TASK1-EMBEDDING

word embedding using Word2Vec

In [1]:
```
!pip install --upgrade gensim
```

```
Collecting gensim
  Downloading gensim-3.8.3-cp37-cp37m-win_amd64.whl (24.2 MB)
Requirement already satisfied, skipping upgrade: six>=1.5.0 in c:\users
\acer\anaconda3\lib\site-packages (from gensim) (1.12.0)
Requirement already satisfied, skipping upgrade: numpy>=1.11.3 in c:\us
ers\acer\anaconda3\lib\site-packages (from gensim) (1.18.2)
Requirement already satisfied, skipping upgrade: scipy>=0.18.1 in c:\us
ers\acer\anaconda3\lib\site-packages (from gensim) (1.4.1)
Collecting smart-open>=1.8.1
  Downloading smart_open-2.0.0.tar.gz (103 kB)
Collecting Cython==0.29.14
  Downloading Cython-0.29.14-cp37-cp37m-win_amd64.whl (1.7 MB)
Requirement already satisfied, skipping upgrade: requests in c:\users\a
cer\anaconda3\lib\site-packages (from smart-open>=1.8.1->gensim) (2.22.
0)
Requirement already satisfied, skipping upgrade: boto in c:\users\acer
\anaconda3\lib\site-packages (from smart-open>=1.8.1->gensim) (2.49.0)
Collecting boto3
  Downloading boto3-1.14.11-py2.py3-none-any.whl (128 kB)
Requirement already satisfied, skipping upgrade: idna<2.9,>=2.5 in c:\u
sers\acer\anaconda3\lib\site-packages (from requests->smart-open>=1.8.1
->gensim) (2.8)
Requirement already satisfied, skipping upgrade: urllib3!=1.25.0,!=1.2
5.1,<1.26,>=1.21.1 in c:\users\acer\anaconda3\lib\site-packages (from r
equests->smart-open>=1.8.1->gensim) (1.24.2)
Requirement already satisfied, skipping upgrade: chardet<3.1.0,>=3.0.2
in c:\users\acer\anaconda3\lib\site-packages (from requests->smart-open
>=1.8.1->gensim) (3.0.4)
```

```
Requirement already satisfied, skipping upgrade: certifi>=2017.4.17 in
c:\users\acer\anaconda3\lib\site-packages (from requests->smart-open>=
1.8.1->gensim) (2019.9.11)
Collecting s3transfer<0.4.0,>=0.3.0
  Downloading s3transfer-0.3.3-py2.py3-none-any.whl (69 kB)
Collecting jmespath<1.0.0,>=0.7.1
  Downloading jmespath-0.10.0-py2.py3-none-any.whl (24 kB)
Collecting botocore<1.18.0,>=1.17.11
  Downloading botocore-1.17.11-py2.py3-none-any.whl (6.3 MB)
Requirement already satisfied, skipping upgrade: docutils<0.16,>=0.10 i
n c:\users\acer\anaconda3\lib\site-packages (from botocore<1.18.0,>=1.1
7.11->boto3->smart-open>=1.8.1->gensim) (0.15.2)
Requirement already satisfied, skipping upgrade: python-dateutil<3.0.0,
>=2.1 in c:\users\acer\anaconda3\lib\site-packages (from botocore<1.18.
0,>=1.17.11->boto3->smart-open>=1.8.1->gensim) (2.8.0)
Building wheels for collected packages: smart-open
  Building wheel for smart-open (setup.py): started
  Building wheel for smart-open (setup.py): finished with status 'done'
  Created wheel for smart-open: filename=smart_open-2.0.0-py3-none-any.
whl size=101346 sha256=bc8212bde31ce3103d94cdafb817259ba8bacc67bc9f900f
c63f6eeb183c967d
  Stored in directory: c:\users\acer\appdata\local\pip\cache\wheels\bb
\1c\9c\412ec03f6d5ac7d41f4b965bde3fc0d1bd201da5ba3e2636de
Successfully built smart-open
Installing collected packages: jmespath, botocore, s3transfer, boto3, s
mart-open, Cython, gensim
  Attempting uninstall: Cython
    Found existing installation: Cython 0.29.13
    Uninstalling Cython-0.29.13:
      Successfully uninstalled Cython-0.29.13
Successfully installed Cython-0.29.14 boto3-1.14.11 botocore-1.17.11 ge
nsim-3.8.3 jmespath-0.10.0 s3transfer-0.3.3 smart-open-2.0.0
```

```
WARNING: You are using pip version 20.1; however, version 20.1.1 is ava
ilable.
You should consider upgrading via the 'c:\users\acer\anaconda3\python.e
xe -m pip install --upgrade pip' command.
```

In [ ]:
```
!pip install Word2Vec
```

In [36]:
```
from gensim.models import Word2Vec
# define training data
sentences = [['this', 'is', 'the', 'first', 'sentence', 'for', 'word2vec'],
            ['this', 'is', 'the', 'second', 'sentence'],
            ['yet', 'another', 'sentence'],
            ['one', 'more', 'sentence'],
            ['and', 'the', 'final', 'sentence']]
```

In [37]:
```
# train model
model = Word2Vec(sentences, min_count=1)
```

In [38]:
```
# summarize the loaded model
print(model)
```

```
Word2Vec(vocab=14, size=100, alpha=0.025)
```

In [39]:
```
# summarize vocabulary
words = list(model.wv.vocab)
print(words)
```

```
['this', 'is', 'the', 'first', 'sentence', 'for', 'word2vec', 'second',
'yet', 'another', 'one', 'more', 'and', 'final']
```

