Mining IMDb reviews

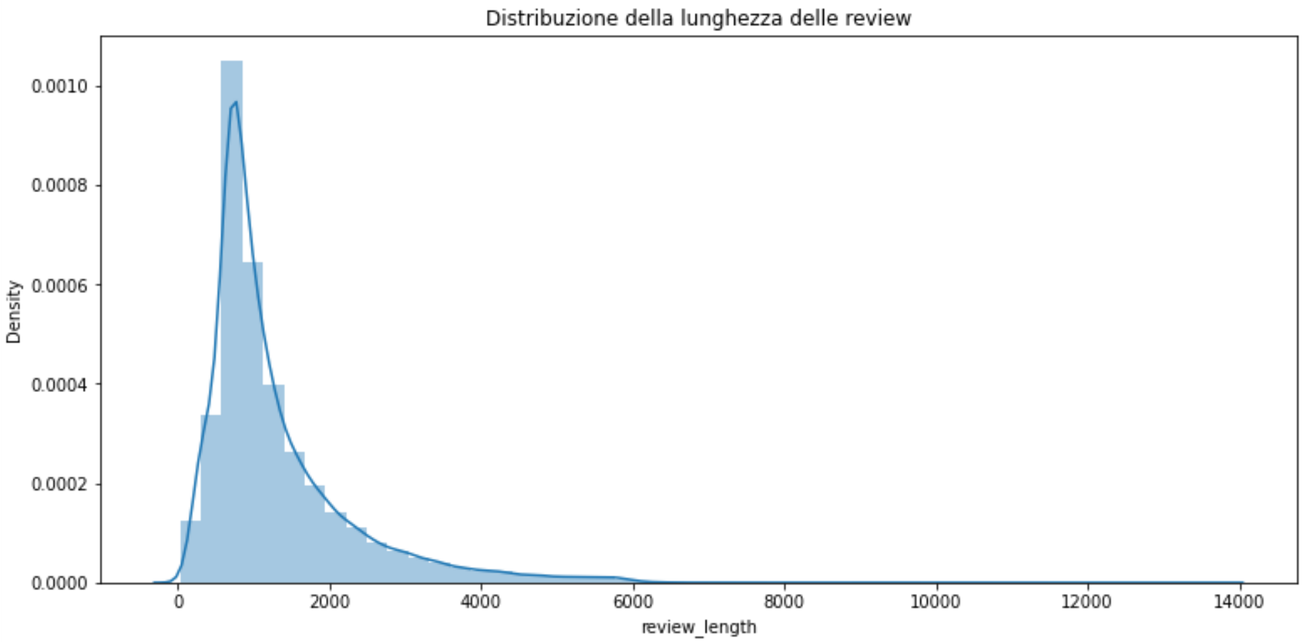
Casula Filippo Maria
Mura Giulia
Pisani Caterina

