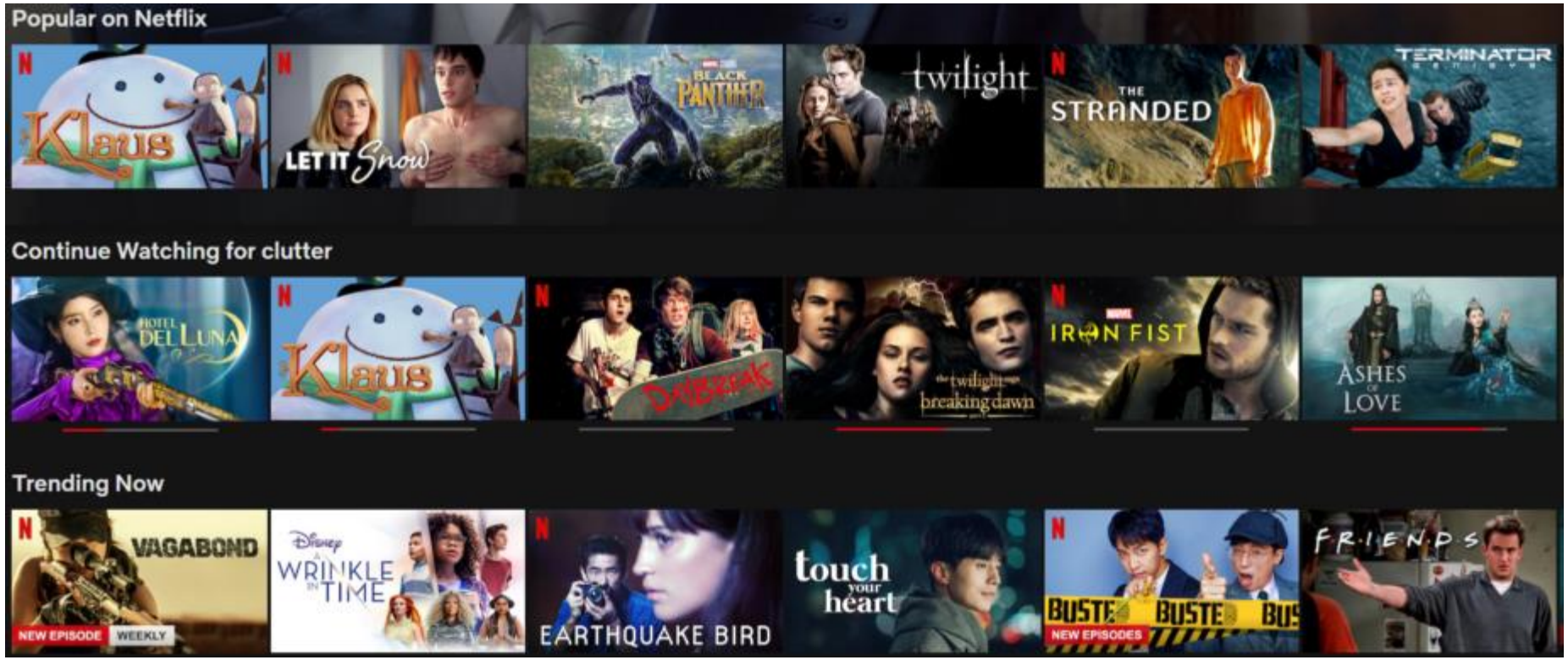
Content-based recommender using Natural Language Processing (NLP)

