Challenge
Data Science

<u>But :</u> à partir des données fournies, déterminer quels films recommander à quelqu'un qui a vu Inferno.

On a à notre disposition les données suivantes :
movies.csv (qui contient movieId, title, genre(s))
ratings.csv (qui contient userId, movieId, rating, timestamp)
tags.csv (qui contient userId, movieId, tag, timestamp)
links.csv (qui contient movieId, imdbId,tmdbId)

Quelques chiffres :
Dans le fichier *movies.csv*, 45844 films sont listés dans 62 genres différents.
Dans le fichier *tags.csv*, 18052 utilisateurs ont taggé 25308 films avec 53481 tags différents.

Pour recommander un film, je me suis demander qu'est ce qui pouvais faire qu'une personne allait voir un film sachant qu'elle en a vu un autre et j'ai combiné deux approches :

- La première consiste à sélectionner des films qui sont très ressemblants au film visionné. Avec les données à notre disposition cela signifie comparer les genres et les tags. Cela pourrait se traduire par : j'ai vu un film de genre X et avec des tags Y, quels sont les autre films de genre X et tags Y qui sont sortis ?
- La seconde consiste à ne regarder que les évaluations des utilisateurs, sélectionner ceux qui ont adoré le film et proposer d'autres films que ces utilisateurs ont adoré. Cette approche pourrait se traduire par : je suis une personne qui a adoré le film, si d'autres personnes ont également adoré le film, cela veut surement dire que nous avons des gouts communs et s'ils ont adoré d'autres films, je les adorerai aussi.

La combinaison de ces deux approches s'est faite de la façon suivante :

1. classement des films par ressemblance au film visionné et filtrage des films moins bien notés en moyenne que Inferno.
2. filtrage du classement pour ne garder que des films que des utilisateurs ayant adoré Inferno ont aussi adoré.

1. <u>Classement par ressemblance :</u>

Dans le fichier *movies.csv*, on peut voir que le film Inferno est catégorisé en deux genres **Mystery** et **Thriller**. On crée une métrique qui associe à chaque film un score entre 0 et 1 qui montre l'adéquation du film en terme de genre avec le film Inferno, elle prend les valeurs suivantes :
1 si le genre comprend Mystery ET Thriller.
0.5 si le genre comprend Mystery OU Thriller (à priori Mystery n'est pas plus important que Thriller).
0 si le genre comprend NI Mystery NI Thriller.

Exemple : un film a comme genre Mystery et Action, la métrique aura pour valeur 0.5.

Il y a 7842 films dont le genre correspond même partiellement à ceux d'Inferno, les résultats sont de plus très similaire car la métrique genre ne peux prendre que 3 valeurs, il faut utiliser d'autres informations pour le classement !

On s'intéresse alors aux tags :
Dans le fichier *tags.csv*, on a sur chaque ligne un tag laissé par un utilisateur sur un film, Inferno est ainsi décrit par l'ensemble des tags suivants :
« prospect maybe », « stupid story », « bioterrorism », « predictable », « action », « tom hanks »

En tenant compte du nombre de fois que le tag a été utilisé et le nombre total de fois que le film a été taggé, on obtient les fréquences des tags suivantes :
prospect maybe : 0.08, stupid story : 0.08, bioterrorism : 0.08, predictable : 0.42, action : 0.16, tom hanks : 0.16

A partir de là on regarde la similitude de répartition des tags utilisés pour les autres films par rapport à la répartition des tags utilisés pour Inferno. D'abord on sélectionne les films qui étaient marqués avec au moins un des tags utilisé pour décrire Inferno, ensuite on calcule la fréquence d'apparition du tag parmi les tags utilisé pour décrire Inferno et enfin on multiplie chaque fréquence par la fréquence de référence d'apparition du tag pour décrire Inferno. La métrique associée au tag est ensuite la somme de chaque terme.
Exemple : parmi tous les tags utilisés pour décrire un film, action et bioterrorism apparaissent 2 et 5 fois.
Leurs fréquences parmi les tags utilisés pour décrire Inferno sont de 2/7=0.28 et 5/7=0.72 respectivement.
En les multipliant par les fréquences des tags dans le film Inferno, on obtient 0.28*0.16 et 0.72*0.08 respectivement.
Au final la métrique vaut : 0.28*0.16 + 0.72*0.08 = 0.1024.

