ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Spectral analysis of 5000 movies network

By:
Macko Vladimir, Novakovic Milica,
Pavué Clément, Roussaky Mehdi

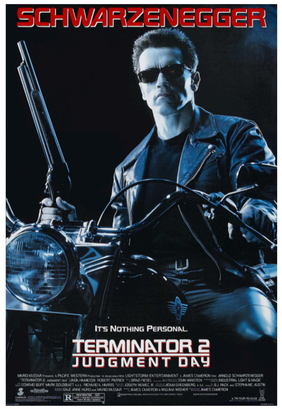
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Goals of the Project

- Build & analyse the movie graph
- Genres classification investigation
- Movie recommender engine

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Data Acquisition & cleaning



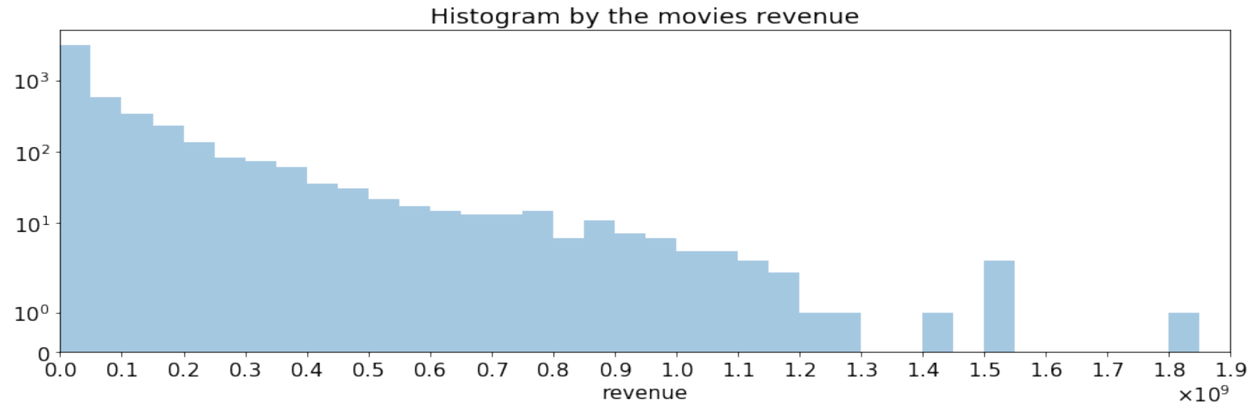
- Starting point: 5000 movies datasets from Kaggle of movies and production crews [in text and JSON format]
- Datasets used to build Pandas data-frame with movies as lines and their attributes as columns.
- Unuseful lines removed: not released movies, movies with no title or no actor,...
- Unuseful attributes removed: movie IDs, webpage,...

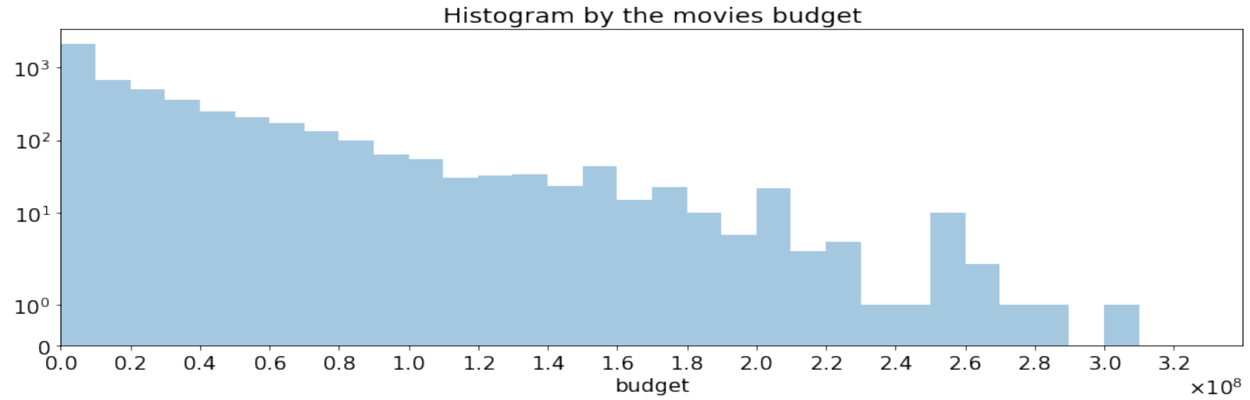
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

