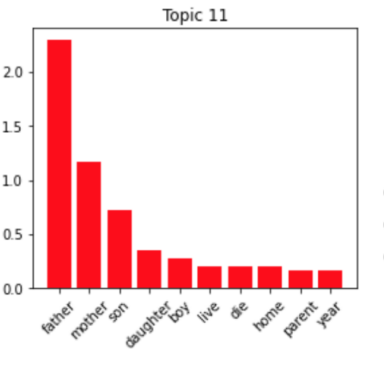

Topic 2
murder
police
killer
officer
case
crime
detective
kill
arrest
suspect

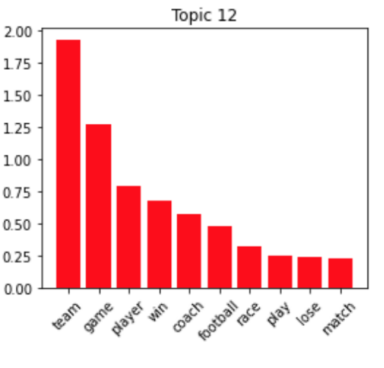

Topic 3
love
fall
girl
marriage
friend
meet
story
college
come
wedding

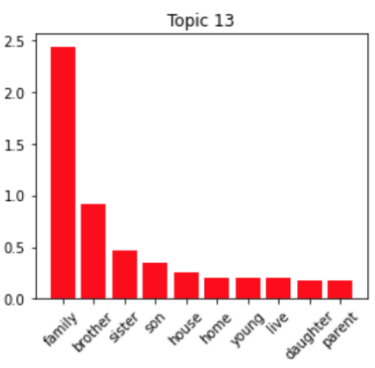

Topic 4
war
army
american
soldier
japanese
british
world
agent
officer
force

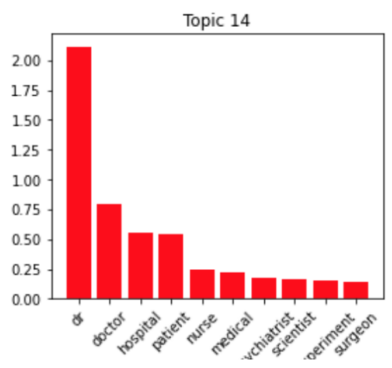

Topic 9
kill
house
shoot
attack
car
escape
body
death
dog
group

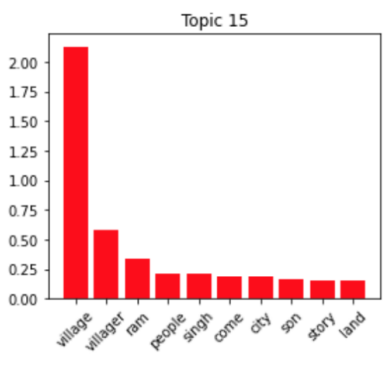

Topic 10
man
woman
young
old
town
husband
lady
life
beautiful
lover

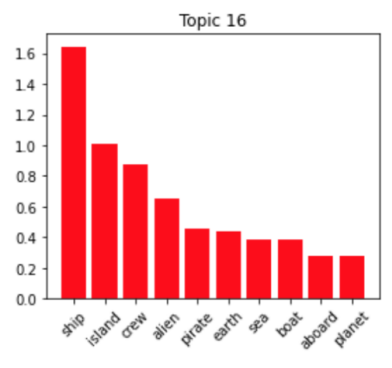

Topic 11
father
mother
son
daughter
boy
live
die
home
parent
year

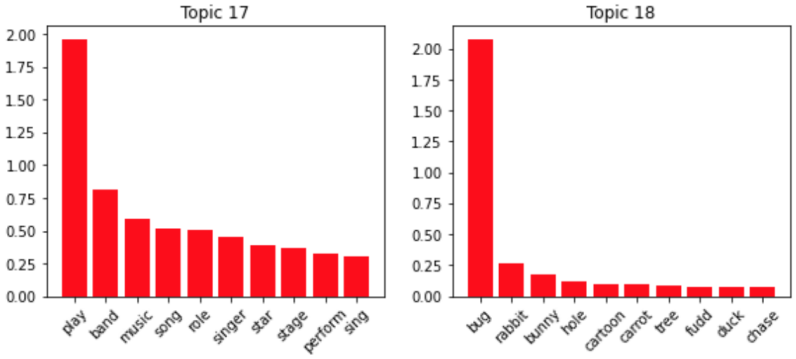

Topic 12
team
game
player
win
coach
football
race
play
lose
match


Topic 13
family
brother
sister
son
house
home
young
live
daughter
parent


Topic 14
dr
doctor
hospital
patient
nurse
medical
surgeon


Topic 15
village
villager
ram
people
singh
come
city
son
story
land


Topic 16
ship
island
crew
alien
pirate
earth
sea
boat
aboard
planet



Vertical Axis: Coefficient for the word in the topic vector

Horizontal axis: Top 10 words per topic

A close look at the topics indicates that the topics qualitatively make sense. An example is the topic -

“love fall girl marriage friend meet story college come wedding”

It is clear that the topic makes sense in itself because we can quite clearly imagine that in a plot where the word love appears, marriage, girl and college and friend would be commonplace because movies of the romance genre generally have these elements of college romance or wedding scenes.

Another example -

“War army american soldier japanese british world agent officer force”