In [40]:
```
# access vector for one word
print(model['sentence'])
```

```
[ 7.6911051e-04  2.6431947e-04 -1.8240251e-03  3.8620243e-03
 -3.5477227e-03  3.2040644e-03  1.6649488e-04  1.7281098e-03
  1.8329378e-03  4.4308160e-03  9.0016407e-04  3.4924310e-03
 -2.1655366e-03  1.0737572e-03 -3.8606529e-03 -1.1830005e-03
  1.3286733e-03  3.2175309e-03 -3.3865373e-03  2.0454843e-04
 -3.3649914e-03 -2.0537258e-03  1.6624051e-03 -1.5510300e-03
 -3.0116865e-03  2.2319752e-04  3.2403120e-03 -2.6152974e-03
 -4.9428041e-03 -4.8097786e-03 -2.5440059e-03  5.4570864e-04
 -1.9992865e-03 -1.1522527e-03  1.2714770e-03 -4.8160178e-04
 -1.4063554e-03  1.8131027e-03 -1.7294649e-03  3.9021755e-03
```

```
 4.4160406e-03  3.5023803e-03  2.1587661e-03  2.8929892e-03
 2.3308585e-03 -2.9569983e-03 -4.8294500e-04  5.7870981e-05
-2.4950744e-03 -2.8349042e-03 -1.9563557e-04  2.2408927e-03
-4.1549485e-03 -4.9416097e-03 -1.4531723e-03 -2.3163480e-03
-2.3125233e-03  2.3783883e-03  3.4458544e-03  3.2365194e-03
 1.1603539e-03 -2.0322362e-03 -3.4471191e-03 -3.5589191e-04
 4.8587685e-03 -1.0439306e-03  1.1716568e-03  3.5842373e-03
 3.8885353e-03  3.1136193e-03 -3.0828745e-03 -1.9476783e-03
 6.4240245e-04 -1.1667300e-03 -1.0525422e-03  1.8118330e-03
 2.7007023e-03  1.9334722e-03  7.2264404e-04  1.2582352e-03
 3.9223747e-04  8.4605790e-04  1.2621116e-03 -3.6125933e-04
-3.6043341e-03 -2.3092814e-03  4.8685982e-03  2.2516809e-03
-3.4046436e-03  3.8587363e-03 -1.0734973e-03 -3.7409554e-03
 4.9976362e-03 -1.9786067e-03  3.6118680e-03  3.2020983e-04
-4.5747054e-03  4.9105287e-03  1.0145215e-03  1.3058506e-03]
```

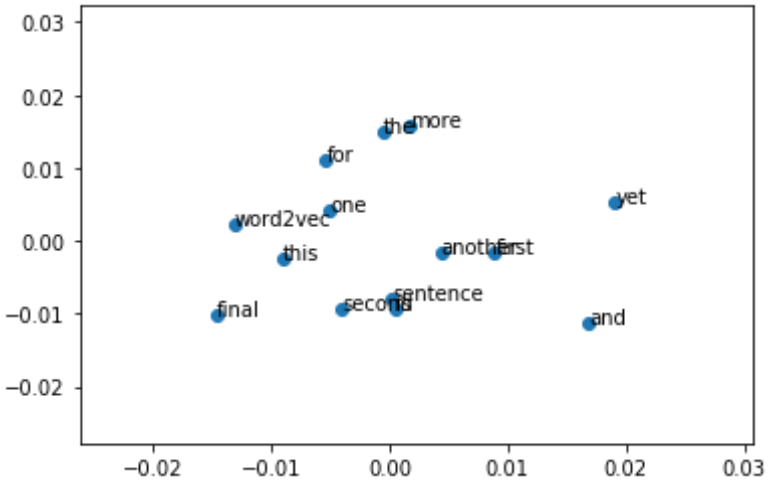
```
C:\Users\ACER\Anaconda3\lib\site-packages\ipykernel_launcher.py:2: Depr
ecationWarning: Call to deprecated `__getitem__` (Method will be remove
d in 4.0.0, use self.wv.__getitem__() instead).
```

In [41]:
```
# save model
model.save('model.bin')
```

In [42]:
```
# load model
new_model = Word2Vec.load('model.bin')
print(new_model)
```

```
Word2Vec(vocab=14, size=100, alpha=0.025)
```

In [43]:
```
# create a scatter plot of the projection
from matplotlib import pyplot
pyplot.scatter(result[:, 0], result[:, 1])
words = list(model.wv.vocab)
for i, word in enumerate(words):
    pyplot.annotate(word, xy=(result[i, 0], result[i, 1]))
pyplot.show()
```



tf idf vectorizer

In [46]:
```
import pandas as pd
import numpy as np
from sklearn.feature_extraction.text import TfidfVectorizer
data=['it was the best of times','it was the worst of times','it was th
e age of wisdome','it was the age of foolishness']
```

In [47]:
```
#stop word removal
tfidf_vectorizer=TfidfVectorizer(stop_words='english')
```

In [49]:
```
#fitting the model
tfidf_feature=tfidf_vectorizer.fit_transform(data)

#creating dataframe of vectors
dataframe=pd.DataFrame(data=tfidf_feature.todense(),columns=tfidf_vecto
rizer.get_feature_names())
dataframe
```

Out[49]:

	age	best	foolishness	times	wisdome	worst

	age	best	foolishness	times	wisdome	worst
0	0.00000	0.785288	0.000000	0.61913	0.000000	0.000000
1	0.00000	0.000000	0.000000	0.61913	0.000000	0.785288
2	0.61913	0.000000	0.000000	0.00000	0.785288	0.000000
3	0.61913	0.000000	0.785288	0.00000	0.000000	0.000000

In [ ]: