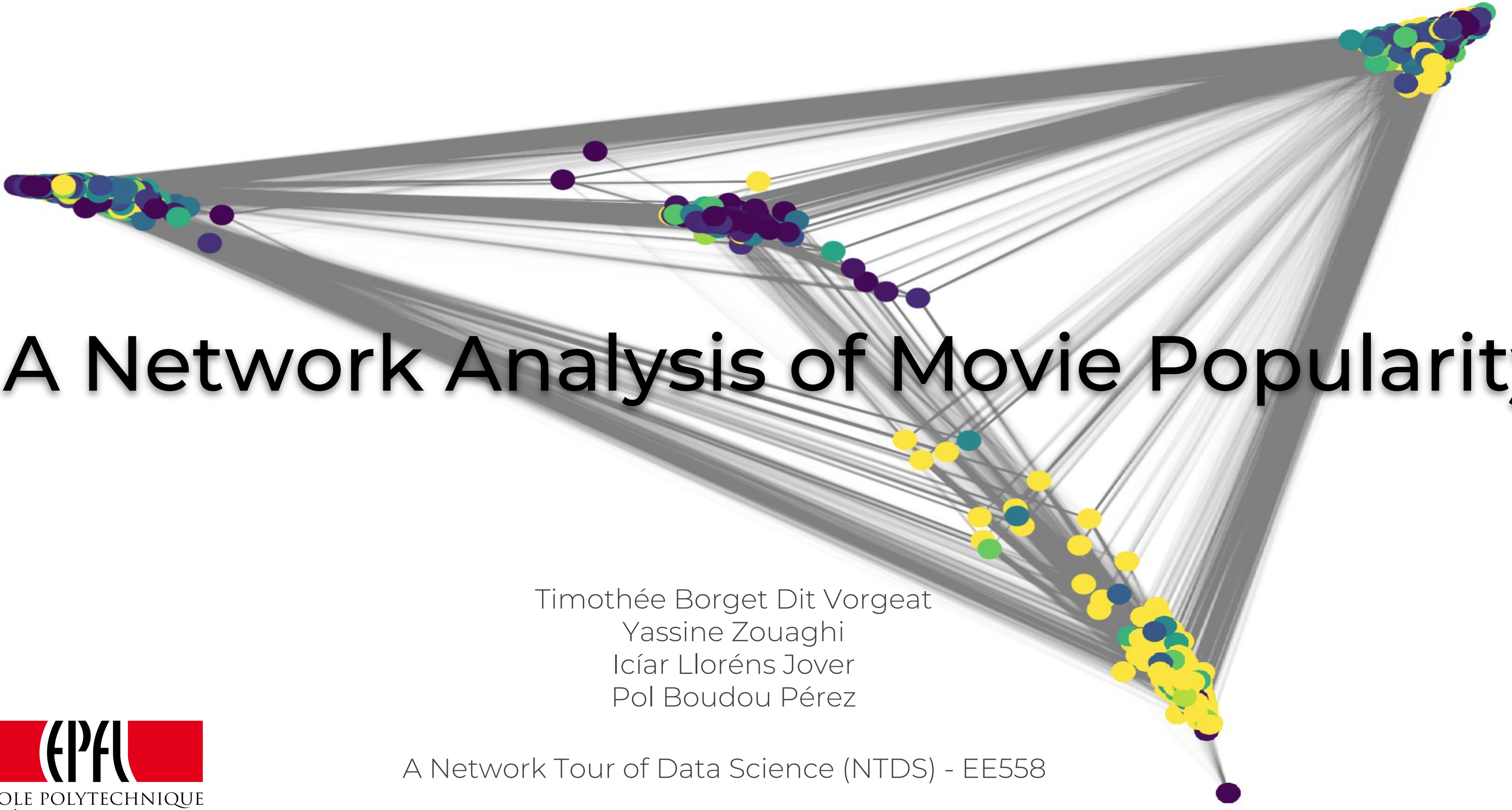


A Network Analysis of Movie Popularity

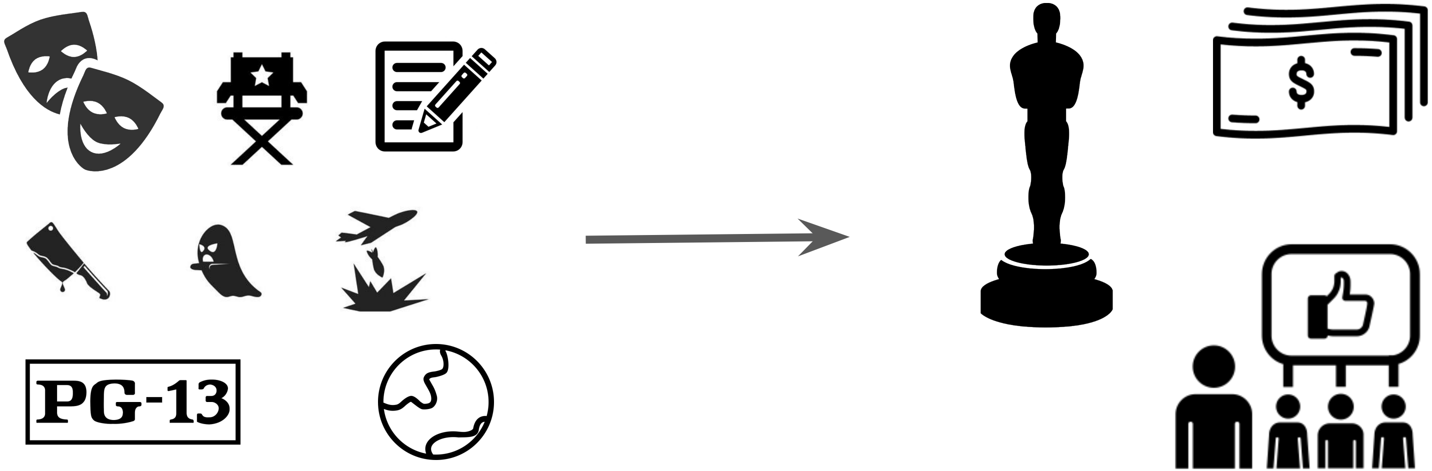
Timothée Borget Dit Vorgeat
Yassine Zouaghi
Icíar Lloréns Jover
Pol Boudou Pérez

A Network Tour of Data Science (NTDS) - EE558

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Aim

Analyze several movie features by creating a film network in order to observe which characteristics lead to popularity.



