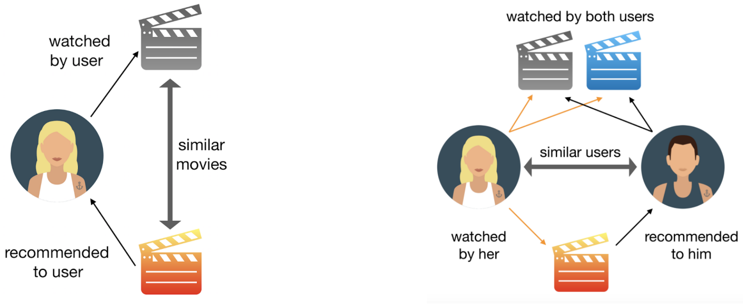
CmpE493 - Assignment 2
A Movie Recommendation System

Gökçe Uludoğan (gokce.uludogan@boun.edu.tr)

Deadline: April 8, 2019, Monday, 23:59

1 Description

Building a recommendation system is a common task in many modern applications. The goal of a recommendation system is to identify relevant data for their users. These systems are generally based on two methods: collaborative filtering and content based filtering. While collaborative filtering exploits similar users' rates, content based filtering considers the contents of the items liked by the same user.



(a) Content based filtering

(b) Collaborative filtering

Figure 1: Recommendation systems [1]

In this assignment, you will build a simple movie recommendation system from scratch. Firstly, you will extract contents and recommendations of movies from IMDB. Then, to represent the contents of movies, you will implement the vector-space model based on TF-IDF weighting. After vectorizing all movies' contents using TF-IDF, you will make recommendations for a given movie by getting the most similar K movies based on cosine similarity. Finally, you will calculate the evaluation metrics considering the IMDB recommendations as the relevant (ground truth) movies. The IMDB ids for the movies and a notebook template that you must follow are available on Moodle.
The libraries you are allowed to use for this assignment are *requests*[2] and the standard Python libraries.

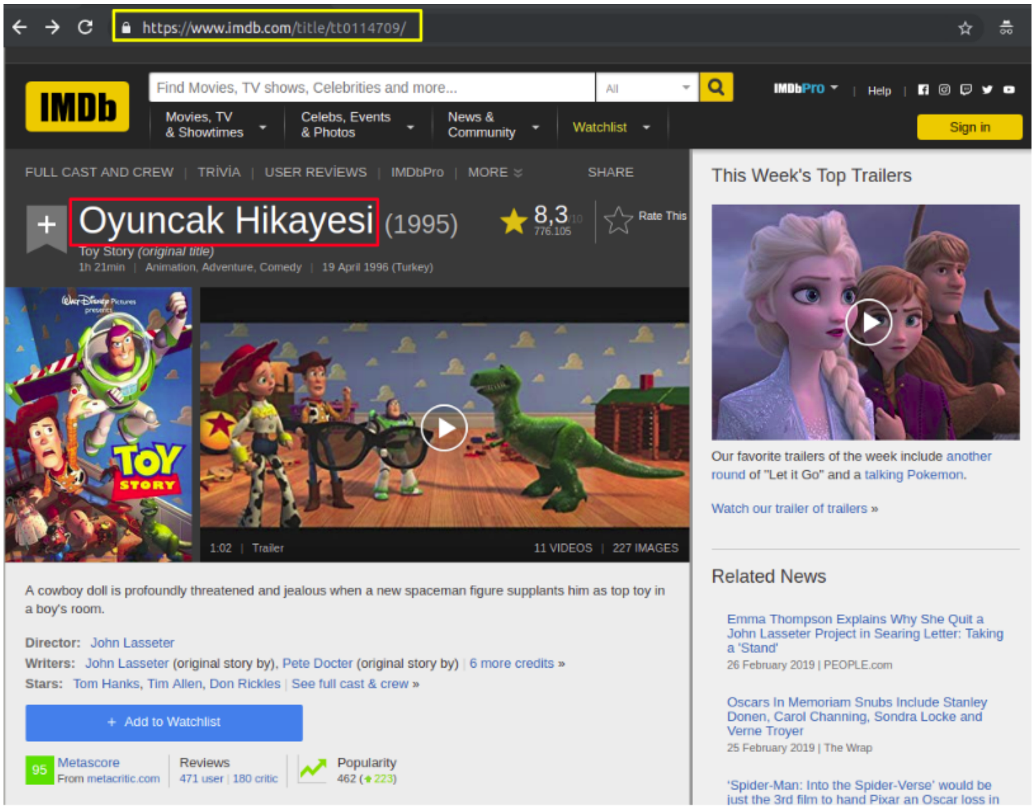
[1] Source: `https://bit.ly/2AncOE4`

[2] Requests is a library to make HTTP requests.

