Bag-of-words model

Represent each **document** as a **bag of words**, ignoring words' ordering. Why? For **simplicity**.

Unstructured text becomes **a vector of numbers**

e.g., docs: "I like visualization", "I like data".

1 : "I"

2 : "like"

3 : "data"

4 : "visualization"

"I like visualization" ➡ [1, 1, 0, 1]

"I like data" ➡ [1, 1, 1, 0]