Dataset

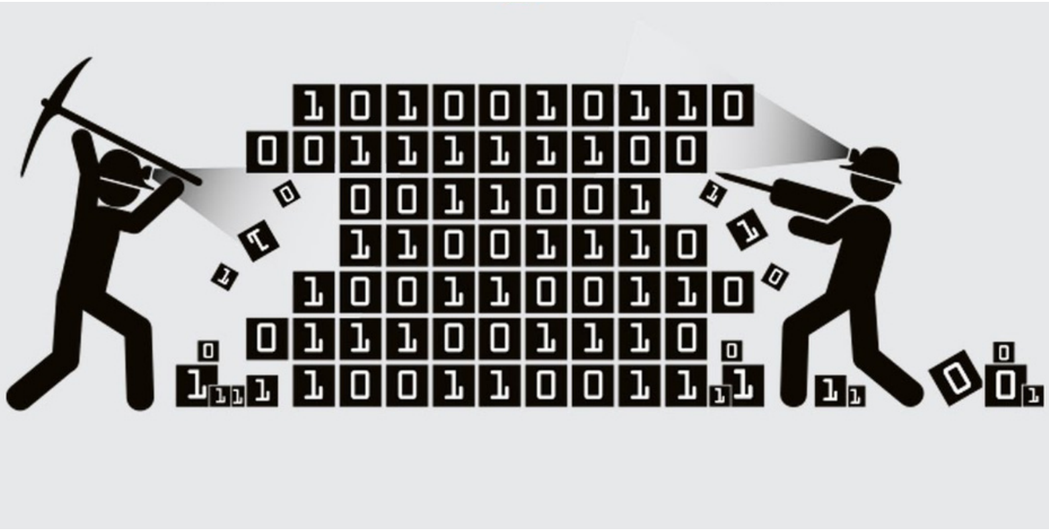
50000 recensioni di film

25000 positive: voto tra 7 e 10
250000 negative: voto tra 1 e 5

DISTRIBUZIONE LUNGHEZZA DEL TESTO



DISCUSSION OUTLINE



Tasks

Preprocessing

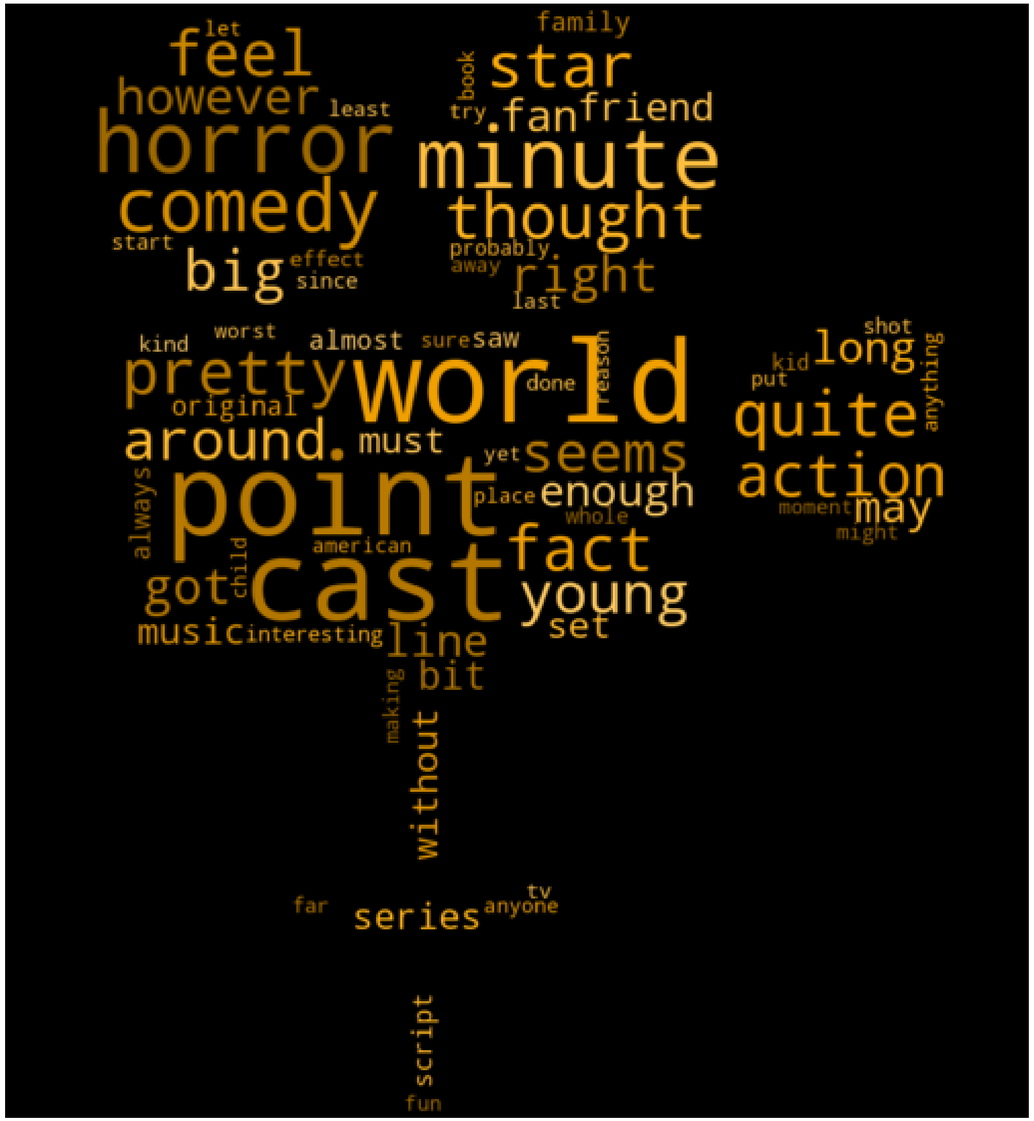
Rappresentazione del testo

Classificazione

Sentiment Analysis

Clustering

Topic modeling



Preprocessing

- **Normalizzazione**
  - Espansione forme contratte
  - Lower case
  - Rimozione \n, \t e
  - Rimozione urls
  - Rimozione punteggiatura
  - Rimozione numeri
  - Rimozione emoji
  - Rimozioni whitespace
- **Rimozione stop words**
- **Lemmatizzazione e tokenizzazione**

MINING IMDB REVIEWS

Text Representation

1. Bag of words
