NLP ASSIGNMENT-2

Problem 1:

Part 1-

For 2 class classifier, on testing for a file using add-1 smoothing:

```
Enter
1) 2 Classes
2) 20 classes
1
Enter
1) for file input
2) for sentence input
1
Enter the file name new.txt
rec.motorcycles  ------>  -1757.60634706
rec.sport.baseball  ------>  -1746.30972512
talk.politics.mideast  have highest log(probability) value  -1557.29923419
```

For 20-class classifier

```
Enter
1) 2 Classes
2) 20 classes
2
Enter
1) for file input
2) for sentence input
1
Enter the file name new.txt
sci.crypt  ------>  -1722.79501176
misc.forsale  ------>  -1792.00160006
talk.religion.misc  ------>  -1681.20229783
soc.religion.christian  ------>  -1674.24359874
rec.motorcycles  ------>  -1757.60634706
comp.graphics  ------>  -1797.51507597
comp.windows.x  ------>  -1810.9647455
comp.sys.ibm.pc.hardware  ------>  -1805.49632586
talk.politics.guns  ------>  -1700.58009768
sci.electronics  ------>  -1767.71600691
sci.med  ------>  -1707.22208927
sci.space  ------>  -1684.52408463
talk.politics.misc  ------>  -1633.6693405
alt.atheism  ------>  -1677.70165691
rec.sport.hockey  ------>  -1739.95037061
rec.sport.baseball  ------>  -1746.30972512
```

talk.politics.mideast ------> -1557.29923419
comp.os.ms-windows.misc ------> -1880.46169992
rec.autos ------> -1748.12355204
comp.sys.mac.hardware ------> -1788.65872898
talk.politics.mideast have highest log(probability) value -1557.29923419

Part 2-

For the 2 classes using different values of k:

For k = 5

Enter
1) for file input
2) for sentence input
1
Enter the file name new.txt
rec.motorcycles ------> -1805.07944546
rec.sport.baseball ------> -1804.76126916
rec.sport.baseball have highest log(probability) value -1804.76126916

For, 20 class classifier, given a file as input

Enter
1) for file input
2) for sentence input
1
Enter the file name new.txt
sci.crypt ------> -1830.16155686
misc.forsale ------> -1805.82677722
talk.religion.misc ------> -1801.94087876
soc.religion.christian ------> -1802.72571025
rec.motorcycles ------> -1805.07944546
comp.graphics ------> -1865.49223084
comp.windows.x ------> -1876.40567603
comp.sys.ibm.pc.hardware ------> -1833.19406855
talk.politics.guns ------> -1814.42119605
sci.electronics ------> -1815.15157538
sci.med ------> -1814.68968507
sci.space ------> -1802.4056832
talk.politics.misc ------> -1782.75638909
alt.atheism ------> -1798.67526118
rec.sport.hockey ------> -1821.75331753
rec.sport.baseball ------> -1804.76126916
talk.politics.mideast ------> -1728.02626635
comp.os.ms-windows.misc ------> -1950.58953211
rec.autos ------> -1805.32408617

comp.sys.mac.hardware ------> -1817.73679005
talk.politics.mideast have highest log(probability) value -1728.02626635

Part 3-

On increasing the value of k in Add-k smoothing , the model starts predicting wrong class . The probability of the the class to which the file belongs decreases and the probability of some other class increases and thus the file is wrongly classified to some other class.

So, Add-1 smoothing is the better one than Add-k smoothing for k=5,10,100

Problem 2:

Part 1: The sentence generated by the 2 classes for different models are:

For class rec.sport.baseball

Enter 1) for class1 and 2) for class2 1

unigram add_one value- -48.4086166184 perflexity 126.578372507
Bigram add_one value- -56.292009021 perflexity 278.439528268
Trgram add_one value- -8.18572176322 perflexity 2.26726027741

unigram add_one value- -59.5089218081 perflexity 1700.39392877
Bigram add_one value- -60.2293147178 perflexity 1860.61868868
Trgram add_one value- -67.5044719355 perflexity 4619.57963546

unigram add_one value- -37.77748214 perflexity 12636.8260811
Bigram add_one value- -30.412550238 perflexity 2004.47519929
Trgram add_one value- -31.4185211695 perflexity 2577.64196354
()
<S> the to in and of is that for he it <E>
<S> and the game in the braves fans are the best in <E>
<S> in article writes in article david <E>

unigram add_one value- -48.4086166184 perflexity 126.578372507
Bigram add_one value- -56.292009021 perflexity 278.439528268
Trgram add_one value- -8.18572176322 perflexity 2.26726027741
Enter the sentence baseball is a good game

unigram add_one value- -27.5576344321 perflexity 981.820604674
Bigram add_one value- -24.3165559642 perflexity 436.653070621
Trgram add_one value- -15.1063588597 perflexity 43.6668710447
Log probability using good turning(unigram) is 0.0

Log probability using good turning(bigram) is -29.1501460422
Log probability using good turning(trigram) is -22.8339901869

For class motorcycles

Enter 1) for class1 and 2) for class2 2

unigram add_one value- -47.9039526251 perflexity 120.348928683
Bigram add_one value- -59.9754568771 perflexity 402.439867312
Trgram add_one value- 0.0 perflexity 1.0

unigram add_one value- -60.8071759473 perflexity 859.50176903
Bigram add_one value- -62.6470615649 perflexity 1054.46058184
Trgram add_one value- -71.0914771027 perflexity 2694.72924175

unigram add_one value- -65.4071610486 perflexity 11428.6672599
Bigram add_one value- -55.9567120506 perflexity 2962.58064564
Trgram add_one value- -60.1054138542 perflexity 5358.75897882
()
<S> the to and of in it is you that on <E>
<S> and the bike is the same thing to the right side <E>
<S> in article writes in article charles parr writes bought it <E>

unigram add_one value- -47.9039526251 perflexity 120.348928683
Bigram add_one value- -59.9754568771 perflexity 402.439867312
Trgram add_one value- 0.0 perflexity 1.0
Enter the sentence baseball is a good game

unigram add_one value- -36.4206394313 perflexity 9001.62015936
Bigram add_one value- -7.98879744646 perflexity 7.36839097606
Trgram add_one value- 0.0 perflexity 1.0
Log probability using good turning(unigram) is -21.1252936194
Log probability using good turning(bigram) is -11.3074158125
Log probability using good turning(trigram) is 0.0