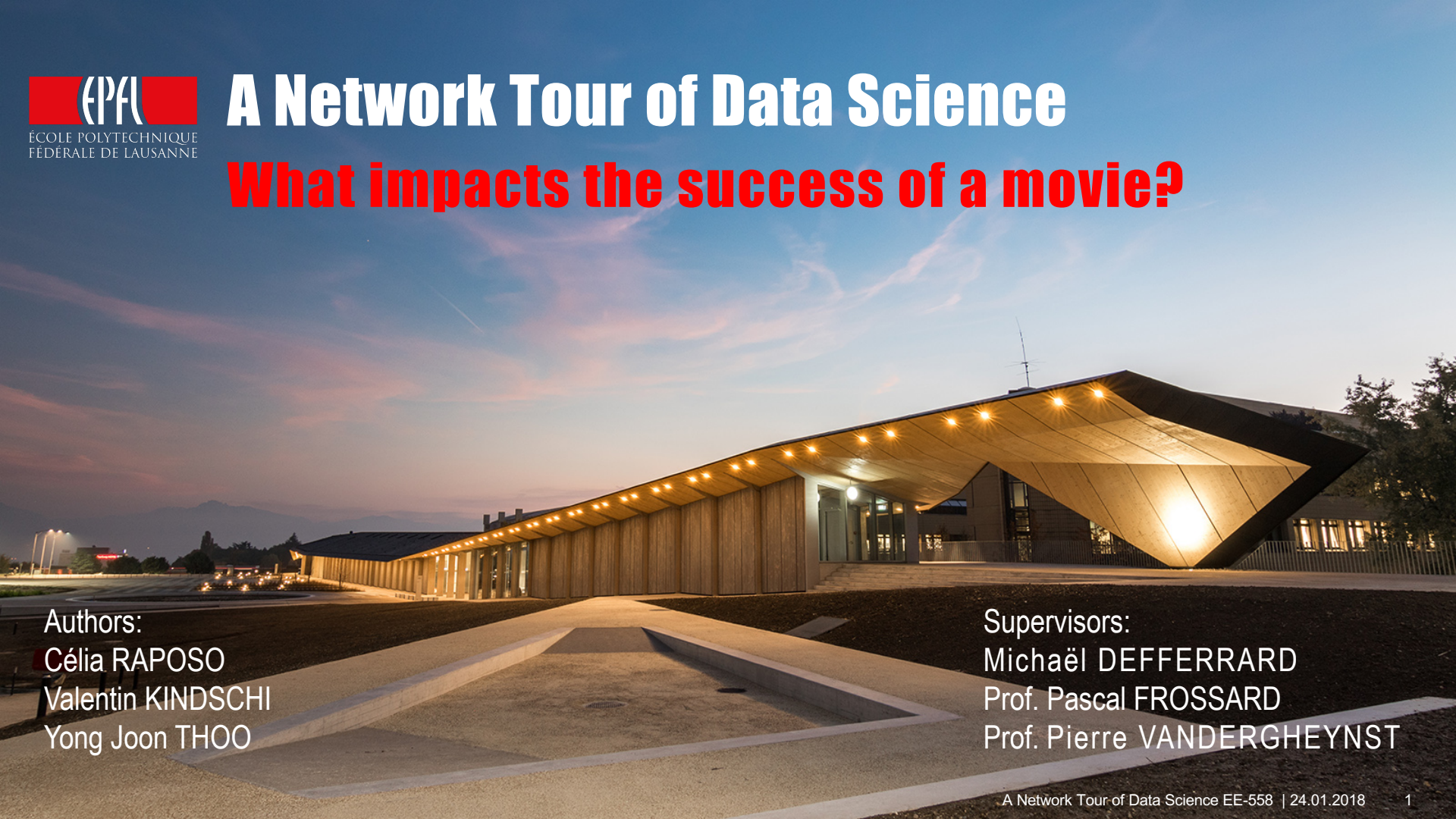


ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

A Network Tour of Data Science

What impacts the success of a movie?

