MOVIE RECOMMENDATION SYSTEM

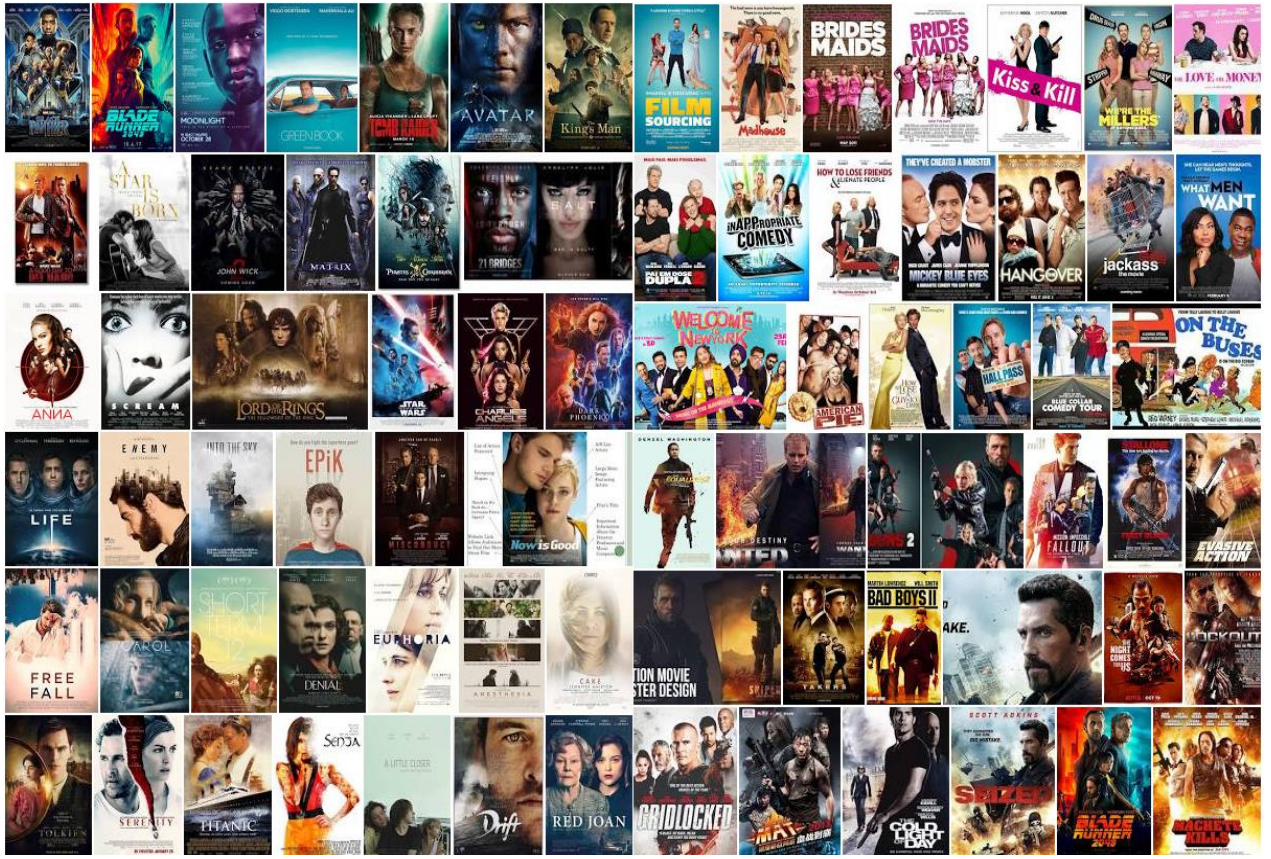
A Network Tour of Data Science (EE-558)

Ecole polytechnique fédérale de Lausanne,
January 2020

Team 16
André Clerc, Aurélien Kinet,
Jules Afresne, Jelena Simeunović

Motivation

- 20.75 minutes average “research” on Netflix
- Growth of online streaming platforms



