ASSIGNMENT 2

Q1.MULTINOMIAL NAIVE BAYES MODE

Assumptions:

- Meta data is considered as a part of data set.
- Convert all the data in the lower form.
- Remove all the words starting with digits.
- Remove punctuation.
- Lemmatization is performed.
- Remove stopwords.

Add -1 Smoothing output:

```
F:\MTECH1\NLP\Assignment3>python Q11.py
Enter your choice of input data2
Enter the filename Development_set.txt
Enter the value of k for task 1 1
probability of rec.motorcyle -78.73620773281229
probability of rec.sport.baseball -82.71776328917578
TASK1 START
rec.motorcycles -78.73620773281229
TASK1 END

Enter the value of k for task 2 1
TASK2 START
alt.atheism -81.50869517344573
comp.graphics -79.67978476232079
comp.os.ms-windows.misc -79.84715759211325
comp.sys.ibm.pc.hardware -78.8816618951236
comp.sys.mac.hardware -69.86117772676882
comp.windows.x -83.34674491041879
misc.forsale -83.3091312743828
rec.autos -86.06529598809595
rec.motorcycles -85.84684331149576
rec.sport.baseball -89.42488008756925
rec.sport.hockey -88.52947973979907
sci.crypt -77.80724585330175
sci.electronics -79.3168564115194
sci.med -82.60313297440455
sci.space -83.79851785541695
soc.religion.christian -85.53541247639872
talk.politics.guns -84.46600468538283
talk.politics.mideast -88.47568572514089
talk.politics.misc -86.47241558594294
talk.religion.misc -85.40890302386016


comp.sys.mac.hardware
-69.86117772676882
TASK2 END
```

Add -k Smoothing for k=5:

```
F:\MTECH1\NLP\Assignment3>python Q11.py
Enter your choice of input data 2
Enter the value of k for task 1 5
probability of rec.motorcyle -77.68231858973354
probability of rec.sport.baseball -78.87830206262346
TASK1 START
rec.motorcycles -77.68231858973354
TASK1 END

Enter the value of k for task 2 5
TASK2 START
alt.atheism -86.67082667765521
comp.graphics -85.342428161958
comp.os.ms-windows.misc -86.12848222299931
comp.sys.ibm.pc.hardware -85.17384400649838
comp.sys.mac.hardware -78.89522424987722
comp.windows.x -87.64868628784338
misc.forsale -89.09247914206435
rec.autos -90.42928596604209
rec.motorcycles -90.78011504648717
rec.sport.baseball -91.60297087400673
rec.sport.hockey -91.51600585558953
sci.crypt -83.72745144142134
sci.electronics -85.79553314035093
sci.med -87.55455577476731
sci.space -88.15158756554072
soc.religion.christian -88.37008601837402
talk.politics.guns -88.72855697401657
talk.politics.mideast -90.00514952020688
talk.politics.misc -89.38521830612123
talk.religion.misc -88.57445921050237


comp.sys.mac.hardware
-78.89522424987722
TASK2 END

F:\MTECH1\NLP\Assignment3>
```

Add –k Smoothing for k=10:

```
F:\MTECH1\NLP\Assignment3>python Q11.py
Enter your choice of input data 2
Enter the value of k for task 1 10
probability of rec.motorcyle -78.03029712992733
probability of rec.sport.baseball -78.38509431934938
TASK1 START
rec.motorcycles -78.03029712992733
TASK1 END

Enter the value of k for task 2 10
TASK2 START
alt.atheism -89.35703780217767
comp.graphics -88.34246581295245
comp.os.ms-windows.misc -89.21544191663973
comp.sys.ibm.pc.hardware -88.25951433075392
comp.sys.mac.hardware -83.04810417573663
comp.windows.x -90.20687435721251
misc.forsale -91.80580077936341
rec.autos -92.50154064665041
rec.motorcycles -93.05376027237655
rec.sport.baseball -93.1391676927801
rec.sport.hockey -93.21802600599378
sci.crypt -86.97290091043217
sci.electronics -88.96686964978012
sci.med -90.1567490760161
sci.space -90.56208295001346
soc.religion.christian -90.2598507404307
talk.politics.guns -91.00226873295945
talk.politics.mideast -91.48895287508653
talk.politics.misc -91.26155583297414
talk.religion.misc -90.66101362748623


comp.sys.mac.hardware
-83.04810417573663
TASK2 END
```

