



Since we are considering eucledian distance hence power is 2

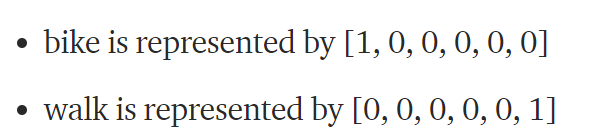
**ONE HOT VECTOR APPROACH**:

1**.Vocabulary**= bike, bicycle, car, jeep, truck, walk

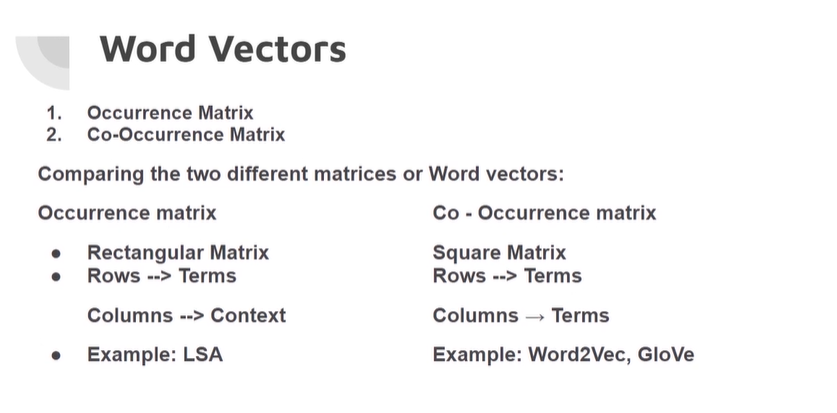
2.**one-hot vector** — a vector that contains 1’s in a single value and 0 in the other values

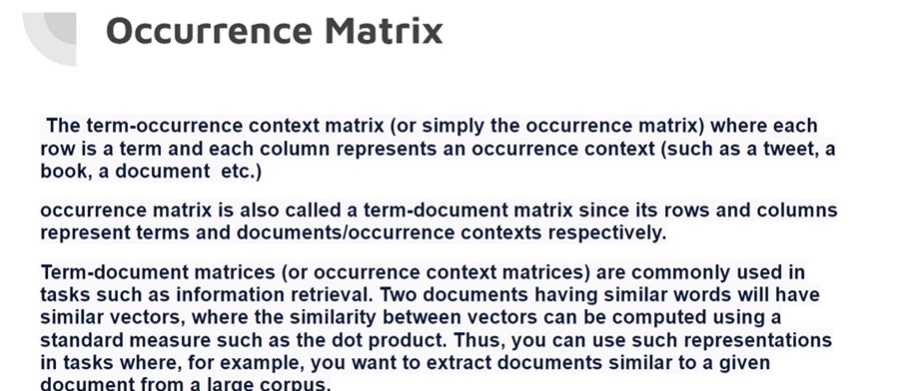
3. **sort our vocabulary**- bike, bicycle, car, jeep, truck, walk

4. substituting our words for 0, converting the elements above to an array format



5.

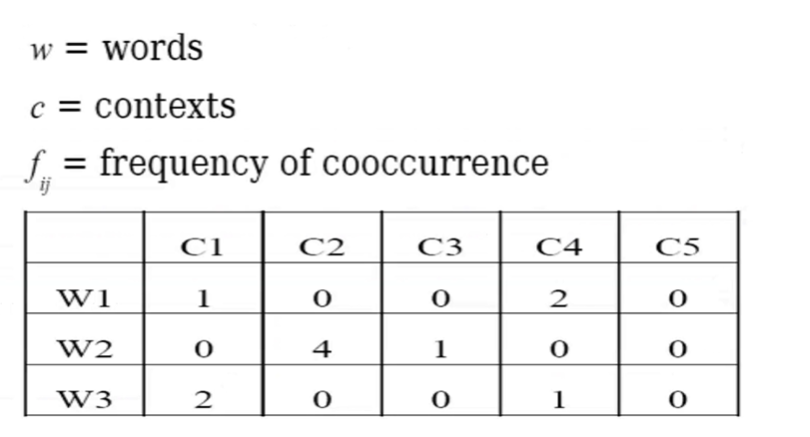




ROWS-🡪 documents TERMS/WORDS

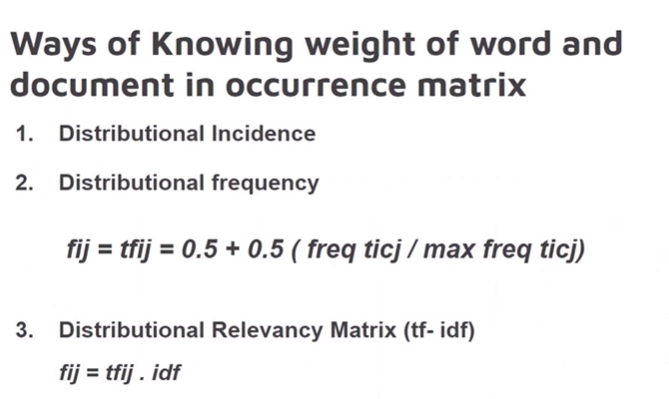
COLUMNS🡪 correspond to DOCUMENTS/CONTEXT