The topic is for films based on war. Again the topic makes sense as we can imagine the plot of a war movie containing the words soldier, agent and officer. Additionally there are a lot of hollywood movies in the dataset which implies that plots would contain an american army fighting other armies such as those of Japan in World War 2 based movies.

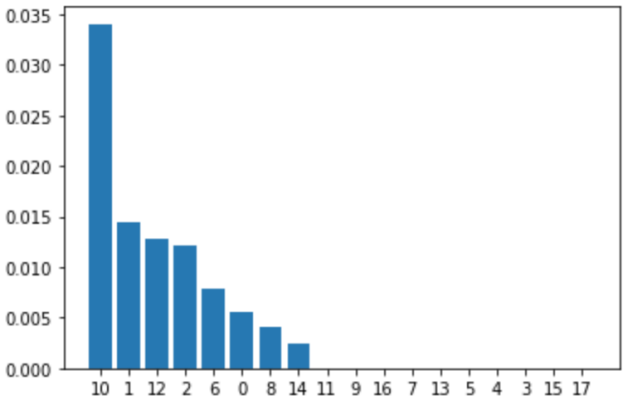
Interpreting W

W enables us to interpret the topics for a particular document. For the data from the training set we obtain a suitable decomposition for the movies as shown below.

Johnny Angel a merchant ship captain george raft find his father ship – of the same line – derelict at sea the entire crew having disappear from an unlisted passenger signe hasso the only survivor of the hijacking of the ship he learn of a plot involve secret gold and search new orlean for his father murderer	Topic: ship island crew alien pirate earth sea boat aboard planet
Shaft (2000) In 1998, NYPD Detective John Shaft II, a relative of legendary private investigator John Shaft I, is called in to investigate the grievous assault of Trey Howard outside a restaurant.	Topic: murder police killer officer case crime detective kill arrest suspect
Rock On Rock On!! begins in Mumbai, with the rock band Magik, in 1998. The band are best friends and enjoy a carefree, freewheeling life together. Aditya Shroff (Farhan Akhtar) is the lead singer who rebelled against his well-to-do family to play music. Joseph (Joe) Mascarhenas (Arjun Rampal) is the lead guitarist who feels the necessity to prove his worth as a musician.	Topic: play band music song role singer star stage perform sing

The movies from the training dataset when analysed for more than one topic in the decreasing order of coefficient value, showed a part based breakup of the plots into various topics.
For instance one such movie example is as shown below.

