

Mario Alessandro Napoli Pietro Valenti Paolo Lindia

## Obiettivi del progetto

Applicazione degli algoritmi di Text Mining ai copioni dei film

#### In particolare:

- Classificazione dei film in base al genere
- 2. Clustering e visualizzazione spaziale dei titoli
- 3. Analisi dei topic principali contenuti nel corpus

## Table of Contents

1

#### Dataset

- Origine dei dati
- Preprocessing
- Representation

2

#### Classificazione

- Modelli impiegati
- Risultati

3

## Clustering

- Modelli impiegati
- Risultati
- Visualizzazioni

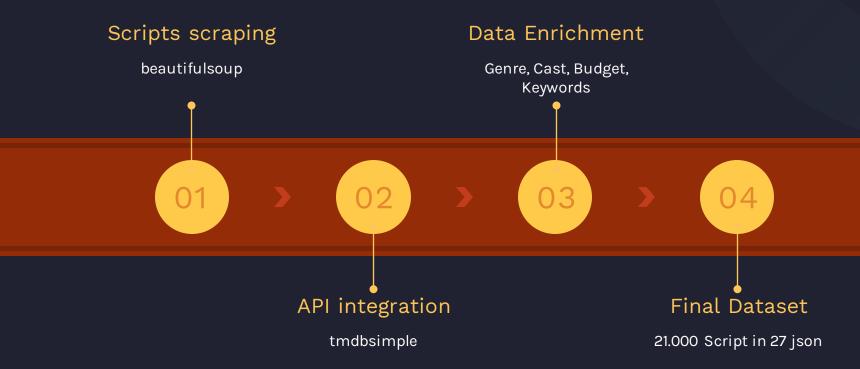
4

## **Topic Modelling**

- Modello
- Topic identificati

Dataset

# Origine dei dati



# Preprocessing

- Rimozione numeri
- Rimozione punteggiatura
- Conversione in minuscolo
- Rimozione testo all'interno delle parentesi
- Tokenizzazione
- Lemmatizzazione
- Rimozione Stop Words
- UnderSampling

# Insights sul corpus:

Numero di documenti Termini totali

Lunghezza media documenti

5735

240400

#### Truncated SVD

2 versioni, basate su DTM con TF e con TFIDF

# Data Representations

#### Doc2Vec

Gensim Doc2Vec modello pre-trained

#### GloVe

Conversione Gensim di un modello GloVe pre-trained

# Classification

### **Metodo**

Split in dati di training e test (67%, 33%)

Utilizzo dell'algoritmo KNN per valutare e selezionare le features: TF, TF-IDF, SVD su TF-IDF, Doc2Vec, GloVe. GloVe risulta il migliore.

Applicazione e confronto tra modelli di classificazione: KNN, RandomForest, SVM (kernel rbf), ANN, CNN. Valutando performance e onerosità computazionale SVM risulta il migliore.

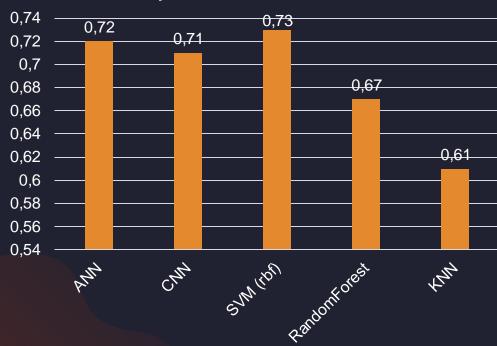
Utilizzo di Cross-Validation per l'ottimizzazione dei parametri della SVM

Valutazione della performance finale (Accuracy, Recall, Precision, F1, ...)

## Scores







### Metriche di performance per SVM

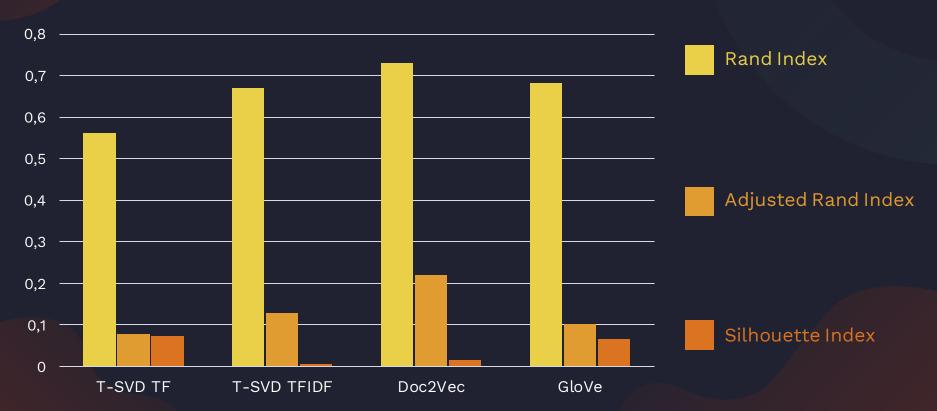
	Precision	Recall	F1	Support
Action	0.70	0.75	0.72	420
Comedy	0.64	0.70	0.67	382
Docum.	0.89	0.88	0.88	382
Drama	0.63	0.53	0.58	380
Horror	0.73	0.76	0.68	384

Accuracy

# Clustering

## Scores





# Topic Modelling

### LDA model



E' un modello di analisi del linguaggio naturale che permette di comprendere il significato semantico del testo analizzando la somiglianza tra la distribuzione dei termini del documento con quella di un argomento specifico (topic) o di un'entità.

Documento Considerato come un insieme di topic

Ognuno dei quali è caratterizzato da una particolare distribuzione di termini e parole Il modello LDA
analizza il contenuto
delle frasi in ogni
documento e assegna
a ciascuna parola la
sua appartenenza ai
topic



# Results



Topic 0		Topic 1	Topic 2	Topic 3	Topic 4
neonle	ir one ent	come wayback one sir take sift right move let	hey right come gonna aint gotta	game 1ts trans to booby 1111 lve	okay yeah youre thats knowhey like
	tell let like	Topic 6 see know yes would mustfather onelike well come	fucking shit man rightyeah fuck goons sure	Topic 8 iike iii seopie teil please only weil aii sould wiii	Topic 9 well right know want thats like one youre see ill

Word cloud of each topic

Word cloud of all token



```
Tight dad Way Job Man body house place shirt some shirt some shirt shirt some shirt way mean house like life day be thought love isnt may mean hell even the come come come remember still made even think people one remember still in the say think people one prothing girls show and problem kind came but wife said first told talking actually brother follows father, are work year year also understand day yes to the father, are work year year also.
```

