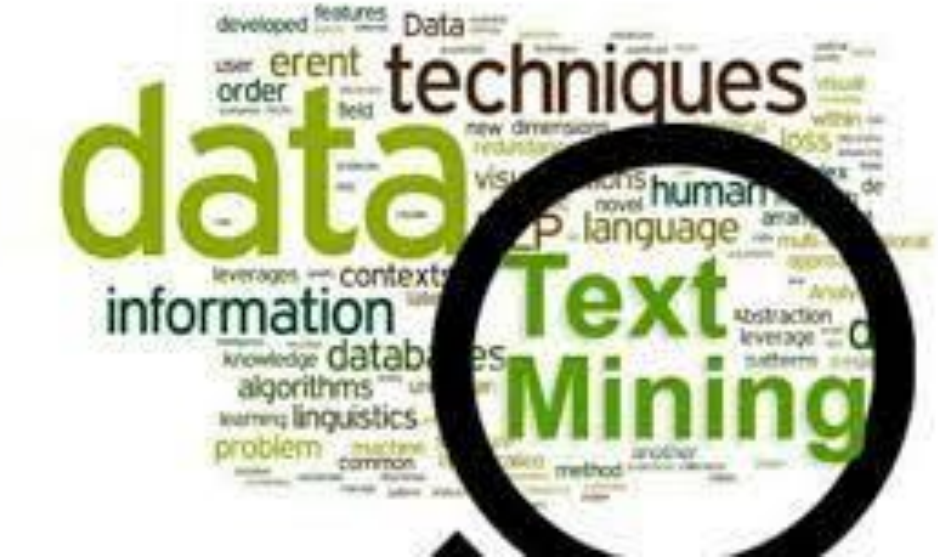
Text Mining

Francesco Pugliese, PhD

neural1977@gmail.com

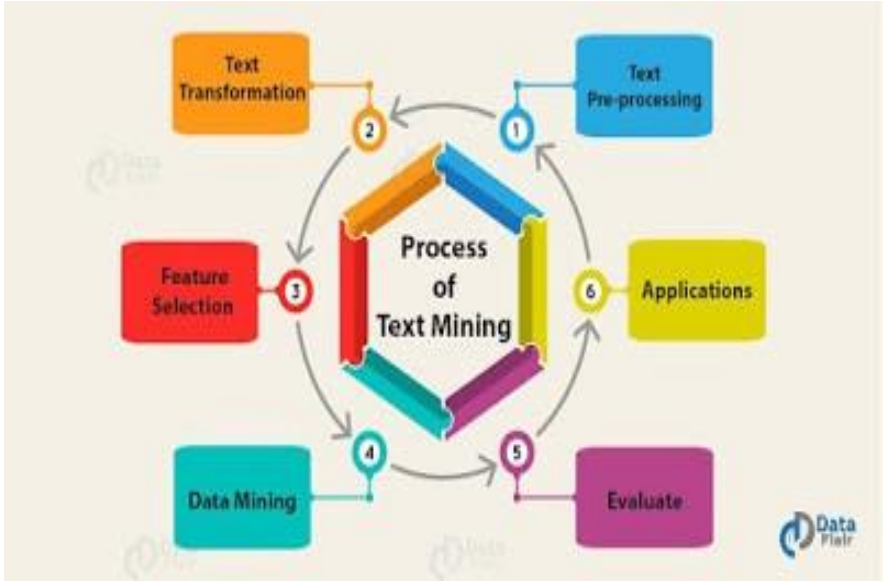
Cosa è il Text Mining

1) Il **Text Mining** (anche detto **Text Analytics**) è una tecnologia dell'Intelligenza Artificiale (AI) che utilizza **l'Elaborazione del Linguaggio Naturale (Natural Language Processing – NLP)** per trasformare il testo (dato non strutturato presente all'interno di documenti o database) in dati normalizzati e strutturati per l'analisi oppure per inserirli come input di algoritmi di **Apprendimento Automatico** (**Machine Learning – ML**).