2. Binary bag of words
3. Tf - Idf

SINGULAR VALUE DECOMPOSITION

TASK 1

Text Classification & Sentiment Analysis

TEXT CLASSIFICATION

- Random Forest
- Logistic Regression
- KNN
- GaussianNB
- Recurrent Neural Network

SENTIMENT ANALYSIS

- Vader
- Text Blob
- Afinn

Text Classification

Tf-Idf

E' stata la tecnica di text representation più performante sui i seguenti modelli

Random Forest
Logistic Regression
KNN
GaussianNB

Recurrent Neural Network

Input layer: Embedding con dimensione vocabolario: 30000, embedding: 100 dimensioni, lunghezza sequenza: 250

LSTM Layer: 100 neuroni

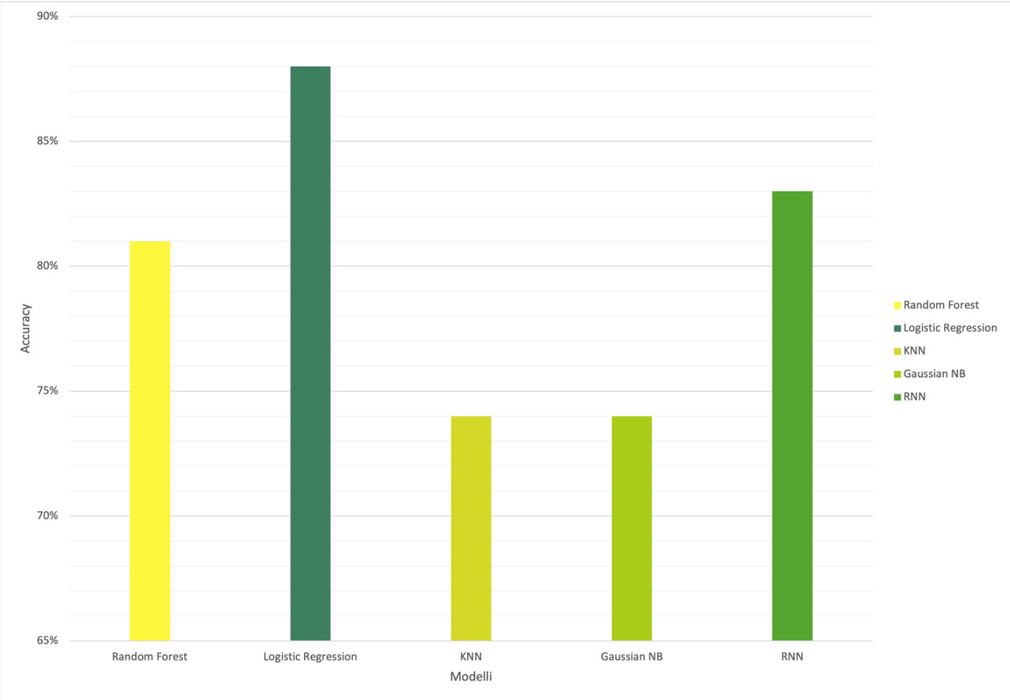
Output Layer: Dense layer 2 neuroni, attivazione: sigmoid