Data

- TMDb Dataset (Kaggle) for the budget
- OMDb API

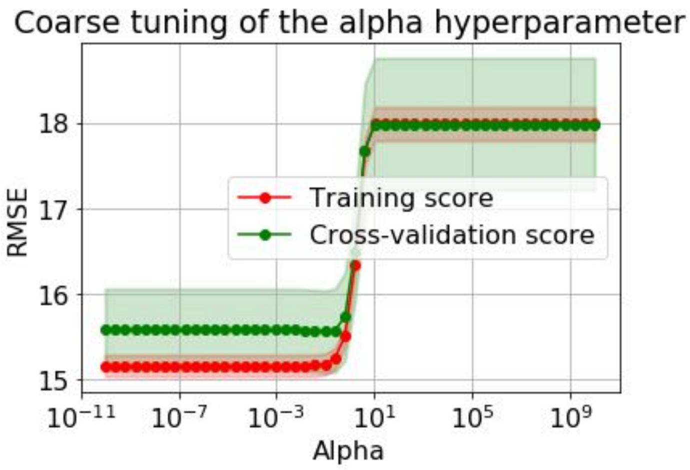
Features

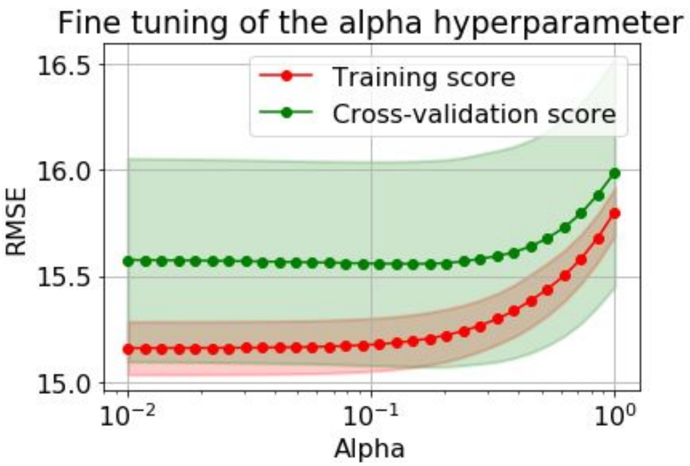
Labels

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

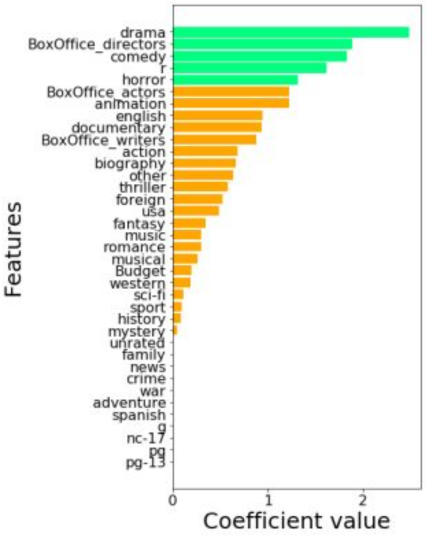
Lasso regression

Minimize $$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|$$

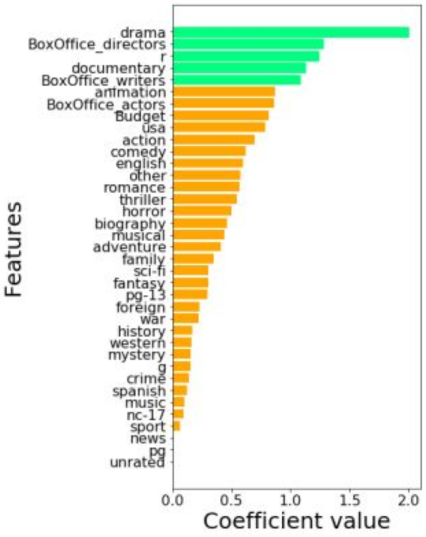




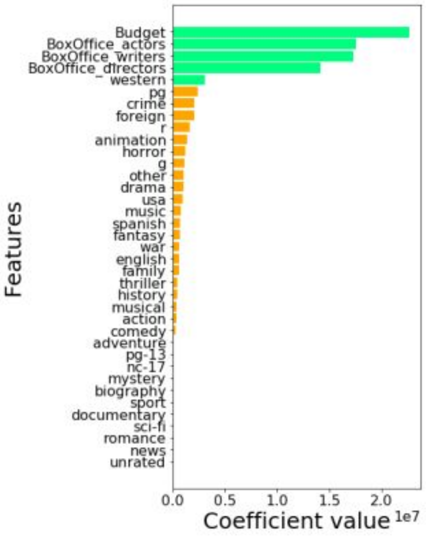
Selected features



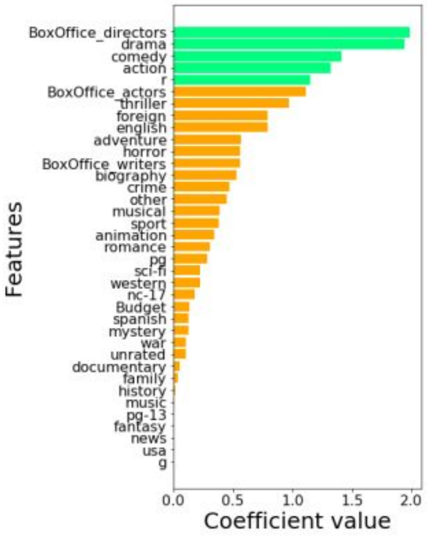
(a) IMDb grades



(b) Combination of Rotten tomatoes and Metacritic grades as targets



(c) BoxOffice generated

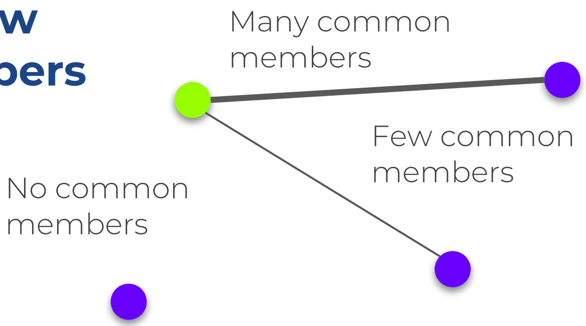


(d) Combination of nominations and wins as targets

