Untitled1

October 26, 2019

1 Homework 2 Question 1 : Word2vec - Word Embeddings

1.0.1 Github repo : https://github.com/anjaliverma96/avw4127_msia414_2019

In [2]:
```
#### WRITE TEXT FILES FROM 5 SUB-FOLDERS THAT WERE PART OF THE NEWSGROUP FOLDER INTO O

import os
os.chdir("20-newsgroups/")

import glob
read_files = glob.glob("*.txt")

with open("final_text.txt", "wb") as outfile:
    i=0
    for f in read_files:
        i+=1
        if(i==5):
            break
        with open(f, "rb") as infile:
            outfile.write(infile.read())

#### READ IN THE FINAL TEXT FILE

with open("final_text.txt",encoding="utf8", errors='ignore') as file:
    test_text = file.read()
```

1.1 Pre-Processing

In [3]:
```
#### Create a list of all emails in the text by splitting on the word 'Newsgroup:' (Si
email_list = test_text.split("Newsgroup:")
```

In [4]:
```
#### Example of an email from the list
email_list[5000]
```

Out[4]:
```
" misc.forsale\nDocument_id: 75856\nFrom: mmm@cup.portal.com (Mark Robert Thorson)\nSu
```

In [41]:
```
#### Import necessary libraries
 import re
```

```
import datetime
import string
import nltk
from nltk.tokenize import sent_tokenize, word_tokenize, RegexpTokenizer
from nltk.corpus import stopwords
from nltk.stem.wordnet import WordNetLemmatizer
from nltk.stem.porter import PorterStemmer
from nltk.stem.wordnet import WordNetLemmatizer
import numpy as np
```

In [6]:
```
#### Download necessary packages
nltk.download()
nltk.download('stopwords')
nltk.download('punkt')
SENT_DETECTOR = nltk.data.load('tokenizers/punkt/english.pickle')
```

```
showing info https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/index.xml


[nltk_data] Downloading package stopwords to
[nltk_data]     /Users/anjaliverma/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to
[nltk_data]     /Users/anjaliverma/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

In [7]:
```
#### Define function that performs the following pre processing steps at once :

def preprocessing_nltk(text):

    now = datetime.datetime.now()

    #### Create tokenizer that only tokenizes alpha-numeric words
    tokenizer = RegexpTokenizer(r'\w+')

    #### Convert text to lower case, tokenize text and remove numeric tokens
    revised_tokens =[word for word in tokenizer.tokenize(text.lower()) if word.isalpha

    #### Remove stopwords
    words = [w for w in revised_tokens if w not in stopwords.words('english')]

    #### Lemmatize tokens obtained after removing stopwords
    wnl = WordNetLemmatizer()
    tagged = nltk.pos_tag(words)
    lem_list = []
    for word, tag in tagged:
        wntag = tag[0].lower()
        wntag = wntag if wntag in ['a', 'r', 'n', 'v'] else None
```

```
                if not wntag:
                    lemma = word
                else:
                    lemma = wnl.lemmatize(word, wntag)
                lem_list.append(lemma)
            #lem_text = " ".join(lemma for lemma in lem_list)

            #print("Took %s"%(datetime.datetime.now()-now))

            return lem_list
```

```
In [8]: #### Example of a pre-processed email within the document
        doc = email_list[4000]
        print(preprocessing_nltk(doc))
```

```
['misc', 'forsale', 'subject', 'diamond', 'stealth', 'svga', 'sale', 'cleveland', 'freenet', 'c
```

```
In [9]: #### Loop through each email in the email list to preprocess each email
        #### Store each list within list processed_emails
        processed_emails = []
        for i in range(len(email_list)):
            processed_emails.append(preprocessing_nltk(email_list[i]))
```

```
In [10]: del processed_emails[0]
```

```
In [11]: print(processed_emails[500])
```

```
['sci', 'crypt', 'rschnapp', 'metaflow', 'com', 'rus', 'schnapp', 'subject', 'tap', 'code', 'g
```

```
In [12]: #### Write the preprocessed emails to a text file
         with open("anjali_verma_preprocessed_emails.txt", "w") as fobj:
             for x in processed_emails:
                 doc = " ".join(lemma for lemma in x)
                 fobj.write(doc + "\n")
```

1.2 Word2Vec : Creating word embeddings

```
In [13]: #### Example of the data in desired format for word embeddings
         #### The dataset is in the form of a list of list of tokens for each document (Newsgr
         print(processed_emails[500:502])
```

```
[['sci', 'crypt', 'rschnapp', 'metaflow', 'com', 'rus', 'schnapp', 'subject', 'tap', 'code', 'g
```

```
In [14]: import gensim, logging
         from gensim.models import Word2Vec
         logging.basicConfig(format='%(asctime)s : %(levelname)s : %(message)s', level=logging
```

```
In [46]: def cosine_distance (model, word,target_list , num) :
             cosine_dict ={}
             word_list = []
             a = model[word]
             for item in target_list :
                 if item != word :
                     b = model [item]
                     cos_sim = np.dot(a, b)/(np.linalg.norm(a)*np.linalg.norm(b))
                     cosine_dict[item] = cos_sim
             dist_sort=sorted(cosine_dict.items(), key=lambda dist: dist[1],reverse = True) ##
             for item in dist_sort:
                 word_list.append((item[0], item[1]))
             return word_list[0:num]

In [35]: target_list= list(set([y for x in processed_emails for y in x ]))

In [37]: print(target_list[0:10])

['esp', 'spiffy', 'pixie', 'supertwist', 'extroverted', 'posteriorly', 'chow', 'repudiate', 'ir
```

1.3 SKIP-GRAM MODEL (model parameter sg = 1)

```
In [66]: #### train word2vec
         model = Word2Vec(processed_emails, min_count=1,size= 50,workers=3, window =3, sg = 1)

