Terms in text preprocessing:

1. Corpus → paragraph.
2. Document → sentences from corpus.
3. vocabulary → set of unique words from document.
4. Words → All the words from corpus

Tokenization :

Corpus → { This is DL class and we are covering NLP.
It is organized by learnbay. Taken by Suresh }

↓

sentence tokenization ⇒ corpus ⇒ document

↓

document → { This is DL class and we are covering NLP. ⇒ document-1
It is organized by learnbay. ⇒ document-2
Taken by Suresh. ⇒ document-3

↓

word tokenization → document ⇒ words.

↓

words

{ The cat drink the milk. The dog chases the cat to drink the milk. } ⇒ corpus

The cat drink the milk. ⇒ document-1 (d-1)

The dog chases the cat to drink the milk. ⇒ document-2 (d-2)

[The, cat, drink, the, milk] ⇒ words from d-1

[The, dog, chases, the, cat, to, drink, the, milk] ⇒ words from d-2

[The, cat, drink, the, milk, dog, chases, to] ⇒ vocabulary