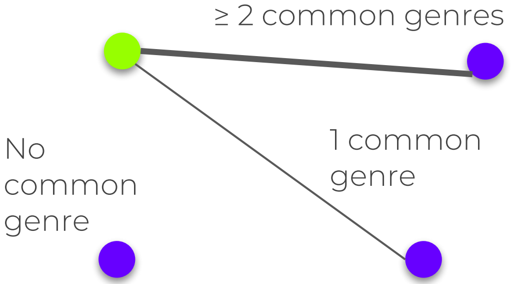
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Graph creation

Crew members

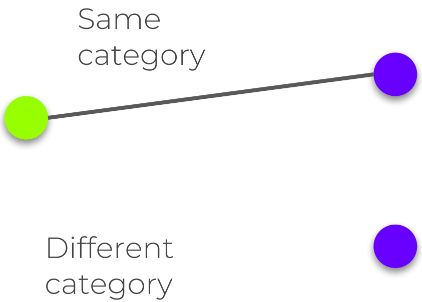


Genre



Budget

- High budget: more than 101 million $
- Medium budget: 41 to 100 million $
- Low budget: 10 to 40 million $
- Independent: 100 000 to 10 million $
- No budget: less than 10 000 $



Clustering

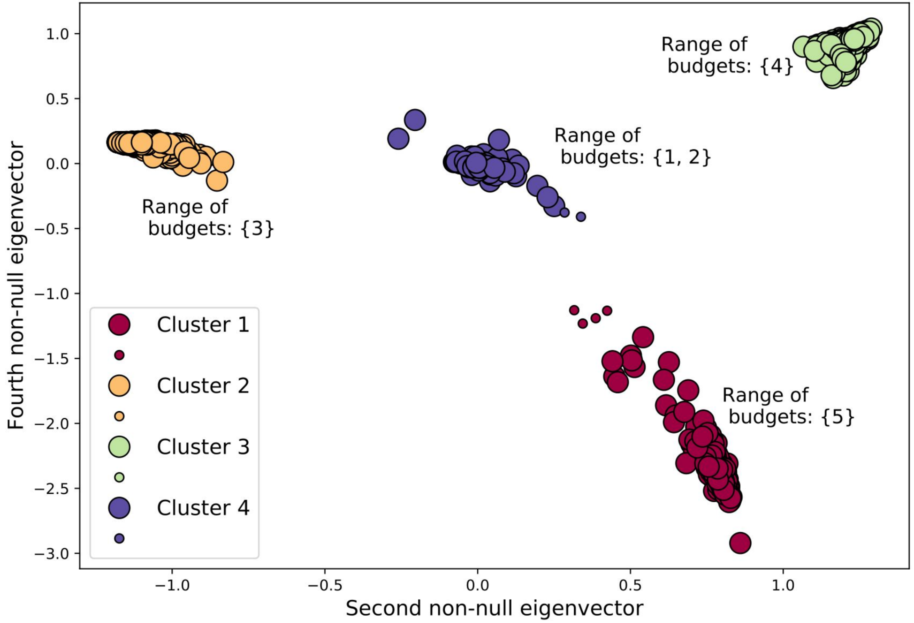
1. Laplacian embedding
2. DBSCAN clustering
   - 4 clusters
   - 0 outliers
   - Silhouette coeff: 0.941

High budget (more than 101 million $)

Medium budget (41 to 100 million $)

Low budget (10 to 40 million $)

Independent + No budget (less than 10 million $)