2019-10-26 22:20:29,259 : WARNING : consider setting layer size to a multiple of 4 for greater
2019-10-26 22:20:29,261 : INFO : collecting all words and their counts
2019-10-26 22:20:29,263 : INFO : PROGRESS: at sentence #0, processed 0 words, keeping 0 word ty
2019-10-26 22:20:29,503 : INFO : collected 30557 word types from a corpus of 1122850 raw words
2019-10-26 22:20:29,503 : INFO : Loading a fresh vocabulary
2019-10-26 22:20:29,739 : INFO : effective_min_count=1 retains 30557 unique words (100% of orig
2019-10-26 22:20:29,740 : INFO : effective_min_count=1 leaves 1122850 word corpus (100% of orig
2019-10-26 22:20:29,840 : INFO : deleting the raw counts dictionary of 30557 items
2019-10-26 22:20:29,841 : INFO : sample=0.001 downsamples 22 most-common words
2019-10-26 22:20:29,842 : INFO : downsampling leaves estimated 1089714 word corpus (97.0% of pr
2019-10-26 22:20:29,941 : INFO : estimated required memory for 30557 words and 50 dimensions: 1
2019-10-26 22:20:29,942 : INFO : resetting layer weights
2019-10-26 22:20:30,253 : INFO : training model with 3 workers on 30557 vocabulary and 50 featu
2019-10-26 22:20:31,269 : INFO : EPOCH 1 - PROGRESS: at 60.21% examples, 606820 words/s, in_qsi
2019-10-26 22:20:31,990 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-10-26 22:20:32,002 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-10-26 22:20:32,012 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-10-26 22:20:32,013 : INFO : EPOCH - 1 : training on 1122850 raw words (1089767 effective w
2019-10-26 22:20:33,030 : INFO : EPOCH 2 - PROGRESS: at 60.21% examples, 606468 words/s, in_qsi
2019-10-26 22:20:33,744 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-10-26 22:20:33,747 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-10-26 22:20:33,754 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-10-26 22:20:33,755 : INFO : EPOCH - 2 : training on 1122850 raw words (1089567 effective w
```

```
2019-10-26 22:20:34,780 : INFO : EPOCH 3 - PROGRESS: at 65.60% examples, 639189 words/s, in_qsi
2019-10-26 22:20:35,455 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-10-26 22:20:35,469 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-10-26 22:20:35,493 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-10-26 22:20:35,494 : INFO : EPOCH - 3 : training on 1122850 raw words (1089755 effective w
2019-10-26 22:20:36,514 : INFO : EPOCH 4 - PROGRESS: at 55.53% examples, 576437 words/s, in_qsi
2019-10-26 22:20:37,283 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-10-26 22:20:37,297 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-10-26 22:20:37,309 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-10-26 22:20:37,310 : INFO : EPOCH - 4 : training on 1122850 raw words (1089753 effective w
2019-10-26 22:20:38,321 : INFO : EPOCH 5 - PROGRESS: at 62.97% examples, 629001 words/s, in_qsi
2019-10-26 22:20:39,004 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-10-26 22:20:39,014 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-10-26 22:20:39,023 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-10-26 22:20:39,024 : INFO : EPOCH - 5 : training on 1122850 raw words (1089867 effective w
2019-10-26 22:20:39,025 : INFO : training on a 5614250 raw words (5448709 effective words) tool
```

```
In [67]: model.save("word2vec1.model")
         model = Word2Vec.load("word2vec1.model")
```

```
2019-10-26 22:20:43,677 : INFO : saving Word2Vec object under word2vec1.model, separately None
2019-10-26 22:20:43,678 : INFO : not storing attribute vectors_norm
2019-10-26 22:20:43,679 : INFO : not storing attribute cum_table
2019-10-26 22:20:43,886 : INFO : saved word2vec1.model
2019-10-26 22:20:43,887 : INFO : loading Word2Vec object from word2vec1.model
2019-10-26 22:20:44,031 : INFO : loading wv recursively from word2vec1.model.wv.* with mmap=Non
2019-10-26 22:20:44,031 : INFO : setting ignored attribute vectors_norm to None
2019-10-26 22:20:44,032 : INFO : loading vocabulary recursively from word2vec1.model.vocabulary
2019-10-26 22:20:44,033 : INFO : loading trainables recursively from word2vec1.model.trainables
2019-10-26 22:20:44,033 : INFO : setting ignored attribute cum_table to None
2019-10-26 22:20:44,035 : INFO : loaded word2vec1.model
```

```
In [49]: w1 = 'databases'
```

```
In [50]: model[w1]
```

```
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:1: DeprecationWa
  """Entry point for launching an IPython kernel.
```

```
Out[50]: array([-0.16538486, -0.22892961, -0.12038412,  0.39569497,  0.10224129,
                 0.11696152,  0.41309643,  0.15227576, -0.0053171 ,  0.329913  ,
                 0.00425125,  0.14172778, -0.04164088, -0.01574311,  0.269511  ,
                 0.2694517 ,  0.2946141 , -0.05229336,  0.02670644, -0.5384178 ,
                 0.2846486 ,  0.00451329, -0.22964086,  0.29623625, -0.12663527,
                -0.36935067, -0.03723044, -0.37925726, -0.34062475, -0.40986732,
                -0.26562893, -0.20815687, -0.15960026, -0.03676778,  0.05217892,
```

```
       0.33948576, -0.3676621 , -0.09999277,  0.06539803, -0.21091425,
       0.15911348,  0.21590903,  0.50502086, -0.06880483, -0.22672614,
      -0.0215673 ,  0.33174634,  0.3337871 ,  0.35224402, -0.2900794 ],
     dtype=float32)
```

```
In [51]: model.similarity(w1, 'computer')
```

```
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:1: DeprecationWa
  """Entry point for launching an IPython kernel.
```

```
Out[51]: 0.5401499
```

```
In [52]: model.wv.most_similar(positive = w1,topn=5)
```

```
Out[52]: [('payware', 0.9192549586296082),
          ('vital', 0.9152900576591492),
          ('adminstrator', 0.9141732454299927),
          ('uncompressor', 0.9132228493690491),
          ('factoring', 0.9109092950820923)]
```

```
In [53]: cosine_distance (model,'databases',target_list,5)
```

```
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:4: DeprecationWa
  after removing the cwd from sys.path.
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:7: DeprecationWa
  import sys
```

```
Out[53]: [('payware', 0.919255),
          ('vital', 0.91529),
          ('adminstrator', 0.9141732),
          ('uncompressor', 0.91322285),
          ('factoring', 0.9109094)]
```

1.4 Changing model parameter window form 3 to 5

```
In [68]: model1 = Word2Vec(processed_emails, min_count=1,size= 50,workers=3, window =5, sg = 1)
         model1.save("word2vec2.model")
         model1 = Word2Vec.load("word2vec2.model")
```

```
2019-10-26 22:21:01,935 : WARNING : consider setting layer size to a multiple of 4 for greater
2019-10-26 22:21:01,937 : INFO : collecting all words and their counts
2019-10-26 22:21:01,937 : INFO : PROGRESS: at sentence #0, processed 0 words, keeping 0 word ty
2019-10-26 22:21:02,194 : INFO : collected 30557 word types from a corpus of 1122850 raw words
2019-10-26 22:21:02,195 : INFO : Loading a fresh vocabulary
2019-10-26 22:21:02,258 : INFO : effective_min_count=1 retains 30557 unique words (100% of orig
2019-10-26 22:21:02,260 : INFO : effective_min_count=1 leaves 1122850 word corpus (100% of orig
2019-10-26 22:21:02,364 : INFO : deleting the raw counts dictionary of 30557 items
```

