Analisi del testo

L'analisi del testo nei social media è un processo potente per estrarre informazioni rilevanti e classificarle in base a vari criteri.

La prima fase dell'analisi del testo coinvolge la raccolta di dati da diverse fonti. Questi dati possono provenire da varie fonti, tra cui sondaggi online, e-mail, piattaforme di e-commerce, tweet e ticket di supporto.

- **Sentiment analysis** : Un aspetto fondamentale dell'analisi del testo nei social media è l'identificazione del sentiment o del tono dei messaggi. Gli algoritmi di analisi del sentiment determinano se un messaggio è positivo, negativo o neutro. Questo aiuta a valutare l'opinione degli utenti rispetto a un prodotto, servizio o argomento specifico.

- **Topic analysis** : Un'altra parte importante dell'analisi del testo è la classificazione dei messaggi in base ai temi o argomenti trattati. Questo può essere fatto utilizzando algoritmi di apprendimento automatico che identificano le parole chiave o le frasi che indicano l'argomento principale di un messaggio.

- **Intent detection** : Questo processo coinvolge la comprensione e l'etichettatura dell'intento che un utente ha quando condivide un messaggio sui social media. Ad esempio, un utente potrebbe esprimere l'intento di ottenere assistenza per un problema, fare una domanda, esprimere una recensione o addirittura effettuare un acquisto. L'identificazione dell'intento è fondamentale per indirizzare in modo appropriato le risposte o le azioni successive, come fornire assistenza clienti, consigliare prodotti o raccogliere feedback.

- **Text extraction** : Questa fase dell'analisi del testo riguarda l'acquisizione e l'estrazione di informazioni rilevanti dai testi. Può comportare l'identificazione di dati chiave, citazioni, o qualsiasi altra informazione di interesse all'interno dei documenti o dei messaggi.

- **Word Frequency** : Questo aspetto dell'analisi del testo coinvolge il calcolo della frequenza con cui le parole specifiche appaiono all'interno dei documenti o dei messaggi. La frequenza delle parole è utile per identificare le parole chiave o i termini più comuni in un testo, il che può aiutare a comprendere i temi principali trattati. Inoltre, la frequenza delle parole può essere utilizzata per valutare l'importanza o la rilevanza di termini specifici all'interno di un determinato contesto.

Struttura del testo

Prima di procedere con l'analisi vera e propria, i dati raccolti vengono sottoposti a un processo di pre-elaborazione. Questo può includere la rimozione di dati non rilevanti o duplicati, la normalizzazione del testo (ad esempio, trasformare tutte le lettere in minuscolo) e la rimozione di caratteri speciali o punteggiatura.

Il testo può essere trovato in una forma **non strutturata** o **strutturata**. Vengono usate le espressioni regolari per analizzare dati in forma strutturata, mentre per i dati non strutturati si necessita di algoritmi che analizzino il testo e quindi usare tecniche di NLP (natural lenguage processing).

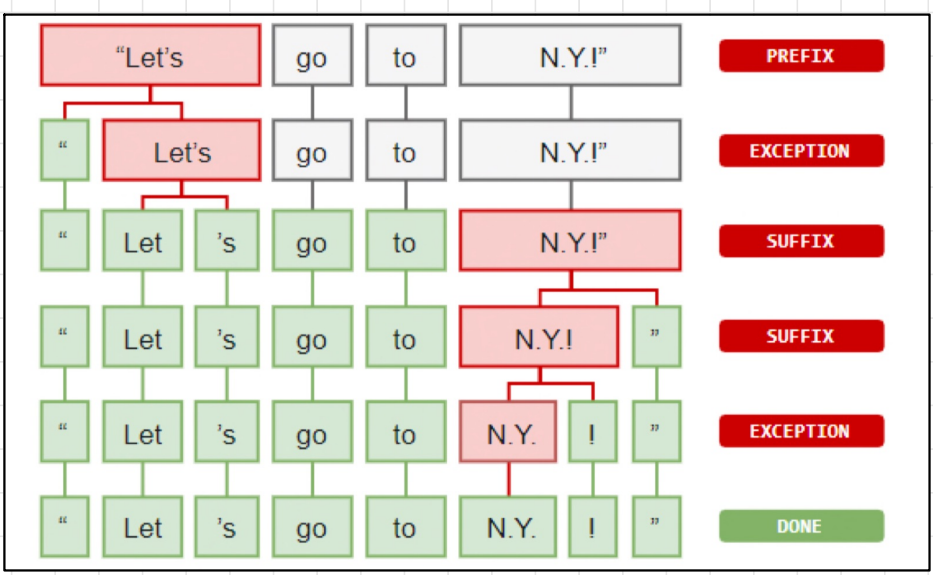
Un' **espressione regolare** è un modello composto da una stringa, un **set** o un **gruppo** di caratteri che definisce un pattern specifico. Questi strumenti sono comunemente impiegati per la ricerca, l'estrazione o la manipolazione di testo o dati strutturati seguendo regole o schemi predefiniti.

