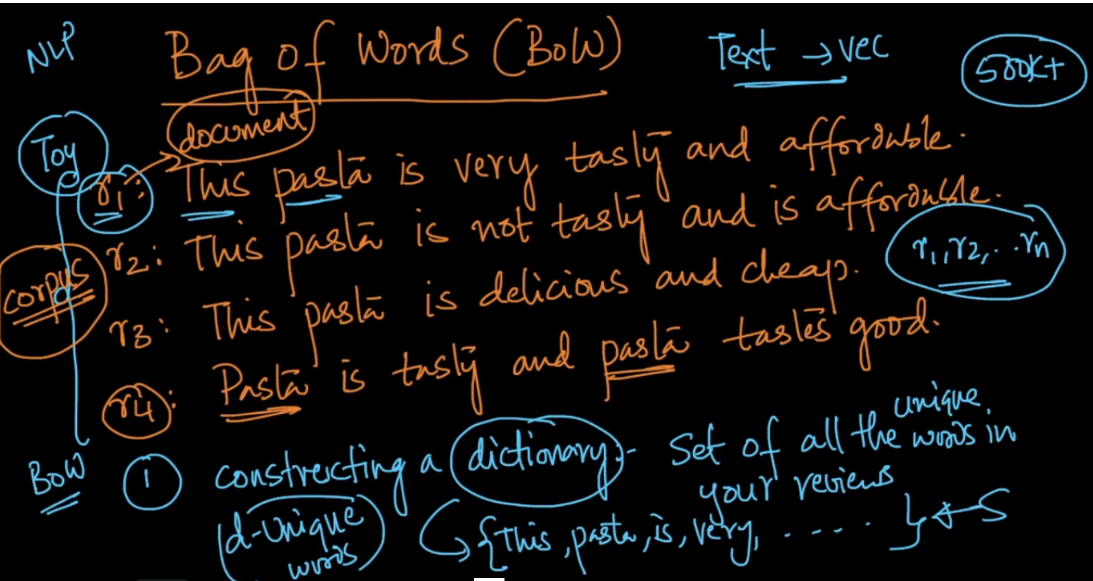
**Bag of Words (BoW)**

In nlp each sentence(in our example reviews) is known as document i.e here r1,r2,r3,r4 are different documents.

In BOW we construct a dictionary which is a set of all the unique words in reviews/document.

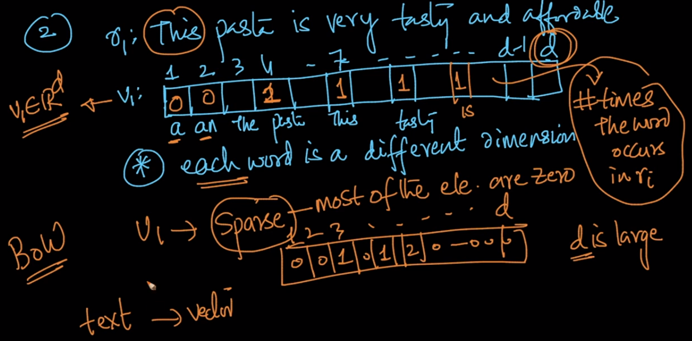
Suppose there are d-unique words in all documents, then there will d dimensions in each vector

Group of documents is known as corpus



Below image shows the vector generation in BOW, where each dimension corresponds to a unique word, and the vector will contains the occurrence of each word.

Generated vector will be sparse vector, as some of them will be empty or contains 0, as that word would not be in corresponding sentence.



