Textual Cinema

A journey in mining movie script

Mario Alessandro Napoli
Pietro Valenti
Paolo Lindia

Obiettivi del progetto

Applicazione degli algoritmi di Text Mining ai copioni dei film

In particolare:

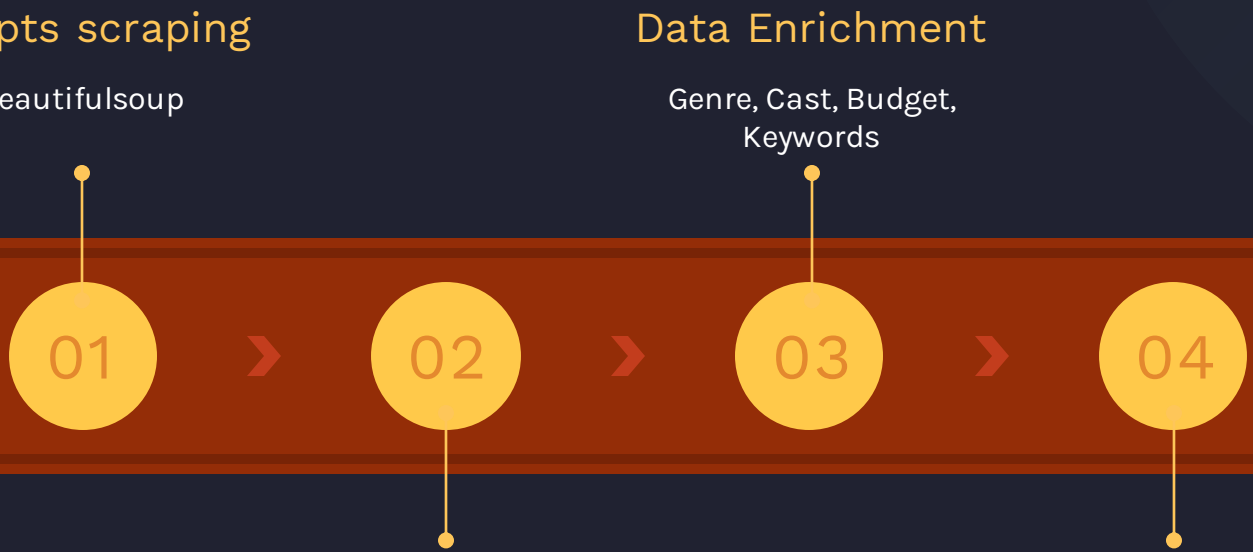
1. Classificazione dei film in base al genere
2. Clustering e visualizzazione spaziale dei titoli
3. Analisi dei topic principali contenuti nel corpus

Table of Contents

Dataset

1

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

5735

Termini totali

240400

Lunghezza media documenti

3570

Data Representations

Truncated SVD

2 versioni, basate su DTM con TF e con TFIDF

Doc2Vec

Gensim Doc2Vec modello pre-trained

GloVe

Conversione Gensim di un modello GloVe pre-trained

Classification

2

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.

↓

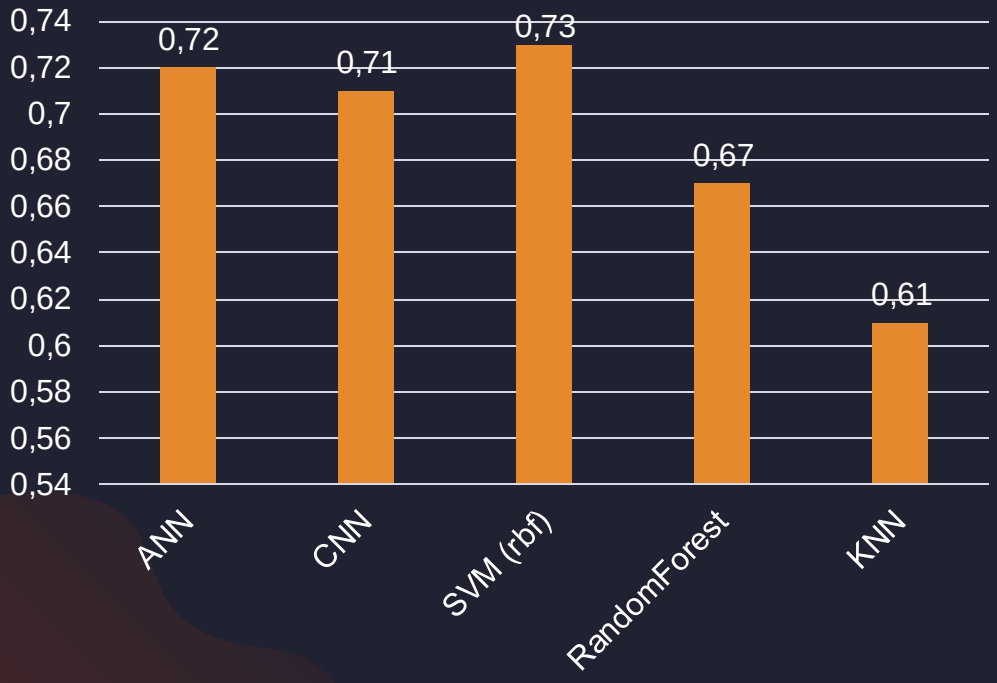
Utilizzo di Cross-Validation per l'ottimizzazione dei parametri della SVM