```
2019-10-26 22:21:02,366 : INFO : sample=0.001 downsamples 22 most-common words
2019-10-26 22:21:02,366 : INFO : downsampling leaves estimated 1089714 word corpus (97.0% of pr
2019-10-26 22:21:02,460 : INFO : estimated required memory for 30557 words and 50 dimensions: 1
2019-10-26 22:21:02,461 : INFO : resetting layer weights
2019-10-26 22:21:02,750 : INFO : training model with 3 workers on 30557 vocabulary and 50 featu
2019-10-26 22:21:03,775 : INFO : EPOCH 1 - PROGRESS: at 40.53% examples, 460818 words/s, in_qsi
2019-10-26 22:21:04,802 : INFO : EPOCH 1 - PROGRESS: at 92.23% examples, 472590 words/s, in_qsi
2019-10-26 22:21:05,080 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-10-26 22:21:05,100 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-10-26 22:21:05,119 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-10-26 22:21:05,120 : INFO : EPOCH - 1 : training on 1122850 raw words (1089889 effective w
2019-10-26 22:21:06,135 : INFO : EPOCH 2 - PROGRESS: at 28.47% examples, 373311 words/s, in_qsi
2019-10-26 22:21:07,148 : INFO : EPOCH 2 - PROGRESS: at 84.28% examples, 421828 words/s, in_qsi
2019-10-26 22:21:07,564 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-10-26 22:21:07,600 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-10-26 22:21:07,618 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-10-26 22:21:07,619 : INFO : EPOCH - 2 : training on 1122850 raw words (1089935 effective w
2019-10-26 22:21:08,641 : INFO : EPOCH 3 - PROGRESS: at 32.67% examples, 405770 words/s, in_qsi
2019-10-26 22:21:09,649 : INFO : EPOCH 3 - PROGRESS: at 87.36% examples, 444913 words/s, in_qsi
2019-10-26 22:21:10,003 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-10-26 22:21:10,015 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-10-26 22:21:10,035 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-10-26 22:21:10,036 : INFO : EPOCH - 3 : training on 1122850 raw words (1089538 effective w
2019-10-26 22:21:11,051 : INFO : EPOCH 4 - PROGRESS: at 31.85% examples, 398562 words/s, in_qsi
2019-10-26 22:21:12,058 : INFO : EPOCH 4 - PROGRESS: at 84.28% examples, 423006 words/s, in_qsi
2019-10-26 22:21:12,541 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-10-26 22:21:12,575 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-10-26 22:21:12,583 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-10-26 22:21:12,584 : INFO : EPOCH - 4 : training on 1122850 raw words (1089916 effective w
2019-10-26 22:21:13,597 : INFO : EPOCH 5 - PROGRESS: at 41.83% examples, 475363 words/s, in_qsi
2019-10-26 22:21:14,610 : INFO : EPOCH 5 - PROGRESS: at 93.51% examples, 487974 words/s, in_qsi
2019-10-26 22:21:14,820 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-10-26 22:21:14,846 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-10-26 22:21:14,865 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-10-26 22:21:14,866 : INFO : EPOCH - 5 : training on 1122850 raw words (1089618 effective w
2019-10-26 22:21:14,866 : INFO : training on a 5614250 raw words (5448896 effective words) took
2019-10-26 22:21:14,875 : INFO : saving Word2Vec object under word2vec2.model, separately None
2019-10-26 22:21:14,876 : INFO : not storing attribute vectors_norm
2019-10-26 22:21:14,877 : INFO : not storing attribute cum_table
2019-10-26 22:21:15,096 : INFO : saved word2vec2.model
2019-10-26 22:21:15,098 : INFO : loading Word2Vec object from word2vec2.model
2019-10-26 22:21:15,236 : INFO : loading wv recursively from word2vec2.model.wv.* with mmap=Non
2019-10-26 22:21:15,236 : INFO : setting ignored attribute vectors_norm to None
2019-10-26 22:21:15,237 : INFO : loading vocabulary recursively from word2vec2.model.vocabulary
2019-10-26 22:21:15,237 : INFO : loading trainables recursively from word2vec2.model.trainables
2019-10-26 22:21:15,238 : INFO : setting ignored attribute cum_table to None
2019-10-26 22:21:15,239 : INFO : loaded word2vec2.model
```

```
In [62]: print("Euclidean Similarity of word 'databases' to computer is: ", model.similarity(wi
         print(" ")
         print("Top 5 words similar to the given word according to euclidean similarity: ",mode
         print(" ")
         print("Top 5 words similar to the given word according to cosine similarity: ", cosine
```

```
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:1: DeprecationWa
  """Entry point for launching an IPython kernel.
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:4: DeprecationWa
  after removing the cwd from sys.path.
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:7: DeprecationWa
  import sys


Euclidean Similarity of word 'databases' to computer is:  0.5401499

Top 5 words similar to the given word according to euclidean similarity:  [('payware', 0.919254

Top 5 words similar to the given word according to cosine similarity:  [('payware', 0.919255),
```

1.5 CONTINUOUS BAG OF WORDS (model parameter sg = 0)

```
In [69]: #### train word2vec
         new_model = Word2Vec(processed_emails, min_count=1,size=50, workers=3, window=3, sg =
```

```
2019-10-26 22:21:30,374 : WARNING : consider setting layer size to a multiple of 4 for greater
2019-10-26 22:21:30,375 : INFO : collecting all words and their counts
2019-10-26 22:21:30,376 : INFO : PROGRESS: at sentence #0, processed 0 words, keeping 0 word ty
2019-10-26 22:21:30,610 : INFO : collected 30557 word types from a corpus of 1122850 raw words
2019-10-26 22:21:30,610 : INFO : Loading a fresh vocabulary
2019-10-26 22:21:30,663 : INFO : effective_min_count=1 retains 30557 unique words (100% of orig
2019-10-26 22:21:30,663 : INFO : effective_min_count=1 leaves 1122850 word corpus (100% of orig
2019-10-26 22:21:30,763 : INFO : deleting the raw counts dictionary of 30557 items
2019-10-26 22:21:30,765 : INFO : sample=0.001 downsamples 22 most-common words
2019-10-26 22:21:30,766 : INFO : downsampling leaves estimated 1089714 word corpus (97.0% of pr
2019-10-26 22:21:30,851 : INFO : estimated required memory for 30557 words and 50 dimensions: 1
2019-10-26 22:21:30,852 : INFO : resetting layer weights
2019-10-26 22:21:31,137 : INFO : training model with 3 workers on 30557 vocabulary and 50 featu
2019-10-26 22:21:31,881 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-10-26 22:21:31,888 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-10-26 22:21:31,890 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-10-26 22:21:31,890 : INFO : EPOCH - 1 : training on 1122850 raw words (1089837 effective w
2019-10-26 22:21:32,692 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-10-26 22:21:32,701 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-10-26 22:21:32,702 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-10-26 22:21:32,703 : INFO : EPOCH - 2 : training on 1122850 raw words (1089652 effective w
2019-10-26 22:21:33,444 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-10-26 22:21:33,452 : INFO : worker thread finished; awaiting finish of 1 more threads
```

```
2019-10-26 22:21:33,453 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-10-26 22:21:33,453 : INFO : EPOCH - 3 : training on 1122850 raw words (1089582 effective w
2019-10-26 22:21:34,273 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-10-26 22:21:34,283 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-10-26 22:21:34,284 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-10-26 22:21:34,285 : INFO : EPOCH - 4 : training on 1122850 raw words (1089859 effective w
2019-10-26 22:21:35,047 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-10-26 22:21:35,056 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-10-26 22:21:35,057 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-10-26 22:21:35,058 : INFO : EPOCH - 5 : training on 1122850 raw words (1089919 effective w
2019-10-26 22:21:35,059 : INFO : training on a 5614250 raw words (5448849 effective words) took
```

```
In [70]: new_model.save("word2vec3.model")
         new_model = Word2Vec.load("word2vec3.model")
```

```
2019-10-26 22:21:44,765 : INFO : saving Word2Vec object under word2vec3.model, separately None
2019-10-26 22:21:44,766 : INFO : not storing attribute vectors_norm
2019-10-26 22:21:44,767 : INFO : not storing attribute cum_table
2019-10-26 22:21:44,980 : INFO : saved word2vec3.model
2019-10-26 22:21:44,981 : INFO : loading Word2Vec object from word2vec3.model
2019-10-26 22:21:45,343 : INFO : loading wv recursively from word2vec3.model.wv.* with mmap=Non
2019-10-26 22:21:45,343 : INFO : setting ignored attribute vectors_norm to None
2019-10-26 22:21:45,344 : INFO : loading vocabulary recursively from word2vec3.model.vocabulary
2019-10-26 22:21:45,345 : INFO : loading trainables recursively from word2vec3.model.trainables
2019-10-26 22:21:45,346 : INFO : setting ignored attribute cum_table to None
2019-10-26 22:21:45,347 : INFO : loaded word2vec3.model
```

```
In [55]: new_model[w1]
```

```
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:1: DeprecationWa
  """Entry point for launching an IPython kernel.
```

```
Out[55]: array([-0.02918519, -0.0337136 ,  0.08795231,  0.02421778,  0.1780647 ,
                 0.0553538 ,  0.12115074,  0.0975095 ,  0.05988046,  0.02197726,
                 0.03644676,  0.10601006, -0.00516792,  0.16420509,  0.12053949,
                 0.07873943,  0.05759237, -0.04323119,  0.0278474 , -0.11028142,
                 0.11981519, -0.0336413 ,  0.00403806,  0.2104455 , -0.03523133,
                -0.12024543, -0.10330275, -0.11238363, -0.06754733, -0.08081181,
                -0.10363467, -0.09096432,  0.00277428, -0.00078453,  0.091377  ,
                 0.20121473, -0.06839076, -0.06283233, -0.07026026, -0.08496032,
                 0.05678556,  0.10152854,  0.11724606,  0.07272658, -0.12158132,
                -0.01087536,  0.11207011,  0.11630771,  0.01763994, -0.06558477],
               dtype=float32)
```

```
In [56]: new_model.similarity(w1, 'computer')
```

```
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:1: DeprecationWa
  """Entry point for launching an IPython kernel.


Out[56]: 0.536373

In [57]: new_model.wv.most_similar(positive = w1,topn=5)

2019-10-26 21:18:49,626 : INFO : precomputing L2-norms of word weight vectors


Out[57]: [('linotronic', 0.9489021301269531),
          ('decoder', 0.9417294859886169),
          ('dma', 0.9396167397499084),
          ('protocols', 0.9371689558029175),
          ('conecting', 0.9359794855117798)]

In [58]: cosine_distance (new_model,'databases',target_list,5)

/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:4: DeprecationWa
  after removing the cwd from sys.path.
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:7: DeprecationWa
  import sys


Out[58]: [('linotronic', 0.94890213),
          ('decoder', 0.9417295),
          ('dma', 0.9396167),
          ('protocols', 0.93716884),
          ('conecting', 0.9359795)]
```

1.6 Changing model parameter window form 3 to 5

```
In [71]: new_model1 = Word2Vec(processed_emails, min_count=1,size= 50,workers=3, window =5, sg
         new_model1.save("word2vec4.model")
         new_model1 = Word2Vec.load("word2vec4.model")

2019-10-26 22:21:59,305 : WARNING : consider setting layer size to a multiple of 4 for greater
2019-10-26 22:21:59,306 : INFO : collecting all words and their counts
2019-10-26 22:21:59,307 : INFO : PROGRESS: at sentence #0, processed 0 words, keeping 0 word ty
2019-10-26 22:21:59,540 : INFO : collected 30557 word types from a corpus of 1122850 raw words
2019-10-26 22:21:59,540 : INFO : Loading a fresh vocabulary
2019-10-26 22:21:59,594 : INFO : effective_min_count=1 retains 30557 unique words (100% of orig
2019-10-26 22:21:59,595 : INFO : effective_min_count=1 leaves 1122850 word corpus (100% of orig
2019-10-26 22:21:59,697 : INFO : deleting the raw counts dictionary of 30557 items
2019-10-26 22:21:59,698 : INFO : sample=0.001 downsamples 22 most-common words
2019-10-26 22:21:59,699 : INFO : downsampling leaves estimated 1089714 word corpus (97.0% of pr
2019-10-26 22:21:59,789 : INFO : estimated required memory for 30557 words and 50 dimensions: 1
2019-10-26 22:21:59,790 : INFO : resetting layer weights
```

