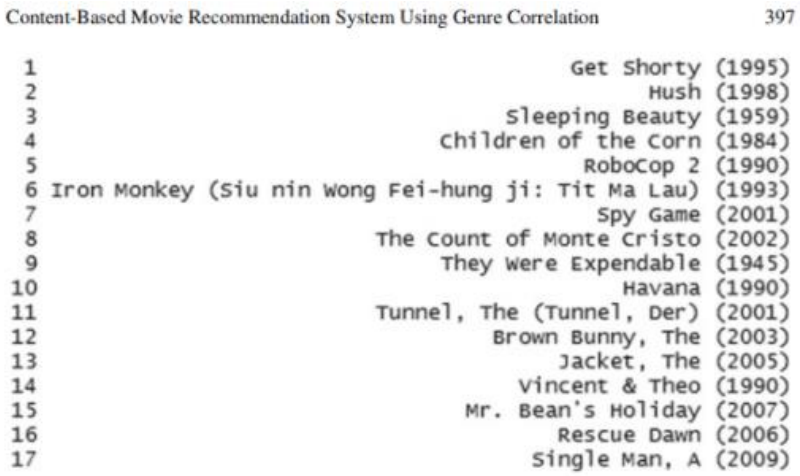
Prototype NLP Tasks

Literature review

As movies are extremely popular there a large amount of recommendation systems for them. From a literature review recommendation system provide suggestions to users for certain resources, Movie recommendation systems usually predict what a movie a user would like. In this study by Subramanyam Kunisetti and his colleagues evaluated a content-based recommended system using Genre Correlation an approach similar to mine. The paper explored the Movie Lens dataset which is provided online free by Movie Lens containing 9000 movies classified by genres and ratings from users. They also discuss the approaches of other NLP systems but decided on a Content-Based system which constructs a genre dataset with the movie ID as rows and genres as columns separated by a pipeline. Further on, they calculate two matrixes assigning a 1 for movies with a rating more than 3 and a -1 for a rating less than 3 thus calculating the dot product of the two matrices giving the resultant matrix. Towards the end a Euclidian distance is calculated between the current user and other users retaining rows with a minimum distance as these are the recommendations. This provided good recommendations for the user with User 1's recommendations and User 2 Recommendations were very similar.



Content-Based Movie Recommendation System Using Genre Correlation 397

1	Get Shorty (1995)
2	Hush (1998)
3	Sleeping Beauty (1959)
4	Children of the Corn (1984)
5	RoboCop 2 (1990)
6	Iron Monkey (Siu nin Wong Fei-hung ji: Tit Ma Lau) (1993)
7	Spy Game (2001)
8	The Count of Monte Cristo (2002)
9	They Were Expendable (1945)
10	Havana (1990)
11	Tunnel, The (Tunnel, Der) (2001)
12	Brown Bunny, The (2003)
13	Jacket, The (2005)
14	Vincent & Theo (1990)
15	Mr. Bean's Holiday (2007)
16	Rescue Dawn (2006)
17	Single Man, A (2009)

Fig. 5 User 1 recommendations

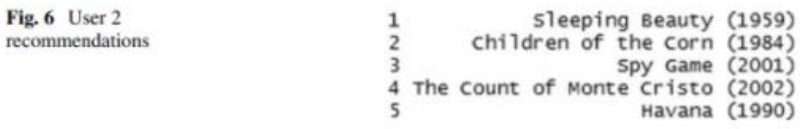


Fig. 6 User 2 recommendations

1	Sleeping Beauty (1959)
2	Children of the Corn (1984)
3	Spy Game (2001)
4	The Count of Monte Cristo (2002)
5	Havana (1990)

Rational selection of the NLP Task

The rational for this NLP task was to recommend similar movies based on a user's entry, to help ease the annoyance of sorting through movies to only choose a movie that would waste an hour of your time which is becoming more important as time moves on. By giving users an easy way to sort through the large dataset of movies would certainly provide an enjoyable experience for the user thus keeping the user coming back to the product you are presenting them. These systems are now the most important parts of movie and tv streaming services like Netflix, Stan, and Amazon thus recommenders will always be around as long as users are using services like these.

Data pre-processing of inputs and outputs

```
df['genre'] = df['genre'].str.replace(',',' ')
df['year'] = df['year'].str.extract('(\d+)').astype(int)
```

```
: df2['keywords'] = df2['about'].str.cat(df2['genre'],sep=" ")
```

```
: df2['bag_of_words'] = df2['keywords'].str.cat(df2['director'],sep=" ")
```

```
: df2.drop(["genre"], axis=1, inplace=True)
  df2.drop(["director"], axis=1, inplace=True)
  df2.drop(["about"], axis=1, inplace=True)
  df2.drop(['keywords'], axis=1, inplace=True)
```

After web scraping, I did a small bit of pre-processing which was removing commas from genres column and brackets from the Year column. For the bag of words model however, I did a bit more removing stop words, joining columns and dropping unused columns.

Specification and justification of hyperparameter

As this was a content-based recommendations system there was no hyperparameter, but this could definitely be explored in the future.

Preliminary assessment of NLP Task performance

To gauge this recommendation system is to decide on whether the recommended movies from the systems are accurate.

NLP Recommendation Engine 1

```
Courses similar to The Shawshank Redemption are:

index                                                     361
movie                                                  Dev.D
genre                                  Drama  Romance
director                                        Anurag Kashyap
about       After breaking up with his childhood sweethear...
Name: 361, dtype: object
index                                                      72
movie                                                 Oldboy
genre                        Action  Drama  Mystery
director                                      Chan-wook Park
about       After being kidnapped and imprisoned for fifte...
Name: 72, dtype: object
index                                                     182
movie                                        The Great Escape
genre                     Adventure  Drama  History
director                                         John Sturges
about       Allied prisoners of war plan for several hundr...
Name: 182, dtype: object
index                                                     258
movie                                  In the Mood for Love
genre                                  Drama  Romance
director                                       Kar-Wai Wong
about       Two neighbors form a strong bond after both su...
Name: 258, dtype: object
index                                                     136
movie                                       Koe no katachi
genre                      Animation  Drama  Family
director                                        Naoko Yamada
about       A young man is ostracized by his classmates af...
Name: 136, dtype: object
```

As seen the recommendation engine 1 using Tfidf Vectorisation and Nearest neighbour has provided an okay set of movies. The first movie Dev.D having similar genres of Drama and Oldboy the second movie recommended having a similar plot description of being imprisoned and the genre of Drama.

NLP Recommendation Engine 2

```
recommendation('The Shawshank Redemption')
```

```
25                    The Green Mile
361                           Dev.D
324    Mr. Smith Goes to Washington
72                           Oldboy
258           In the Mood for Love
Name: movie, dtype: object
```

NLP Recommendation engine uses a bag of words model also with Tfidf Vectorisation however using cosine similarity as the recommender. This has also given good recommendations providing a list of movies that are quite similar with The Green Mile also being set in prison and having the same genres. This also has provided similar movies to the previous recommender.

Conclusion

In conclusion the movie recommender has completed the task provided web scraping data and assigning two NLP Tasks being two NLP recommendation engines.

Bibliography

Reddy, S., Nalluri, S., Kunisetti, S., Ashok, S., & Venkatesh, B. (2018). Content-Based Movie Recommendation System Using Genre Correlation. Smart Intelligent Computing and Applications, 391-397. doi: 10.1007/978-981-13-1927-3_42