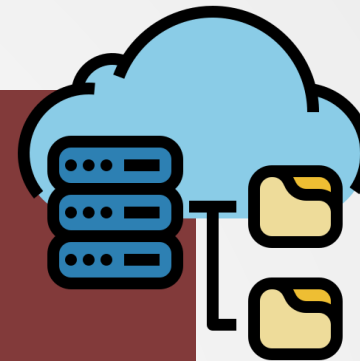


Progetto di Text Mining



Wish Upon a Star: predire le stelle di una review dal suo contenuto.

- Lorenzo Camaione 850380, Davide Toniolo 800458800458
 - Università di Milano-Bicocca

Introduzione



Gli obiettivi

Dataset

Text preprocessing

Modelli a confronto

Conclusioni

Gli obiettivi



E' possibile predire il numero di stelle di una review dal suo testo?

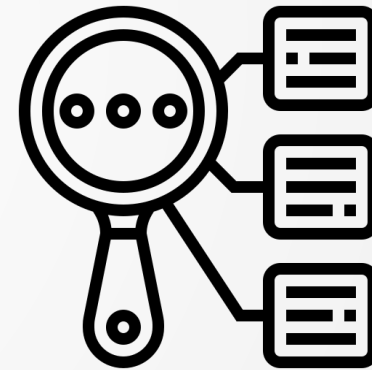
Dataset

Esplorazione

280 000 review in formato JSON

Filtro dei campi interessanti

Discretizzazione della variabile target



Text preprocessing

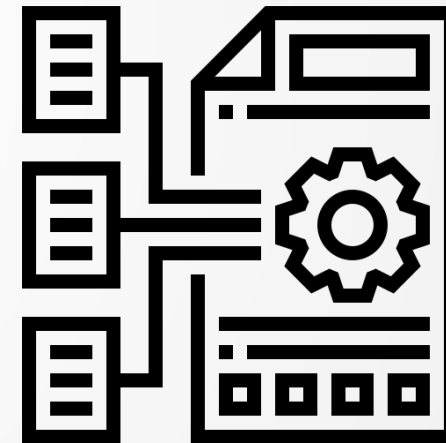
Text representation

Ottimizzazione

Tokenization

Stop words removal

Stemming o lemmatization



Text preprocessing

Text representation

Ottimizzazione

Bag of words

Td-idf



Text preprocessing

Text representation

Ottimizzazione

Ottimizzazione dell'utilizzo di memoria

Matrici Sparse

Decomposizione



Modelli

Gaussian NB

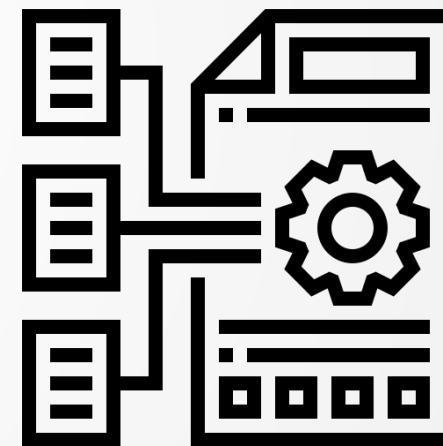
Reti neurali

Knn

Random forest

Linear SVC

Confronto tra modelli



Performance dei vari modelli misurati in termini di precision e recall

	Metric	Gaussian NB	Random Forest	kNN	Linear SVM	ANN
Class 0	Precision	0.18	0.75	0.24	0.32	0.37
	Recall	0.74	0.10	0.34	0.84	0.82
	F1-measure	0.28	0.18	0.28	0.46	0.51
Class 1	Precision	0.95	0.90	0.84	0.98	0.97
	Recall	0.58	1.00	0.87	0.79	0.83
	F1-measure	0.72	0.95	0.86	0.87	0.90
Macro Avg	Precision	0.56	0.82	0.43	0.65	0.67
	Recall	0.66	0.55	0.43	0.81	0.83
	F1-measure	0.50	0.56	0.42	0.67	0.70
Accuracy		60%	90%	73%	79%	83%



Conclusioni

**Grazie per
l'attenzione!**