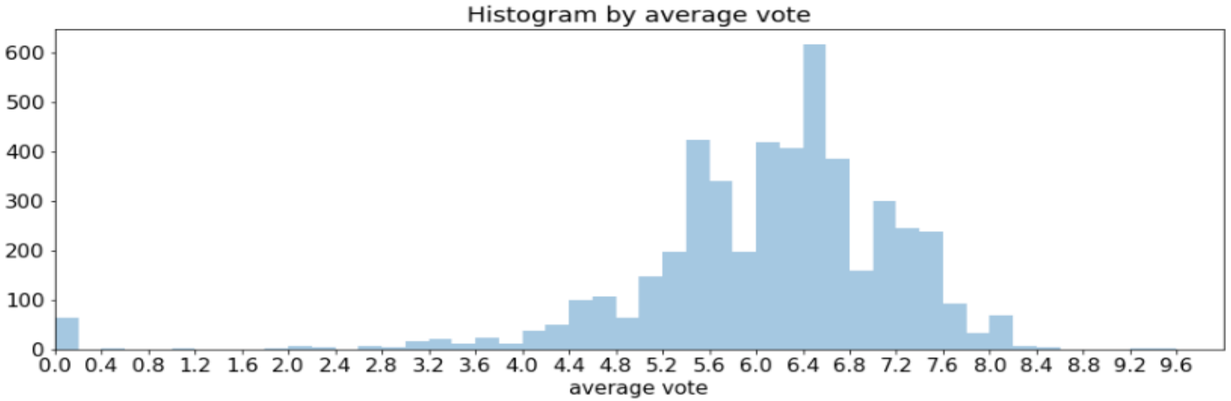
Data Exploration

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

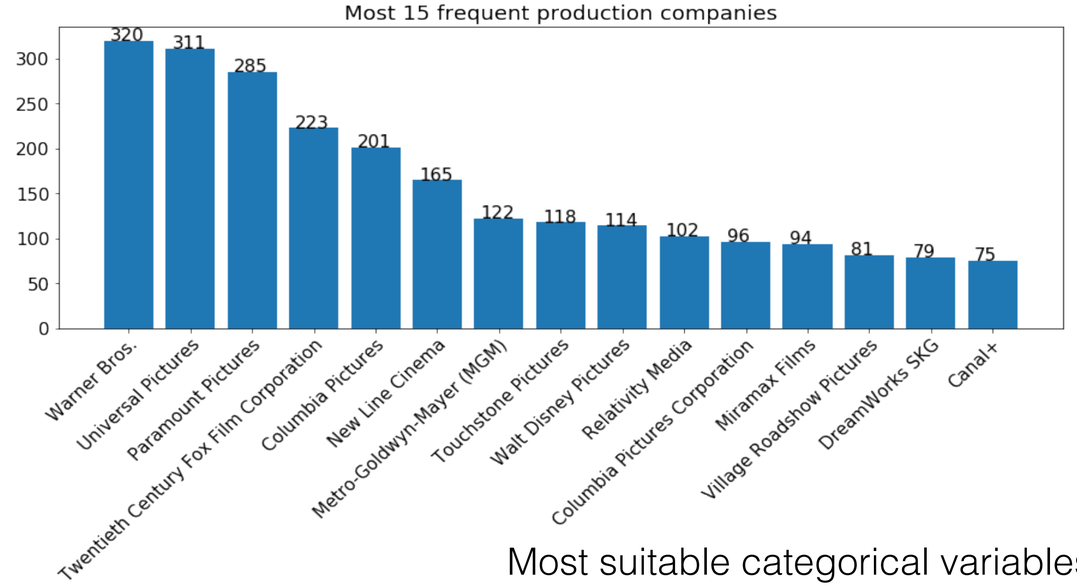
Numerical attributes

