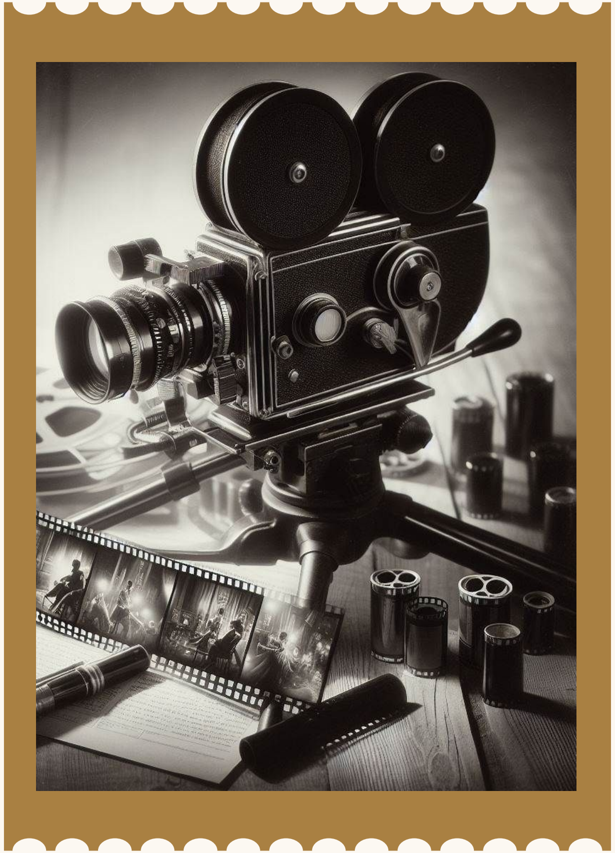
Analysis of Movie Reviews & Tags

Text mining project

Giulia Beccaria
Roberta Di Santo
Lorenzo Lobosco



01 Introduction

Overview

- Project goal: Develop an advanced model predicting film ratings using textual reviews.
- Technique: Apply clustering to group films based on similar reviews and tags.
- Enhancement: Improve recommendation experience by identifying film groups aligned with user preferences.

Dataset

- 100,000 Ratings
- 3,600 Tag Applications
- 9,000 Movies
- 600 Users

02 GOALS & STRATEGY

PRE PROCESSING

- Cleaning dataset
- Tokenization
- Normalization
- Stopwords
- Lemmatization

TEXT REPRESENTATION

- Tf idf
- Word2vec
- T-sne

TEXT CLASSIFICATION

- Negative and Positive
- Rating between 1 and 5

TEXT CLUSTERING

- K-means
- Hierarchical
- DBSCAN
- Agglomerative

03 PRE PROCESSING

Libraries used	NLTK, re, emoji, Text Blob
Tokenization	Breaking a character sequence into individual word tokens
Normalization	Aligning text and query terms to a consistent form
Stop-words removal	Optional exclusion of highly common words from the analysis (may be included or excluded)
Stemming	Matching different forms of a word to its root for improved consistency
Lemmatization	Achieving a "proper" reduction to the dictionary headword form for more accurate analysis

04 TEXT REPRESENTATION

GOAL

Converting textual data into a machine - readable format, in this context we use a **vectorized form**

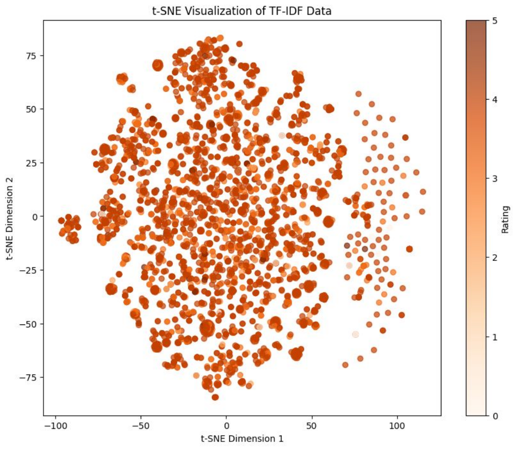
APPROACH

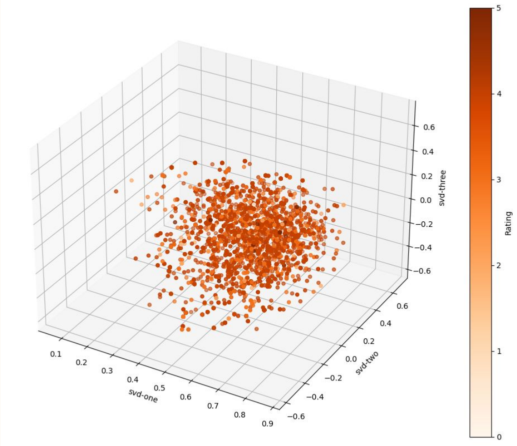
- **Word2Vec** → unsatisfactory for our model
- **TF-IDF** → assigning weights to words based on their importance in a document relative to a collection

RESULTS

We employed **t-SNE** for representation
(2 and 3 dimensions)

The majority of values cluster around 4 and 5.