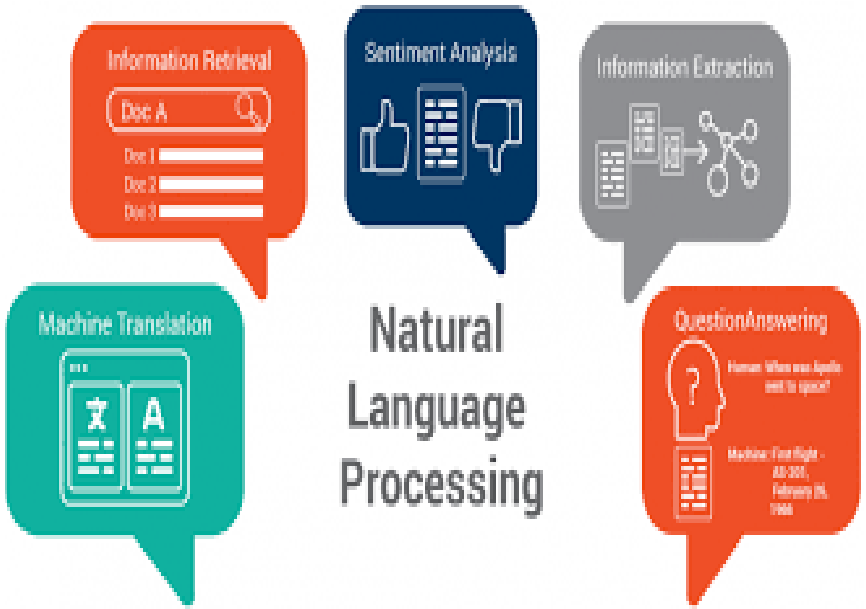
Text Mining

- ✓ Ampiamente utilizzato nelle organizzazioni **knowledge-driven**, il **Text Mining** rappresenta il processo di analisi di grandi **collezioni** di documenti per scoprire nuova informazione. Il Text Mining identifica fatti, relazioni e asserzioni che altrimenti rimarrebbero nascoste nell'enorme massa dei **Big Data Testuali**.
- ✓ Una volta estratta l'informazione viene convertita in una **forma strutturata** che può essere analizzata ulteriormente o presentata usando tabelle HTML clusterizzate, mappe mentali (mind map), grafici ecc.



Text Mining ed NLP

- ✓ I dati strutturati creati dal **Text Mining** possono essere integrati in database, **data warehouse** o **dashboard** di business intelligence e usati per analisi
- ✓ Il **Text Mining** impiega una varietà molto estesa di metodologie per l'elaborazione del testo, uno dei più importanti dei quali è il **Natural Language Processing (NLP)**
- ✓ Il **Natural Language Understanding** permette alle macchine di «leggere» il testo simulando l'abilità umana di comprendere un linguaggio **naturale** come l'Inglese, Spagnolo o Cinese.