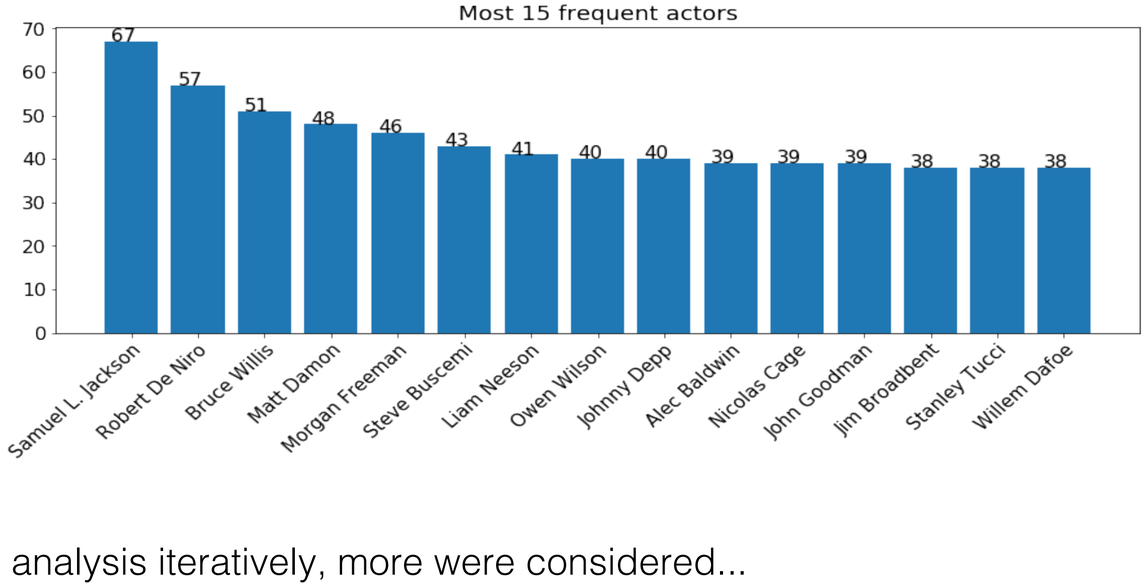




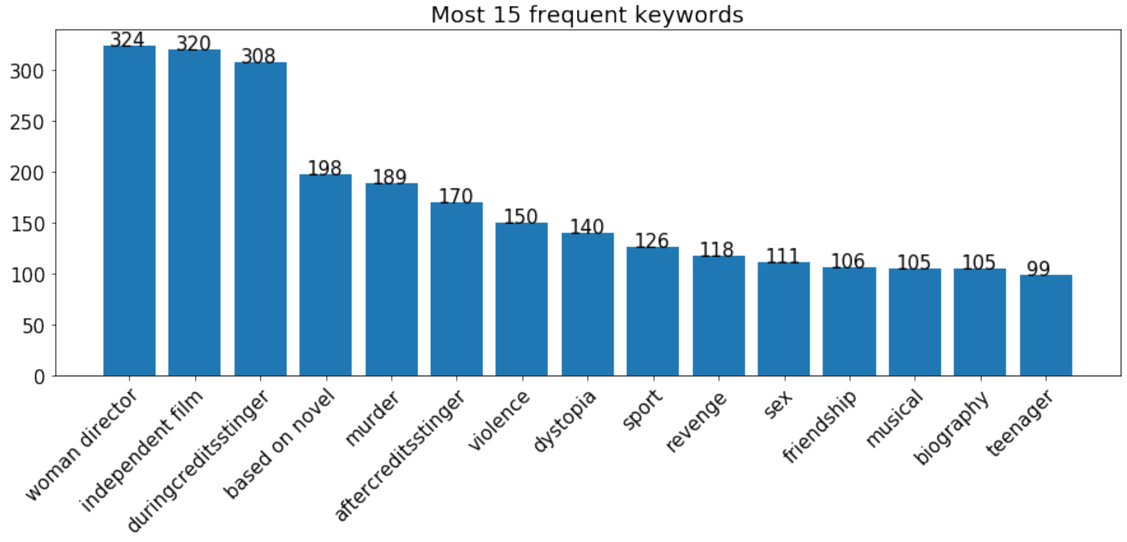


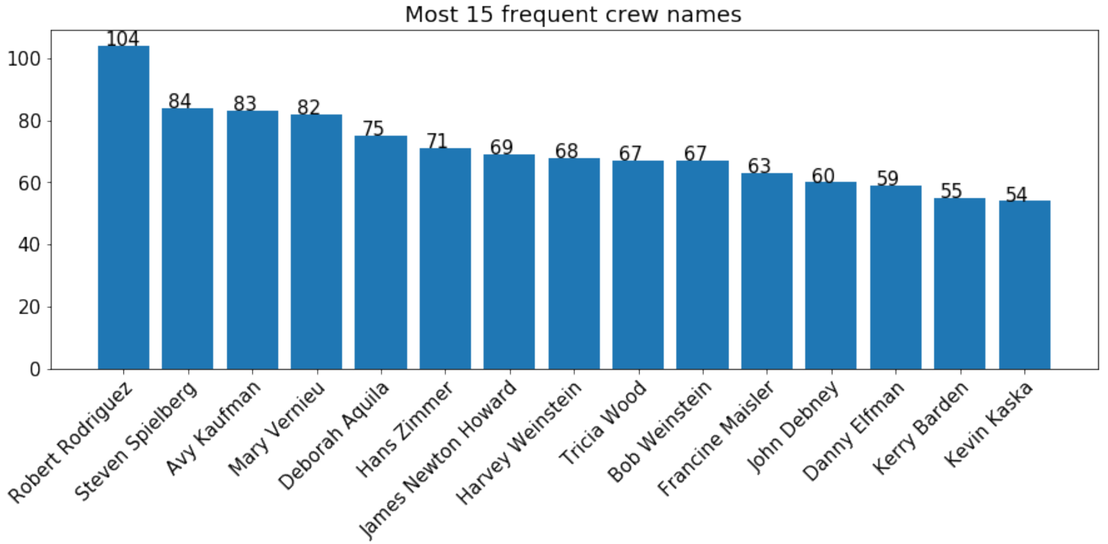
Categorical attributes