2 Implementation

2.1 Extracting contents from IMDB

For a given IMDB ID, the web page for the corresponding movie can be found at `https://www.imdb.com/title/<imdb-id>` as shown in the yellow box in Figure 2.



A cowboy doll is profoundly threatened and jealous when a new spaceman figure supplants him as top toy in a boy's room.

Director: John Lasseter
Writers: John Lasseter (original story by), Pete Docter (original story by) | 6 more credits »
Stars: Tom Hanks, Tim Allen, Don Rickles | See full cast & crew »

Our favorite trailers of the week include another round of "Let it Go" and a talking Pokemon.

Watch our trailer of trailers »

Emma Thompson Explains Why She Quit a John Lasseter Project in Searing Letter: Taking a 'Stand'
26 February 2019 | PEOPLE.com

Oscars In Memoriam Snubs Include Stanley Donen, Carol Channing, Sondra Locke and Verne Troyer
25 February 2019 | The Wrap

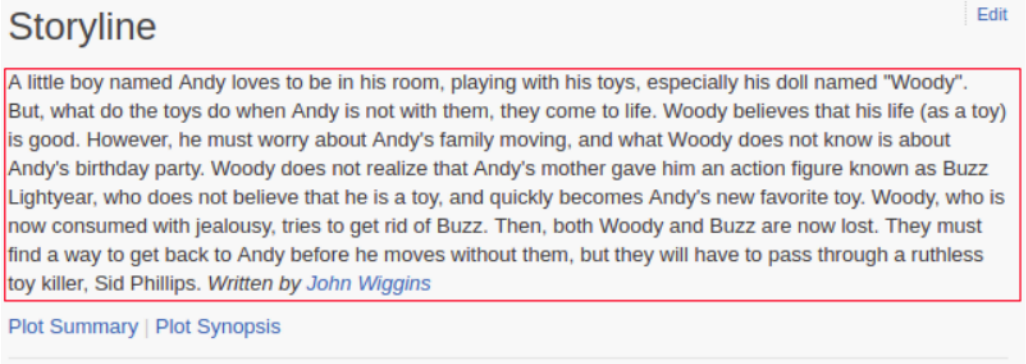
'Spider-Man: Into the Spider-Verse' would be just the 3rd film to hand Pixar an Oscar loss in

Figure 2: IMDB web page for Toy Story

The data which must be collected for each movie are:

1. Title of a movie (seen in a red box in Figure 2)
2. Storyline of a movie (seen in a red box in Figure 3)
3. IMDB ids of all recommended movies[3] (seen in a red box in Figure 4)

[3] Not only the ones seen in the current panel, but also those in the next panels.