CONTENT

- Recommender Algorithms
- Data acquisition
- Data exploration
- Data exploitation:
- Data exploitation: K-means clustering
- Data exploitation: Cluster analysis
- Data exploitation: Recommendation system
- Data exploitation: Comparison of results

EPFL

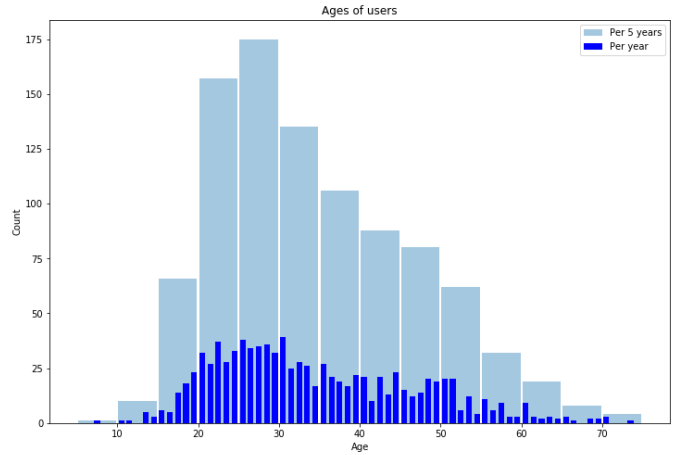
Recommender Algorithms

- Content – based recommender algorithms:
  - Uses attributes (i.e. genre) to find similarity between items to make recommendation

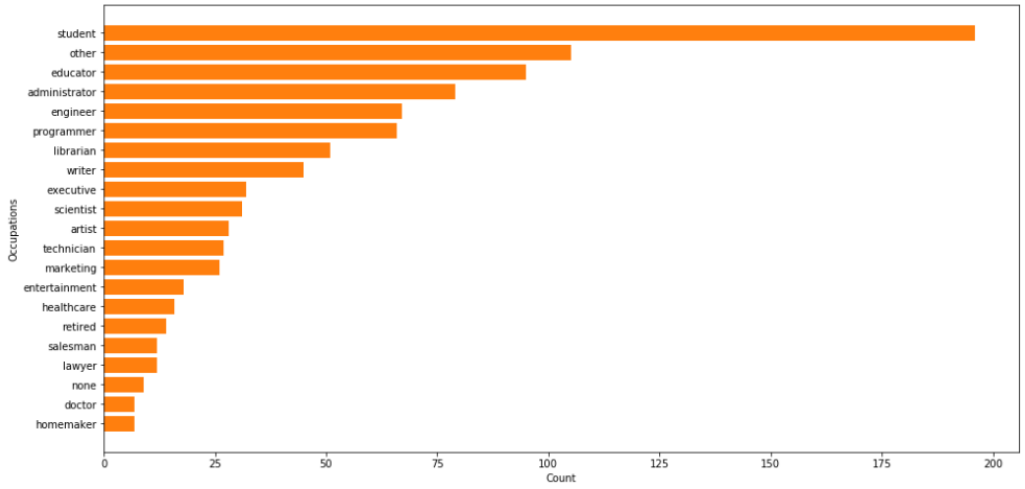
- Collaborative filtering:
  - Uses historical preferences of user

Data acquisition

- Movielens 100k Dataset
- 943 users (attributes: gender, occupation, age and zipcode)
- 1682 movies (attributes: genre, release year)

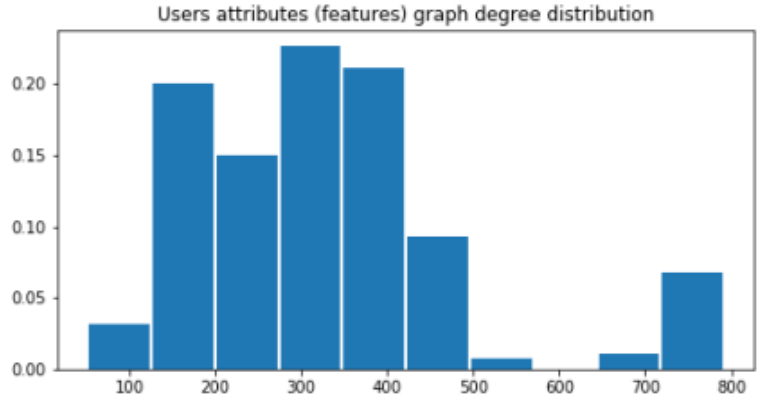


- Three graphs constructed:
    1. User graph
    2. Movie graph
    3. User-movie bipartite graph
       (edges are user ratings)
    4. User – category graph