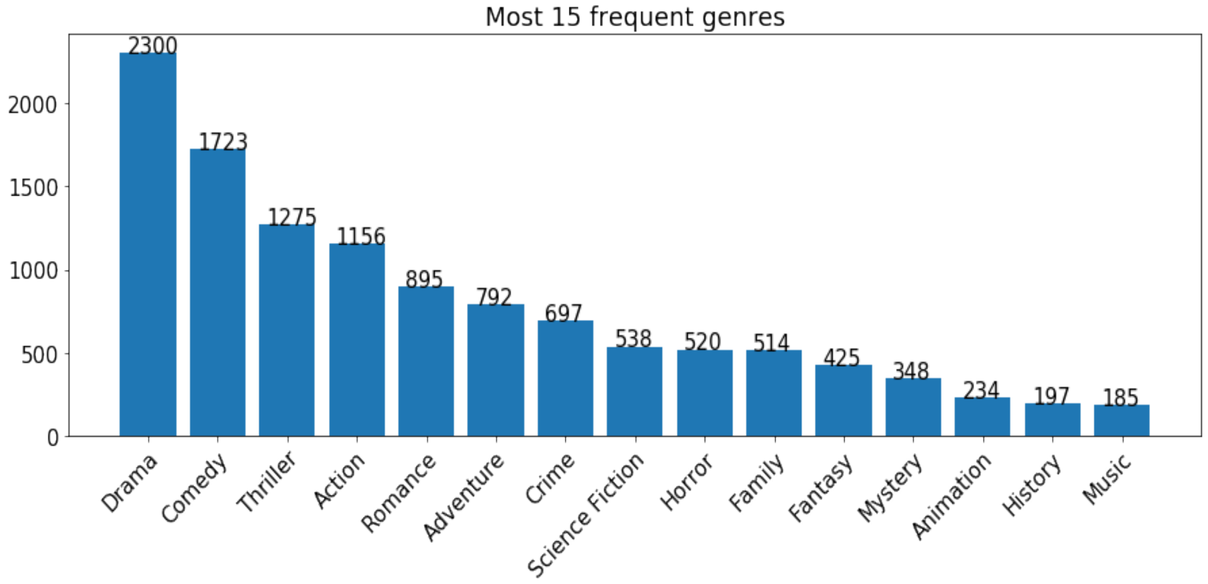


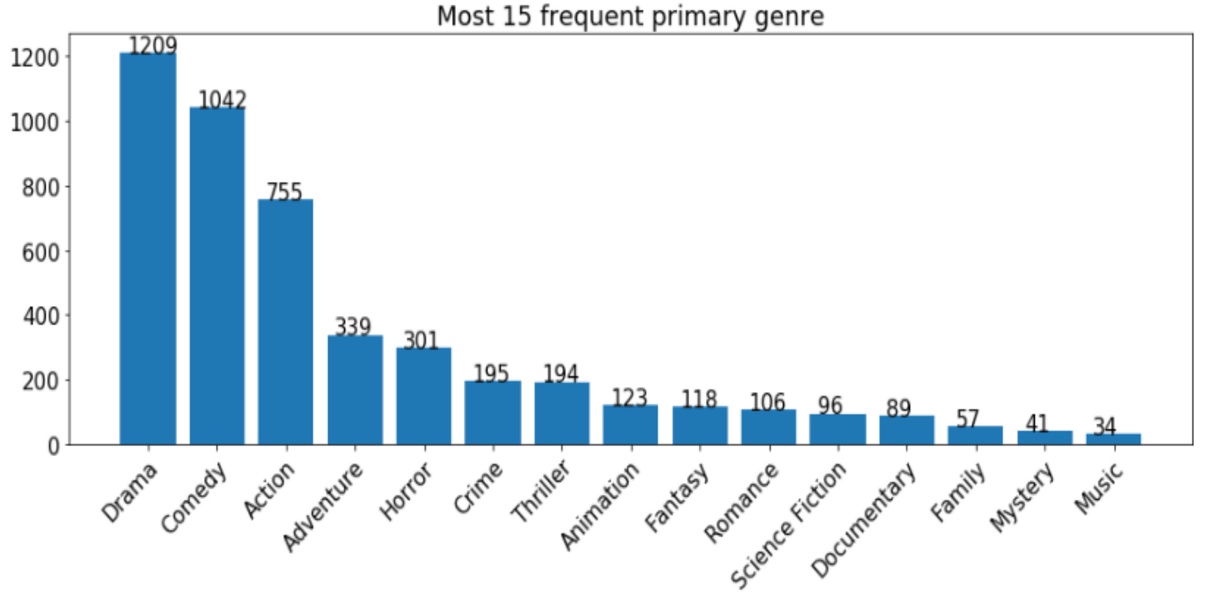
Most suitable categorical variables selected for analysis iteratively, more were considered...