```
2019-10-26 22:22:00,054 : INFO : training model with 3 workers on 30557 vocabulary and 50 featu
2019-10-26 22:22:00,834 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-10-26 22:22:00,842 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-10-26 22:22:00,843 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-10-26 22:22:00,843 : INFO : EPOCH - 1 : training on 1122850 raw words (1089876 effective w
2019-10-26 22:22:01,677 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-10-26 22:22:01,691 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-10-26 22:22:01,692 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-10-26 22:22:01,693 : INFO : EPOCH - 2 : training on 1122850 raw words (1089591 effective w
2019-10-26 22:22:02,467 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-10-26 22:22:02,473 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-10-26 22:22:02,474 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-10-26 22:22:02,475 : INFO : EPOCH - 3 : training on 1122850 raw words (1089643 effective w
2019-10-26 22:22:03,329 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-10-26 22:22:03,339 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-10-26 22:22:03,341 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-10-26 22:22:03,342 : INFO : EPOCH - 4 : training on 1122850 raw words (1089645 effective w
2019-10-26 22:22:04,077 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-10-26 22:22:04,082 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-10-26 22:22:04,084 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-10-26 22:22:04,084 : INFO : EPOCH - 5 : training on 1122850 raw words (1089722 effective w
2019-10-26 22:22:04,085 : INFO : training on a 5614250 raw words (5448477 effective words) took
2019-10-26 22:22:04,092 : INFO : saving Word2Vec object under word2vec4.model, separately None
2019-10-26 22:22:04,094 : INFO : not storing attribute vectors_norm
2019-10-26 22:22:04,095 : INFO : not storing attribute cum_table
2019-10-26 22:22:04,290 : INFO : saved word2vec4.model
2019-10-26 22:22:04,291 : INFO : loading Word2Vec object from word2vec4.model
2019-10-26 22:22:04,433 : INFO : loading wv recursively from word2vec4.model.wv.* with mmap=Non
2019-10-26 22:22:04,434 : INFO : setting ignored attribute vectors_norm to None
2019-10-26 22:22:04,434 : INFO : loading vocabulary recursively from word2vec4.model.vocabulary
2019-10-26 22:22:04,435 : INFO : loading trainables recursively from word2vec4.model.trainables
2019-10-26 22:22:04,436 : INFO : setting ignored attribute cum_table to None
2019-10-26 22:22:04,437 : INFO : loaded word2vec4.model
```

```
In [64]: print("Euclidean Similarity of word 'databases' to computer is: ", new_model1.similari
         print(" ")
         print("Top 5 words similar to the given word according to euclidean similarity: ",new_
         print(" ")
         print("Top 5 words similar to the given word according to cosine similarity: ", cosine
```

```
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:1: DeprecationWa
  """Entry point for launching an IPython kernel.
2019-10-26 21:59:25,281 : INFO : precomputing L2-norms of word weight vectors
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:4: DeprecationWa
  after removing the cwd from sys.path.
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:7: DeprecationWa
  import sys
```

```
Euclidean Similarity of word 'databases' to computer is:  0.5324575

Top 5 words similar to the given word according to euclidean similarity:  [('multi', 0.9284030

Top 5 words similar to the given word according to cosine similarity:  [('multi', 0.9284031),
```

1.7 Comparison of 10 handpicked words

```
In [65]: w2 = "popular"
         w3 = "technology"
         w4 = "question"
         w5 = "information"
         w6 = "letter"
         w7 = "digit"
         w8 = "data"
         w9 = "prohibition"
         w10 = "faith"
```

1.8 Model : Parameter sg = 1, window = 3

```
In [83]: print(w2)
         print("")
         print("Euclidean Similarity of word " + w2 + " to computer is: ", model.similarity(w2
         print(" ")
         print("Top 5 words similar to the given word according to euclidean similarity: ",mode
         print(" ")
         print("Top 5 words similar to the given word according to cosine similarity: ", cosine
         print(" ")
         print(w3)
         print("")
         print("Euclidean Similarity of word " + w3 + " to computer is: ", model.similarity(w3
         print(" ")
         print("Top 5 words similar to the given word according to euclidean similarity: ",mode
         print(" ")
         print("Top 5 words similar to the given word according to cosine similarity: ", cosine
         print(" ")
         print(w4)
         print("")
         print("Euclidean Similarity of word " + w4 + " to computer is: ", model.similarity(w4
         print(" ")
         print("Top 5 words similar to the given word according to euclidean similarity: ",mode
         print(" ")
         print("Top 5 words similar to the given word according to cosine similarity: ", cosine
         print(" ")
         print(w5)
         print("")
         print("Euclidean Similarity of word " + w5 + " to computer is: ", model.similarity(w5
```

```
print(" ")
print("Top 5 words similar to the given word according to euclidean similarity: ",mode
print(" ")
print("Top 5 words similar to the given word according to cosine similarity: ", cosine
print(" ")
print(w6)
print("")
print("Euclidean Similarity of word " + w6 + " to computer is: ", model.similarity(w6
print(" ")
print("Top 5 words similar to the given word according to euclidean similarity: ",mode
print(" ")
print("Top 5 words similar to the given word according to cosine similarity: ", cosine
print(" ")
print(w7)
print("")
print("Euclidean Similarity of word " + w7 + " to computer is: ", model.similarity(w7
print(" ")
print("Top 5 words similar to the given word according to euclidean similarity: ",mode
print(" ")
print("Top 5 words similar to the given word according to cosine similarity: ", cosine
print(" ")
print(w8)
print("")
print("Euclidean Similarity of word " + w8 + " to computer is: ", model.similarity(w8
print(" ")
print("Top 5 words similar to the given word according to euclidean similarity: ",mode
print(" ")
print("Top 5 words similar to the given word according to cosine similarity: ", cosine
print(" ")
print(w9)
print("")
print("Euclidean Similarity of word " + w9 + " to computer is: ", model.similarity(w9
print(" ")
print("Top 5 words similar to the given word according to euclidean similarity: ",mode
print(" ")
print("Top 5 words similar to the given word according to cosine similarity: ", cosine
print(" ")
print(w10)
print("")
print("Euclidean Similarity of word " + w10 + " to computer is: ", model.similarity(w
print(" ")
print("Top 5 words similar to the given word according to euclidean similarity: ",mode
print(" ")
print("Top 5 words similar to the given word according to cosine similarity: ", cosine
```

```
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:3: DeprecationWa
  This is separate from the ipykernel package so we can avoid doing imports until
2019-10-26 22:43:35,166 : INFO : precomputing L2-norms of word weight vectors
```

```
popular

Euclidean Similarity of word popular to computer is:  0.47004202

Top 5 words similar to the given word according to euclidean similarity:  [('revision', 0.8567

/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:4: DeprecationWa
  after removing the cwd from sys.path.
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:7: DeprecationWa
  import sys

Top 5 words similar to the given word according to cosine similarity:  [('revision', 0.8567767)

technology

Euclidean Similarity of word technology to computer is:  0.3986991

Top 5 words similar to the given word according to euclidean similarity:  [('telecommunication

/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:11: DeprecationW
  # This is added back by InteractiveShellApp.init_path()

