Word 2 vec

- when we have a sentense we convert it into BOW and TF-IDF ### problems with them
- both bow and tf-idf approach simantic information is not stored ,tf-idf gives importance to uncommonn words
- thre is defiently chance of overfitting ### to over come this problem we use Word2Vec
- in this specific model each word is basically represented as a vector of 32 or more dimenstions insted of single num
- here the semantic info and relation b/w diff words is also presented
- steps to create bOW
- tokenization of sents
- create hist
- take most frequent words
- create matrix with all unique words, it also represents the occurance relation b/w the words
- gensim is the meathod

In [1]:
```
pip install gensim
```

```
Collecting gensim
  Downloading gensim-3.8.1-cp36-cp36m-macosx_10_6_intel.macosx_10_9_intel.macosx_10_9_x86_64.
macosx_10_10_intel.macosx_10_10_x86_64.whl (24.7 MB)
     |████████████████████████████████| 24.7 MB 1.8 MB/s eta 0:00:01    |███████████▏
| 9.0 MB 636 kB/s eta 0:00:25     |████████████                    | 9.9 MB 1.4 MB/s eta 0:0
0:11     |██████████████████████|          | 17.1 MB 1.2 MB/s eta 0:00:07
Requirement already satisfied: scipy>=0.18.1 in ./Downloads/anaconda3/lib/python3.6/site-pack
ages (from gensim) (1.4.1)
Requirement already satisfied: numpy>=1.11.3 in ./Downloads/anaconda3/lib/python3.6/site-pack
ages (from gensim) (1.18.1)
Collecting smart-open>=1.8.1
  Downloading smart_open-1.10.0.tar.gz (99 kB)
     |████████████████████████████████| 99 kB 1.6 MB/s eta 0:00:01
Requirement already satisfied: six>=1.5.0 in ./Downloads/anaconda3/lib/python3.6/site-package
s (from gensim) (1.14.0)
Requirement already satisfied: requests in ./Downloads/anaconda3/lib/python3.6/site-packages
(from smart-open>=1.8.1->gensim) (2.22.0)
Collecting boto3
  Downloading boto3-1.12.30-py2.py3-none-any.whl (128 kB)
     |████████████████████████████████| 128 kB 1.0 MB/s eta 0:00:01
Collecting google-cloud-storage
  Downloading google_cloud_storage-1.26.0-py2.py3-none-any.whl (75 kB)
     |████████████████████████████████| 75 kB 2.8 MB/s eta 0:00:011
Requirement already satisfied: chardet<3.1.0,>=3.0.2 in ./Downloads/anaconda3/lib/python3.6/s
ite-packages (from requests->smart-open>=1.8.1->gensim) (3.0.4)
Requirement already satisfied: certifi>=2017.4.17 in ./Downloads/anaconda3/lib/python3.6/site
-packages (from requests->smart-open>=1.8.1->gensim) (2019.11.28)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in ./Downloads/anacond
a3/lib/python3.6/site-packages (from requests->smart-open>=1.8.1->gensim) (1.25.8)
Requirement already satisfied: idna<2.9,>=2.5 in ./Downloads/anaconda3/lib/python3.6/site-pac
kages (from requests->smart-open>=1.8.1->gensim) (2.8)
Collecting botocore<1.16.0,>=1.15.30
  Downloading botocore-1.15.30-py2.py3-none-any.whl (6.0 MB)
     |████████████████████████████████| 6.0 MB 35 kB/s eta 0:00:012
Collecting s3transfer<0.4.0,>=0.3.0
  Downloading s3transfer-0.3.3-py2.py3-none-any.whl (69 kB)
     |████████████████████████████████| 69 kB 632 kB/s eta 0:00:01
Collecting jmespath<1.0.0,>=0.7.1
  Downloading jmespath-0.9.5-py2.py3-none-any.whl (24 kB)
Collecting google-resumable-media<0.6dev,>=0.5.0
  Downloading google_resumable_media-0.5.0-py2.py3-none-any.whl (38 kB)
Collecting google-auth<2.0dev,>=1.11.0
  Downloading google_auth-1.12.0-py2.py3-none-any.whl (83 kB)
     |████████████████████████████████| 83 kB 902 kB/s eta 0:00:011
Collecting google-cloud-core<2.0dev,>=1.2.0
