Name : Arul Kumar ARK

Roll No. : 225229103

LAB : 4

```
                                        Computing Documents Similarity Using Doc2Vec Model
```

Exercise : 1

Import Dependencies

In [24]:

```
from gensim.models.doc2vec import Doc2Vec , TaggedDocument
from nltk.tokenize import word_tokenize
from sklearn import utils
```

In [25]:

```
import nltk
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\arulk\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping tokenizers\punkt.zip.
```

Out[25]:

```
True
```

Create Dataset

In [26]:

```
data = ['I love machine learing. Its awesome.','I love coding python','I love buliding chatbots','they chat amagingly well']
```

Create TaggedDocument

In [27]:

```
tagged_data = [TaggedDocument(words=word_tokenize(d.lower()),
                              tags =[str(i)] )for i,d in enumerate(data)]
```

Traing Model

In [28]:

```
vec_size=20
alpha=0.025
model  = Doc2Vec(vector_size=vec_size,alpha=alpha,min_alpha=0.00025,min_count=1,dm=1)
model.build_vocab(tagged_data)
tagged_data=utils.shuffle(tagged_data)
model.train(tagged_data,
            total_examples=model.corpus_count,
            epochs=30)

model.save("d2v.model")
print("Model Saved")
```

```
Model Saved
```

Find Similar document for the given document

In [29]:

```
from gensim.models.doc2vec import Doc2Vec

model=Doc2Vec.load("d2v.model")

test_data=word_tokenize("I love chatbots".lower())
v1=model.infer_vector(test_data)
print("v1_infer",v1)

similar_doc=model.dv.most_similar('1')
print(similar_doc)

print(model.dv["1"])
```

```
v1_infer [ 0.00059448  0.00966144  0.01951231  0.01481926 -0.00131763  0.01376816
 -0.00051552  0.01128929  0.01356268 -0.00446643  0.00918196  0.01555888
  0.01014236 -0.00794917  0.01224201  0.00184336 -0.0143489  -0.01296302
  0.01350386 -0.01675778]
[('2', 0.3254072666168213), ('0', 0.2771632969379425), ('3', 0.21568474173545837)]
[-0.01906641  0.01296389 -0.02857265  0.01302118  0.02940756 -0.04099821
 -0.04192136 -0.05011528  0.02455854 -0.04616568  0.0291213   0.03425032
 -0.0331385  -0.02327707 -0.00666657  0.00835614 -0.00746358 -0.04276369
 -0.01852638  0.00886416]
```

Exercise : 2

In [30]:

```
docs= ['the house had a tiny little mouse','the cat saw the mouse','the mouse ran away from the house','the end of the mouse'
```

In [31]:

```
tagged_doc = [TaggedDocument(words=word_tokenize(dc.lower()),tags =[str(i)] )for i,dc in enumerate(docs)]
```

In [32]:

```
vec_size1=20
alpha1=0.025
model1 = Doc2Vec(vector_size=vec_size,alpha=alpha,min_alpha=0.00025,min_count=1,dm=1)
model1.build_vocab(tagged_doc)
tagged_doc=utils.shuffle(tagged_doc)
model1.train(tagged_doc,total_examples=model.corpus_count,epochs=30)
model1.save("d3v.model")
print("Model1 Saved")
```

```
Model1 Saved
```

In [42]:

```
model1=Doc2Vec.load("d2v.model")
test_doc=word_tokenize("cat stayed in the house".lower())
v2=model.infer_vector(test_doc)
print("v2_infer",v2)
similar_doc1=model.dv.most_similar('2')
print(similar_doc1)
print(model.dv)
```

```
v2_infer [ 0.01366494  0.00260812  0.01637502  0.0209187  -0.0210546   0.01154996
  0.00684811  0.00077518  0.00268406  0.00021969 -0.00479817  0.02059207
  0.00165972 -0.02398996 -0.02053437 -0.01026757  0.01913402  0.0218966
 -0.01392823  0.01641821]
[('3', 0.3388660252094269), ('1', 0.3254072964191437), ('0', -0.11331021040678024)]
<gensim.models.keyedvectors.KeyedVectors object at 0x000001432E05A850>
```

In [ ]: