Basic Terminology in NLP

1. Corpus
2. Documents
3. Vocabulary
4. Words

Corpus

1. Corpus is large collection of text documents
2. Example:- If we are analysing customer reviews for product

```
Corpus = [ "The product is great",
           "I love this product",
           "product did.... expectation"
         ]
```

Document

A document is an individual piece of text within the corpus

```
Corpus = [ R1 ,
           R2 ,
           R3
         ]
```

Vocabulary

It is the set of unique words or token in the entire corpus without duplicates

```
Vocabulary = {
    "The", "product", "is", "great" . . .
    . . .  .  .
}
```

Words (Tokens)

It is the individual components of document