  Downloading google_cloud_core-1.3.0-py2.py3-none-any.whl (26 kB)
Requirement already satisfied: python-dateutil<3.0.0,>=2.1 in ./Downloads/anaconda3/lib/pytho
n3.6/site-packages (from botocore<1.16.0,>=1.15.30->boto3->smart-open>=1.8.1->gensim) (2.8.1)
Collecting docutils<0.16,>=0.10
  Downloading docutils-0.15.2-py3-none-any.whl (547 kB)
     |████████████████████████████████| 547 kB 574 kB/s eta 0:00:01
Collecting cachetools<5.0,>=2.0.0
  Downloading cachetools-4.0.0-py3-none-any.whl (10 kB)
Collecting rsa<4.1,>=3.1.4
  Downloading rsa-4.0-py2.py3-none-any.whl (38 kB)
Collecting pyasn1-modules>=0.2.1
  Downloading pyasn1_modules-0.2.8-py2.py3-none-any.whl (155 kB)
     |████████████████████████████████| 155 kB 1.5 MB/s eta 0:00:01
Requirement already satisfied: setuptools>=40.3.0 in ./Downloads/anaconda3/lib/python3.6/site
-packages (from google-auth<2.0dev,>=1.11.0->google-cloud-storage->smart-open>=1.8.1->gensim)
(45.2.0.post20200210)
Collecting google-api-core<2.0.0dev,>=1.16.0
  Downloading google_api_core-1.16.0-py2.py3-none-any.whl (70 kB)
     |████████████████████████████████| 70 kB 1.4 MB/s eta 0:00:011
Collecting pyasn1>=0.1.3
  Downloading pyasn1-0.4.8-py2.py3-none-any.whl (77 kB)
     |████████████████████████████████| 77 kB 6.4 MB/s  eta 0:00:01
Requirement already satisfied: pytz in ./Downloads/anaconda3/lib/python3.6/site-packages (fro
m google-api-core<2.0.0dev,>=1.16.0->google-cloud-core<2.0dev,>=1.2.0->google-cloud-storage->
smart-open>=1.8.1->gensim) (2019.3)
Collecting protobuf>=3.4.0
  Downloading protobuf-3.11.3-cp36-cp36m-macosx_10_9_x86_64.whl (1.3 MB)
     |████████████████████████████████| 1.3 MB 40 kB/s eta 0:00:013
Collecting googleapis-common-protos<2.0dev,>=1.6.0
  Downloading googleapis-common-protos-1.51.0.tar.gz (35 kB)
Building wheels for collected packages: smart-open, googleapis-common-protos
  Building wheel for smart-open (setup.py) ... done
  Created wheel for smart-open: filename=smart_open-1.10.0-py3-none-any.whl size=90632 sha256
=d1a8db64193879872846e8ca4e26b08d844eae1d6e4b4de8156fd749c9d6349b
  Stored in directory: /Users/abhilashavadhanula/Library/Caches/pip/wheels/a1/f3/8a/c46924c44
ee290a85840f98cab840a65669cd377175eaf6850
  Building wheel for googleapis-common-protos (setup.py) ... done
  Created wheel for googleapis-common-protos: filename=googleapis_common_protos-1.51.0-py3-no
ne-any.whl size=77592 sha256=8089ec434737c44351f3501df90cbb79b5491a6e1f76f6d9db91481d7d883379
  Stored in directory: /Users/abhilashavadhanula/Library/Caches/pip/wheels/35/8d/af/a922cb188
00b31fadac3523cadf6c1efdf233b788fe7a4da70
Successfully built smart-open googleapis-common-protos
Installing collected packages: docutils, jmespath, botocore, s3transfer, boto3, google-resuma
ble-media, cachetools, pyasn1, rsa, pyasn1-modules, google-auth, protobuf, googleapis-common-
protos, google-api-core, google-cloud-core, google-cloud-storage, smart-open, gensim
  Attempting uninstall: docutils
    Found existing installation: docutils 0.16
    Uninstalling docutils-0.16:
      Successfully uninstalled docutils-0.16
Successfully installed boto3-1.12.30 botocore-1.15.30 cachetools-4.0.0 docutils-0.15.2 gensim
-3.8.1 google-api-core-1.16.0 google-auth-1.12.0 google-cloud-core-1.3.0 google-cloud-storage
-1.26.0 google-resumable-media-0.5.0 googleapis-common-protos-1.51.0 jmespath-0.9.5 protobuf-
3.11.3 pyasn1-0.4.8 pyasn1-modules-0.2.8 rsa-4.0 s3transfer-0.3.3 smart-open-1.10.0
Note: you may need to restart the kernel to use updated packages.
```

