Bag of Words: (BOW) ⇒ frequency of words

Tom is travelling in bus
John is travelling in bus
Emma took train and bus

D1 ⇒ Tom travelling bus.
D2 ⇒ John travelling bus.
D3 ⇒ Emma took train bus.

Vocabulary:

Tom - 1
travelling - 2
bus - 3
John - 1
Emma - 1
took - 1
train - 1

[Words are arranged in descending order of frequency and kept in the bag.]

[bus, travelling, tom, john, Emma, took, train]

[Bus travelling tom John Emma took train] → took

1 1 1 0 0 0 0 ⟶ D-1

1 1 0 1 0 0 0 ⟶ D-2

1 0 0 0 1 1 1 ⟶ D-3

New sentence → Tom took bus then train then bus

→ OOV ←

2 0 1 0 0 1 1

Disadvantage:

* Sparse matrix
* Words are shuffled based on frequency.
* semantic meaning is unclear.
* OOV.

Advantage:

* No need of fixed length document.
* Easy implementation & cost effective.

⇒ TF-IDF :

↳ [Term Frequency - Inverse Document Frequency]

① Term Frequency: Focused on words in each sentence

$$TF : \frac{\#\text{ repetation of words in sentence}}{\#\text{ words in the sentence}}$$

② Inverse document Frequency: Focused on the sentences in the corpus.

$$IDF : \log_e\left(\frac{\#\text{ sentences}}{\#\text{ sentences containing the word}}\right)$$

D-1 : Good boy

D-2 : Good girl

D-3 : Boy girl good

[Good, boy, girl] ⇒ vocabulary

Term Frequency — step:1

	D1	D2	D3
Good	1/2	1/2	1/3
boy	1/2	0	1/3
girl	0	1/2	1/3

IDF : — step:2

- Good → $\log_e\left(\frac{3}{3}\right)$
- boy → $\log_e(3/2)$
- girl → $\log_e(3/2)$

step-3

	Good	boy	girl
D-1	0	$1/2 \times \log_e(3/2)$	0
D-2	0	0	$1/2 \times \log_e(3/2)$
D-3	0	$1/3 \times \log_e(3/2)$	$1/3 \times \log_e(3/2)$

Disadvantage

- → OOV
- → Sparse matrix

Advantage:

- → word importance is captured.
- → length of sentences can be dynamic.
- → Easy to implement.

Word Frequency and encoding:

- → one hot encoding
- → Bag of Words (BOW)
- → TF-IDF

} Common Issue

- ✗ sparse matrix
- ✗ OOV

To fix the above issue Google come up with the idea 

of deep learning trained library in 2013

↳ ANN

 ⇒ Word2Vec – CBOW (Continious bag of words)

vocabulary ⇒	Boy	Girl	king	Queen	Mango	Apple	Women
Gender	1	−1	0.98	−0.97	0.01	0.02	−0.99
Royal	0.01	0.03	0.99	0.98	0.02	−0.01	0.01
Age	0.9	0.91	0.9	0.89	0.4	0.45	0.86
Food	0.01	0.01	0.02	0.02	0.99	0.96	0.01

⇓

dimensions

⇒ Gender [king − boy + Queen] = Woman

⇒ Gender [0.98 − 1 − 0.97] ⇒ 0.99 (Coor is fixed)

	Mango	Apple	Orange	Grapes
Gender	0.01	−0.03	0.02	0.06
Food	0.97	0.96	0.94	0.92

→ Gender [Mango − Apple + orange]

Gender [0.01 + 0.03 + 0.02] ⇒ 0.06 ⇒ Grapes doesnot belong to Gender

→ Food [Mango − Apple + orange]

Food [0.97 − 0.96 + 0.94] ⇒ 0.92 ⇒ Grapes belongs to Food