Top 5 words similar to the given word according to cosine similarity:  [('telecommunication', 0

question

Euclidean Similarity of word question to computer is:  0.33207655

Top 5 words similar to the given word according to euclidean similarity:  [('answer', 0.818938

/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:19: DeprecationW

Top 5 words similar to the given word according to cosine similarity:  [('answer', 0.8189388),

information

Euclidean Similarity of word information to computer is:  0.36308715
```

```
Top 5 words similar to the given word according to euclidean similarity:  [('omission', 0.70929

/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:27: DeprecationW

Top 5 words similar to the given word according to cosine similarity:  [('omission', 0.7092919
letter
Euclidean Similarity of word letter to computer is:  0.23286408
Top 5 words similar to the given word according to euclidean similarity:  [('morris', 0.732562

/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:35: DeprecationW

Top 5 words similar to the given word according to cosine similarity:  [('morris', 0.73256224)
digit
Euclidean Similarity of word digit to computer is:  0.46139246
Top 5 words similar to the given word according to euclidean similarity:  [('megabuck', 0.9212

/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:43: DeprecationW

Top 5 words similar to the given word according to cosine similarity:  [('megabuck', 0.9212495
data
Euclidean Similarity of word data to computer is:  0.37617567
Top 5 words similar to the given word according to euclidean similarity:  [('acquisition', 0.76

/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:51: DeprecationW

Top 5 words similar to the given word according to cosine similarity:  [('acquisition', 0.76253
prohibition
```

```
Euclidean Similarity of word prohibition to computer is:  0.30014408

Top 5 words similar to the given word according to euclidean similarity:  [('levitical', 0.889

/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:59: DeprecationW

Top 5 words similar to the given word according to cosine similarity:  [('levitical', 0.8893965

faith

Euclidean Similarity of word faith to computer is:  0.1316617

Top 5 words similar to the given word according to euclidean similarity:  [('reject', 0.736490

/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:67: DeprecationW

Top 5 words similar to the given word according to cosine similarity:  [('reject', 0.73648995)
```

1.9 Model : Parameter sg = 1, window = 5

```
In [84]: print(w2)
         print("")
         print("Euclidean Similarity of word " + w2 + " to computer is: ", model1.similarity(w2
         print(" ")
         print("Top 5 words similar to the given word according to euclidean similarity: ",mode
         print(" ")
         print("Top 5 words similar to the given word according to cosine similarity: ", cosine
         print(" ")
         print(w3)
         print("")
         print("Euclidean Similarity of word " + w3 + " to computer is: ", model1.similarity(w3
         print(" ")
         print("Top 5 words similar to the given word according to euclidean similarity: ",mode
         print(" ")
         print("Top 5 words similar to the given word according to cosine similarity: ", cosine
         print(" ")
         print(w4)
         print("")
         print("Euclidean Similarity of word " + w4 + " to computer is: ", model1.similarity(w4
         print(" ")
         print("Top 5 words similar to the given word according to euclidean similarity: ",mode
```

```
print(" ")
print("Top 5 words similar to the given word according to cosine similarity: ", cosine
print(" ")
print(w5)
print("")
print("Euclidean Similarity of word " + w5 + " to computer is: ", model1.similarity(w5
print(" ")
print("Top 5 words similar to the given word according to euclidean similarity: ",mode
print(" ")
print("Top 5 words similar to the given word according to cosine similarity: ", cosine
print(" ")
print(w6)
print("")
print("Euclidean Similarity of word " + w6 + " to computer is: ", model1.similarity(w6
print(" ")
print("Top 5 words similar to the given word according to euclidean similarity: ",mode
print(" ")
print("Top 5 words similar to the given word according to cosine similarity: ", cosine
print(" ")
print(w7)
print("")
print("Euclidean Similarity of word " + w7 + " to computer is: ", model1.similarity(w7
print(" ")
print("Top 5 words similar to the given word according to euclidean similarity: ",mode
print(" ")
print("Top 5 words similar to the given word according to cosine similarity: ", cosine
print(" ")
print(w8)
print("")
print("Euclidean Similarity of word " + w8 + " to computer is: ", model1.similarity(w8
print(" ")
print("Top 5 words similar to the given word according to euclidean similarity: ",mode
print(" ")
print("Top 5 words similar to the given word according to cosine similarity: ", cosine
print(" ")
print(w9)
print("")
print("Euclidean Similarity of word " + w9 + " to computer is: ", model1.similarity(w9
print(" ")
print("Top 5 words similar to the given word according to euclidean similarity: ",mode
print(" ")
print("Top 5 words similar to the given word according to cosine similarity: ", cosine
print(" ")
print(w10)
print("")
print("Euclidean Similarity of word " + w10 + " to computer is: ", model1.similarity(w
print(" ")
print("Top 5 words similar to the given word according to euclidean similarity: ",mode
```

```
        print(" ")
        print("Top 5 words similar to the given word according to cosine similarity: ", cosine
```

```
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:3: DeprecationWa
  This is separate from the ipykernel package so we can avoid doing imports until
2019-10-26 22:45:27,782 : INFO : precomputing L2-norms of word weight vectors
```

```
popular

Euclidean Similarity of word popular to computer is:  0.43820477

Top 5 words similar to the given word according to euclidean similarity:  [('halfway', 0.800319
```

```
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:4: DeprecationWa
  after removing the cwd from sys.path.
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:7: DeprecationWa
  import sys
```

```
Top 5 words similar to the given word according to cosine similarity:  [('halfway', 0.80031914)

technology

Euclidean Similarity of word technology to computer is:  0.3719428

Top 5 words similar to the given word according to euclidean similarity:  [('prodigy', 0.732673
```

```
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:11: DeprecationW
  # This is added back by InteractiveShellApp.init_path()
```

```
Top 5 words similar to the given word according to cosine similarity:  [('prodigy', 0.7326734)

question

Euclidean Similarity of word question to computer is:  0.21110871

Top 5 words similar to the given word according to euclidean similarity:  [('answer', 0.859965
```

```
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:19: DeprecationW
```

```
Top 5 words similar to the given word according to cosine similarity:  [('answer', 0.85996604)
information
Euclidean Similarity of word information to computer is:  0.4516124
Top 5 words similar to the given word according to euclidean similarity:  [('herein', 0.716072

/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:27: DeprecationW

Top 5 words similar to the given word according to cosine similarity:  [('herein', 0.7160725),
letter
Euclidean Similarity of word letter to computer is:  0.18696895
Top 5 words similar to the given word according to euclidean similarity:  [('moody', 0.7129958

/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:35: DeprecationW

Top 5 words similar to the given word according to cosine similarity:  [('moody', 0.7129996),
digit
Euclidean Similarity of word digit to computer is:  0.43333927
Top 5 words similar to the given word according to euclidean similarity:  [('uncompressed', 0.8