Authors:
Célia RAPOSO
Valentin KINDSCHI
Yong Joon THOO

Supervisors:
Michaël DEFFERRARD
Prof. Pascal FROSSARD
Prof. Pierre VANDERGHEYNST

Overview

- Introduction
- Data acquisition
- Definition of Success
- Presentation of Features:
  - Normalization
  - Genres, Actors, Storyline, Popularity…
- Contribution of all features
- Results and Discussion
- Personnal Thoughts

Introduction

- Which factors impact the success of a movie?
- Is it possible to predict the success of a new movie?

- **Goal:**
  - Identify which features/attributes have the most impact on the success of a movie
  - Potentially evaluate the success of an upcoming film based on its similarity with past movies

Data acquisition

- Two datasets containing information of 350'000 movies:
  - Genre, actors, budget, etc: aspects of the movie

- Merged and cut down to ~2'620 movies:
  - Missing or NaN information
  - Movies from the 21st century
  - Selection of only «english» films

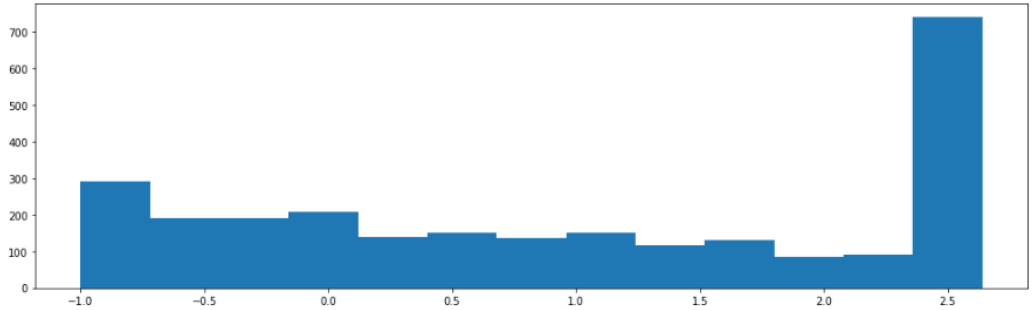
- Additional data collection: popularity of the movie

Definition of success

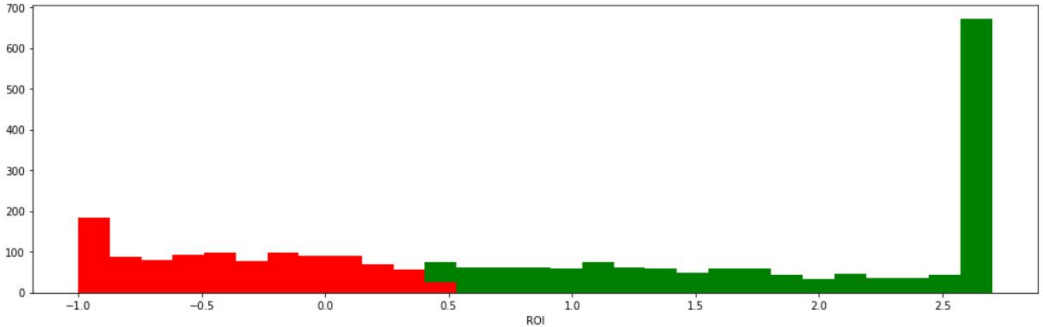
- How do we define the success?
  - ROI: Return on Investment

$$ROI = \frac{revenue\ - budget}{budget}$$

  - Labels: after saturation, we consider the lowest 40% as unsuccessful



Number of movies with ROIs between -1 and 2.6



Success labels

Note: Correction of the report, where successful movies could have a negative ROI.

Features

Directors

Production companies

Press Reviews

Actors

Trailer views

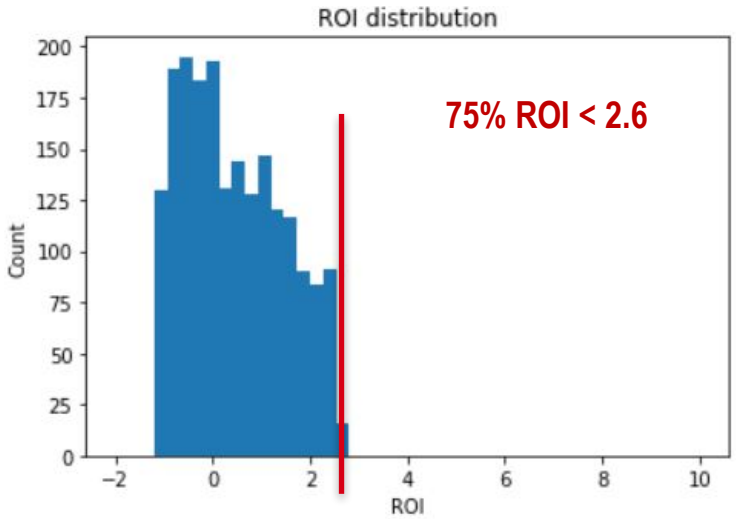
Genres

Storyline

Budget

Data Normalization

Example: ROI



Motivation:

In most of our data, we have a lot of "outliers" with values much higher than the mean

- **Saturation**: all values saturated to the 3rd quartile
- **Normalization**: all values divided by the 3rd quartile

$$X_{norm} = \frac{X_{max=3^{rd}\ quartile}}{3^{rd}\ quartile}$$

Actors

Actor's profitability: How much revenue actors have generated in the dataset

Total Actors profitability in movie i ($prof_i$) : Sum of the profitability of the actors present in movie i

$$\Delta \text{prof}_{ij} = \frac{|\text{prof}_i - \text{prof}_j|}{\text{3rd quartile}}$$

$$W_{ij} = 1 - \Delta prof_{ij}$$

Actor's tenure: Time difference between the first movie an actor appeared in and last movie in which he appeared

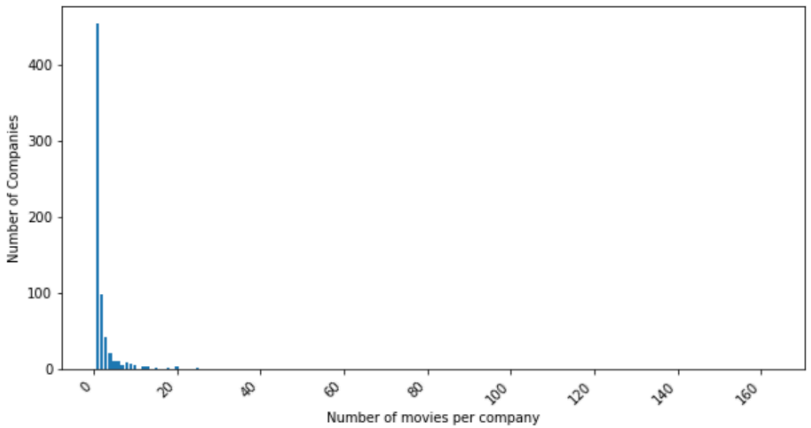
Total tenure of movie i (ten_i): Sum of the tenures of the actors present in movie i

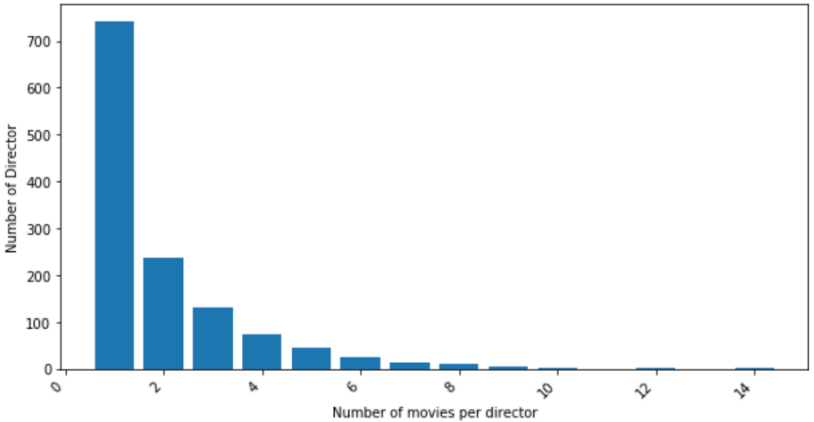
$$\Delta \text{ten}_{ij} = \frac{|\text{ten}_i - \text{ten}_j|}{3rd\ quartile}$$

$$W_{ij} = 1 - \Delta \text{ten}_{ij}$$

$$W_{ij} = \begin{cases} 0, \Delta x = 1 \\ 1 - \Delta x, \text{otherwise} \end{cases}$$

Production companies / Directors





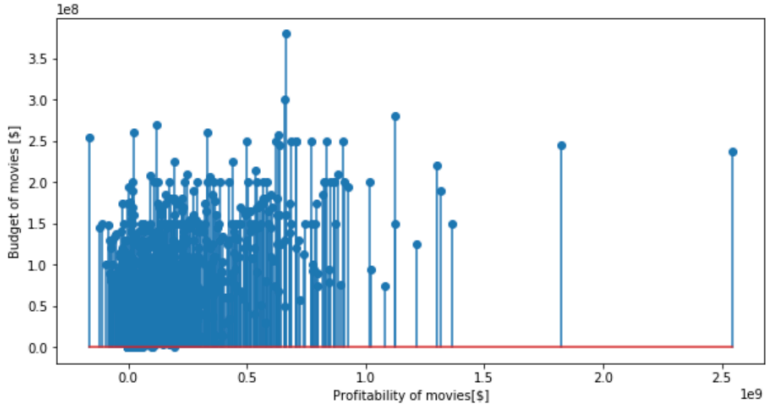
$$\Delta \text{prod\_}mov_{ij} = \frac{\left|\text{prod\_mov}_i - \text{prod\_mov}_j\right|}{3\text{rd } quartile}$$

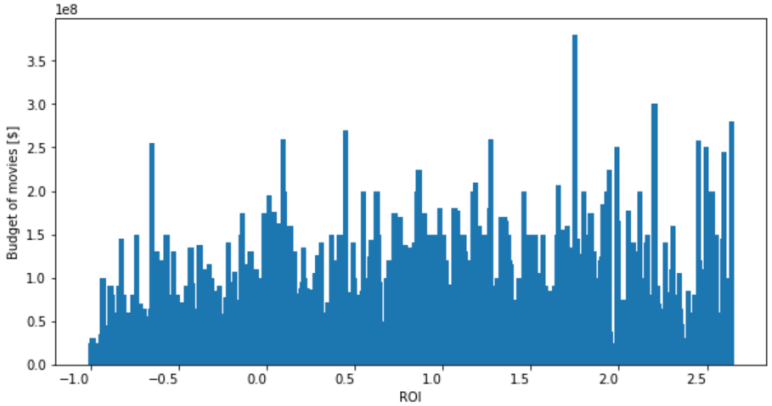
$$\Delta \text{dir\_}mov_{ij} = \frac{\left|\text{dir\_mov}_i - \text{dir\_mov}_j\right|}{3\text{rd } quartile}$$

Budget

- No correlation between Budget and Profitability
- No correlation between Budget and ROI

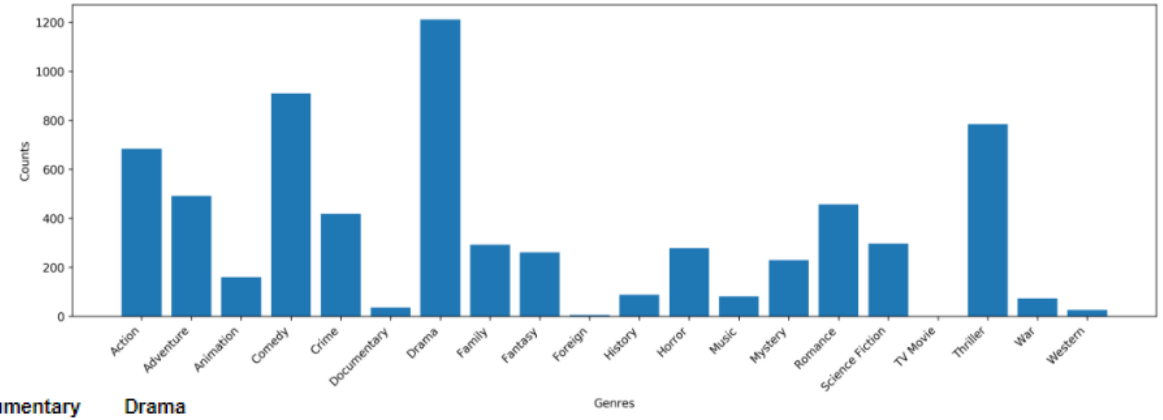
$$\Delta \text{budg}_{ij} = \frac{\left|\text{budg}_i - \text{budg}_j\right|}{\max(\text{budg})}$$