Compilazione : adam e loss binary crossentropy

Epoche: 3

Batch Size: 2048

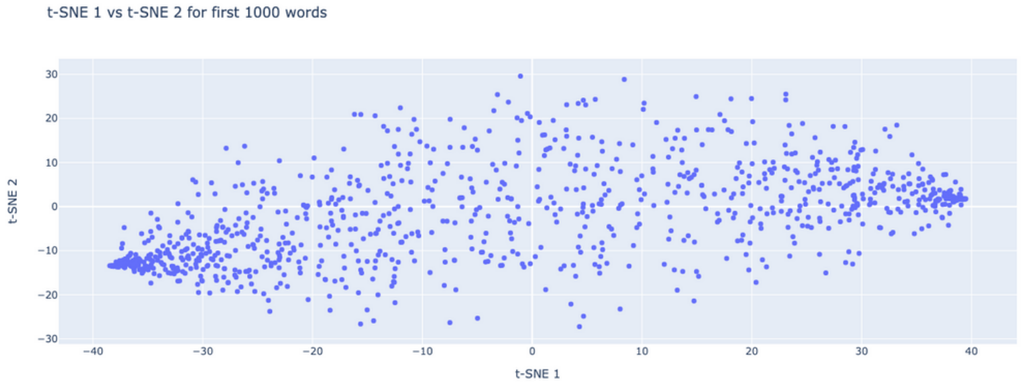
Text Classification Results



Come si può notare dal grafico, i modelli che hanno avuto un'accuratezza più elevata sono stati:

- Logistic Regression : 88%
- Recurrent Neural Network : 83%

Nel grafico sottostante è presente una visualizzazione dei pesi ottenuti dal word embedding



Sentiment Analysis

Vader

Mapping delle caratteristiche lessicali ed emozioni
Analisi della frase
Polarità

Text Blob

Analisi della frase
Polarità

Afinn

Analisi della parola
Sentiment Score

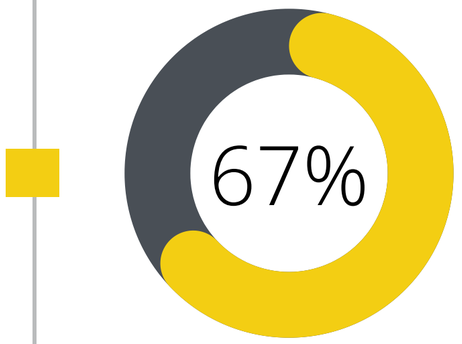
Step Principali

Applicazione della sentiment
Normalizzazione dei risultati
Label
Calcolo accuracy