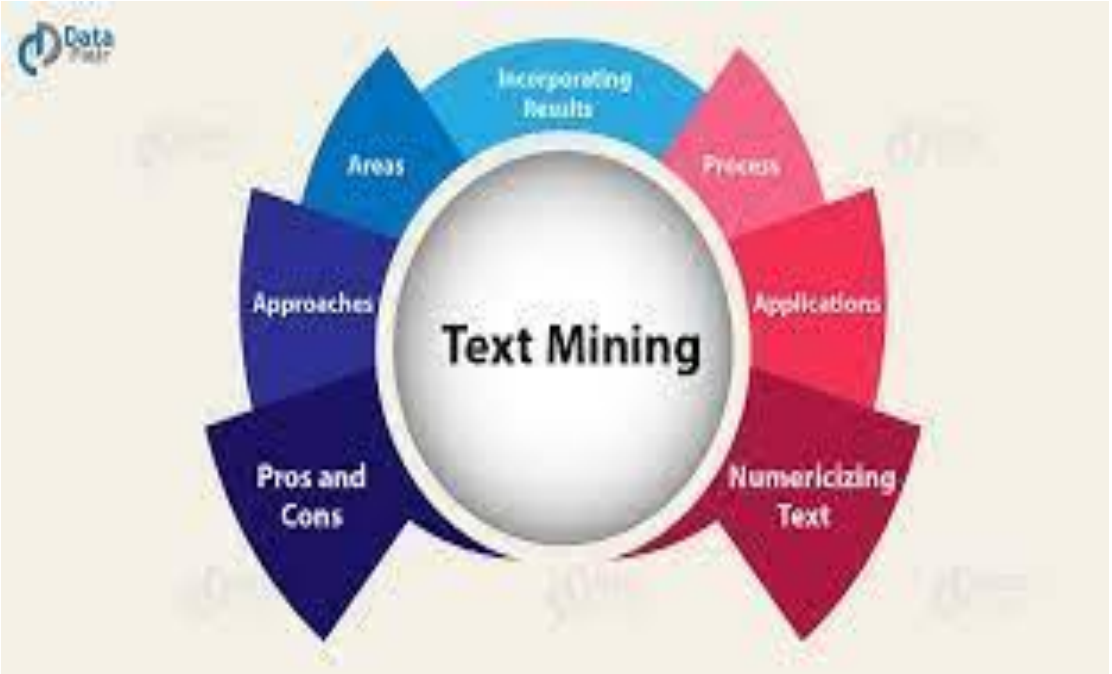


Text Mining ed NLP

- Il **Natural Language Processing** include sia il **Natural Language Understanding** (Comprensione del Linguaggio Naturale) che il **Natural Language Generation**, il quale invece simula la capacità degli esseri umani di creare testo in linguaggio naturale come per esempio nella summarization dell'informazione (sintesi del testo) o prendere parte ad un dialogo.
- I sistemi **NLP** di oggi possono analizzare una quantità illimitata di dati testuali senza ovviamente avere fatica e i un modo coerente e senza bias. Data l'enorme quantità di dati non strutturati prodotta ogni giorno dai social o da dispositivi elettronici, questo tipo di automazione è diventata essenziale.

Scopi principali del Text Mining

- ✓ Individuare i principali gruppi tematici (**topic modeling**)
- ✓ Classificare i documenti in categorie predefinite
- ✓ Scoprire associazioni nascoste (legami tra argomenti, o tra autori, trend temporali, …)
- ✓ Estrarre informazioni specifiche (es. nomi di geni, nomi di aziende, …)
- ✓ Addestrare **motori di rice**rca
- ✓ Estrarre concentti per la creazione di ontologie (**ontology learning**)



Scopi principali del Text Mining

- ✓ Con l'aumento del testo disponibile online, in un formato digitale accessibile attramersc mezzi tecnologici, il **Text Mining** sta assumendo sempre più importanza.
- ✓ Il **Data Mining** si occupa di estratte informazione significativa e di valore da grandi dataset.
- ✓ Nel **Text Mining** esistono due processi che possono essere applicati in generale:

1) L'Information Retrieval (IR): che si occupa di recuperare informazione dai documenti attraverso delle query logiche basate su un insieme di parole chiave.

2) L'Information Extraction (IE): si occupa di estrarre infromazione specifica da documenti strutturati o non strutturati utilizzando speciali algoritmi di Machine Learning.

Scopi principali del Text Mining

- ✓ Al fine di raggiungere i sopra citati processi, è necessario eseguire una sorta di summarization dei documenti e metterli in un ordine logico e accessibile. Hai dunque la necessità di classificarli, clusterizzarli ed etichettare ciascun pezzo di testo con delle parole chiave.
- ✓ Da una prospettiva moderna del Text Mining, l'Information Extraction (IE) è più utile dell'Information Retrieval (IR), perchè l'IE cerca di identificare concetti e raggiunge l'informazione necessaria, in altre parole l'IE cerca di derivare informazione strutturata accurata da testo non strutturato.

Bibliografia

https://www.linguamatics.com/what-text-mining-text-analytics-and-natural-language-processing