In [1]:
```
import nltk

from gensim.models import Word2Vec
from nltk.corpus import stopwords

import re

paragraph = """I have three visions for India. In 3000 years of our history, people from all
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"""



# Preprocessing the data
text = re.sub(r'\[[0-9]*\]',' ',paragraph)
text = re.sub(r'\s+',' ',text)
text = text.lower()
text = re.sub(r'\d',' ',text)
text = re.sub(r'\s+',' ',text)

# Preparing the dataset
sentences = nltk.sent_tokenize(text)

sentences = [nltk.word_tokenize(sentence) for sentence in sentences]

for i in range(len(sentences)):
    sentences[i] = [word for word in sentences[i] if word not in stopwords.words('english')]


# Training the Word2Vec model
model = Word2Vec(sentences, min_count=1)


words = model.wv.vocab

# Finding Word Vectors
vector = model.wv['war']

# Most similar words
similar = model.wv.most_similar('vikram')
```

In [2]:
```
similar
```

Out[2]:
```
[('history', 0.26455605030059814),
 ('gdp', 0.2242116630077362),
 ('third', 0.21682527661323547),
 ('self-reliant', 0.20927684009075165),
 ('people', 0.19927534461021423),
 ('done', 0.19690413773059845),
 ('minds', 0.19126096367835999),
 ('dr.', 0.17590805888175964),
 ('believe', 0.1676056981086731),
 ('respect', 0.15933313965797424)]
```

In [3]:
```
vector
```

Out[3]:
```
array([-4.5318590e-03,  3.9077303e-03,  4.9251230e-03, -7.3886500e-04,
       -2.7530566e-03, -1.7321950e-03,  2.8165595e-03, -4.0270106e-04,
        1.0794778e-03,  3.7667805e-03,  7.9818092e-05,  2.2256102e-03,
       -1.5923602e-03, -2.8615631e-03, -2.2300135e-03,  4.1088308e-03,
       -4.8616785e-04, -4.5985845e-03,  3.3402930e-03,  4.1668382e-03,
       -4.2176475e-03, -4.3101367e-03,  4.5713536e-03, -2.4332015e-03,
        2.0950753e-03, -2.5563878e-03,  1.8732026e-03, -2.3130984e-03,
       -6.4515212e-04, -2.0438337e-03,  1.1258164e-03,  3.7262372e-03,
        3.7458553e-03, -4.3479649e-06,  4.7960710e-03, -2.7725238e-03,
        1.5816654e-03,  3.1074921e-03,  9.4872325e-05, -3.2664370e-03,
        1.0409479e-03, -2.3774337e-03, -2.9540593e-03, -4.7876020e-03,
        1.1785326e-03, -1.9264962e-03,  3.5236487e-03, -7.7402154e-05,
        5.7641073e-04,  1.4490491e-03, -4.8192483e-03,  4.2212061e-03,
        4.1004941e-03,  2.9684880e-03,  1.9664699e-03, -1.1106387e-03,
       -1.9017430e-03, -8.8893395e-04, -4.6617552e-03, -4.6043387e-03,
        9.5787289e-04, -2.9777535e-03,  1.7592536e-03,  4.9229972e-03,
        3.2170434e-03,  4.1672643e-03, -2.1173768e-03,  4.7129500e-03,
        4.8544756e-03,  4.5516109e-03,  2.1604808e-04,  4.7104419e-03,
       -2.4606988e-03,  2.2666468e-03,  7.3137000e-04,  3.2101059e-03,
       -7.0529192e-04,  2.9802339e-03,  4.9997196e-03,  4.9064485e-03,
        7.0919783e-04,  2.5308217e-04, -3.4874890e-03, -2.6927161e-04,
       -4.5297756e-03, -4.7746766e-03,  1.2374456e-03, -1.3823274e-03,
        3.9448626e-03,  3.1529733e-03, -1.8025811e-03, -1.2900915e-03,
       -4.8636994e-03,  3.5871994e-03, -2.6756686e-03,  5.6660030e-04,
        8.3965564e-04,  4.8806081e-03, -2.9548877e-03, -4.1412874e-04],
      dtype=float32)
```