A term-document matrix **represents the relationship between terms and documents**, where each row stands for a term and each column for a document, and an entry is the number of occurrences of the term in the document

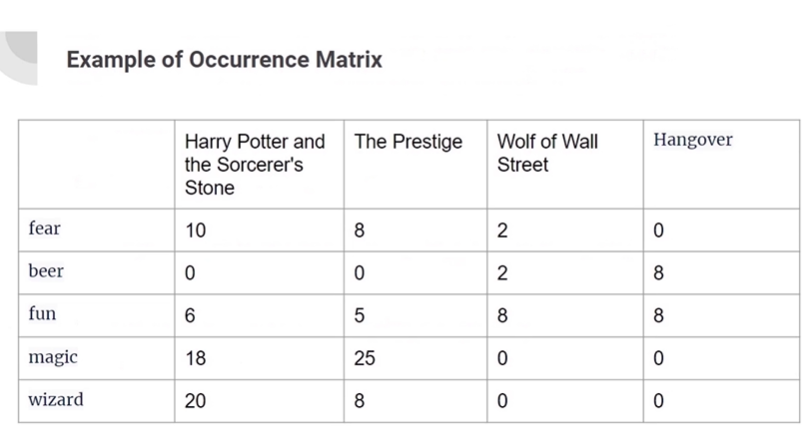


W1 is occurring in c1 1 time

W2 is occurring in C2 4 times



* **DISTRIBUTIONAL INCIDENCE**🡪W2 is having high weight as its 4 in C2



Movie reviews taken—words taken—occurrence matrix created

**Latent Semantic Analysis**

1.LSA is **an unsupervised learning technique** that rests on two pillars:

* The **distributional hypothesis**, which states that words with similar meanings appear frequently together. This is best summarised by JR Firth’s quote “You shall know a word by the company it keeps”
* **Singular Value Decomposition (SVD)**

2.Like all Machine Learning concepts, LSA can be broken down into 3 parts: the intuition, the maths and the code.

**Problem for occurrence matrix:**

1.For Large number of documents there will be large number of matrix-the matrix will become more larger and larger and will become sparce.

“Curse of dimensionality”

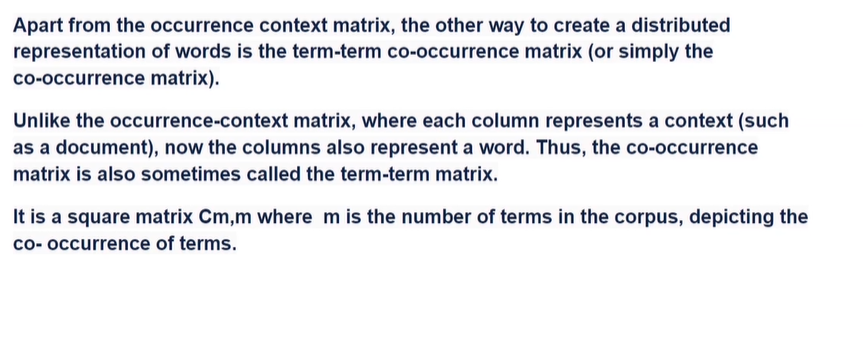
2.Polysemic word-A given word has different meaning based on the kind of sentence a word is used.

