NATURAL LAUNGUAGE PROCESS(NLP)

mercoledì 25 ottobre 2023 17:59

Conduttura della PNL
Una pipeline PNL comune implica generalmente quanto segue passaggi:
• Tokenizzazione delle parole;
• Interrompere la rimozione delle parole;
• Lemmatizzazione;
• Etichettatura POS (a seconda dell'applicazione);
• Etichettatura NER (a seconda dell'applicazione);
• Segmentazione delle frasi.

-Work tokenization
Questo operazione consente di dividere in più token una stringa

-Token: È un entità che può essere attribuita a uno o più caratteri

I token sono di 4 tipi:
• Prefisso: caratteri all'inizio;
• Suffisso: caratteri finali;
• Infisso: caratteri intermedi;
• Eccezione: regole per casi speciali in cui dividere una stringa diversi token o impedire che un token venga creato e diviso quando vengono applicate le regole di punteggiatura

-Lemmatization
Analisi morfologica del testo, consente di risolvere l'ambiguità e quindi serve per disambiguare il testo.
Associo delle parole a una singola parola di riferimento.

-Stop words
Sono tutte quelle parole molto comuni nei testi che vanno eliminare per
non danno informazioni rilevanti
Es: i, me, my, myself, we, our

-Part of Speech(POS) Tagging
Consente di associare alle parole un concetto ben preciso
Es: matteo->nome

-Named entity recognitin(NER)
Consente di associare parole o gruppi di parole a un concetto più specifi
infatti si parla di entità
Es: ue->organizzazione

-Sentence segmentation

Bag of words representation
Algoritmi basati sul machine learning che pero utilizzano dei dati a
lunghezza fissa.
Prende tutte le parole e le considera come parole sconnesse , come se
fossero in una borsa, quindi non si considera la sintassi e conta il numero
occorrenze per ogni parola
Fasi:
-Prendo il corpo di tutti i documento(corpus of document)
-Pre-processing, fase in cui si utilizzano tutte le procedure viste prima
-Ricavo il vocabolario
-Conto per ogni documento quante volte è presente ogni parola del
vocaboli

Il risultato sarà una vettore di uguale dimensione per ogni documento

TF-IDF (term frequency, inverse document frequency)

ché

co,

o di

TF-IDF (term frequency, inverse document frequency)

Il Term Frequency-Inverse Document Frequency (TF-IDF) è una tecnica utilizzata nell'elaborazione del linguaggio naturale per valutare l'importa di una parola in un documento rispetto all'intero corpus di documenti. E una spiegazione breve:
Term Frequency (TF): Rappresenta la frequenza con cui una parola speci appare in un documento. Più una parola compare frequentemente in un documento, più alta sarà la sua TF per quel documento.
TF(t, d) = (Numero di volte che la parola "t" appare nel documento "d") / (Numero totale di parole nel documento "d")
Inverse Document Frequency (IDF): Misura l'importanza di una parola nell'intero corpus. Le parole comuni che compaiono in molti documenti avranno un IDF basso, mentre le parole più rare e specifiche avranno un più alto.
IDF(t, D) = log(Numero totale di documenti nel corpus "D" / Numero di documenti che contengono la parola "t")
Infine TF*IDF

La formula TF-IDF combina queste due metriche per calcolare un punteg che riflette l'importanza di una parola in un documento specifico rispetto all'intero corpus. Le parole con punteggi TF-IDF elevati sono generalmen considerate più rilevanti per quel documento.

nza
cco

fica

/

IDF

ggio
o
nte