Term Frequency and Inverse Document Frequency

- there is a problem with BOF so we choose TFIDf the probelm is that
- all the words will have same importance
- no semantic information is preserved
- semantic =relating to meaning in language or logic.

TFIDF Solution

- some semantic info is preserves as uncommon words are given more importance than common words
- ex : she is beautiful here beautiful is given more importance than she or is
- this is also same as bag of words like
- same steps
- tfidf formula
- num of occurances of the words in the doc/
- num of words in the doc ## IDF Formula:
- log(num of docs)->sentenses/
- (num of docs containing words) ## TF * IDF

in to the programm

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

In [2]:
```
paragraph="""I have three visions for India. In 3000 years of our history, people from all o
ver
               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 [3]:
```
## cleaning the text
```

In [5]:
```
import re
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from nltk.stem import WordNetLemmatizer
```

In [8]:
```
# intializing the objects
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 = [wordnet.lemmatize(word) for word in review if not word in set(stopwords.words(
'english'))]
    review = ' '.join(review)
    corpus.append(review)

# Creating the TF-IDF model
from sklearn.feature_extraction.text import TfidfVectorizer
cv = TfidfVectorizer()
X = cv.fit_transform(corpus).toarray()
```

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

Out[9]:
```
array([[0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.25883507, 0.30512561,
        0.        ],
       [0.        , 0.28867513, 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]:
```
TfidfVectorizer(analyzer='word', binary=False, decode_error='strict',
                dtype=<class 'numpy.float64'>, encoding='utf-8',
                input='content', lowercase=True, max_df=1.0, max_features=None,
                min_df=1, ngram_range=(1, 1), norm='l2', preprocessor=None,
                smooth_idf=True, stop_words=None, strip_accents=None,
                sublinear_tf=False, token_pattern='(?u)\\b\\w\\w+\\b',
                tokenizer=None, use_idf=True, vocabulary=None)
```

In [11]:
```
sentences
```

Out[11]:
```
['I have three visions for India.',
 'In 3000 years of our history, people from all over \n               the world have come and
invaded us, captured our lands, conquered our minds.',
 'From Alexander onwards, the Greeks, the Turks, the Moguls, the Portuguese, the British,\n
the French, the Dutch, all of them came and looted us, took over what was ours.',
 'Yet we have not done this to any other nation.',
 'We have not conquered anyone.',
 'We have not grabbed their land, their culture, \n               their history and tried to
enforce our way of life on them.',
 'Why?',
 'Because we respect the freedom of others.That is why my \n               first vision is th
at of freedom.',
 'I believe that India got its first vision of \n               this in 1857, when we started
the War of Independence.',
 'It is this freedom that\n               we must protect and nurture and build on.',
 'If we are not free, no one will respect us.',
 'My second vision for India’s development.',
 'For fifty years we have been a developing nation.',
 'It is time we see ourselves as a developed nation.',
 'We are among the top 5 nations of the world\n               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-confidence to\n               see ourselves as a developed nation, sel
f-reliant and self-assured.',
```