Comme ici la métrique n'est pas nécessairement comprise entre 0 et 1 strictement, on normalise toutes les métriques comme ça on a une métrique tag qui a le même poids que la métrique genre.

La métrique de ressemblance est alors la somme de la métrique tag et de la métrique genre. Puisque chaque métrique est comprise entre 0 et 1, en faisant la somme, elles apportent le même poids dans la métrique finale. Cette métrique est ensuite utilisée pour classer les films du plus ressemblant au moins ressemblant.

On filtre ensuite les films du classement en enlevant ceux qui ont eu une note moyenne inférieure à la note moyenne (3.02) obtenue par Inferno (cela ne sert à rien de proposer des nanars).

2. <u>Filtrage du classement :</u>

A partir du fichier *rating.csv*, on a accès à la note qu'un utilisateur a donné à un film.
On peut ainsi filtrer les utilisateurs pour ne garder que ceux qui ont donné une note de 5/5 au film Inferno. On a alors une liste de personnes qui ont adoré ce film. Puis dans un second temps, on reprend le fichier de base en filtrant cette fois-ci les notes dont l'utilisateur n'est pas dans la liste et ensuite on ne garde que les notes de 5/5. Cela nous donne une liste de films associés.

On peut alors filtrer le classement établi précédemment pour ne garder que les films qui se trouvent dans cette liste de films.

Résultats:

Les 10 premiers films sont :

Format: (movieId,title,score, average rating)
(165347,Jack Reacher: Never Go Back (2016),1.0909090909090908,3.0526960784313726)
(68554,Angels & Demons (2009),1.0575,3.271150814503416)
(1620,Kiss the Girls (1997),1.0465116279069768,3.3777092675635276)
(114246,Walk Among the Tombstones, A (2014),1.0377358490566038,3.1659663865546217)
(5418,Bourne Identity, The (2002),1.0334246575342465,3.911517001789662)
(102903,Now You See Me (2013),1.0306532663316583,3.6643631724431214)
(7445,Man on Fire (2004),1.0292682926829269,3.8061813186813187)
(78088,Buried (2010),1.029126213592233,3.3575697211155378)
(45447,Da Vinci Code, The (2006),1.0284584980237155,3.144982464079647)
(648,Mission: Impossible (1996),1.0280851063829788,3.400149588631264)

On voit qu'on retrouve Ange et Démon en 2ème position et Da Vinci Code en 9ème position. Ceux sont des films similaires à Inferno, du même réalisateur, avec le même acteur principal et basés sur les livres d'un même auteur, Dan Brown.
Compte tenu du peu d'information que l'on a sur les films et du fait que les informations ne soient pas très pertinentes (le tag le plus utilisé pour Inferno est « prédictable »), réussir à avoir ces deux films dans le top 10 montre que la procédure marche bien !

Variations possibles :

- J'ai aussi extrait l'année de sortie du film et en ai fait une métrique à partir de la différence entre 2016 et la sortie du film mais cette métrique a repoussé les films ange et Démon et Da Vinci Code à des positions plus éloignées donc je ne l'ai pas gardé.

- Pour une question de temps (challenge de 2-4h) et parce que le film Inferno est récent je n'ai pas du tout considéré les timestamps dans les données et je n'ai pas non plus utilisé le fichier *links.csv* pour miner les pages de IMDB par exemple. Mais cela aurait été intéressant de récupérer d'autres informations à partir des liens et de ne considérer que les tags ou les notes récents sur un film pour prendre en compte le fait qu'un film ai bien vieilli ou pas.

- Dans le calcul de la métrique tag, j'aurais pu ne considérer que des tags postés par des utilisateurs fréquents. On a le mot action qui reviens, j'aurais pu m'en servir comme genre aussi.

- Le calcul des notes moyennes aurait aussi pu se faire en ne prenant en compte que les notes des utilisateurs qui ont posté un certain nombre de fois.

- On aurait aussi pu ne pas simplement utiliser les films bien notés par les utilisateurs qui ont adoré Inferno a des fins de filtrage mais par exemple récupérer les tags associés et s'en servir pour améliorer le classement des films.

- J'ai considéré que les deux genres associés à Inferno ont le même poids, on aurait pu en privilégier un par rapport à l'autre, de même pour le calcul de la métrique finale, on aurait pu privilégier la métrique tag par rapport à la métrique genre car elle est plus descriptive.

Malgré les approximations que j'ai fait, l'algorithme de scoring semble marcher.