NLP (Natural Lenguage Processing)

Word tokenizzation

La tokenizzazione delle parole è un processo fondamentale nell'analisi del testo e nel Natural Language Processing (NLP). Consiste nel suddividere una sequenza di testo in singole parole o "token", in modo che ciascun token rappresenti un'entità separata. Nell'ambito della tokenizzazione, è possibile considerare i seguenti elementi:

- **Prefix (Prefisso)**: I caratteri di inizio parola sono quelli che si trovano all'inizio di una parola. Ad esempio, nell'inglese, "un" in "unhappy" è un prefisso.
- **Suffix (Suffisso)**: I caratteri di fine parola sono quelli che si trovano alla fine di una parola. Ad esempio, "-ly" in "quickly" è un suffisso.
- **Infix (In mezzo)**: I caratteri infix sono quelli che si trovano all'interno di una parola. Ad esempio, nella parola "co-operate," il trattino è un carattere infix.
- **Eccezione (Exception)**: In alcuni casi, ci sono parole o sequenze di caratteri che costituiscono casi speciali e richiedono una tokenizzazione specifica. Ad esempio, "I've" potrebbe essere trattato come una singola parola invece di essere diviso in "I" e "ve."



Lemmatizzazione

La lemmatizzazione è un processo di analisi morfologica al testo che mira a ridurre le parole di un testo alla loro forma base o radice, nota come "lemma". L'obiettivo della lemmatizzazione è quello di raggruppare le parole che condividono la stessa radice, in modo da ridurre l'ambiguità e semplificare l'analisi del testo. Ecco alcuni punti chiave relativi alla lemmatizzazione:

- **Lemma**: Un lemma è la forma di base di una parola dalla quale derivano tutte le forme grammaticali. Ad esempio, il lemma del verbo "corro" è "correre", e il lemma dell'aggettivo "veloci" è "veloce".
- **Ambiguità**: La lemmatizzazione è utile per risolvere l'ambiguità nelle parole. Ad esempio, la parola "correndo" può riferirsi a "correre" nel suo lemma anziché a "corro" o "corri".
- **Normalizzazione**: La lemmatizzazione aiuta a normalizzare il testo. Ciò significa che le parole simili vengono ridotte alla stessa forma base, semplificando l'analisi del testo. Ad esempio, "corro" e "correndo" vengono entrambi ridotti a "correre".
- **Vantaggi nell'analisi testuale**: La lemmatizzazione è spesso utilizzata in NLP per migliorare l'analisi del testo, compresa l'analisi del sentiment, l'identificazione degli intenti e la ricerca di informazioni.

Stop words

Le "stop words" (parole di arresto) sono parole che vengono comunemente rimosse durante la fase di analisi del testo nel campo del Natural Language Processing (NLP) e dell'information retrieval. Queste parole sono considerate comuni, non portatrici di significato distintivo o troppo frequenti nel linguaggio naturale per essere utili in alcune applicazioni di analisi del testo. Di solito, vengono rimosse per migliorare le prestazioni delle applicazioni di NLP e ridurre la dimensione dei dati senza perdita significativa di informazioni.

Le stop words includono parole comuni come "e", "il", "la", "in", "che", "un", "per", "su", "con", "non", "è", "ma", ecc. Queste parole sono spesso molto frequenti e non portano un significato distintivo in un testo.

Part-of-speach (POS) Tagging

Il Part-of-Speech (POS) tagging, noto anche come etichettatura grammaticale o disambiguazione delle categorie delle parole, è un'attività fondamentale nell'elaborazione del linguaggio naturale (NLP). Essa consiste nell'assegnare una specifica categoria grammaticale o parte del discorso a ciascuna parola in una frase o in un testo. Questa categorizzazione aiuta a comprendere la struttura sintattica e il significato di una frase.