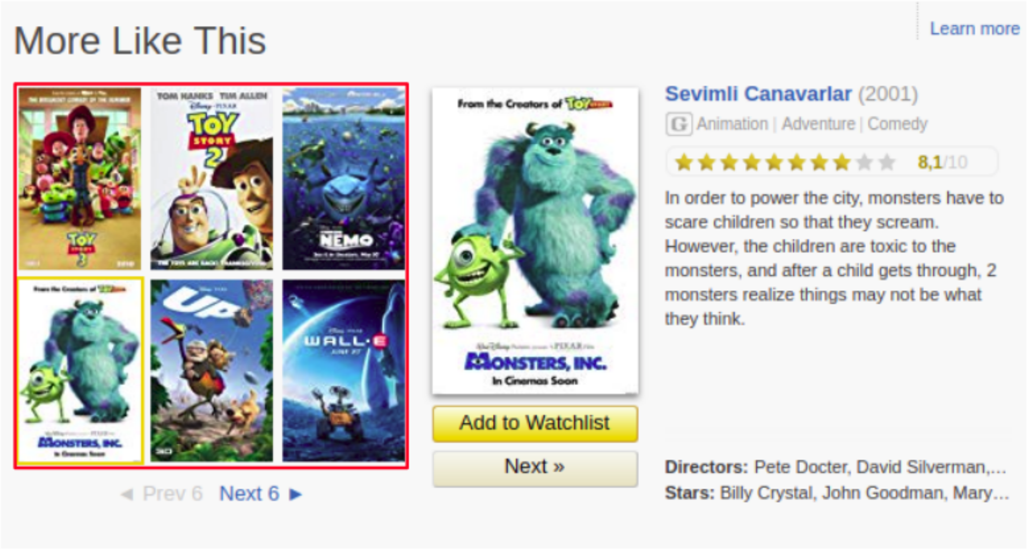
Storyline

Edit

A little boy named Andy loves to be in his room, playing with his toys, especially his doll named "Woody". But, what do the toys do when Andy is not with them, they come to life. Woody believes that his life (as a toy) is good. However, he must worry about Andy's family moving, and what Woody does not know is about Andy's birthday party. Woody does not realize that Andy's mother gave him an action figure known as Buzz Lightyear, who does not believe that he is a toy, and quickly becomes Andy's new favorite toy. Woody, who is now consumed with jealousy, tries to get rid of Buzz. Then, both Woody and Buzz are now lost. They must find a way to get back to Andy before he moves without them, but they will have to pass through a ruthless toy killer, Sid Phillips. *Written by* John Wiggins

Plot Summary | Plot Synopsis

Figure 3: Storyline for Toy Story



More Like This

Learn more

Sevimli Canavarlar (2001)

G Animation | Adventure | Comedy

★★★★★★★★☆☆ 8,1/10

In order to power the city, monsters have to scare children so that they scream. However, the children are toxic to the monsters, and after a child gets through, 2 monsters realize things may not be what they think.

Add to Watchlist

Next »

Directors: Pete Docter, David Silverman,...
Stars: Billy Crystal, John Goodman, Mary...

◄ Prev 6 Next 6 ►

Figure 4: Recommendations for Toy Story

2.2 TF-IDF Model

To represent the storylines of movies in a vector space, you will implement the TF-IDF based vector space model. Firstly, you will process the storylines of movies, identify the vocabulary, as well as the term and inverse document frequencies. Then, you will get the words with highest TF-IDF scores[4] and encode each movie's storyline by using the occurrences and scores of these words.

[4] You can either set a threshold for TF-IDF score or simply take the highest scoring N words. Explain clearly how you decide it in your notebook.

2.3 Recommendation

To recommend movies, a simple content based strategy will be followed. Given a movie, the system will recommend the most similar K movies based on cosine similarity.

2.4 Evaluation Metrics

To evaluate the system, precision, recall and F1 scores must be calculated for $K = 1, 2, 3, 10$ when recommendations are requested for a movie. The definitions of these metrics in this setting are as follows:

- **Precision** is the fraction of IMDB recommendations among your system's recommendations.
- **Recall** is the fraction of your system's recommendations among the IMDB recommendations.
- **F1 score** is the harmonic average of the precision and recall.

Figure 5 might help to understand what these metrics represents. The search results in this figure corresponds to the recommendations of your system, while the relevant documents corresponds to the IMDB recommendations. [5]

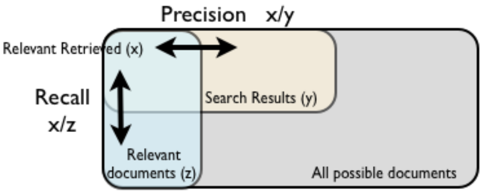


Figure 5: Precision and Recall[6]

[5] Nice post explaining metrics for recommendation: http://sdsawtelle.github.io/blog/output/mean-average-precision-MAP-for-recommender-systems.html

[6] Source: http://aimotion.blogspot.com/2011/05/evaluating-recommender-systems.html

3 Submission & Grading

You are expected to submit a single .ipynb file that is runnable. Name your notebook with your student id (e.g. 2018700100.ipynb) and do not write your name inside the notebook or anywhere else for the sake of blind review.

Note that this notebook will be your report as well, so explain your work in the related sections of the notebook.

1. IMDB Scraping: 30 pts

2. Recommendation: 35 pts

3. Evaluation Metrics: 20 pts

4. Notebook Organization & Explanations: 15 pts

5. Bonus[7]: 10 pts

Late Submission

You are allowed a total of 5 late days on homeworks with no late penalties applied. You can use these 5 days as you wish (e.g., 2 days for homework 1, 2 days for homework 2, and 1 day for homework 3). After using these 5 extra days, 10 points will be deducted for each late day.

[7] You are encouraged to improve the recommendation system by getting more valuable data while scraping and integrating them to the system or enhancing the models in the pipeline.