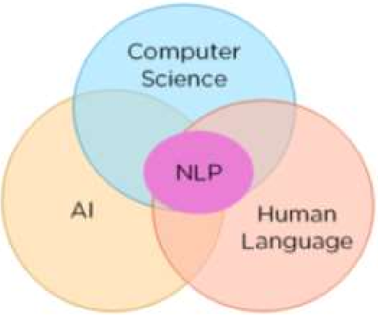
NLP

16 December 2023 20:30

Natural Language Processing

Natural Language Processing (NLP) is the branch of data science primarily concerned with dealing with textual data. It is the intersection of linguistics, artificial intelligence, and computer science.



Syntactic Analysis Vs Semantic Analysis

SYNTACTIC ANALYSIS

In syntactic analysis, we use rules of formal grammar to validate a group of words. With syntactic analysis, we validate the structure of our sentences.

SEMANTIC ANALYSIS

Semantic analysis deals with the part where we try to understand the meaning conveyed by sentences. This will allow computers to understand natural language better.

✓ NLTK ✓ SPACY ✓

spaCy

Hugging Face

(NLP)

NLTK

✓ (NLP)

Data Cleaning in NLP

NLP

i/p data

↳ textual format

The data cleaning process largely depends on the problem that we are working on. But normally, what we do is remove any special characters such as $, %, #, @, <, >, etc.

Data Preprocessing in NLP

LOWERCASE

YES | yes

If any character in our text is in uppercase we convert it to lowercase. Otherwise, our model will perceive the uppercase and lowercase characters as different from each other.

TOKENIZATION

In tokenization, we take our text from the documents and break them down into individual words.

STOPWORDS REMOVAL

We remove words from our text data that don’t add much information to the document. So, they add noise to the data.

STEMMING

'Caring'

↳ Car ×

↳ Care ✓

In stemming we reduce a word to its root word. It transforms the word back to its original form i.e reduces inflection.

LEMMATIZATION

Lemmatization does the same thing as stemming but in lemmatization, we get a root word that has some meaning.

N GRAMS

N Grams are used to preserve the sequence of information which is present in the document.

When N = 1, they are called Unigrams. When N = 2, they are called bigrams. When N = 3, they are called trigrams. And so on.

Word Vectorization

ONE HOT VECTOR ENCODING

In one-hot vector encoding, we made embeddings of the entire corpus. In these types of word vectors, all the words are independent of each other.

WORD2VEC

With word2vec, we were able to form a dependence of words with other words. These were a considerable improvement over One Hot Vector.

CBOW Model

In the CBOW (continuous bag of words) model, we predict the target (center) word using the context (neighboring) words.

The CBOW model is faster than the skip-gram model because it requires fewer computations and it is great at representing less frequent words.

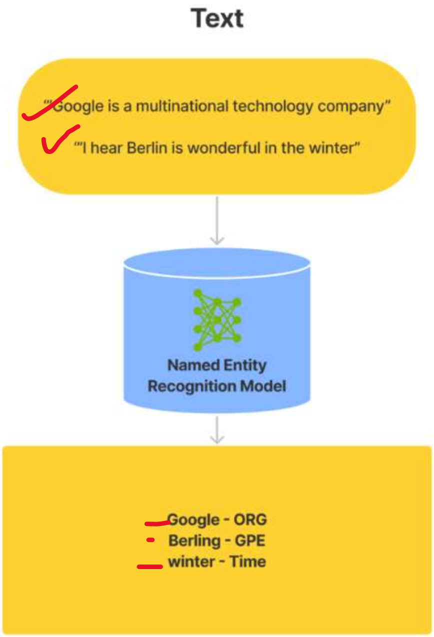
Skip Gram Model

With Skip Gram, we predict the context words using the target word.

Even though the skip-gram model is a bit slower than the CBOW model, it is still great at representing rare words.

Named Entity Recognition

Named Entity Recognition is an important information retrieval technique.


Text
"Google is a multinational technology company"
"I hear Berlin is wonderful in the winter"
Named Entity
Recognition Model
Google - ORG
Berling - GPE
winter - Time