05 TEXT CLASSIFICATION

GOAL

Define a classification model of the ratings due to the tags

APPROACH

- Negative ratings < 3, Positive ratings ≥ 3
- Round the ratings and categorize in 6 classes: [0,1,2,3,4,5]

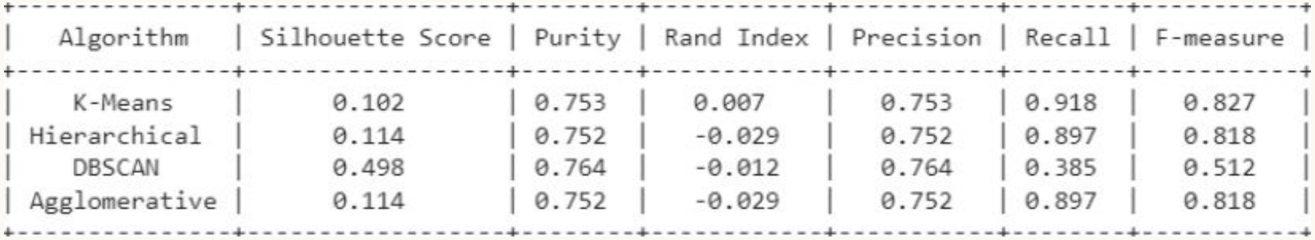
The others models exhibit higher accuracy, the Ordinal Ridge approach proves more valuable in achieving precise categorization within the 6-class rating system.

RESULTS

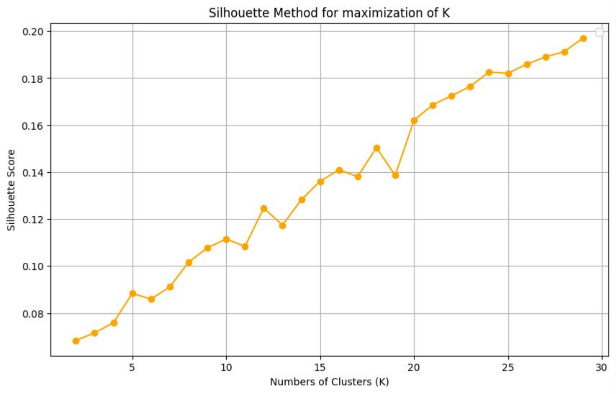
Model type	Accuracy
Logistic Regression	0.805
Random Forest	0.807
Gradient Boosting	0.87
Ordinal Ridge	0.719

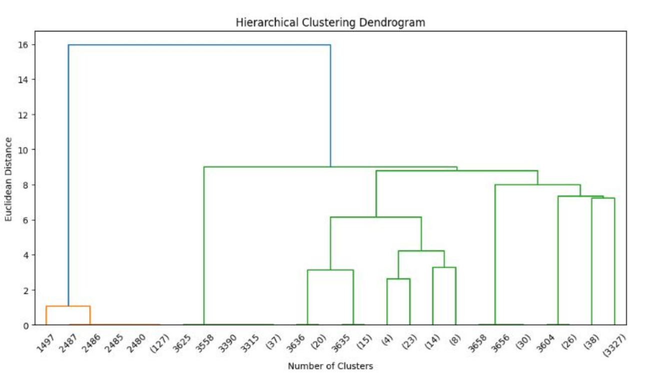
- K-means
- Hierarchical
- DBSCAN
- Agglomerative

06 TEXT CLUSTERING



```
+---------------+-------------------+---------+------------+------------+---------+------------+
|   Algorithm   | Silhouette Score  | Purity  | Rand Index | Precision  | Recall  | F-measure  |
+---------------+-------------------+---------+------------+------------+---------+------------+
|    K-Means    |       0.102       | 0.753   |   0.007    |   0.753    | 0.918   |   0.827    |
| Hierarchical  |       0.114       | 0.752   |   -0.029   |   0.752    | 0.897   |   0.818    |
|    DBSCAN     |       0.498       | 0.764   |   -0.012   |   0.764    | 0.385   |   0.512    |
| Agglomerative |       0.114       | 0.752   |   -0.029   |   0.752    | 0.897   |   0.818    |
+---------------+-------------------+---------+------------+------------+---------+------------+
```





Most representative tag for each cluster

Cluster 0	Awesom
Cluster 1	Atmoshper
Cluster 2	Psycholog
Cluster 3	Alien
Cluster 4	Comedi
Cluster 5	Netflix
Cluster 6	Scifi
Cluster 7	School

07 CONCLUSION

POTENTIAL FUTURE IDEAS

1. **Incorporating more demographic information**: The current dataset used in the project did not include specific demographic information about the users..

2. Exploring **additional text representation** techniques.

3. Incorporating **additional data sources:** The current dataset used in the project included ratings and tags applied to movies by users. However, additional data sources such as movie reviews or social media posts could be incorporated to provide a more comprehensive understanding of user sentiments towards films.

4. Developing a **more sophisticated predictive model**: While the project developed a predictive model for film ratings, there is potential to develop a more sophisticated model that incorporates additional features and techniques. For example, deep learning techniques such as recurrent neural networks (RNNs) could be used to capture temporal dependencies within the text data.