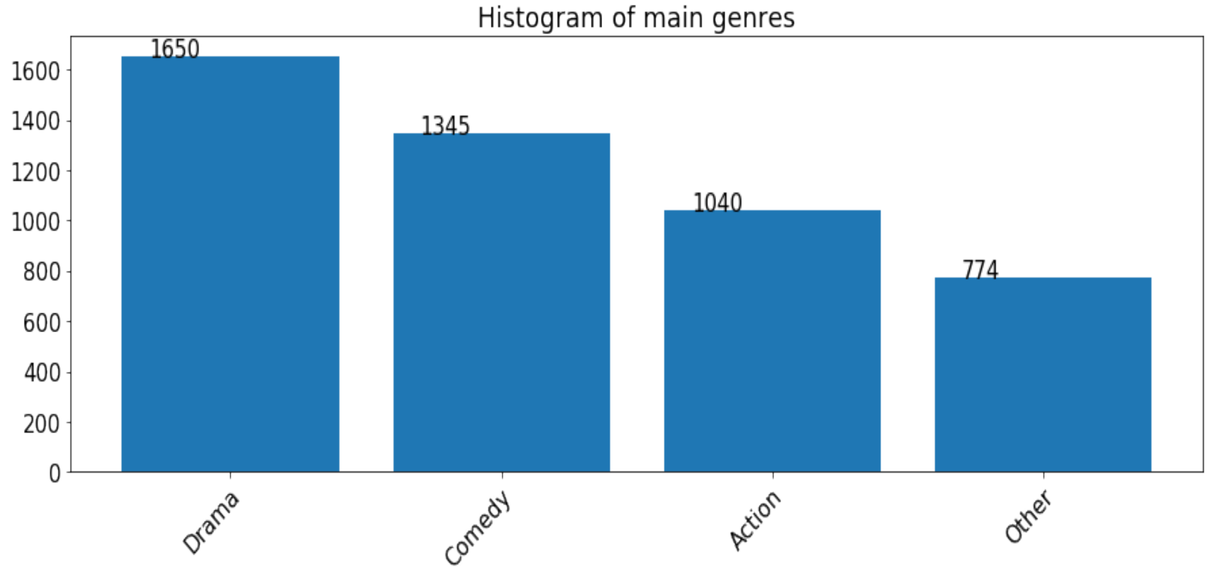


Genres





- For classification simplification we include all genres with low population into combined category «other»



ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Data exploitation

1. Investigation of possibility of genre classification using movie graph.

2. Movie recommender system.

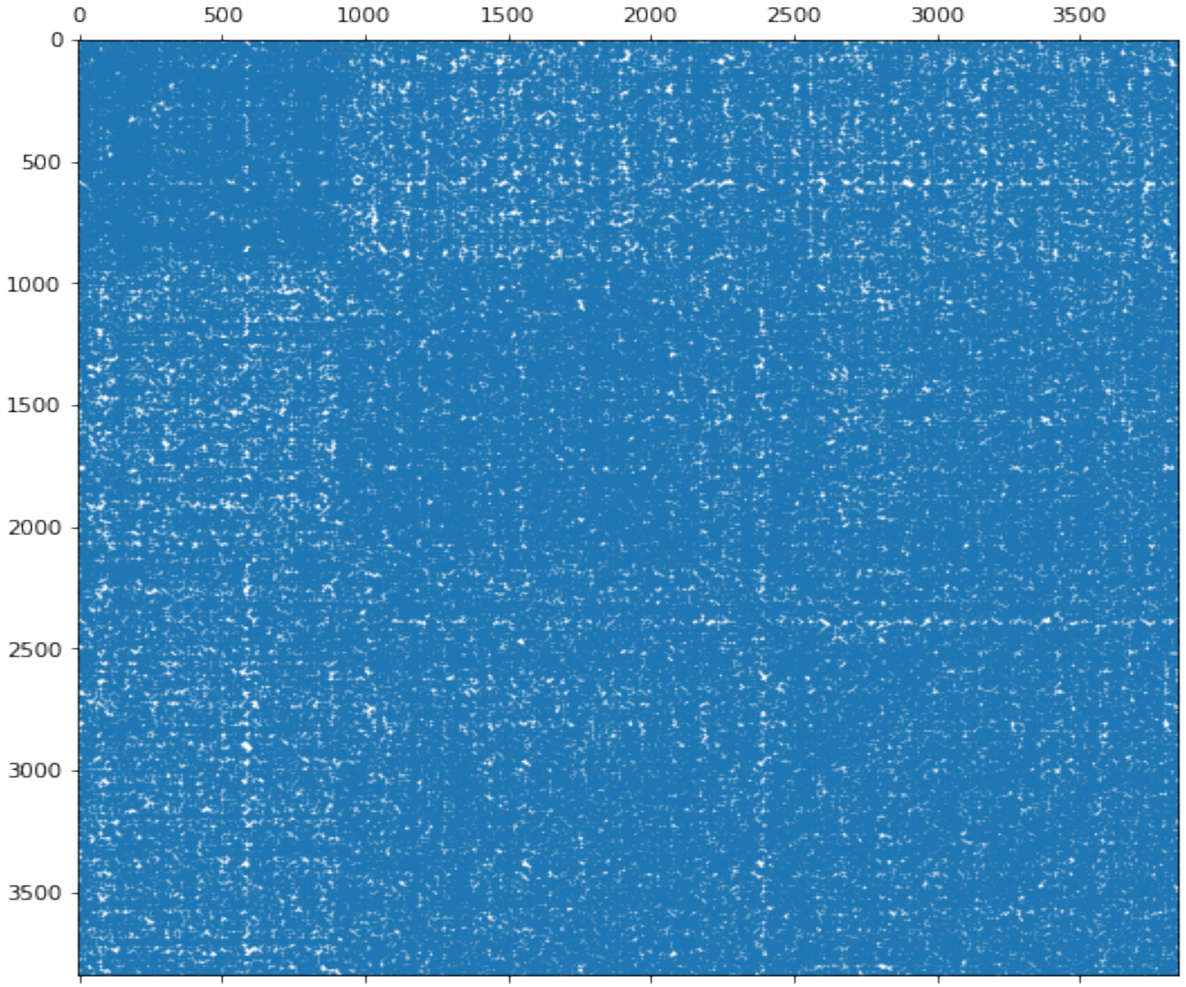
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