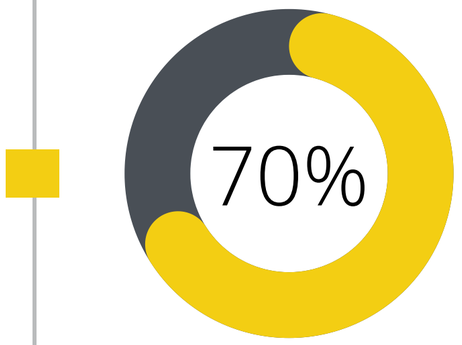
Results

Sentiment Analysis

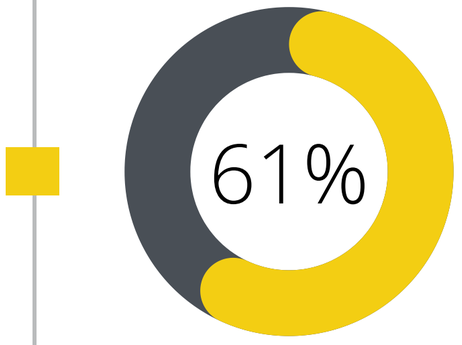
	Vader	TextBlob	Afinn
False	16706	15094	19938
True	33294	34906	30062



Vader



TextBlob



Afinn

TASK 2

Text Clustering

OVERVIEW

TEXT REPRESENTATION

- Bag of Words
- Binary
- Tf-Idf

K-MEANS

- K = 2
- Metrica di valutazione: Mutual Information

GRAFICO CLUSTER E WORD CLOUD

RISULTATI

Mutual Information Normalizzato risulta essere in ogni caso basso, ma la migliore rappresentazione risulta il Tf-Idf

	Bag of Words	Binary	Tf-Idf
Normalized Mutual Information	9.56e-05	0.00016	0.01734

Tabella 1. Punteggio della metrica normalizzata Mutual Information per ogni rappresentazione

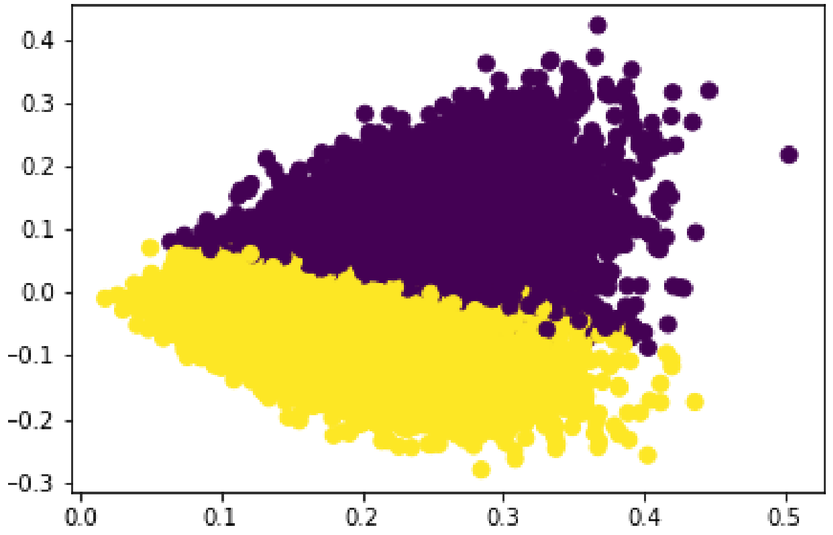
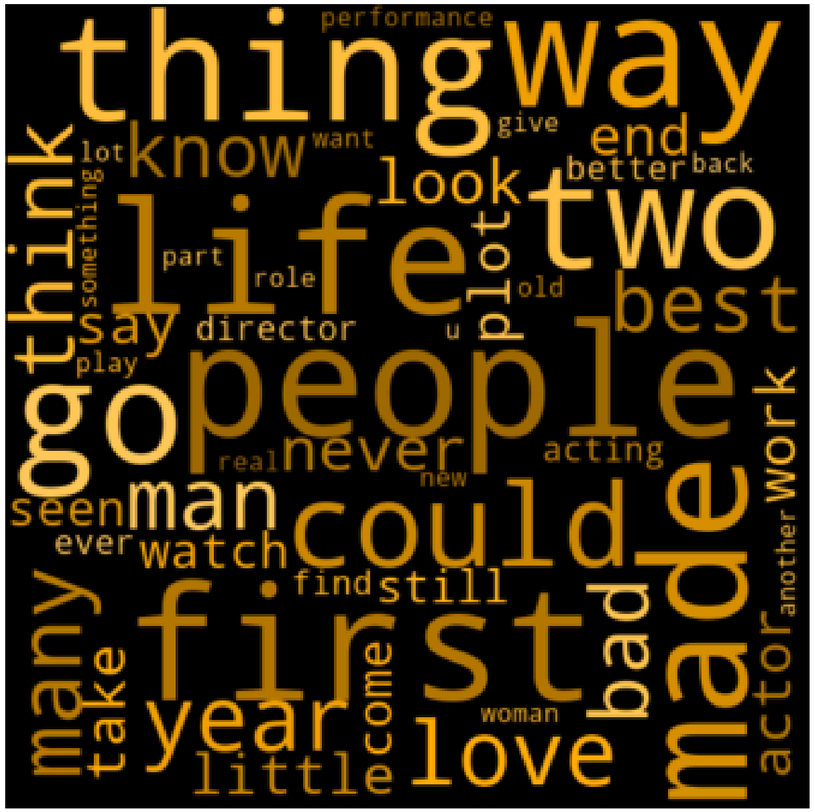