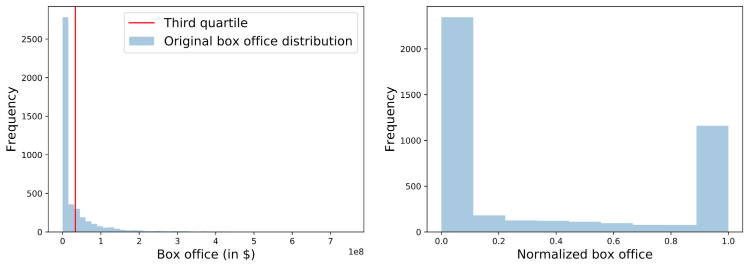


Epsilon =0.35, min samples = 10

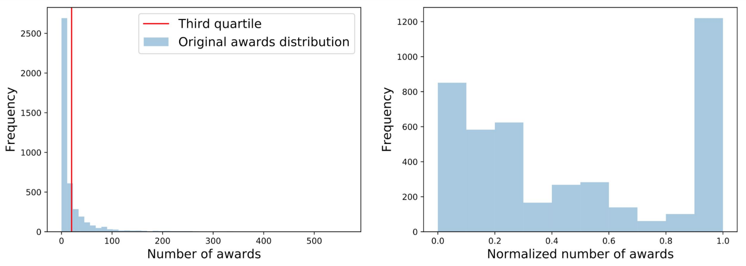
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Labels

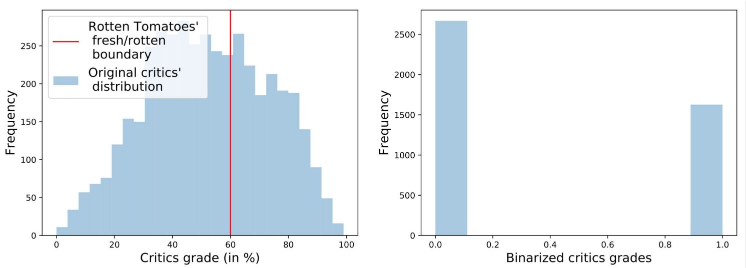
Box office: clipped and normalized



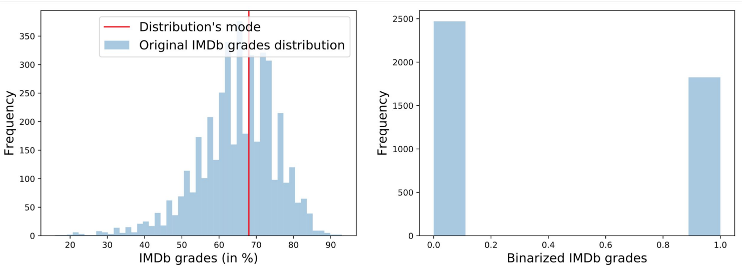
Awards: clipped and normalized



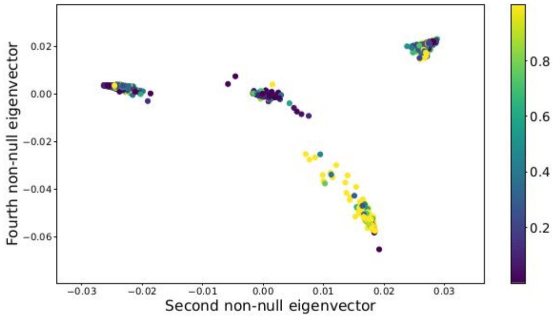
Critics’ grades: thresholded at 60%



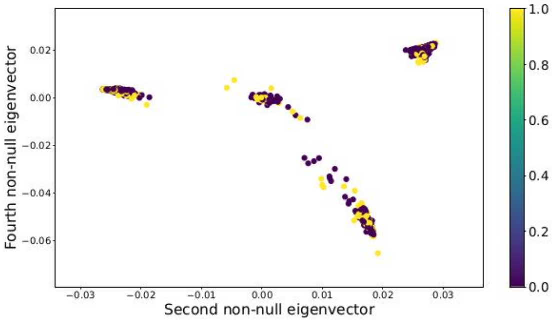
Users’ grades (IMDb): thresholded at 68%