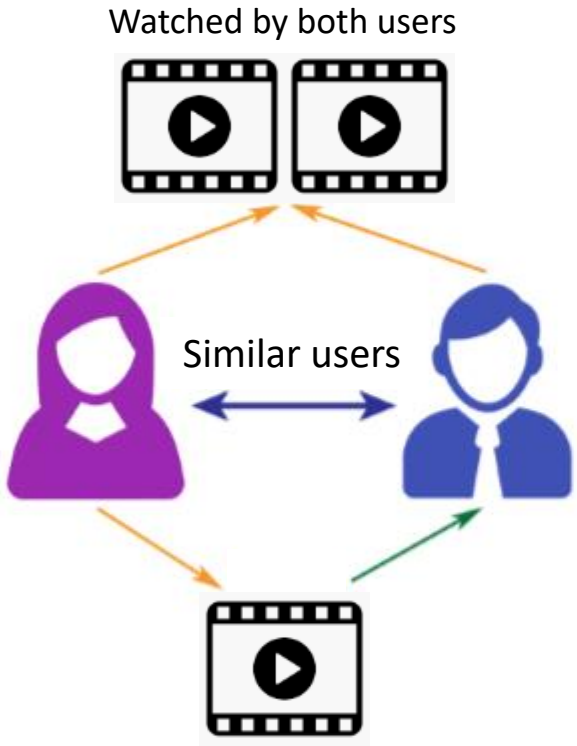
A guide to build a content-based movie recommender model based on NLP



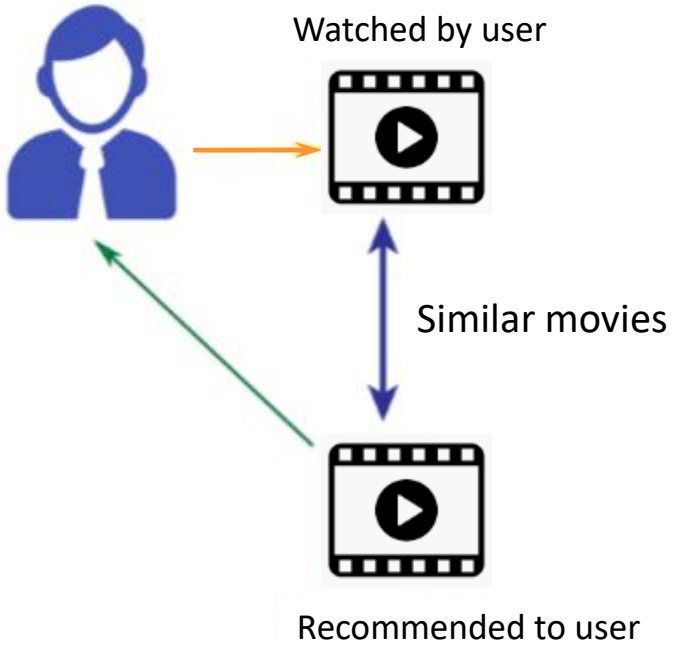
IMDB Dataset 2017, 250 movies

2 types of recommender systems

Collaborative Filtering



Content-Based Filtering



Data pre-processing

Director	Actors	Genre	Plot
Frank Darabont	Tim Robbins, Morgan Freeman, Bob Gunton, William Sadler	Crime, Drama	Two imprisoned men bond over a number of years, finding ...
Francis Ford Coppola	Marlon Brando, Al Pacino, James Caan, Richard S. Castellano	Crime, Drama	The aging patriarch of an organized crime dynasty transf...
Francis Ford Coppola	Al Pacino, Robert Duvall, Diane Keaton, Robert De Niro	Crime, Drama	The early life and career of Vito Corleone in 1920s New ...
Christopher Nolan	Christian Bale, Heath Ledger, Aaron Eckhart, Michael Caine	Action, Crime, Drama	When the menace known as the Joker emerges from his myst...
Sidney Lumet	Martin Balsam, John Fiedler, Lee J. Cobb, E.G. Marshall	Crime, Drama	A jury holdout attempts to prevent a miscarriage of just...