Genre

- 20 different kinds of distinct genres
  - ❑ ≥ 1 genre per movie
- Which genres are often associated with each other?

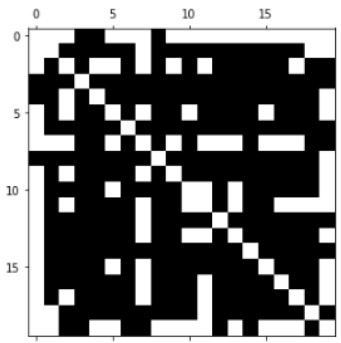


(Above) Number of times each genre appears in our dataset

	Action	Adventure	Animation	Comedy	Crime	Documentary	Drama
0	Thriller	Action	Family	Drama	Thriller	Music	Thriller
1	Adventure	Family	Comedy	Romance	Drama	Comedy	Romance
2	Crime	Comedy	Adventure	Family	Action	Family	Comedy

Top 3 associations for the first 7 genres

- Most films (75%) had 2 or less similar genres in common



(Right) Adjacency matrix of the genres

For a pair of films i and j:

$$G_{ij} = Nb\ genres\ in\ common\ for\ i\ and\ j$$

$$W_{ij} = \begin{cases} \frac{G_{ij}}{2} & if\ G_{ij} \leq 1 \\ 1 & if\ G_{ij} \geq 2 \end{cases}$$

Storyline analysis

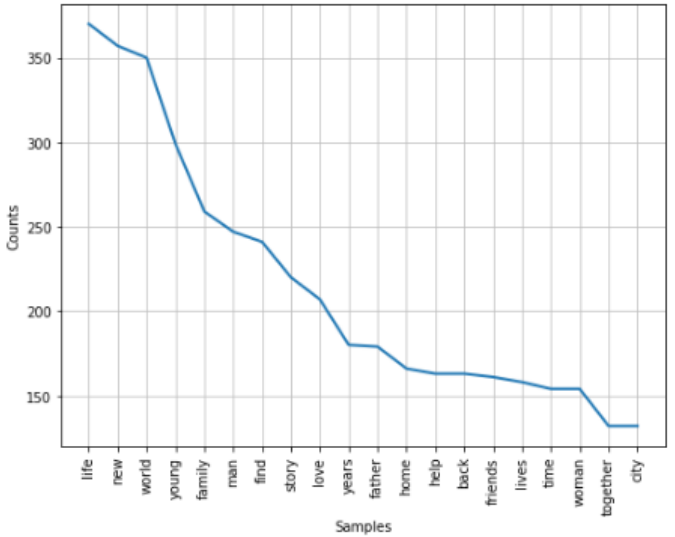
- Natural Language Toolkit (NLTK) [2]
- Tokenization of the plot summaries

```
This is a test that isn't so simple: 1.23.
"This" "is" "a" "test" "that" "is" "n't"
"so" "simple" ":" "1.23" "."
```

- Removal of stop words

Sample text with Stop Words	Without Stop Words
GeeksforGeeks – A Computer Science Portal for Geeks	GeeksforGeeks , Computer Science, Portal ,Geeks
Can listening be exhausting?	Listening, Exhausting
I like reading, so I read	Like, Reading, read

- Strange characters still remain: *'ii', ';', etc*
- Similar words:
  'europe' & *'european'*, *'inventor'* & *'creator'*

