NLP Assignment 6

Shreya Sharma
2015096

Question 1

Used most_similar function to get the answers

model.wv.most_similar(positive=['China', 'Delhi'], negative=['India'])[0]
model.wv.most_similar(positive=['USA', 'ISRO'], negative=['India'])[0]

a) If Delhi is the capital of India then what is the capital of China?
('Beijing', 0.7975110411643982)

b) If ISRO is related to India then what is related to USA.
('STScI', 0.42706042528152466)

Question 2

Trained Doc2Vec on the entire dataset and used it to convert new documents into vectors and then used **Cosine Similarity** to calculate the similarity between the 2 documents.

Score of comp.graphics with files of other classes:
alt.atheism : 0.48609670996665955
comp.os.ms-windows.misc : 0.5277971029281616
comp.sys.ibm.pc.hardware : 0.4836440682411194
comp.sys.mac.hardware : 0.5062006711959839
comp.windows.x : 0.27190542221069336
misc.forsale : 0.5189643502235413
rec.autos : 0.6048580408096313
rec.motorcycles : 0.49248719215393066
rec.sport.baseball : 0.4755517244338989
rec.sport.hockey : 0.4719200134277344
sci.crypt : 0.3368961215019226
sci.electronics : 0.3051990568637848
sci.med : 0.5349641442298889
sci.space : 0.2626318335533142
soc.religion.christian : 0.38013920187950134
talk.politics.guns : 0.398502916097641

talk.politics.mideast : 0.09206746518611908
talk.politics.misc : 0.46515798568725586
talk.religion.misc : 0.35496023297309875

Score of comp.graphics with files of same class:
0 : 0.395177960395813
1 : 0.5227679014205933
2 : 0.4819713234901428
3 : 0.4869748055934906
4 : 0.29186519980430603
5 : 0.552078366279602
6 : 0.4898204207420349
7 : 0.5236486196517944
8 : 0.38389429450035095
9 : 0.4844417870044708
10 : 0.3403414189815521
11 : 0.5106260776519775
12 : 0.5614442825317383
13 : 0.438974529504776
14 : 0.46276289224624634
15 : 0.543857216835022
16 : 0.6490538120269775
17 : 0.6278152465820312
18 : 0.42681437730789185

MEAN (other classes) = 0.4194707501875727
MEAN (same class) = 0.48285950171320063

MEAN for 'same class' is more as expected