Il POS tagging assegna a una parola un'etichetta che indica il suo ruolo sintattico all'interno di una frase. Ad esempio, classifica le parole come sostantivi, verbi, aggettivi, avverbi, pronomi, preposizioni, congiunzioni e così via.
Molte parole in una lingua possono avere significati e funzioni diverse a seconda del contesto. L'POS tagging aiuta a eliminare l'ambiguità di queste parole classificandole in base al loro ruolo in una specifica frase.

In italiano, per esempio, l'POS tagging può assegnare la categoria "NOM" a una parola come "cane" (sostantivo) o "VER" a "corre" (verbo).

Name Entity Recognition (NER)

Il riconoscimento di entità nominative (NER), noto anche come identificazione di entità o estrazione di entità, è una fondamentale attività di elaborazione del linguaggio naturale (NLP). Consiste nell'individuare e classificare le entità nominative all'interno di un testo. Le entità nominative sono entità specifiche che hanno nomi, come nomi di persone, organizzazioni, luoghi, date, valori monetari e altro.

Il NER è il processo di individuare e categorizzare le entità nominate in un testo in categorie predefinite come nomi di persone, nomi di organizzazioni, luoghi geografici, date e altre.

Le entità nominate possono essere suddivise in vari tipi, inclusi ma non limitati a :

- Nomi di Persone: Identificazione di nomi di individui.
- Nomi di Organizzazioni: Riconoscimento dei nomi di aziende, istituzioni o altre organizzazioni.
- Nomi di Luoghi: Individuazione di nomi di città, paesi e altri luoghi geografici.
- Date e Orari: Riconoscimento di date, orari ed espressioni temporali.
- Valori Monetari: Identificazione di importi di denaro.
- Nomi di Prodotti: Riconoscimento dei nomi di prodotti o servizi.

Il NER viene utilizzato in una vasta gamma di applicazioni, tra cui il recupero delle informazioni, l'estrazione di informazioni, i sistemi di risposta a domande, i chatbot e la traduzione automatica. È anche prezioso per estrarre dati strutturati da testi non strutturati e migliorare i sistemi di ricerca e raccomandazione.

Sentence Segmentation

La segmentazione delle frasi è il processo di individuare i limiti tra le frasi all'interno di un testo o documento. Una frase è generalmente una sequenza di parole che esprime un concetto completo o un'idea. La segmentazione delle frasi è fondamentale in NLP poiché consente ai sistemi di trattare ciascuna frase separatamente. Questo è essenziale per svolgere analisi grammaticali, riconoscere le entità nominate e analizzare il sentiment a livello di frase.

La segmentazione delle frasi è utile in applicazioni come la traduzione automatica, il riassunto del testo, l'analisi del sentiment, la trascrizione del discorso e l'indicizzazione dei motori di ricerca.

NLP Pipeline

Per riassumere, una pipeline NLP comune, in genere, prevede le seguenti fasi :

• Word tokenization;

• Stop words removal;

• Lemmatization;

• POS tagging (depending on the application);

• NER tagging (depending on the application);

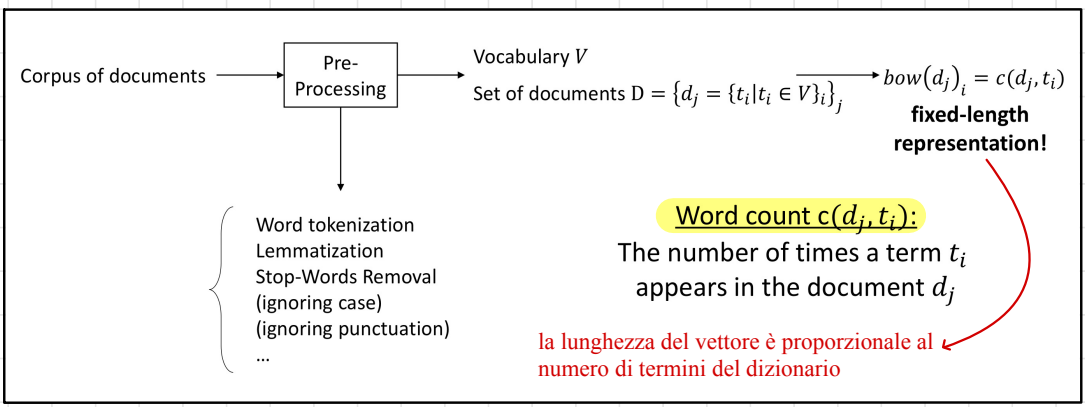
• Sentence segmentation.