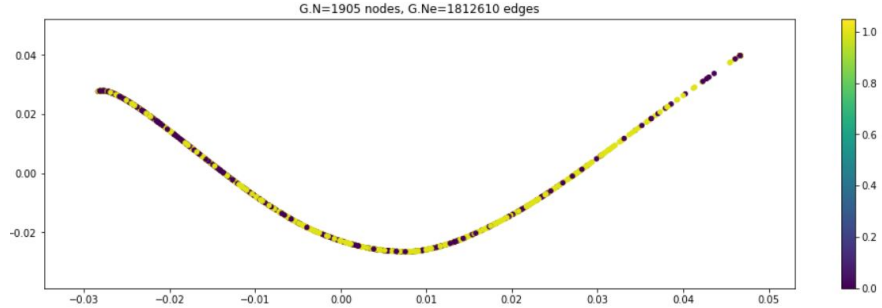


First 20 most common words in the plot summaries of our dataset and the number of times they appear

For a pair of films i and j

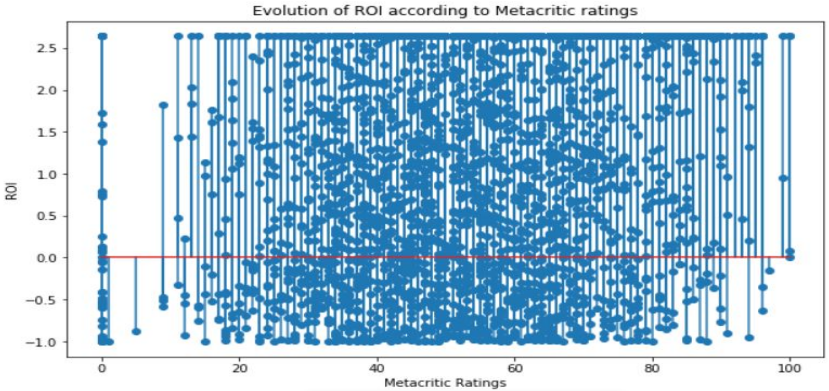
$$W_{ij} = \begin{cases} 0, & Nb\ of\ similar\ common\ words\ between\ i\ and\ j = 0 \\ 1, & Nb\ of\ similar\ common\ words\ between\ i\ and\ j \geq 1 \end{cases}$$

Popularity: Youtube & Metacritic



YouTube

- Collections of videos on 5 different channels
- Found only 75% of dataset's movies
- Express the movies' trend on Internet



metacritic

- Weighted sum of several professional critics
- Rating before a movie comes out
- Give a qualitative idea of the success

These features don't explain as well as expected the financial success of movies

Contribution of all features: Weight matrix

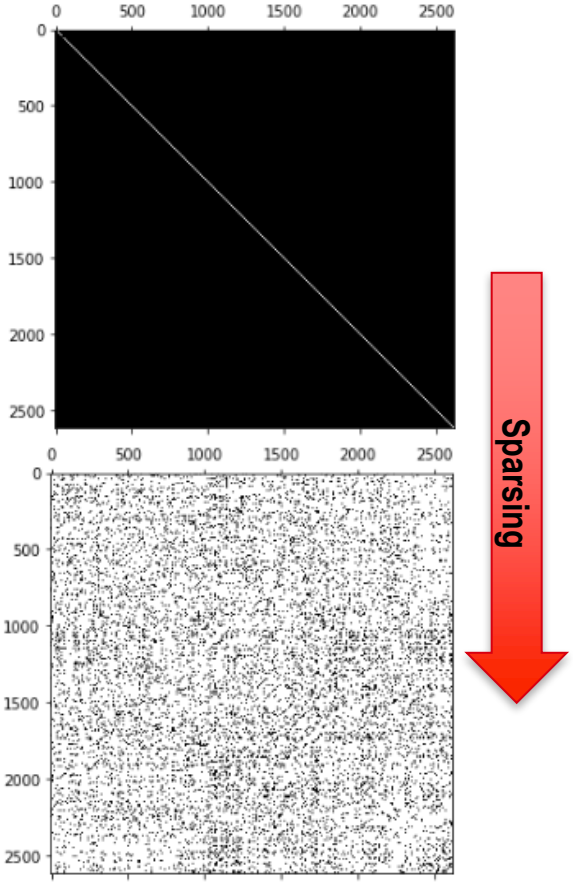
- No lone feature seems to impact the success (cf. report)

