In [1]:

```
import nltk
import numpy as np
import re
from nltk import punkt
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize, sent_tokenize
from nltk.stem import PorterStemmer
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.preprocessing import OneHotEncoder
```

In [2]:

```
file1 = open("bag.txt","r")
text = file1.read()
print("The text is : ",text)
text1 = sent_tokenize(text)
i = 0
filteredText = []
visited = []
stop_words = set(stopwords.words('english'))
ps = PorterStemmer()
```

```
The text is :  I am eating orange icecream but I love chocolate icecream.
```

In [3]:

```
while(i<len(text1)):
    tokens = word_tokenize(text1[i])
    for token in tokens:
        if token not in (stop_words and visited) and token.isalpha():
            visited.append(token)
            filteredText.append(ps.stem(token))
        i += 1
print('The preprocessed text is:\n', filteredText)
doc = np.array(text1).reshape(-1,1)
print("The documents are: \n",doc)

vectorizer = CountVectorizer()
bow = vectorizer.fit_transform(text1)
print("The vocabulary is :\n",vectorizer.get_feature_names_out())
print("\nThe bag of Words is:\n",bow.toarray())
```

```
The preprocessed text is:
 ['i', 'am', 'eat', 'orang', 'icecream', 'but', 'love', 'chocol']
The documents are:
 [['I am eating orange icecream but I love chocolate icecream.']]
The vocabulary is :
 ['am' 'but' 'chocolate' 'eating' 'icecream' 'love' 'orange']

The bag of Words is:
 [[1 1 1 1 2 1 1]]
```

In [57]:

```
encoder = OneHotEncoder()
filteredText = np.array(filteredText).reshape(-1,1)
ohe = encoder.fit_transform(filteredText)
print("The One Hot Encoding : \n", ohe.toarray())
```

```
The One Hot Encoding :
 [[0. 1. 0.]
 [1. 0. 0.]
 [0. 0. 1.]]
```