Matrix entry vs Topic
Movie : **En Uyir Nee Thaane** (Tamil)



Topic 10:
father mother son daughter boy live die home parent year

Topic 1:
murder police killer officer case crime detective kill arrest suspect

Topic 12:
family brother sister son house home young live daughter parent

Plot :
Vasu (Prabhu) is a rich industrialist who raises his father's illegitimate son after the death of his mother. His father opposes this. Vasu leaves home to live alone with the child. Janaki (Devayani) witnesses a murder and testifies against the murderer. He is convicted. The murderer's brother rapes Janaki in revenge, and Janaki is kicked out of her house by her brother. She goes to her friend's house, but even there, she is kicked out. Vasu comes to her rescue. Janaki becomes a mother for Vasu's illegitimate brother and manager in Vasu's mill. Vasu is drawn towards her and wants to marry her.

Testing on a new sample

Test samples consisting of new movies can also be used to obtain topics, provided they are of a similar nature as the trained samples. The new plot is vectorised using the vocabulary obtained during training and then the *nmf.transform* function is used for the object fitted on the training data.

The document topic distribution can be found by solving the problem of $\min_{W>=0} ||X-WH||$ for the fixed H (topic word distribution vectors) which was obtained from the decomposition of the training data. Here, X is the test dataset (matrix) and W is the document topic distribution matrix corresponding to the test data.

One such example of a test sample is shown below:

Johnny Angel a merchant ship captain george raft find his father ship – of the same line – derelict at sea the entire crew having disappear from an unlisted passenger signe hasso the only survivor of the hijacking of the ship he learn of a plot involve secret gold and search new orlean for his father murderer	Topic: ship island crew alien pirate earth sea boat aboard planet

Observations

The dataset used was directly obtained from wikipedia. The spelling errors for the words were not fixed, as a result of which, the dataset was of slightly lower quality. The mix of movies used was very vast, extending from American to Tamil to Turkish. Due to this vast variety, the names of actors and figures in the movies may have been considered in the vocabulary which should have been dropped as they don't contribute to the storyline theme directly. As a result of this, some stories showed high values for topics that were not necessarily close to the theme.

Coherency Scores

The evaluation of a bag-of-words model output is tricky because it is in an unsupervised learning paradigm. We don't know exactly how the movie themes should look once we get the decomposition. Quantitative measurement of the output is achieved through the coherence score. This score indicates whether the output of the algorithm makes sense semantically. Intuitively the mechanism of calculation can be explained as follows -

- We have a heuristic of semantic correlation between any two words already available to us in the dataset. Whenever two words appear a lot of times together, it implies that they might be semantically correlated, such as (game, player), (love, marriage).
- We use the dataset to plot the words in a word vector space. The word vectors have a similarity score. If the two words appear together frequently in the dataset then they will have a high similarity score and vice-versa.

The sci-kit learn package does not have any implementation of the standard correlation scores available. We wrote our code for calculating the coherence scores. Since calculating the standard coherence scores is difficult, we built our own topic coherence score which tells us the quality of the decomposed topics by scoring all the top word pairs in a topic.

(Code file: Topic_coherence_calculator.ipynb, Input Data: Coherence input folder)

The coherence scores can help us in choosing the optimal number of topics for our dataset decomposition. But due to limited computational resources, we could not decompose our dataset by considering different k values (k = number of topics to decompose) and then decide the best.

Prospects

- The ability to represent a document by a few keywords can be further used in dimensionality reduction for the application of other algorithms on top of the work. Such as using the keywords as features for a movie plot and then feeding this to a neural network.

- The work can also be extended by applying k-means clustering on the coordinates of all the plots in the topic subspace. The output clusters can be labeled as genres with the most frequently occurring keyword across the cluster as the genre title. Such clusters can help us in building a recommendation system where the user can be shown suggestions based on their past movie choices. If a user shows the trend of choosing movies in the genre where romance topics have the highest weightage, then other movies in the same cluster can be recommended to the user by the system. Also, if a user inputs a genre keyword then movies from the cluster which have the keyword most frequently occurring in the top topics can be recommended to the user.