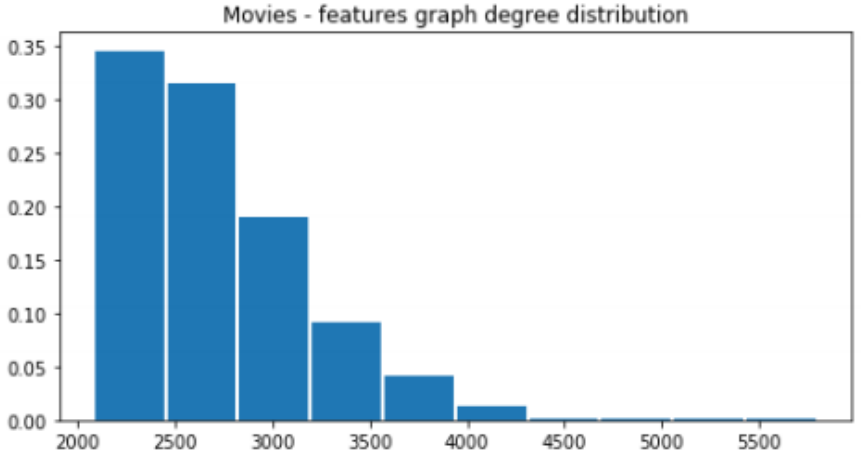


EPFL

Data exploration



- Poisson-like distribution
- Clustering coefficient <C> = 0.68 ,
- Small-world phenomena
- Watts-Strogatz model
- Graph is connected and matrix is sparse



- Right-skewed degree distribution
- Power-low distribution
- Barabási-Albert network
- Graph is connected if adjacency matrix is not too sparse

EPFL

Data exploitation

- Constructed bipartite Movie_user graph
- K-means algoritham did not give expected insights on previously constructed graph
- Created User-Category graph:
  - Find total rating per user according to genre of movies and movie's ratings

