1 What is Bag of Words

Bag of Words is a method to extract features from text documents. These features can be used for training machine learning algorithms. In BoW model, a sentence or a document is considered as a 'Bag' containing words. It takes into account the words and their frequency of occurrence in the sentence or the document disregarding semantic relationship in the sentences.

It creates a vocabulary of all the unique words occurring in all the documents in the training set.

For example, if you have 3 documents

- D1 - "I am feeling very happy today"
- D2 - "I am not well today"
- D3 - "I wish I could go to play"

First, it creates a vocabulary using unique words from all the documents
Unique list of words -
I am feeling very happy today not well wish could go to play
Then, the frequency of the word in the corresponding document is inserted.

	I	am	feeling	very	happy	today	not	well	wish	could	go	to	play
D1	1	1	1	1	1	1	0	0	0	0	0	0	0
D2	1	1	0	0	0	1	1	1	0	0	0	0	0
D3	2	0	0	0	0	0	0	0	1	1	1	1	1

The above table depicts the training features containing frequencies of each word in each document. This is called bag-of-words approach since the number of occurrence and not sequence or order of words is what matters in this approach.

Bag of words has two major issues:

- It has the dimensionality issue as the total dimension is the vocabulary size. It can easily overfit your model. The remedy is to use some well-known dimensionality reduction technique to your input data.
- Bag of words representation doesn't consider the semantic relation between words. Generally, the neighbor words in a sentence should be useful for predicting your target word.