↓

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

Scores

Accuracy modelli di classificazione



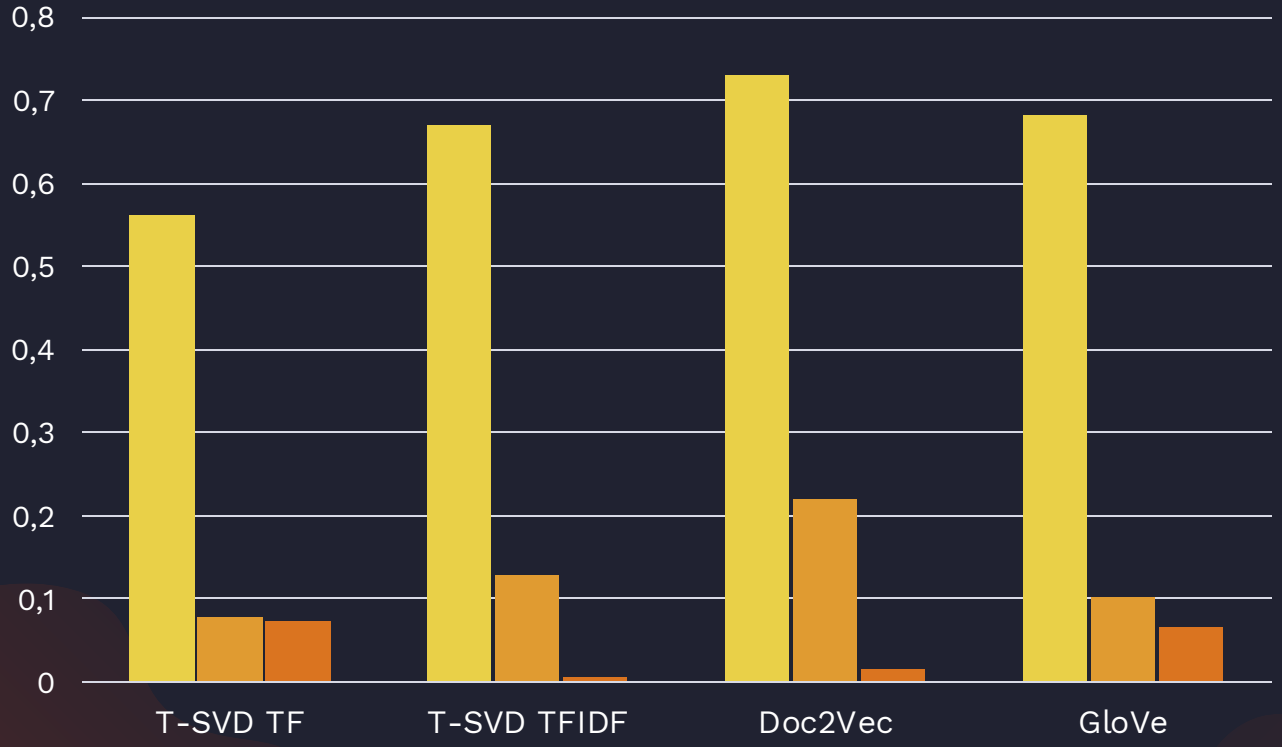
Metriche di performance per SVM

	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

Clustering

3

Scores


0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
T-SVD TF
T-SVD TFIDF
Doc2Vec
GloVe
Rand Index
Adjusted Rand Index
Silhouette Index

Topic Modelling

4

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

Word cloud of each topic

Word cloud of all token

GRAZIE PER L’ATTENZIONE