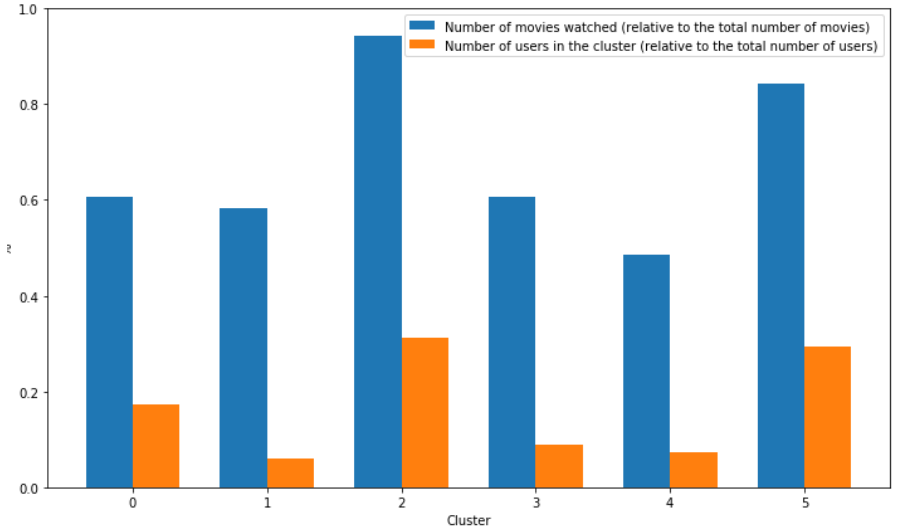
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
0	0.001704	0.127768	0.071550	0.020443	0.042589	0.155026	0.042589	0.008518	0.182283	0.003407	0.001704	0.022147	0.022147	0.008518	0.074957	0.073254	0.088586	0.042589	0.010221
1	0.000000	0.081301	0.024390	0.008130	0.032520	0.130081	0.073171	0.000000	0.284553	0.008130	0.016260	0.016260	0.008130	0.032520	0.130081	0.032520	0.097561	0.024390	0.000000
2	0.000000	0.114754	0.032787	0.000000	0.000000	0.098361	0.081967	0.008197	0.180328	0.000000	0.016393	0.040984	0.016393	0.090164	0.040984	0.065574	0.172131	0.040984	0.000000
3	0.000000	0.142857	0.071429	0.000000	0.000000	0.071429	0.071429	0.017857	0.107143	0.000000	0.000000	0.017857	0.017857	0.089286	0.053571	0.107143	0.196429	0.035714	0.000000
4	0.002604	0.145833	0.085938	0.036458	0.075521	0.213542	0.023438	0.000000	0.070312	0.005208	0.002604	0.072917	0.031250	0.007812	0.049479	0.085938	0.049479	0.036458	0.005208
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
938	0.000000	0.174757	0.097087	0.009709	0.009709	0.145631	0.029126	0.000000	0.174757	0.009709	0.000000	0.009709	0.009709	0.009709	0.097087	0.077670	0.116505	0.029126	0.000000
939	0.000000	0.098291	0.051282	0.008547	0.021368	0.179487	0.029915	0.000000	0.205128	0.000000	0.012821	0.017094	0.029915	0.012821	0.102564	0.081197	0.085470	0.064103	0.000000
940	0.000000	0.185185	0.129630	0.055556	0.037037	0.129630	0.018519	0.000000	0.092593	0.000000	0.000000	0.000000	0.018519	0.018519	0.018519	0.148148	0.129630	0.018519	0.000000
941	0.000000	0.102857	0.062857	0.022857	0.074286	0.125714	0.000000	0.000000	0.177143	0.011429	0.005714	0.017143	0.028571	0.051429	0.097143	0.034286	0.114286	0.057143	0.017143
942	0.000000	0.169312	0.092593	0.005291	0.023810	0.148148	0.047619	0.000000	0.150794	0.005291	0.000000	0.037037	0.010582	0.007937	0.087302	0.058201	0.092593	0.039683	0.023810

