COMPUTING DOCUMENT SIMILARITY USINGDOC2VEC MODEL]

Import dependencies

In [1]:

```
Requirement already satisfied: gensim in
c:\users\elcot\anaconda3\lib\site-packages (4.1.2)
Requirement already satisfied: scipy>=0.18.1 in c:\users\elcot\anaconda3\lib
\site-packages (from gensim) (1.7.3)
Requirement already satisfied: numpy>=1.17.0 in c:\users\elcot\anaconda3\lib
\site-packages (from gensim) (1.21.5)
Requirement already satisfied: smart-open>=1.8.1 in c:\users\elcot\anaconda3
\lib\site-packages (from gensim) (5.1.0)
```

In [2]:

In [3]:

Create dataset

In [4]:

Create TaggedDocument

In [5]:

Train Model

model parameters

In [8]:

create model

In [11]:

build vocabulary

In [12]:

shuffle data

In [13]:

train Doc2Vec model

In [14]:

```
Model Saved
```

Find Similar documents for the given document

In [15]:

To find the vector of a document which is not in training data

In [16]:

```
V1_infer [-0.01256738 -0.01497919 -0.0024978  -0.02138223 -0.02458959 -0.011
00514
 -0.0168768  -0.010017    0.00843245 -0.02090312 -0.01303108  0.01636144
  0.00612498  0.02392749  0.01615989 -0.00515045 -0.01087133  0.01027928
  0.00359414 -0.00996092]
```

To find most similar doc using tags

In [17]:

```
C:\Users\elcot\AppData\Local\Temp\ipykernel_9348\2066422884.p
y:1: Deprecatio nWarning: Call to deprecated `docvecs` (The
`docvecs` property has been renamed `dv`.).
  similar_doc=model.docvecs.most_similar('1')

[('2', 0.31398797035217285), ('0', 0.2646177411079407), ('3', 0.205407455563
54523)]
```

```
[-0.01885638  0.01303607 -0.02846214  0.01310304  0.02902067 -0.04054232
 -0.04165549 -0.0497813   0.02466379 -0.04562103  0.02921205  0.03400678
 -0.03254311 -0.022611   -0.00628033  0.00823286 -0.00740639 -0.04271377
 -0.01802378  0.00866232]
```

```
C:\Users\elcot\AppData\Local\Temp\ipykernel_9348\2066422884.p
y:3: Deprecatio nWarning: Call to deprecated `docvecs` (The
`docvecs` property has been renamed `dv`.).
  print(model.docvecs['1'])
```

EXERCISE - 2

Question 1

In [18]:

Create TaggedDocument

In [19]:

model parameters

In [20]:

create model

In [21]:

build vocabulary

In [23]:

shuffle data

In [24]:

train Doc2Vec model

In [25]:

```
Model Saved
```

Question 2

Find thel most similar TWO documents for the query doument " cat stayed in thehouse".

```
In [26]:
```

to find the vector of a document which is not in training data

```
In [27]:
```

```
v1_infer [ 0.02001861  0.00136964  0.01491838 -0.00677093  0.01567641 -0.025
51127
 -0.00934195  0.02010441 -0.00598909  0.00585111  0.01769784 -0.01855669
 -0.01889012 -0.01871531 -0.02211943  0.01268214  0.00623796 -0.02299893
 -0.02508585 -0.0209804 ]
```

to find most similar doc using tags

```
In [28]:
```

```
[('3', 0.3427582383155823), ('1', 0.325040340423584), ('0', -0.1131285503506
6605)]
[-0.01061909 -0.03621923  0.02069    -0.04284193  0.01396234 -0.02348765
  0.00293782 -0.01052751  0.02685227 -0.0404647  -0.01060271 -0.00022251
 -0.03386855 -0.03325163 -0.01003217  0.04401389 -0.00632442  0.01772947
 -0.02934732  0.04424063]
```

```
In [ ]:
```