Eg: Java🡪 programming language/coffee type/ island name

**CO-OCCURRENCE MATRIX/TERM-TERM MATRIX**

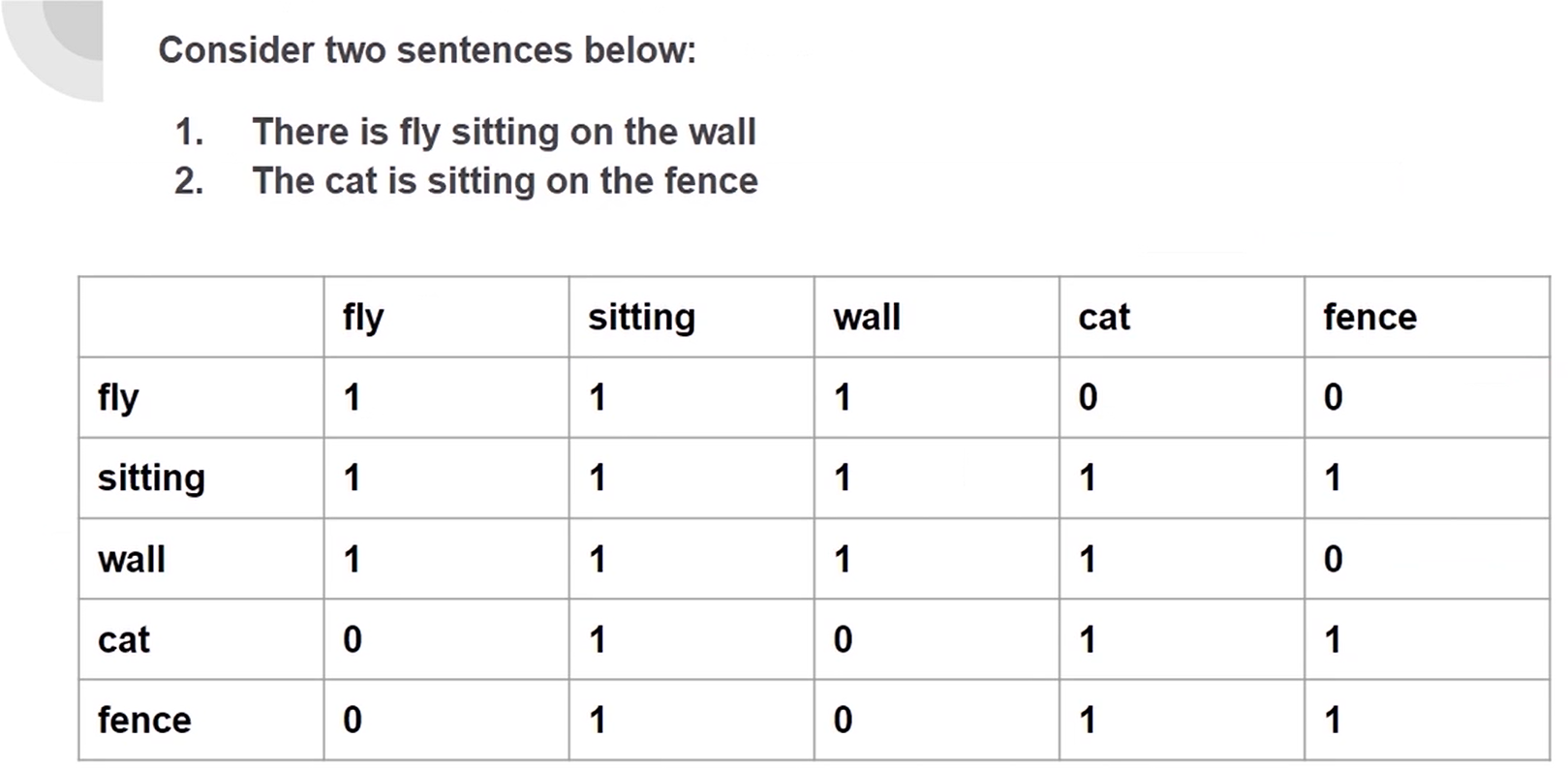
Columns🡪words

Rows🡪words



Cm,m🡪m is the terms in the corpus

METHOD 1:



Thumb rule:a given term is said to co-occur if It occurring in the similar occurrence context

Step 1:removing the stop words-

Here It is :there,is on the

Step 2:if same context-then 1 is given

**METHOD 2:**

1**.Vocabulary**= I’m riding in my car to the beach.

I’m riding in my jeep to the beach.

My car is a jeep.

My jeep is a car.

I ate a banana yesterday.

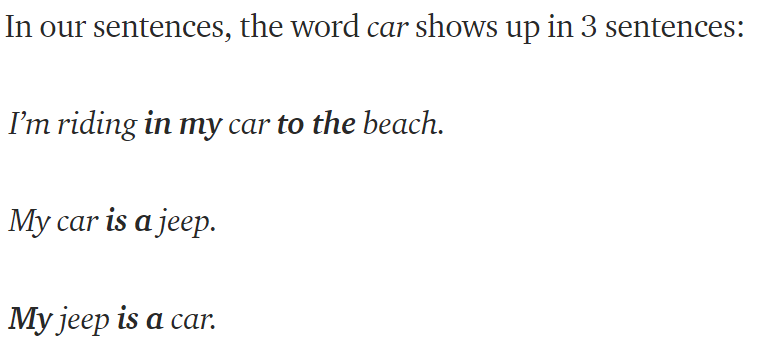
I ate a peach yesterday.