Grafico 1. Rappresentazione cluster ottenuti da Tf-Idf con K = 2

CLUSTER 1



Rappresentazione 1. Word Cloud del cluster 1 rappresentante 50 top words

CLUSTER 2

Rappresentazione 2. Word Cloud del cluster 2 rappresentante 50 top words

TASK 3

Topic Modeling

OVERVIEW

LSA

- Tf-Idf
- Dimensionality Reduction SVD
- 5 topic

LDA

- Gensim
- 5 topic
- Metrica di valutazione: Coherence Measure

LDA

Coherence measure: 0.33

Topic 0	0.007 * ‘great’ + 0.006 * ‘best’ + 0.006 * ‘role’ + 0.005 * ‘play’ + 0.005 * ‘also’ + 0.005 * ‘year’ + 0.005 * ‘comedy’ + 0.005 * ‘performance’ + 0.005 * ‘cast’ + 0.004 * ‘well’
Topic 1	0.008 * ‘story’ + 0.006 * ‘life’ + 0.005 * ‘world’ + 0.005 * ‘time’ + 0.005 * ‘character’ + 0.005 * ‘war’ + 0.004 * ‘u’ + 0.004 * ‘people’ + 0.004 * ‘many’ + 0.004 * ‘well’
Topic 2	0.007 * ‘woman’ + 0.007 * ‘man’ + 0.006 * ‘character’ + 0.005 * ‘life’ + 0.004 * ‘performance’ + 0.004 * ‘scene’ + 0.004 * ‘wife’ + 0.004 * ‘young’ + 0.004 * ‘story’ + 0.004 * ‘two’
Topic 3	0.014 * ‘like’ + 0.011 * ‘good’ + 0.010 * ‘really’ + 0.009 * ‘time’ + 0.008 * ‘would’ + 0.008 * ‘see’ + 0.008 * ‘even’ + 0.008 * ‘character’ + 0.007 * ‘make’ + 0.007 * ‘bad’
Topic 4	0.010 * ‘horror’ + 0.006 * ‘scene’ + 0.005 * ‘get’ + 0.005 * ‘like’ + 0.005 * ‘effect’ + 0.004 * ‘action’ + 0.004 * ‘look’ + 0.004 * ‘bad’ + 0.003 * ‘even’ + 0.003 * ‘kill’

Tabella 2. Cinque topic rappresentati ognuno da 10 top words e le rispettive percentuali.

CONCLUSIONI E SVILUPPI FUTURI

- La classificazione "pura" risulta essere l'analisi migliore per il dataset in questione

- Migliorie nei parametri e/o utilizzo di un dataset più ampio per cercare di ottimizzare la performance dei modelli

GRAZIE PER L'ATTENZIONE!