bag of words will help us for sentiment analysis

-Bag of words steps

- 1:Lowering the words
    - tokenization: converting the paragraphs into sentenses
        - histogram: to count the frequency of the words
            - sort the hist to convert high num to low number words
            - filtering the words
            - take more frequntly occuring words
                - creating the matrix , in this sentence will be provided in the columns ## bag of words is nothing but the document matrix #### jump in to the problem

In [1]:
```
import nltk
```

In [2]:
```
paragraph =  """I have three visions for India. In 3000 years of our history, people from al
l over
               the world have come and invaded us, captured our lands, conquered our minds.
               From Alexander onwards, the Greeks, the Turks, the Moguls, the Portuguese, th
e British,
               the French, the Dutch, all of them came and looted us, took over what was our
s.
               Yet we have not done this to any other nation. We have not conquered anyone.
               We have not grabbed their land, their culture,
               their history and tried to enforce our way of life on them.
               Why? Because we respect the freedom of others.That is why my
               first vision is that of freedom. I believe that India got its first vision of
               this in 1857, when we started the War of Independence. It is this freedom tha
t
               we must protect and nurture and build on. If we are not free, no one will res
pect us.
               My second vision for India’s development. For fifty years we have been a deve
loping nation.
               It is time we see ourselves as a developed nation. We are among the top 5 nat
ions of the world
               in terms of GDP. We have a 10 percent growth rate in most areas. Our poverty
 levels are falling.
               Our achievements are being globally recognised today. Yet we lack the self-co
nfidence to
               see ourselves as a developed nation, self-reliant and self-assured. Isn’t thi
s incorrect?
               I have a third vision. India must stand up to the world. Because I believe th
at unless India
               stands up to the world, no one will respect us. Only strength respects streng
th. We must be
               strong not only as a military power but also as an economic power. Both must
 go hand-in-hand.
               My good fortune was to have worked with three great minds. Dr. Vikram Sarabha
i of the Dept. of
               space, Professor Satish Dhawan, who succeeded him and Dr. Brahm Prakash, fath
er of nuclear material.
               I was lucky to have worked with all three of them closely and consider this t
he great opportunity of my life.
               I see four milestones in my career"""
```

In [4]:
```
##cleaning the text
import re
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
from nltk.stem import WordNetLemmatizer
```

In [7]:
```
ps=PorterStemmer()
wordnet=WordNetLemmatizer()
sentences = nltk.sent_tokenize(paragraph)
corpus = []
for i in range(len(sentences)):
    review = re.sub('[^a-zA-Z]', ' ', sentences[i])
    review = review.lower()
    review = review.split()
    review = [ps.stem(word) for word in review if not word in set(stopwords.words('english'
))]
    review = ' '.join(review)
    corpus.append(review)

# Creating the Bag of Words model
from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer(max_features = 1500)
X = cv.fit_transform(corpus).toarray()
```

In [9]:
```
X
```

Out[9]:
```
array([[0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 1, 1, 0],
       [0, 1, 0, ..., 0, 0, 0],
       ...,
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0]])
```

In [10]:
```
cv
```

Out[10]:
```
CountVectorizer(analyzer='word', binary=False, decode_error='strict',
                dtype=<class 'numpy.int64'>, encoding='utf-8', input='content',
                lowercase=True, max_df=1.0, max_features=1500, min_df=1,
                ngram_range=(1, 1), preprocessor=None, stop_words=None,
                strip_accents=None, token_pattern='(?u)\\b\\w\\w+\\b',
                tokenizer=None, vocabulary=None)
```

In [11]:
```
corpus
```

Out[11]:
```
['three vision india',
 'year histori peopl world come invad us captur land conquer mind',
 'alexand onward greek turk mogul portugues british french dutch came loot us took',
 'yet done nation',
 'conquer anyon',
 'grab land cultur histori tri enforc way life',
 '',
 'respect freedom other first vision freedom',
 'believ india got first vision start war independ',
 'freedom must protect nurtur build',
 'free one respect us',
 'second vision india develop',
 'fifti year develop nation',
 'time see develop nation',
 'among top nation world term gdp',
 'percent growth rate area',
 'poverti level fall',
 'achiev global recognis today',
 'yet lack self confid see develop nation self reliant self assur',
 'incorrect',
 'third vision',
 'india must stand world',
 'believ unless india stand world one respect us',
 'strength respect strength'
```