Bag of Words Representation

Utile per ottenere dati a lunghezza fissa necessarie ad addestrare algoritmi di machine learning.

La rappresentazione BoW consiste nell'analizzare un insieme di documenti (come frasi o testi) per creare un vocabolario di parole uniche all'interno di questi documenti. Successivamente, ogni documento viene rappresentato come un vettore in cui ogni elemento rappresenta il numero di volte in cui una determinata parola appare nel documento stesso. La frequenza di ogni termine nel documento quindi, è registrata in questo vettore, ottenendo così una distribuzione di frequenza delle parole all'interno del documento.

Prima di applicare BoW, i dati di testo sono sottoposti a una serie di operazioni di pre-processing, secondo una delle tecniche descritte precedentemente, per ottenere una rappresentazione più pulita e coerente dei testi.



Feature normalizzation e Term Frequency (TF)

La rappresentazione BoW può comportare una frequenza diversa di termini nei vettori per documenti di lunghezze diverse. Questo può portare a una disparità nelle scale tra i documenti. Per affrontare questo problema, viene applicata una normalizzazione delle feature.

In questo processo, la frequenza dell'i-esimo termine viene diviso per il numero totale di termini presenti nell'intero documento (l'insieme di documenti analizzati). Ciò permette di ottenere una frequenza normalizzata per ciascuna parola in ogni documento. Questo processo di normalizzazione garantisce che i vettori di BoW abbiano una scala uniforme, indipendentemente dalla lunghezza dei documenti.

Dopo la normalizzazione delle feature, si ottiene una rappresentazione normalizzata della frequenza delle parole nei documenti. Questa rappresentazione è spesso denominata "**Term Frequency**" (TF) ed è un vettore di dimensione fissa per ciascun documento in cui ciascuna dimensione rappresenta la frequenza normalizzata di una parola all'interno del documento.

$$\frac{c(d_j, t_i)}{\sum_i c(d_j, t_i)} = tf(d_j, t_i) \quad \textbf{tf: term frequency}$$

Inverse Document Frequency (IDF)

L'IDF considera quante volte una parola è presente in tutti i documenti e quindi quanto quella parola è rara o comune. Viene calcolata come logaritmo del numero di documenti in rapporto al numero di documenti che contengono il termine i-esimo.

$$idf(t_j) = -\log\frac{c(t_j)}{n} = \log\frac{n}{c(t_j)} \text{ (self information)}$$

$$df(t_j) = \log\frac{c(t_j)}{n}$$

con **n** = numero di documenti

Vogliamo calcolare la IDF della parola "dog" all'interno di 10 documenti.

- Se un solo documento contiene la parola "dog", il termine sarà più raro e specifico

$$idf(t_j) = \log\frac{n}{c(t_j)} = \log\frac{10}{1} = \mathbf{1}$$ More rare and specific

- Se tutti e 10 i documenti contengono la parola "dog", il termine sarà più comune e generale

$$idf(t_j) = \log\frac{n}{c(t_j)} = \log\frac{10}{10} = \mathbf{0}$$ More common and general

Term Frequency - Inverse Document Frequency (TF-IDF)

Si necessita di dare un peso alle parole in base all'informazione che danno. L'Inverse Document Frequency è una tecnica utilizzata per dare un peso alle parole in un insieme di documenti. L'obiettivo è individuare le parole chiave o le parole più informative all'interno dei documenti.

Per ottenere il peso complessivo TF-IDF di una parola in un documento, moltiplichiamo la sua TF per il suo IDF:

$$tf(d_i, t_j) \cdot idf(t_j) = \frac{c(d_i, t_j)}{\sum_j c(d_i, t_j)} \log\frac{n}{c(t_j)}$$

Il risultato è un valore che indica l'importanza della parola nel documento, tenendo conto sia della sua frequenza all'interno del documento che della sua rarità nell'intero corpus.

L'uso di TF-IDF consente di assegnare un peso alle parole in modo informato, evidenziando le parole chiave o informative all'interno di ciascun documento