TEXT ANALISYS

mercoledì 25 ottobre 2023 17:14

L'analisi del testo è il processo da cui estrarre e classificare le informazioni rilevanti
utilizzando tecniche diverse rispetto ai testi non strutturati

Estrazione argomento, Analisi del sentiment, Classificazione degli aspetti, Entità denominata
estrazione

Esempi d'uso:
Survey(questionari)
Email
Recezioni
Twit
Forum per aiuto

L'analisi del testo consente di capire:
Sentimenti
Emozioni
Topic
Nome dell'identità

Si fa per vari motivi:
- Classificazione del testo
- Analisi del sentimento
- Analisi dell'argomento
- Rilevamento delle intenzioni
- Estrazione del testo

• Frequenza delle parole

Il testo può essere in due forme:
Forma strutturata: si usano espressioni regolari(REGEX)
Forma non strutturata: generalmente linguaggio non strutturato(Natural Language Processing)

-espressioni regolari:
Le stringhe sono delle espressioni regolari
'the'->trovo tutti i the presenti nel testo
'I'->trova tutte le I presenti nel testo

-sets
Consentono di trovare pattern specifici
'[Tt]'->trovo tutte le t maiuscole e minuscole, quindi tutte le t
'[Tt]he'->tutti i the con t maiuscola o minuscola

-negazioni
Utilizzando ^ si può negare l'espressione(equivale al not)
'[^A-Za-z]'->tutto tranne le lettere
Day|and ->equivale all'or

-composti
'[0-9][0-9]->cifre di due numeri compresi tra 0 e 9

Quantificatori:

- *: Corrisponde a zero o più occorrenze del carattere o del gruppo precedente. Ad esempio, a* corrisponde a "a", "aa", "aaa", ecc.
- +: Corrisponde a una o più occorrenze del carattere o del gruppo precedente. Ad esempio, a+ corrisponde a "a", "aa", "aaa", ecc., ma non a una stringa vuota.
- ?: Corrisponde a zero o una occorrenza del carattere o del gruppo precedente. Ad esempio, a? corrisponde a "a" o a una stringa

vuota.

- {n}: Corrisponde esattamente a "n" occorrenze del carattere o del gruppo precedente. Ad esempio, a{3} corrisponde solo a "aaa".
- {n, m}: Corrisponde da "n" a "m" occorrenze del carattere o del gruppo precedente. Ad esempio, a{2,4} corrisponde a "aa", "aaa" o "aaaa".
- "Punto" (.) rappresenta un carattere jolly e corrisponde a qualsiasi carattere singolo, ad eccezione di un carattere di nuova riga. Quindi, se si utilizza il punto in una regex, esso farà il matching con qualsiasi carattere tranne il newline (\n).

Sequenze di Escape:

- Le regex utilizzano il carattere \ per "escapare" i caratteri speciali e trattarli come caratteri letterali. Ad esempio, per cercare un punto (.), che normalmente ha un significato speciale nelle regex, puoi utilizzare \. per farlo corrispondere solo al carattere "." reale.
- Alcuni caratteri comuni che richiedono l'uso di una sequenza di escape includono ., *, +, ?, (,), [,], {, }, \, |, ecc..

In espressioni regolari, i gruppi sono utilizzati per raggruppare parti di un pattern in modo da poter applicare quantificatori o operazioni su di esse come un'unità singola. I gruppi sono racchiusi tra parentesi tonde "( )" e possono essere utilizzati per vari scopi.

Esempio di gruppo e cattura in una regex:

- (abc)+ corrisponde a "abc", "abcabc", ecc., dove (abc) rappresenta un gruppo che corrisponde a "abc" e il "+" indica che il gruppo può ripetersi una o più volte.
- (a(b)c) corrisponde a "abc" e cattura due gruppi: uno corrisponde a "abc" e l'altro a "b".

I gruppi sono uno strumento potente nelle regex che consentono di creare pattern complessi e di estrarre informazioni rilevanti dalle stringhe corrispondenti.