Natural Language Processing

What is NLP?

NLP is a language that’s computer program can understand human language as it is spoken and written.

**Tokenization:**

Combination of the words sentence. We separate each words from sentence that is called tokenization.

**Example:  
 This is a sample**

**“This” “is” “a” “sample”**

There are some libraries use to tokenization.

* Spacy
* Natural Language Toolkit (NLTK)

**Spacy:**

Resource library https://spacy.io/ .

By using spacy we can do tokenization easily without spacy we will have to write a lot of code to tokenization separately like with punctuation or without punctuation, comma etc.

So there are many codes I have done while following the tutorial.

**Lemmatization:**

Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological **(relating to the form or structure of things)** analysis of words.

|  |  |
| --- | --- |
| Text-Word | Lemma |
| Help | Help (v) |
| Helps | Help (v) |
| Helped | Help (v) |
| Helping | Help (v) |

Lemmatization refers you where correct words you will use like I draw a table before this heading help word has types like helps, helping, helped, and help wo lemmatization will refer you where you should use which type of words.

**Stop Words:**

Those words that do not contribute to the deeper meaning of the phrase. They are the most common words such as: the, a, and is. For some applications like documentation classification, it may make sense to remove stop words. NLTK provides a list of the commonly agreed upon words for a variety of languages, such as English.

**Vocabulary and Phrase Matching in NLP:**

**Rule-based Matching:**

Spacy offer a rule-matching tool the called Matcher that allows you to build a library of token patterns, then match those pattern against a Document object to return a list of found matches. You can match on any part of the token including. You can match any part of token including text and annotations and you can add multiple patterns to the same matcher.

**Name Entity Recognition (NER):**

Entity means some special kind of place name or any organization name so using spacy we find the entity from text, paragraph.

**Sentence Segmentation:**

We divide the raw data according to the requirement. We did un sentence segment for un raw and un structured data.

**Stemming:**

Stemming is the process reduce words to its words stem that affix to suffixes and prefixes or to the roots of words known as lemma. It is important in NLU and NLP

* Prefix Character(s) at the beginning $ ( “ .
* Suffix Character(s) at the end km ) ” , . !
* Infix Character(s) in between - -- / …
* Exception Special-case rule to split a string into several tokens or prevent rules are applied St. U.S.

**Bag Of Words:**

It is suggests you words like your comments and tell you. You should give answer this.

**Features Extraction in NlP:**

* **Frequency:**

In whole documents words how much time is repeating that is called frequency.

* **Document Frequency:**

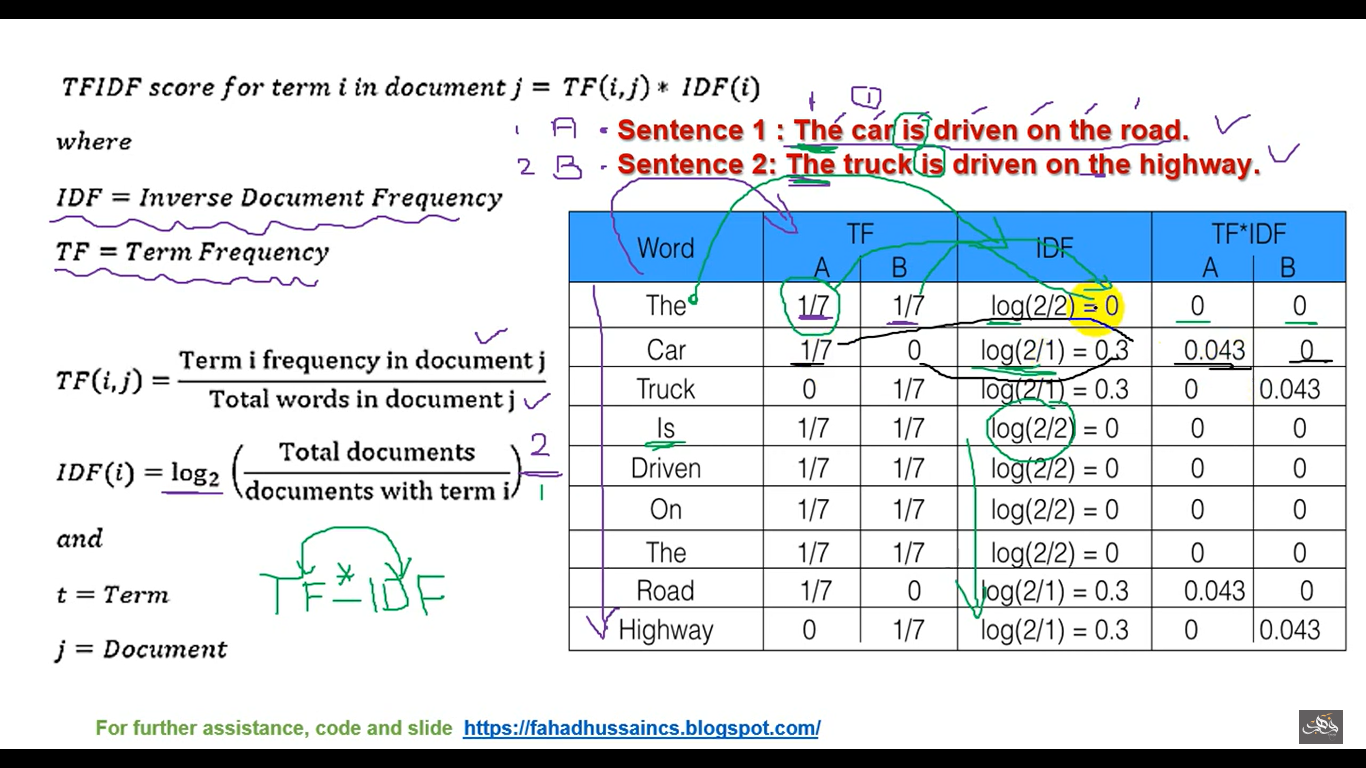
In this if any words repeat again and again so in document frequency will decrease/downscale those word.

* **Inverse Document Frequency (IDF)**

In this we find the elements repeating weight how much time repeat on high level and low level.

* **TF-IDF**

Is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.



**N-grams:**

In this we break the sentence according to condition mean how much words you want to break sentence.

e.g.

['The quick brown fox jumped over the lazy dog.']

Now if you want to break it into single or twice or thrice etc

Split in 1

['The', 'quick', 'brown', 'fox', 'jumped', 'over', 'the', 'lazy', 'dog.']

Split in 2:

['The quick',

'quick brown',

'brown fox',

'fox jumped',

'jumped over',

'over the',

'the lazy',

'lazy dog.']

Split in 3

['The quick brown',

'quick brown fox',

'brown fox jumped',

'fox jumped over',

'jumped over the',

'over the lazy',

'the lazy dog.']

**Sentiment Analysis:**

Sentiment analysis (or opinion mining) is a [natural language processing (NLP)](https://monkeylearn.com/natural-language-processing/) technique used to determine whether data is positive, negative or neutral. Sentiment analysis is often performed on textual data to help businesses monitor brand and product sentiment in [customer feedback](https://monkeylearn.com/customer-feedback/), and understand customer needs.