Natural Language Processing

Tokenization in NLP

1. Corpus- Paragraphs
2. Documents- Sentences
3. Vocabulary- Unique words in a paragraph
4. Words- All the words in the corpus

Tokenization

“My name is Sim and I have an interest in teaching ML, NLP and DL. I am also a Youtuber.”

The above is a corpus

Tokenization is process where we take either a paragraph or sentences and creates tokens from it.Its the process of converting corpus into sentences and then converting to vocabulary

The documents or sentences formed from this corpus are:

1. My name is Sim and I have an interest in teaching ML, NLP and DL.
2. I am also a Youtuber.

These are the 2 tokens formed from a paragraph

1. My
2. name
3. is
4. Sim
5. and
6. I
7. have
8. an
9. interest
10. in
11. teaching
12. ML
13. ,
14. NLP
15. and
16. DL

These are the unique tokens formed from the first sentence.

Vocabulary are unique words in a document

E.g.:

“I like to drink Apple juice. My friend likes mango juice.”

1. I like to drink Apple juice.
2. My friend likes mango juice.

Total number of words: 11

Unique words:10 (Juice is getting repeated)

Difference between spacey and NLTK.