In [4]:
```
sentences
```

Out[4]:
```
[['three', 'visions', 'india', '.'],
 ['years',
  'history',
  ',',
  'people',
  'world',
  'come',
  'invaded',
  'us',
  ',',
  'captured',
  'lands',
  ',',
  'conquered',
  'minds',
  '.'],
 ['alexander',
  'onwards',
  ',',
  'greeks',
  ',',
  'turks',
  ',',
  'moguls',
  ',',
  'portuguese',
  ',',
  'british',
  ',',
  'french',
  ',',
  'dutch',
  ',',
  'came',
  'looted',
  'us',
  ',',
  'took',
  '.'],
 ['yet', 'done', 'nation', '.'],
 ['conquered', 'anyone', '.'],
 ['grabbed',
  'land',
  ',',
  'culture',
  ',',
  'history',
  'tried',
  'enforce',
  'way',
  'life',
  '.'],
 ['?'],
 ['respect', 'freedom', 'others.that', 'first', 'vision', 'freedom', '.'],
 ['believe',
  'india',
  'got',
  'first',
  'vision',
  ',',
  'started',
  'war',
  'independence',
  '.'],
 ['freedom', 'must', 'protect', 'nurture', 'build', '.'],
 ['free', ',', 'one', 'respect', 'us', '.'],
 ['second', 'vision', 'india', '’', 'development', '.'],
 ['fifty', 'years', 'developing', 'nation', '.'],
 ['time', 'see', 'developed', 'nation', '.'],
 ['among', 'top', 'nations', 'world', 'terms', 'gdp', '.'],
 ['percent', 'growth', 'rate', 'areas', '.'],
 ['poverty', 'levels', 'falling', '.'],
 ['achievements', 'globally', 'recognised', 'today', '.'],
 ['yet',
  'lack',
  'self-confidence',
  'see',
  'developed',
  'nation',
  ',',
  'self-reliant',
  'self-assured',
  '.'],
 ['’', 'incorrect', '?'],
 ['third', 'vision', '.'],
 ['india', 'must', 'stand', 'world', '.'],
 ['believe',
  'unless',
  'india',
  'stands',
  'world',
  ',',
  'one',
  'respect',
  'us',
  '.'],
 ['strength', 'respects', 'strength', '.'],
 ['must', 'strong', 'military', 'power', 'also', 'economic', 'power', '.'],
 ['must', 'go', 'hand-in-hand', '.'],
 ['good', 'fortune', 'worked', 'three', 'great', 'minds', '.'],
 ['dr.', 'vikram', 'sarabhai', 'dept', '.'],
 ['space',
  ',',
  'professor',
  'satish',
  'dhawan',
  ',',
  'succeeded',
  'dr.',
  'brahm',
  'prakash',
  ',',
```