2. vocabulary of our group of sentences from the example above,removing stop words and considering a single grama, ate, banana, beach, car, in, is, I’m, jeep, my, riding, to, the, yesterday

3. **sort our vocabulary-** ate, banana, beach, car, in, is, I’m, jeep, my, riding, to, the, yesterday

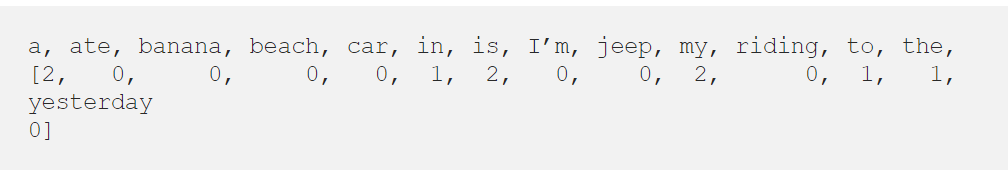
**4 CAR**

* initialize an example vector for the word car 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

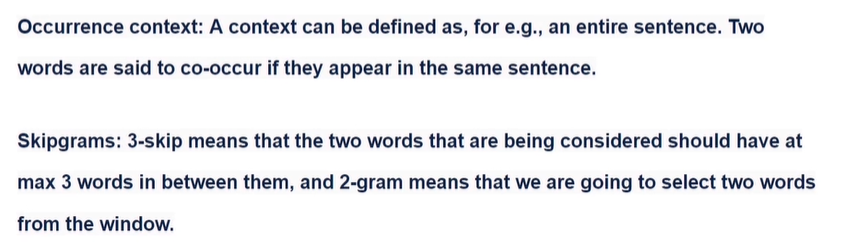


highlighted all the words that co-occur with the word car in the sentences.

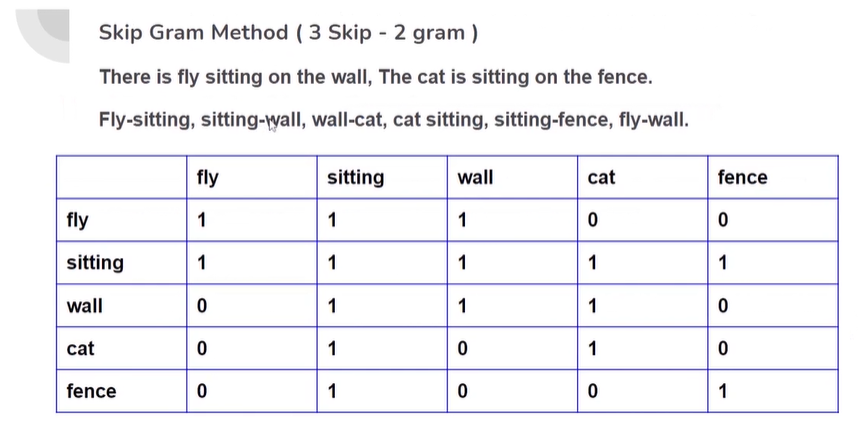
* Updating our vector for the word car by adding 1 to highlighted words



**Methods to creat co-occurrence matrix**







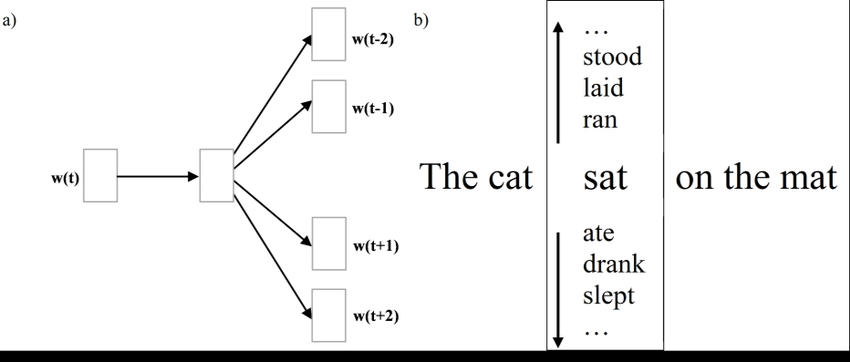
Target word🡪input

Context word🡪output

**Word representation** represents the word in vector space so that if the word vectors are close to one another means that those words are related to one other

As the vocabulary of any language is large and cannot be labeled by human and hence we require unsupervised learning techniques that can learn the context of any word on its own. Skip-gram is one of the unsupervised learning techniques **used to find the most related words for a given word**.

Skip-gram is **used to predict the context word for a given target word**. It’s **reverse of CBOW** algorithm. Here**, target word is input while context words are output**. As there is more than one context word to be predicted which makes this problem difficult.



The **word sat will be given** and we’ll try to predict words cat, mat at position -1 and 3 respectively given sat is at position 0 . We do not predict common or stop words such as the .