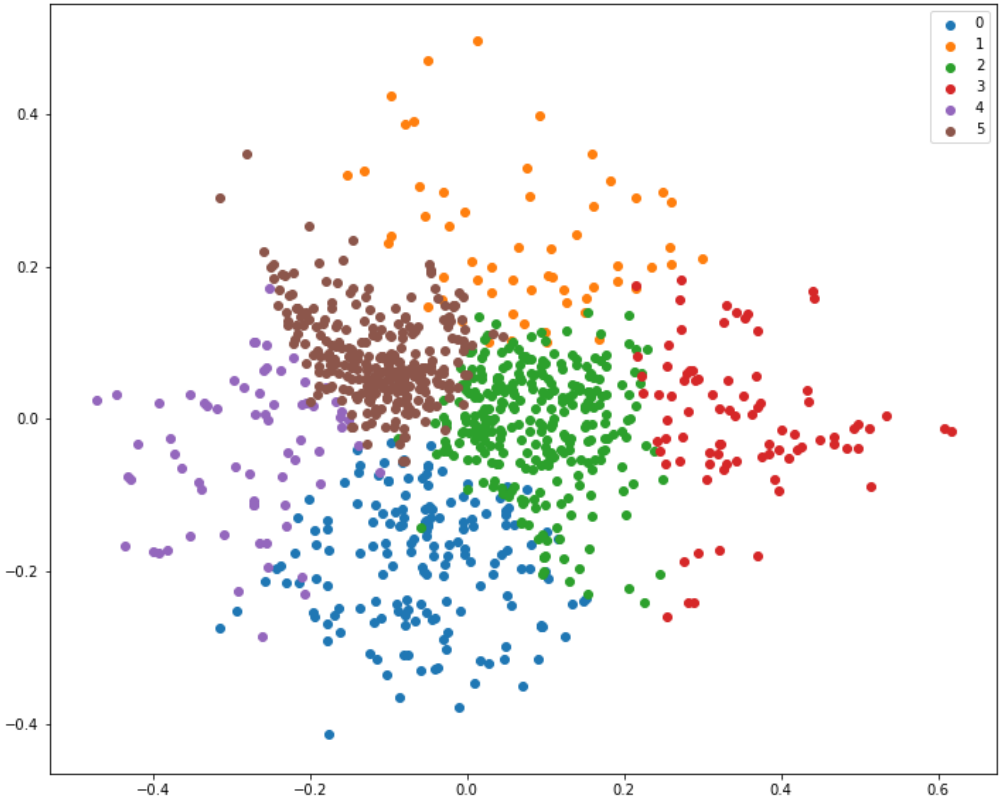
EPFL

Data exploitation: K-means clustering

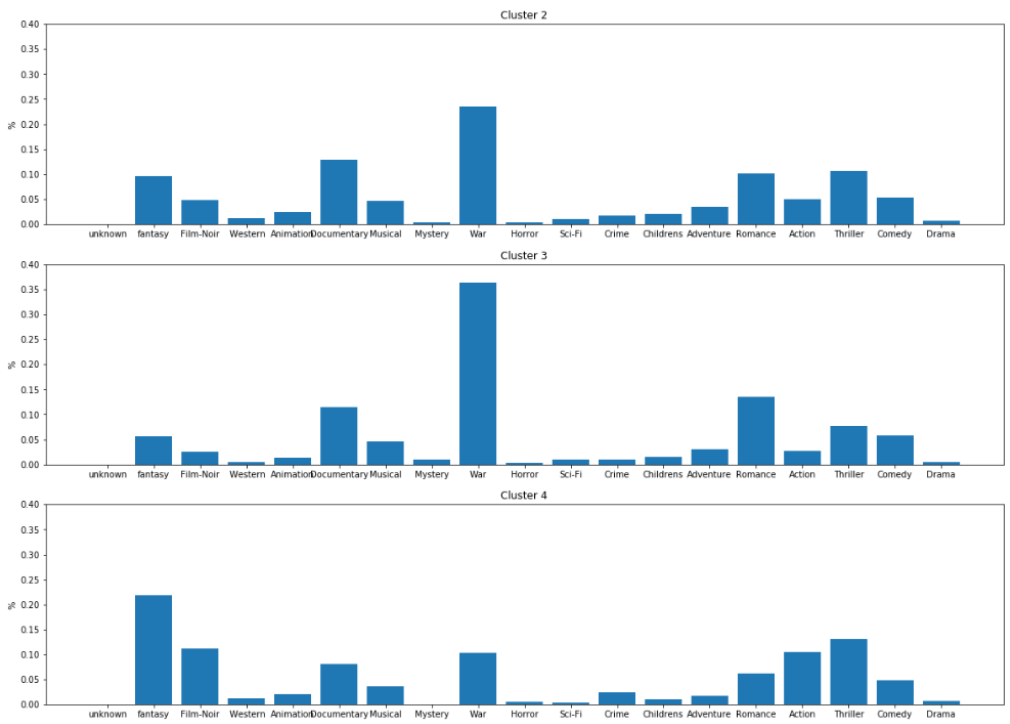
- K-means clustering was performed on User-Category graph
- Elbow method applied and 6 clusters are chosen
- Isomaps are used for dimensionality reduction