/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:43: DeprecationW

Top 5 words similar to the given word according to cosine similarity:  [('uncompressed', 0.887
data
Euclidean Similarity of word data to computer is:  0.29505494
Top 5 words similar to the given word according to euclidean similarity:  [('acquisition', 0.78
```

```
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:51: DeprecationW

Top 5 words similar to the given word according to cosine similarity:  [('acquisition', 0.7813(

prohibition

Euclidean Similarity of word prohibition to computer is:  0.22733936

Top 5 words similar to the given word according to euclidean similarity:  [('rum', 0.869230508

/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:59: DeprecationW

Top 5 words similar to the given word according to cosine similarity:  [('rum', 0.8692305), ('c

faith

Euclidean Similarity of word faith to computer is:  0.2571972

Top 5 words similar to the given word according to euclidean similarity:  [('believer', 0.7753

/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:67: DeprecationW

Top 5 words similar to the given word according to cosine similarity:  [('believer', 0.77539295
```

1.10 Model : Parameter sg = 0, window = 3

```
In [85]: print(w2)
         print("")
         print("Euclidean Similarity of word " + w2 + " to computer is: ", new_model.similarity
         print(" ")
         print("Top 5 words similar to the given word according to euclidean similarity: ",new_
         print(" ")
         print("Top 5 words similar to the given word according to cosine similarity: ", cosine
         print(" ")
         print(w3)
         print("")
         print("Euclidean Similarity of word " + w3 + " to computer is: ", new_model.similarity
         print(" ")
         print("Top 5 words similar to the given word according to euclidean similarity: ",new_
         print(" ")
         print("Top 5 words similar to the given word according to cosine similarity: ", cosine
```

```
print(" ")
print(w4)
print("")
print("Euclidean Similarity of word " + w4 + " to computer is: ", new_model.similarity
print(" ")
print("Top 5 words similar to the given word according to euclidean similarity: ",new_
print(" ")
print("Top 5 words similar to the given word according to cosine similarity: ", cosine
print(" ")
print(w5)
print("")
print("Euclidean Similarity of word " + w5 + " to computer is: ", new_model.similarity
print(" ")
print("Top 5 words similar to the given word according to euclidean similarity: ",new_
print(" ")
print("Top 5 words similar to the given word according to cosine similarity: ", cosine
print(" ")
print(w6)
print("")
print("Euclidean Similarity of word " + w6 + " to computer is: ", new_model.similarity
print(" ")
print("Top 5 words similar to the given word according to euclidean similarity: ",new_
print(" ")
print("Top 5 words similar to the given word according to cosine similarity: ", cosine
print(" ")
print(w7)
print("")
print("Euclidean Similarity of word " + w7 + " to computer is: ", new_model.similarity
print(" ")
print("Top 5 words similar to the given word according to euclidean similarity: ",new_
print(" ")
print("Top 5 words similar to the given word according to cosine similarity: ", cosine
print(" ")
print(w8)
print("")
print("Euclidean Similarity of word " + w8 + " to computer is: ", new_model.similarity
print(" ")
print("Top 5 words similar to the given word according to euclidean similarity: ",new_
print(" ")
print("Top 5 words similar to the given word according to cosine similarity: ", cosine
print(" ")
print(w9)
print("")
print("Euclidean Similarity of word " + w9 + " to computer is: ", new_model.similarity
print(" ")
print("Top 5 words similar to the given word according to euclidean similarity: ",new_
print(" ")
print("Top 5 words similar to the given word according to cosine similarity: ", cosine
```

```
        print(" ")
        print(w10)
        print("")
        print("Euclidean Similarity of word " + w10 + " to computer is: ", new_model.similarit
        print(" ")
        print("Top 5 words similar to the given word according to euclidean similarity: ",new_
        print(" ")
        print("Top 5 words similar to the given word according to cosine similarity: ", cosine
```

```
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:3: DeprecationWa
  This is separate from the ipykernel package so we can avoid doing imports until
2019-10-26 22:48:21,882 : INFO : precomputing L2-norms of word weight vectors


popular

Euclidean Similarity of word popular to computer is:  0.50274193

Top 5 words similar to the given word according to euclidean similarity:  [('actively', 0.9416


/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:4: DeprecationWa
  after removing the cwd from sys.path.
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:7: DeprecationWa
  import sys


Top 5 words similar to the given word according to cosine similarity:  [('actively', 0.9416484)

technology

Euclidean Similarity of word technology to computer is:  0.67519844

Top 5 words similar to the given word according to euclidean similarity:  [('telecommunication


/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:11: DeprecationW
  # This is added back by InteractiveShellApp.init_path()


Top 5 words similar to the given word according to cosine similarity:  [('telecommunication', 0

question

Euclidean Similarity of word question to computer is:  0.21732914

Top 5 words similar to the given word according to euclidean similarity:  [('clergymen', 0.8189
```

```
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:19: DeprecationW

Top 5 words similar to the given word according to cosine similarity:  [('clergymen', 0.818915)

information

Euclidean Similarity of word information to computer is:  0.62731767

Top 5 words similar to the given word according to euclidean similarity:  [('correction', 0.827

/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:27: DeprecationW

Top 5 words similar to the given word according to cosine similarity:  [('correction', 0.827533

letter

Euclidean Similarity of word letter to computer is:  0.20623244

Top 5 words similar to the given word according to euclidean similarity:  [('suggestion', 0.831

/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:35: DeprecationW

Top 5 words similar to the given word according to cosine similarity:  [('suggestion', 0.832242

digit

Euclidean Similarity of word digit to computer is:  0.38559967

Top 5 words similar to the given word according to euclidean similarity:  [('lite', 0.986712931

/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:43: DeprecationW

Top 5 words similar to the given word according to cosine similarity:  [('lite', 0.986713), ('c

data
```

```
Euclidean Similarity of word data to computer is:  0.425191

Top 5 words similar to the given word according to euclidean similarity:  [('transfer', 0.8559

/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:51: DeprecationW

Top 5 words similar to the given word according to cosine similarity:  [('transfer', 0.8559536

prohibition

Euclidean Similarity of word prohibition to computer is:  0.4032716

Top 5 words similar to the given word according to euclidean similarity:  [('outright', 0.9688

/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:59: DeprecationW

Top 5 words similar to the given word according to cosine similarity:  [('outright', 0.9688137)

faith

Euclidean Similarity of word faith to computer is:  0.15602455

Top 5 words similar to the given word according to euclidean similarity:  [('salvation', 0.864

/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:67: DeprecationW

Top 5 words similar to the given word according to cosine similarity:  [('salvation', 0.864400
```

1.11 Model : Parameter sg = 0, window = 5