- Combination of the weight matrices:

$$W_{tot} = \frac{1}{N_{graphs}} \sum_{i=1}^{N_{graphs}} W_i$$

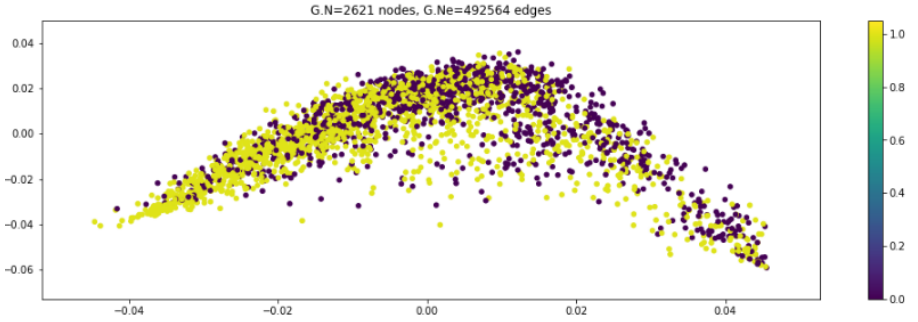
Where W_i is the weighted graph constructed for a certain feature and N_{graphs} is the number of retained graphs

- Sparsification of the matrix (Neighbors = 300)
  - ✓ Reduce computational costs
  - ✓ Give more importance to most similar pairs
  - X Loss of information

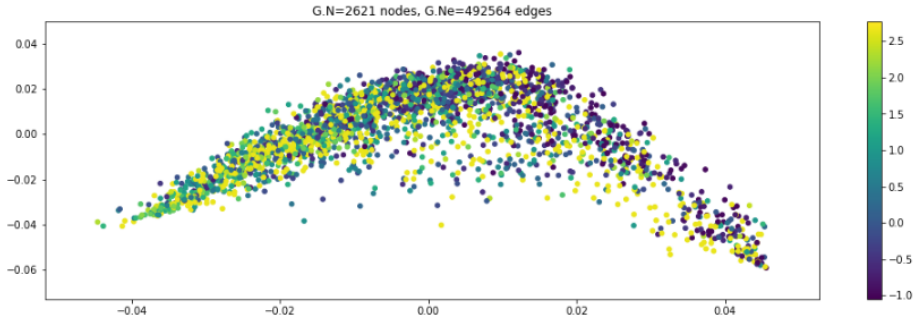


Contribution of all features: Graph embedding

- Computation of the graph Laplacian and its eigenvectors and eigenvalues
- Graph embedding in 2D with Laplacian eigenmaps (2nd and 3rd eigenvectors) with:
  - Successful and unsuccessful labels
  - The value of the ROI
- Tendency to have more successful movies for negative values along the 2nd eigenvector (x-axis).
  - 40-60% classification error with Fiedler vector



Graph embedding with labels: 1 = 'successful' and 0 = 'unsuccessful'



Graph embedding where the signals are the ROI

Results and Discussion

- No lone feature seems to impact the success

- Possible improvements:
  - Find dataset of successful and unsuccessful movies
  - Text analysis: filtering + synonyms and antonyms
  - Further exploration of data outside of dataset
  - Take in account all revenues of movies (not only Box office)

- Encountered problems
  - YouTube channels
  - Wrong/missing data on websites

- Other factors could impact the success:
  - Trends, economy, etc

Personal Thoughts

- Success couldn't be predicted by the data we extracted
- Too many features would need to be taken into account
- Complex but interesting subject

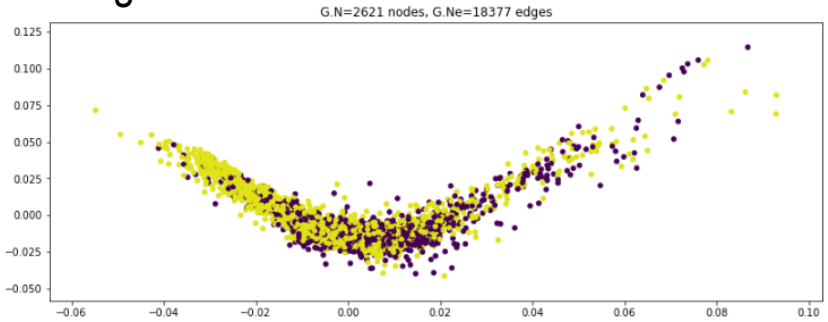
Thank you for your attention!