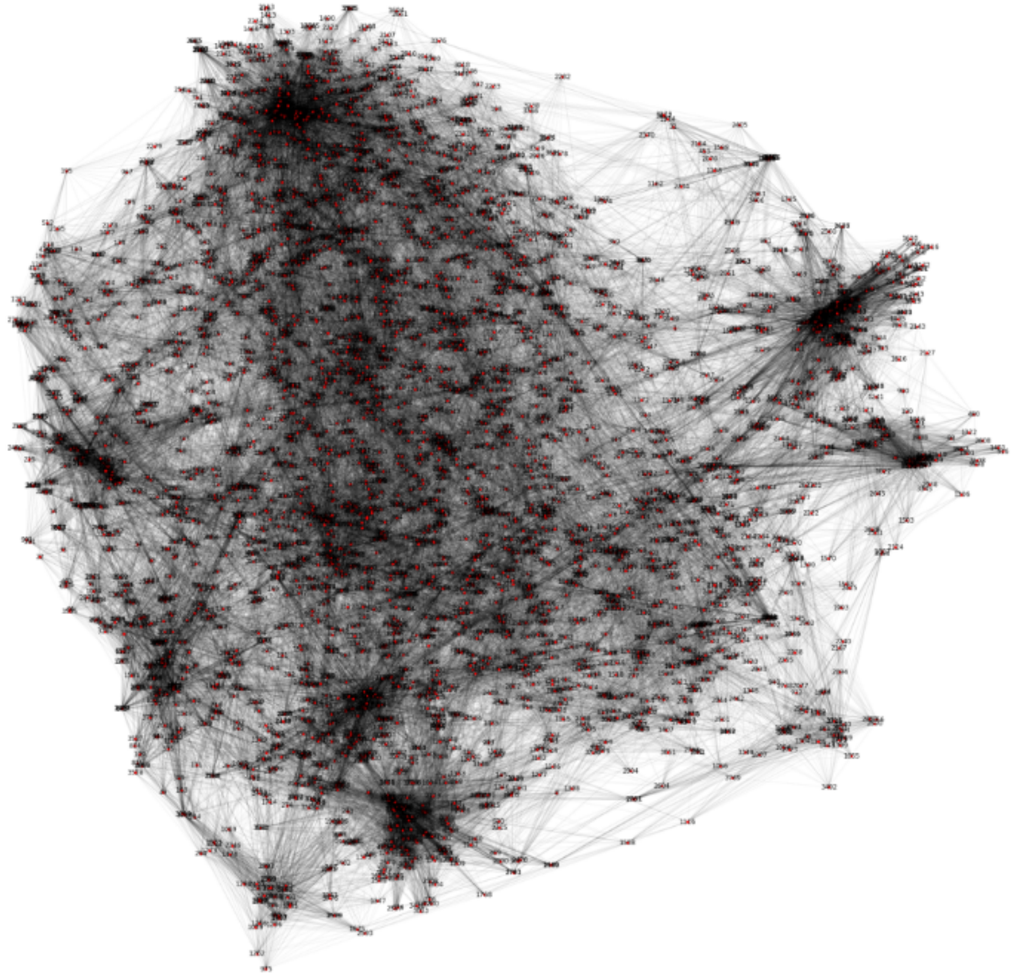
1. Genre classification graph

- Feature selection
- Reducing extremely high dimensionality with applying PCA
- Distance calculation using cosine metric
- Weight matrix calculated from distance matrix using gaussian kernel and knn=50
- Number of selected PCA components and value of knn optimized manually

Movies graph

- Weights matrix shows at least 2 squares representing genres, however, they are not well distinguished





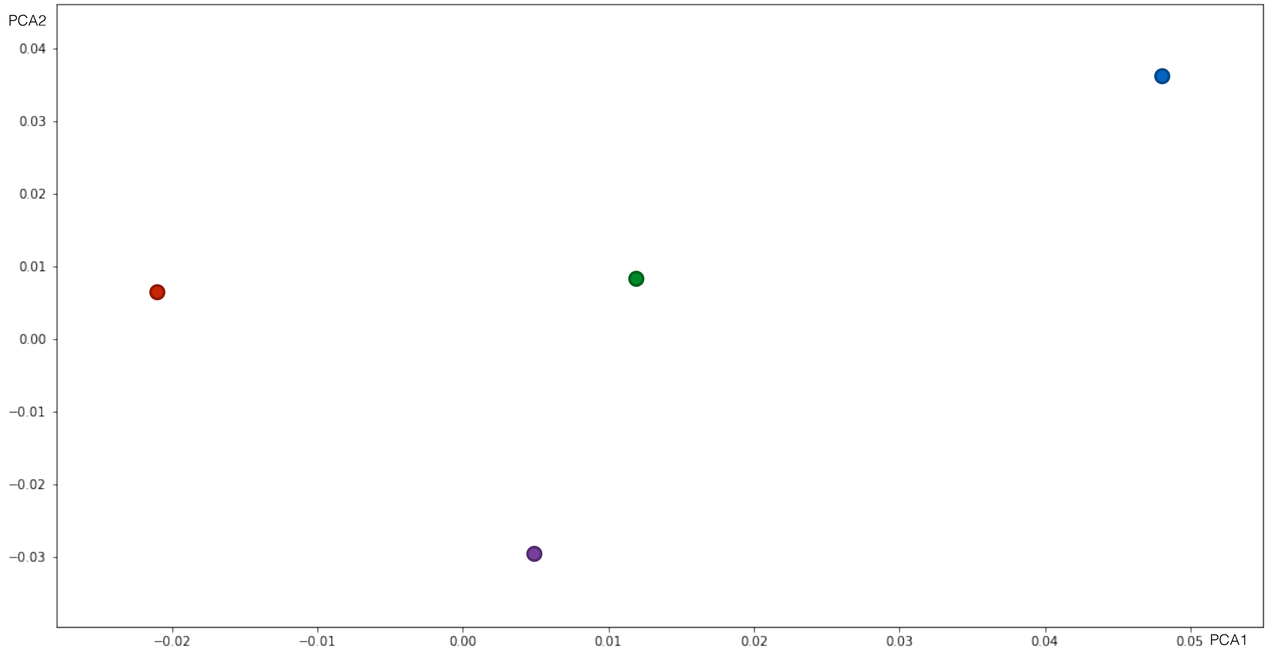
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Unique movies suggestion

- Exploring graph structure one can notice that some nodes are fully disconnected
- These nodes represent unique movies and for the purpose of classification they are dropped
- Some of them are:
  “Antarctica: A Year on Ice”
  “Sound of My Voice”
  “The Sound and the Shadow”
  “Wind Walkers”

Unnormalized spectral clustering

- Distances Matrix –Gaussian Kernel & Knn=50 --> Weight Matrix
- Compute unnormalized laplacian L
- Compute the first k eigenvectors of L and put them as columns in the matrix U
- Cluster with k-means (each row of U can be assigned to a cluster)



