Feature Generation for Inferential Statistics

Use Case : SMS Spam Classification

BOW and TF-IDF

Bag of words will convert the text into the word vector by finding out the occurrence of the words in the document.

TF-IDF(Term frequency — Inverse document frequency (TFIDF) : - TF-IDF measures the importance of the word in the document

Given the dataset below,

Message (Feature)	Label
Welcome to world of AI	spam
NLP is interesting	ham
NLG is more interesting	ham
Welcome to world of NLP	spam

Document

Corpus

Step 1: Extraction of all unique words from the corpus.

Extraction of all unique words from the corpus

[welcome, to, world, of, AI, NLP, is, interesting, NLG, more]

Step 2: Removal of stop words from the unique words

Removal of the stopwords

[Welcome, world, AI, NLP, interesting, NLG]

Vocabulary

Once the stopwords from the unique words are removed, Vocabulary is developed.

Step 3 : Creation of Bag of Words

Vocabulary

	Welcome	**world**	**AI**	**NLP**	**interesting**	**NLG**
Welcome to world of AI	**1**	**1**	**1**	**0**	**0**	**0**
NLP is interesting	**0**	**0**	**0**	**1**	**1**	**0**
NLG is more interesting	**0**	**0**	**0**	**0**	**1**	**1**
Welcome to NLP	**1**	**0**	**0**	**1**	**0**	**0**

Frequency : The occurrence of the word in the document

Document Term Matrix/Bag of Words

Step 4: Calculation of Term Frequency

TF(word) = The occurrence of word in the document /Total no of words in the document.

	Welcome	world	AI	NLP	interesting	NLG
Welcome to world of AI	1/5 = 0.2	1/5 =0.2	1/5 = 0.2	0/5 = 0	0/5 =0	0/5 =0
NLP is interesting	0/3 = 0	0/3 =0	0/3 = 0	1/3 =0.33	1/3 =0.33	0/3 =0
NLG is more interesting	0/4 = 0	0/4 =0	0/4 =0	0/4 =0	¼ =0.25	¼ =0.25
Welcome to NLP	1/3 =0.33	0/3 =0	0/3 =0	1/3 =0.33	0/3 =0	0/3 =0

Step 4: Calculation of Inverse Document Frequency

IDF(word) = log[Total number of documents/number of documents which will contain the particular word]

IDF 0.3 0.6 0.6 0.3 0.3 0.6

	Welcome	world	AI	NLP	interesting	NLG
Welcome to world of AI	0.06	0.12	0.12	0	0	0
NLP is interesting	0	0	0	0.09	0.09	0
NLG is more interesting	0	0	0	0	0.07	0.15
Welcome to NLP	0.09	0	0	0.09	0	0

TF-IDF(Word)   = TF(word) * IDF(word)

The value in the document term matrix shows the importance of that word in the document.

Step 5 : Consider the TF-IDF object as feature for building ML/DL Model

BOW will convert your text features into word vectors. But It will only find out the occurrence of word in the document .

On the other hand, TF-IDF will find out the importance of word in the document.