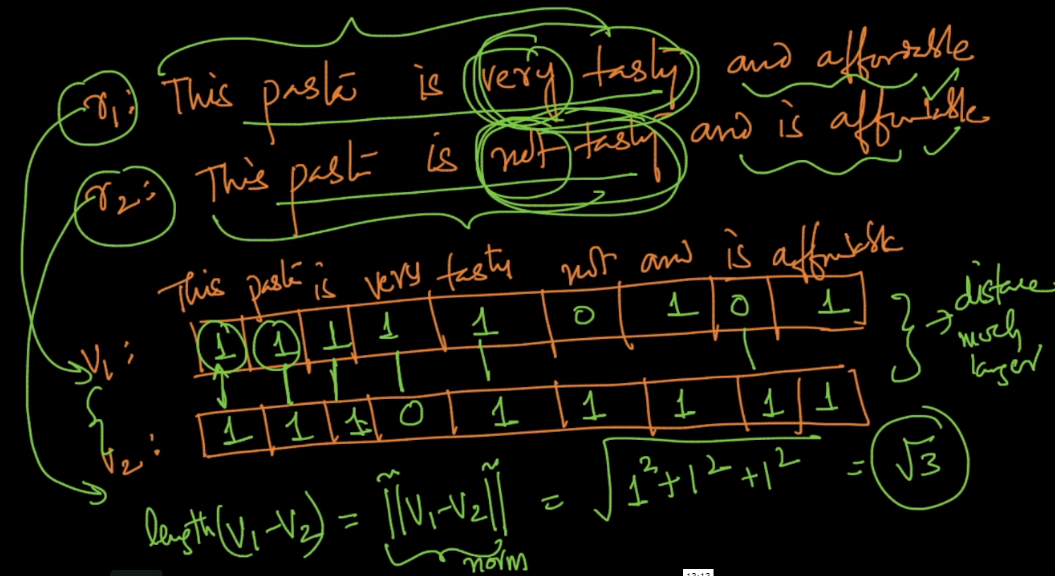
Below image shows the vector generation, given two sentences.

Correction: here ‘is’ are repeated which should be only one time, and therefore

for v1, ‘is’ will contain 1

for v2 ,‘is’ will contain 2.

Distance between both the vectors will be calculated as given in below figure.



**Limitation of BOW:**

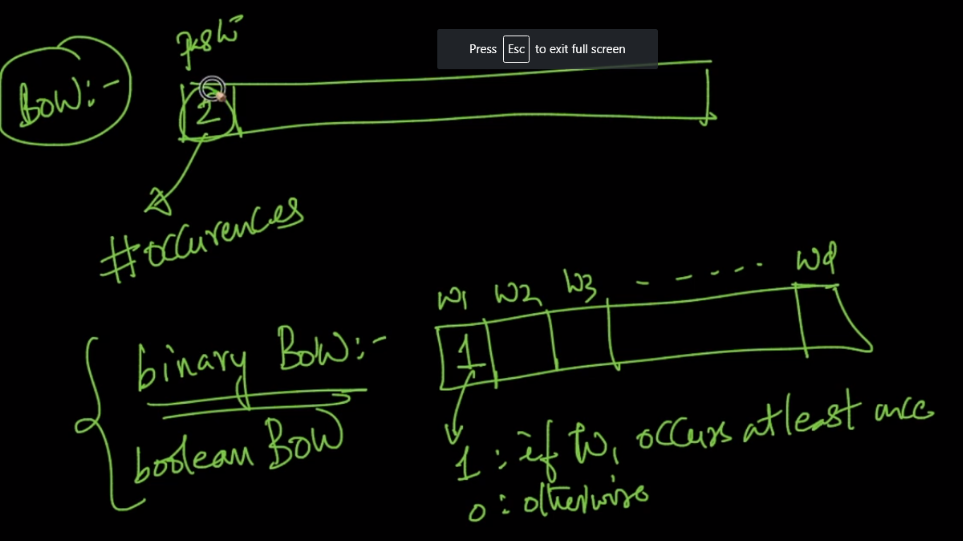
As you can see in above the distance calculate between those two vectors is very samll that mean they are very close to each other, but while reading those in english they both are giving different meaning, hence if there are slight changes in statements which make them different then BOW would not be able to recognize them.

**Binary BOW:**

In this each dimension will contain only two values 0 or 1

1 if value is present, no matter how many time it’s there in sentence.

0 if value is not present.



In Binary BOW, distance between two vectors are calculated as square root of no. of differing words.