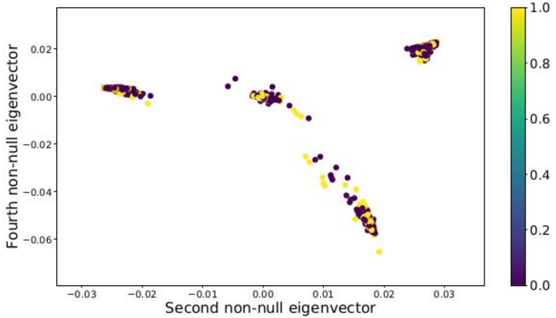
Labels as graph signals



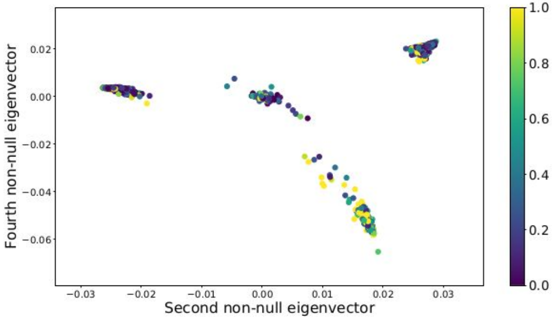
(a) Box office normalized and saturated to the 3rd quartile as signal



(b) Binarized critics grade as signal



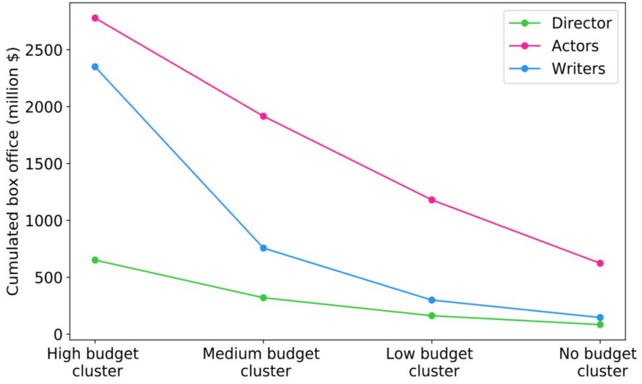
(c) Binarized IMDb grade as signal



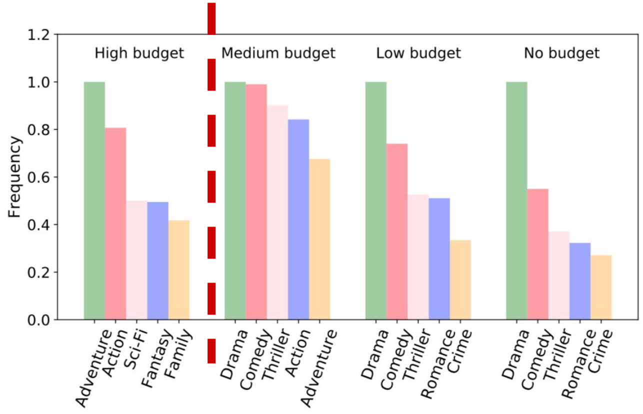
(d) Awards normalized and saturated to the 3rd quartile as signal