```
In [74]: print("Euclidean Similarity of word " + w2 + " to computer is: ", new_model1.similarit
         print(" ")
         print("Top 5 words similar to the given word according to euclidean similarity: ",new_
         print(" ")
         print("Top 5 words similar to the given word according to cosine similarity: ", cosine

/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:1: DeprecationWa
  """Entry point for launching an IPython kernel.
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:4: DeprecationWa
```

```
  after removing the cwd from sys.path.
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:7: DeprecationWa
  import sys


Euclidean Similarity of word popular to computer is:  0.48639983

Top 5 words similar to the given word according to euclidean similarity:  [('worthwhile', 0.934

Top 5 words similar to the given word according to cosine similarity:  [('worthwhile', 0.934532
```

```
In [75]: print("Euclidean Similarity of word " + w3 + " to computer is: ", new_model1.similarit
         print(" ")
         print("Top 5 words similar to the given word according to euclidean similarity: ",new_
         print(" ")
         print("Top 5 words similar to the given word according to cosine similarity: ", cosine
```

```
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:1: DeprecationWa
  """Entry point for launching an IPython kernel.
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:4: DeprecationWa
  after removing the cwd from sys.path.
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:7: DeprecationWa
  import sys


Euclidean Similarity of word technology to computer is:  0.6641607

Top 5 words similar to the given word according to euclidean similarity:  [('specialize', 0.878

Top 5 words similar to the given word according to cosine similarity:  [('specialize', 0.878390
```

```
In [76]: print("Euclidean Similarity of word " + w4 + " to computer is: ", new_model1.similarit
         print(" ")
         print("Top 5 words similar to the given word according to euclidean similarity: ",new_
         print(" ")
         print("Top 5 words similar to the given word according to cosine similarity: ", cosine
```

```
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:1: DeprecationWa
  """Entry point for launching an IPython kernel.
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:4: DeprecationWa
  after removing the cwd from sys.path.
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:7: DeprecationWa
  import sys


Euclidean Similarity of word question to computer is:  0.061645284
```

```
Top 5 words similar to the given word according to euclidean similarity:  [('adnausium', 0.7308

Top 5 words similar to the given word according to cosine similarity:  [('adnausium', 0.7308908
```

```
In [77]: print("Euclidean Similarity of word " + w5 + " to computer is: ", new_model1.similarit
         print(" ")
         print("Top 5 words similar to the given word according to euclidean similarity: ",new_
         print(" ")
         print("Top 5 words similar to the given word according to cosine similarity: ", cosine
```

```
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:1: DeprecationWa
  """Entry point for launching an IPython kernel.
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:4: DeprecationWa
  after removing the cwd from sys.path.
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:7: DeprecationWa
  import sys


Euclidean Similarity of word information to computer is:  0.5705979

Top 5 words similar to the given word according to euclidean similarity:  [('corrects', 0.80256

Top 5 words similar to the given word according to cosine similarity:  [('corrects', 0.8025624)
```

```
In [78]: print("Euclidean Similarity of word " + w6 + " to computer is: ", new_model1.similarit
         print(" ")
         print("Top 5 words similar to the given word according to euclidean similarity: ",new_
         print(" ")
         print("Top 5 words similar to the given word according to cosine similarity: ", cosine
```

```
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:1: DeprecationWa
  """Entry point for launching an IPython kernel.
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:4: DeprecationWa
  after removing the cwd from sys.path.
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:7: DeprecationWa
  import sys


Euclidean Similarity of word letter to computer is:  0.10323852

Top 5 words similar to the given word according to euclidean similarity:  [('nico', 0.787011682

Top 5 words similar to the given word according to cosine similarity:  [('nico', 0.7870118), (
```

```
In [79]: print("Euclidean Similarity of word " + w7 + " to computer is: ", new_model1.similarit
         print(" ")
```

```
        print("Top 5 words similar to the given word according to euclidean similarity: ",new
        print(" ")
        print("Top 5 words similar to the given word according to cosine similarity: ", cosine
```

```
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:1: DeprecationWa
  """Entry point for launching an IPython kernel.
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:4: DeprecationWa
  after removing the cwd from sys.path.
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:7: DeprecationWa
  import sys
```

```
Euclidean Similarity of word digit to computer is:  0.2912973

Top 5 words similar to the given word according to euclidean similarity:  [('rx', 0.9688238501

Top 5 words similar to the given word according to cosine similarity:  [('rx', 0.9688239), ('mu
```

```
In [80]: print("Euclidean Similarity of word " + w8 + " to computer is: ", new_model1.similarit
         print(" ")
         print("Top 5 words similar to the given word according to euclidean similarity: ",new
         print(" ")
         print("Top 5 words similar to the given word according to cosine similarity: ", cosine
```

```
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:1: DeprecationWa
  """Entry point for launching an IPython kernel.
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:4: DeprecationWa
  after removing the cwd from sys.path.
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:7: DeprecationWa
  import sys
```

```
Euclidean Similarity of word data to computer is:  0.40867683

Top 5 words similar to the given word according to euclidean similarity:  [('transmission', 0.8

Top 5 words similar to the given word according to cosine similarity:  [('transmission', 0.870
```

```
In [82]: print("Euclidean Similarity of word " + w9 + " to computer is: ", new_model1.similarit
         print(" ")
         print("Top 5 words similar to the given word according to euclidean similarity: ",new
         print(" ")
         print("Top 5 words similar to the given word according to cosine similarity: ", cosine
```

```
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:1: DeprecationWa
  """Entry point for launching an IPython kernel.
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:4: DeprecationWa
```

```
  after removing the cwd from sys.path.
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:7: DeprecationWa
  import sys


Euclidean Similarity of word prohibition to computer is:  0.34777611

Top 5 words similar to the given word according to euclidean similarity:  [('observation', 0.9

Top 5 words similar to the given word according to cosine similarity:  [('observation', 0.96754
```

```
In [81]: print("Euclidean Similarity of word " + w10 + " to computer is: ", new_model1.similari
         print(" ")
         print("Top 5 words similar to the given word according to euclidean similarity: ",new_
         print(" ")
         print("Top 5 words similar to the given word according to cosine similarity: ", cosine
```

```
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:1: DeprecationWa
  """Entry point for launching an IPython kernel.
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:4: DeprecationWa
  after removing the cwd from sys.path.
/Users/anjaliverma/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:7: DeprecationWa
  import sys


Euclidean Similarity of word faith to computer is:  -0.0008394681

Top 5 words similar to the given word according to euclidean similarity:  [('truth', 0.8152226

Top 5 words similar to the given word according to cosine similarity:  [('truth', 0.8152227),
```

```
In [ ]:
```