Tokenizing and Text Processing ¶

1. Tokenization

In NLP, in broader terms we are dealing with sentences. A machine will not understand a sentence in one go. So we breakdown a long sentence into a shorter sentence.

Using separators like space, tab, period, comma

"Hello, How are you?" - so the input will look like "hello", "how", "are", "you"

2. Stop Word removal

Words that do not carry a lot of significance in a sentence. Example: and, the, a, is

3. Root word extraction (Stemming)

"Eating" - Eat

"Climbing" - Climb

In [1]:
```
import nltk
```

In [2]:
```
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer, WordNetLemmatizer
```

In [4]:
```
nltk.download('punkt') #tokenizing
nltk.download('stopwords')
nltk.download('wordnet')
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\ahana\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping tokenizers\punkt.zip.
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\ahana\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping corpora\stopwords.zip.
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\ahana\AppData\Roaming\nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
```

Out[4]:
```
True
```

In [10]:
```
def tokenize_and_preprocess(text):
    tokens = word_tokenize(text)

    #remove stopwords
    stop_words = set(stopwords.words('english'))
    tokens = [token for token in tokens if token.lower() not in stop_words]

    #stemming
    stemmer = PorterStemmer()
    stemmed_tokens = [stemmer.stem(token) for token in tokens]

    #Lemmatization (grammar)
    lemmatizer = WordNetLemmatizer()
    lemmatized_tokens = [lemmatizer.lemmatize(token) for token in tokens]

    return tokens, stemmed_tokens, lemmatized_tokens
```

In [15]:
```
#Example Usage

text = "What were they eating?"
tokens, stemmed_tokens, lemmatized_tokens = tokenize_and_preprocess(text)
print("Original tokens: ", tokens)
print("Stemmed Tokens: ", stemmed_tokens)
print("Lemmatized tokens: ", lemmatized_tokens)
```

```
Original tokens:  ['eating', '?']
Stemmed Tokens:  ['eat', '?']
Lemmatized tokens:  ['eating', '?']
```