References

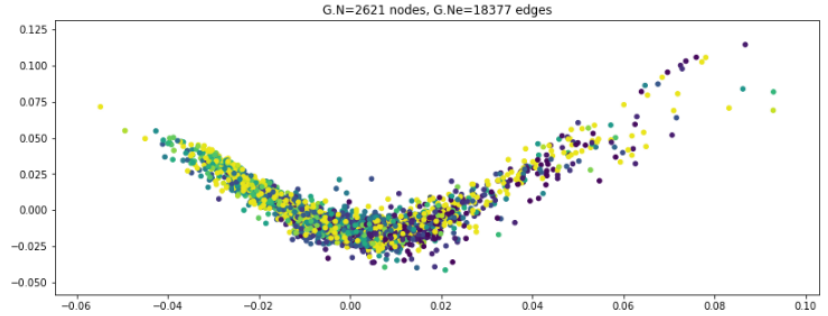
1. http://ew.com/movies/2017/09/04/summer-box-office-fail/
2. Bird, Steven, Edward Loper and Ewan Klein (2009), "Natural Language Processing with Python". O'Reilly Media Inc. (http://www.nltk.org/#natural-language-toolkit)

Appendix A: Graph embedding with other values for sparsification

Neighbors = 10

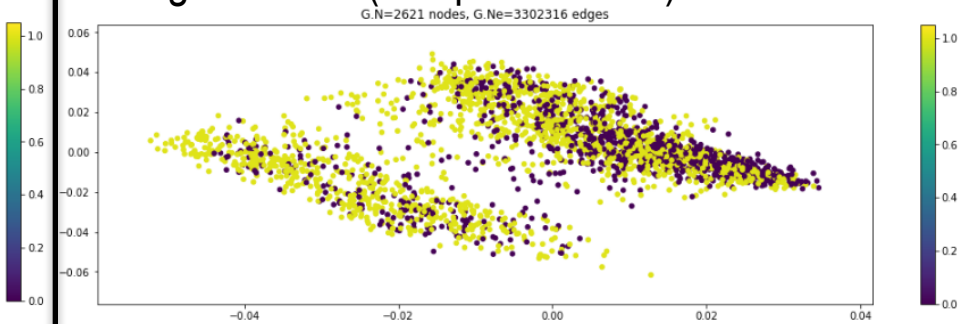


Graph embedding with labels: 1 = ‘successful’ and 0 = ‘unsuccessful’

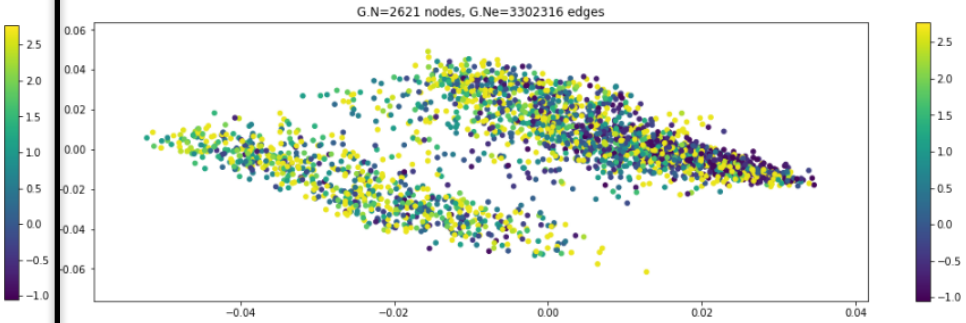


Graph embedding where the signals are the ROI

Neighbors = 0 (no sparsification)



Graph embedding with labels: 1 = ‘successful’ and 0 = ‘unsuccessful’



Graph embedding where the signals are the ROI