Clustering accuracy

- Final accuracy is 33%

Reasons for low accuracy:

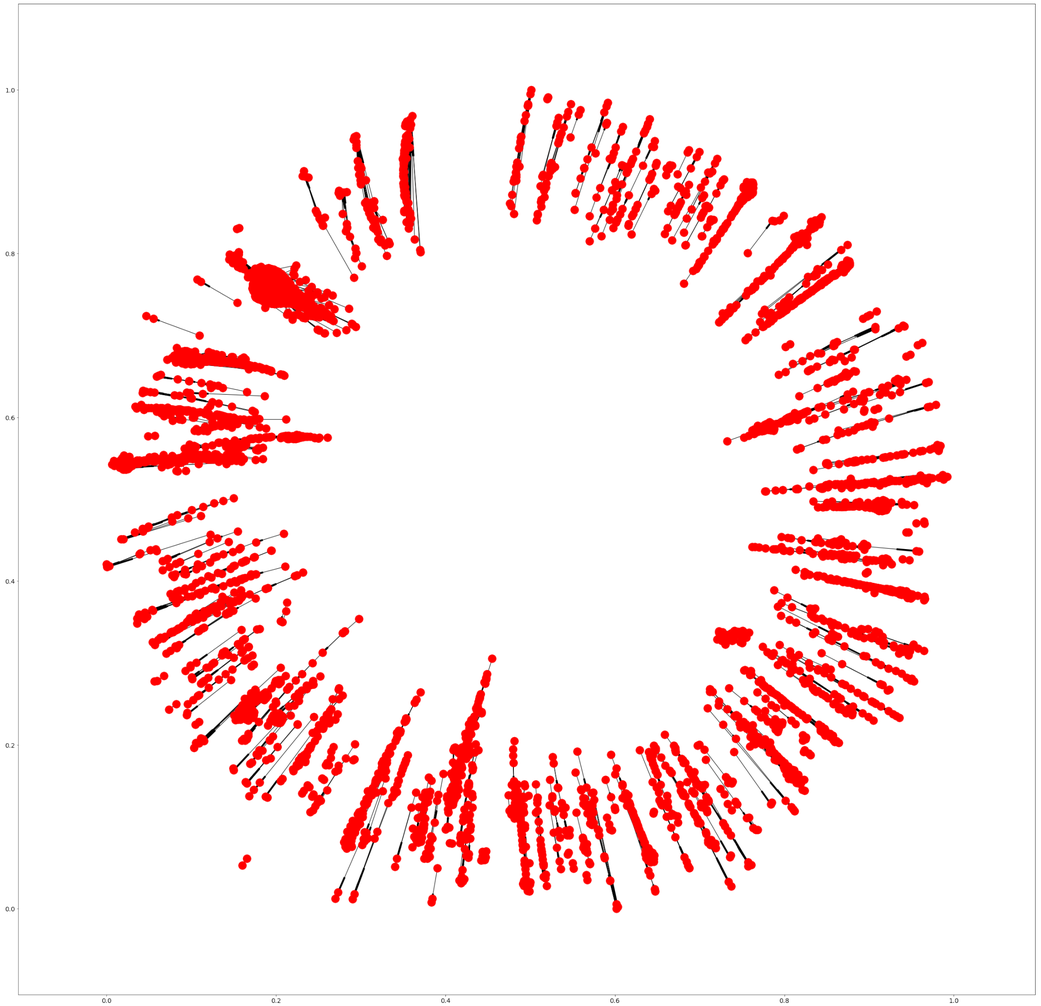
- Our model is based only on main genre, however movies are associated with multiple genres.
- We cumulate poorly populated genre in other.

Solutions to optimize accuracy:

- Tune hyper-parameters: numbers of actors, crew members and keywords, number of neighbors: plays a crucial role in visibility of the squares in the ordered weight matrix
- Advanced features: audiovisual samples
- Other spectral clustering algorithms

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

2. Recommender engine



New graph constructed with a cost function based on genres, keywords and votes of the movies.

Over input was tried such as text mining on the overview but it didn’t give exploitable results.

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Some predictions:

Original Movie	Predicted Movie
Avatar	Star Trek Into Darkness
Avengers: Age of Ultron	Captain America: The Winter Soldier
Suicide Squad	The Dark Knight Rises
Star Wars: Episode III - Revenge of the Sith	Star Trek
300	Gladiator

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Conclusions

- Two graphs has been constructed and studied

- Low accuracy of classification tells us that spectral classification applied on this dataset does not give reliable result

- Sample movie suggestion engine leading to reasonable suggestion based on personal review of suggestions