- Analysis of movies watched in each cluster vs.
  Number of users per cluster

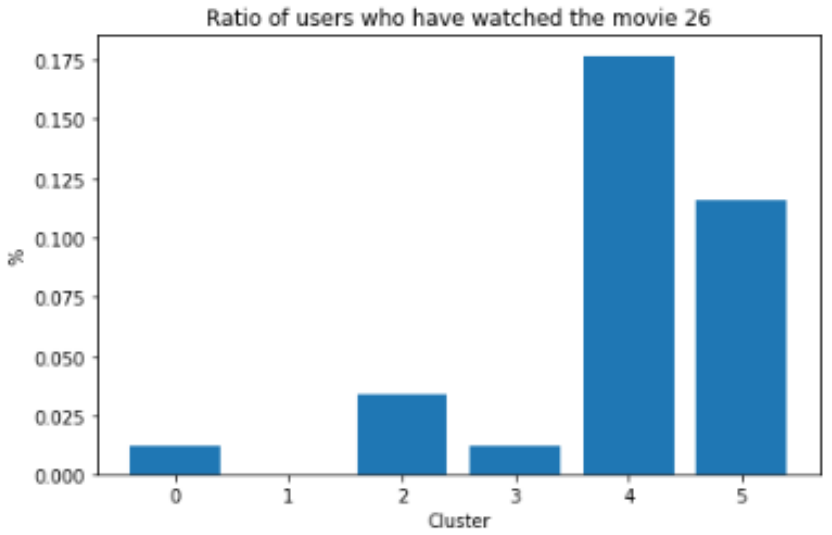




Data exploitation: Cluster analysis



- Ratio of ratings per category
- User preferences in 3 categories towards certain genres



- Analysis: movie (id= 26) Brothers McMullen (1995) is an action movie
- Users that watched and rated it highest are in fifth and sixth cluster

Data exploitation: Recommendation system

- Determine in which category is user
- Find watched and unwatched movies for user, taking into account its relative category (cluster)
- Rating of movie based on arithmetic mean of cluster rating
- Offer movie with the highest rating

- Example: User 189 with 3 recommendations

```
-- Recommandation  1 --
        Title:  Schindler's List (1993)
        Release date:  01-Jan-1993
        Cluster rating:  4.666666666666667 /5
        Rating of all users:  4.466442953020135 /5
        Categorie(s):  ['Drama', 'War']

-- Recommandation  2 --
        Title:  Henry V (1989)
        Release date:  01-Jan-1989
        Cluster rating:  5.0 /5
        Rating of all users:  4.137096774193548 /5
        Categorie(s):  ['Drama', 'War']

-- Recommandation  3 --
        Title:  Thin Man, The (1934)
        Release date:  01-Jan-1934
        Cluster rating:  5.0 /5
        Rating of all users:  4.15 /5
        Categorie(s):  ['Mystery']
```

Data exploitation: Comparison of results

- Collaborative filtering based recommender:
  - Keras Embedding
  - Neural Network

- Tested against Collaborative filtering:
  1. ``Schindler's list (1993)`` predicted rate: 3.91
  2. ``Henry V (1989)`` predicted rate: 3.26
  3. `` The Thin Man (1934)`` predicted rating: 1.74

```
-- Recommandation  1 --
        Title:  Schindler's List (1993)
        Release date:  01-Jan-1993
        Cluster rating:  4.666666666666667 /5
        Rating of all users:  4.466442953020135 /5
        Categorie(s):  ['Drama', 'War']

-- Recommandation  2 --
        Title:  Henry V (1989)
        Release date:  01-Jan-1989
        Cluster rating:  5.0 /5
        Rating of all users:  4.137096774193548 /5
        Categorie(s):  ['Drama', 'War']

-- Recommandation  3 --
        Title:  Thin Man, The (1934)
        Release date:  01-Jan-1934
        Cluster rating:  5.0 /5
        Rating of all users:  4.15 /5
        Categorie(s):  ['Mystery']
```

EPFL

Conclusion

- Beneficial for movie production by targeting majority of users through genre combination
- Our recommendation system seems relevant, although there is a place for improvement
- Hindrance: if user is incorrectly classified in cluster, the predictions will be widely affected
- Possible improvements: application of Graph Convolution Neural Network and compare results

Thank you for attention!