Director	Actors	Genre	Key_words
[frankdarabont]	[timrobbins, morganfreeman, bobgunton]	[crime, drama]	[finding, solace, years, acts, number, eventual, redempt...
[francisfordcoppola]	[marlonbrando, alpacino, jamescaan]	[crime, drama]	[aging, patriarch, organized, crime, dynasty, transfers,...
[francisfordcoppola]	[alpacino, robertduvall, dianekeaton]	[crime, drama]	[portrayed, 1920s, new, york, family, crime, syndicate, ...
[christophernolan]	[christianbale, heathledger, aaroneckhart]	[action, crime, drama]	[mysterious, past, gotham, ability, dark, knight, must, ...
[sidneylumet]	[martinbalsam, johnfiedler, leej.cobb]	[crime, drama]	[miscarriage, jury, holdout, attempts, colleagues, justi...

Title	Bag_of_words
The Shawshank Redemption	crime drama frankdarabont timrobbins morganfreeman bobgu...
The Godfather	crime drama francisfordcoppola marlonbrando alpacino jam...
The Godfather: Part II	crime drama francisfordcoppola alpacino robertduvall dia...
The Dark Knight	action crime drama christophernolan christianbale heathl...
12 Angry Men	crime drama sidneylumet martinbalsam johnfiedler leej.co...

Similarity Matrix

	Movie0	Movie1	Movie2	…	Movie247	Movie248	Movie249
Movie0	1.	0.15789474	0.13764944	...	0.05263158	0.05263158	0.05564149
Movie1	0.15789474	1.	0.36706517	...	0.05263158	0.05263158	0.05564149
Movie2	0.13764944	0.36706517	1.	...	0.04588315	0.04588315	0.04850713
…	...						
Movie247	0.05263158	0.05263158	0.04588315	...	1.	0.05263158	0.05564149
Movie248	0.05263158	0.05263158	0.04588315	...	0.05263158	1.	0.05564149
Movie249	0.05564149	0.05564149	0.04850713	...	0.05564149	0.05564149	1.

$$similarity = cos(\theta) = \frac{\boldsymbol{u} \cdot \boldsymbol{v}}{\|\boldsymbol{u}\| \|\boldsymbol{v}\|} = \frac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\sum_{i=1}^{n} u_i^2} \sqrt{\sum_{i=1}^{n} v_i^2}}$$

$$\boldsymbol{u} \cdot \boldsymbol{v} = [u_1\ u_2\ \ldots\ u_n] \cdot \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = u_1 v_1 + u_2 v_2 + \ldots + u_n v_n = \sum_{i=1}^{n} u_i v_i$$

Run and test the recommender model

```python
# this function takes in a movie title as input and returns the top 10 recommended (similar) movies

def recommend(title, cosine_sim = cosine_sim):
    recommended_movies = []
    idx = indices[indices == title].index[0]   # to get the index of the movie title matching the input movie
    score_series = pd.Series(cosine_sim[idx]).sort_values(ascending = False)   # similarity scores in descending order
    top_10_indices = list(score_series.iloc[1:11].index)   # to get the indices of top 10 most similar movies
    # [1:11] to exclude 0 (index 0 is the input movie itself)

    for i in top_10_indices:   # to append the titles of top 10 similar movies to the recommended_movies list
        recommended_movies.append(list(df['Title'])[i])

    return recommended_movies

recommend('The Dark Knight')
```

```
['The Dark Knight Rises',
 'Batman Begins',
 'The Green Mile',
 'Witness for the Prosecution',
 'Out of the Past',
 'Rush',
 'The Prestige',
 'The Godfather',
 'Reservoir Dogs',
 'V for Vendetta']
```

Run and test the recommender model

```
recommend('Fargo')
```

```
['No Country for Old Men',
 'The Departed',
 'Rope',
 'The Godfather',
 'Reservoir Dogs',
 'The Godfather: Part II',
 'On the Waterfront',
 'Goodfellas',
 'Touch of Evil',
 'The Big Lebowski']
```

```
recommend('The Avengers')
```

```
['Guardians of the Galaxy Vol. 2',
 'Aliens',
 'Guardians of the Galaxy',
 'The Martian',
 'Terminator 2: Judgment Day',
 'The Terminator',
 'The Thing',
 'Interstellar',
 'Spider-Man: Homecoming',
 'The Matrix']
```