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

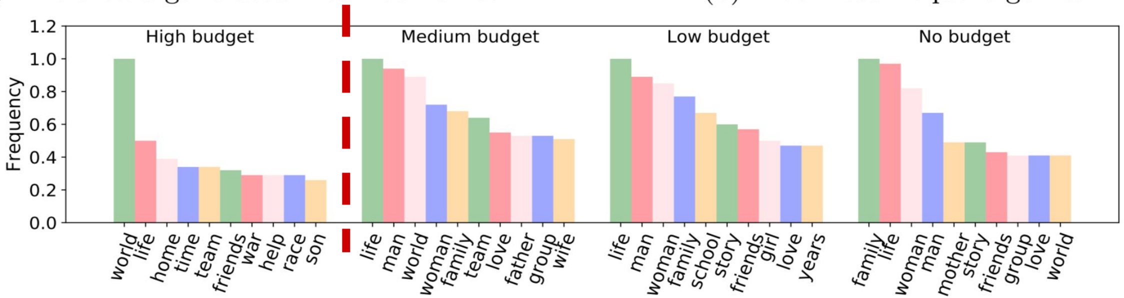
Feature analysis



(a) The crew's generated mean box office



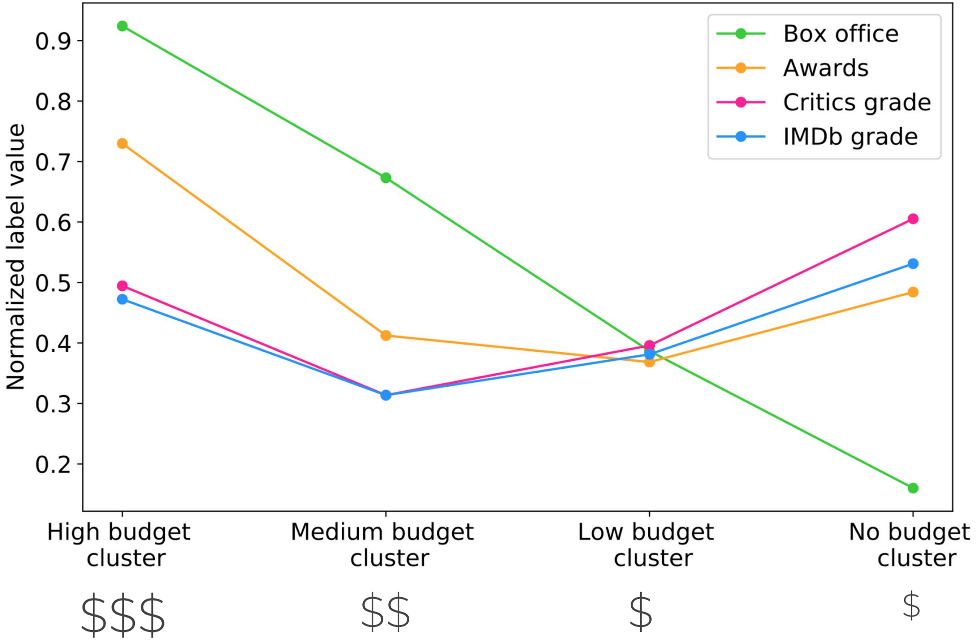
(b) Five most frequent genres



(c) Ten most frequent plot words

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

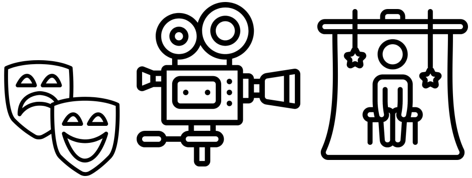
Label analysis



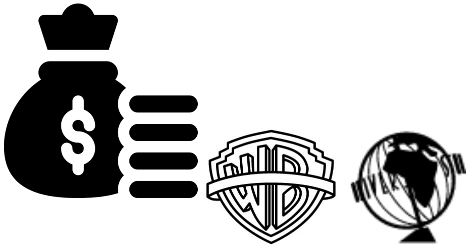
- Academical, critical and fan popularity has a U-shape trend
- Financial success is proportional to the film's budget

Conclusion

- One way for high box office → high budget
- Two ways for user popularity:



Independent route



Blockbuster way

- Future work:

  → Bigger dataset (2000 → 40’000 movies)

References

[1] Omdbapi.com. (n.d.). OMDb API - The Open Movie Database. [online] Available at: http://www.omdbapi.com.

[2] Dane, S. (2018). TMDB 5000 Movie Dataset. [online] Kaggle.com. Available at: https://www.kaggle.com/tmdb/tmdb-movie-metadata [Accessed 17 Jan. 2019].

[3] Computer vision for dummies. (2019). The Curse of Dimensionality in Classification. [online] Available at: http://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/.

[4] wikipedia.org. (2019). Silhouette (clustering). [online] Available at: https://en.wikipedia.org/wiki/Silhouette_(clustering)

[5] Belkin, M. and Niyogi, P. (2003). Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. Neural Computation, 15(6), pp.1373-1396.