Add- k Smoothing for k=100:

```
F:\MTECH1\NLP\Assignment3>python Q11.py
Enter your choice of input data 2
Enter the value of k for task 1 100
probability of rec.motorcyle -81.09341371498967
probability of rec.sport.baseball -80.53381988072191
TASK1 START
rec.sport.baseball -80.53381988072191
TASK1 END

Enter the value of k for task 2 100
TASK2 START
alt.atheism -96.71516447558021
comp.graphics -96.63486009981204
comp.os.ms-windows.misc -97.3407178711065
comp.sys.ibm.pc.hardware -96.6740900633785
comp.sys.mac.hardware -94.31096787472318
comp.windows.x -97.401428218859
misc.forsale -98.64764000224882
rec.autos -97.99764531270718
rec.motorcycles -98.67211727630654
rec.sport.baseball -98.07045751075557
rec.sport.hockey -98.18637966292385
sci.crypt -96.14370875789919
sci.electronics -97.22809244595175
sci.med -97.20051010475228
sci.space -97.31194863459861
soc.religion.christian -96.4213137698529
talk.politics.guns -97.34525457407464
talk.politics.mideast -96.91091004450219
talk.politics.misc -97.03118516141521
talk.religion.misc -97.01346054250672


comp.sys.mac.hardware
-94.31096787472318
TASK2 END
```

Answer1.3:

In add-1 smoothing the expected classification of a test data is biased to one class i.e. the probability of that class is much greater than other classes of the classification.

In add-k smoothing the expected classification of test data is less biased to one class in comparison to add-1 smoothing because it will take less probabilities for OOV from the known words probabilities.

The Add- k Smoothing is better than add-1 smoothing for medium value of k i.e. not very large and not very small values of k.

On increasing the value of k slowly let's say k=5 the data will be smoothened slowly. On k=10 the data will close little bit more. On k=100 they are much closer sometimes it is difficult to distinguish to classify the class. So we select medium value k. So it may reduce the accuracy.

So in terms of accuracy k=1 gives better result.

Answer: 2

Assumptions

- Meta Data is removed from the given dataset.
- Remove the words starting with digits.
- Remove the punctuation.
- We are not considering <s> and</s> before and after the senetence.
- Remove the stopwords only in unigram.
- In good Turing assume k=5.

1. **Print the sentences with maximum probability**

```
F:\MTECH1\NLP\Assignment3>python Q2_1.py
 article bike dod like one get would
 article bike dod like one get would dont
 article bike dod like one get would dont writes
 article bike dod like one get would dont writes know
***unigram***
 article would year game one dont think
 article would year game one dont think team
 article would year game one dont think team last
 article would year game one dont think team last good
***unigram***
 i was a bike and then
 i was a bike and then the
 i was a bike and then the day
 i was a bike and then the day is
***bigram***
 i think that it would not
 i think that it would not so
 i think that it would not so that
 i think that it would not so that in
***bigram***


 it is a good bike to
 it is a good bike to learn
 it is a good bike to learn on
***trigram***


 i dont know what it takes
 i dont know what it takes a
***trigram***
```

2&3. **Log probability and perplexity of a given sentence.**

```
F:\MTECH1\NLP\Assignment3>python Q2_3.py
Enter the sentence There is a group of motorcyclists that gets together and does all the normal
The unigram log probability of the class motorcyc is -38.422369
The unigram perplexity of the given sentence in motorcyc class is 555.206998
*******
The unigram log probability of the class baseball is -40.043802
The unigram perplexity of the given sentence in baseball class is 724.889073
*******
The bigram log probability of the class motorcyc is -46.647838
The bigram perplexity of the given sentence in motorcyc class is 2147.773231
*******
The bigram log probability of the class baseball is -50.381132
The bigram perplexity of the given sentence in baseball class is 3968.736385
*******
The trigram log probability of the class motorcyc is -59.474402
The trigram perplexity of the given sentence in motorcyc class is 17708.083196
*******
The trigram log probability of the class baseball is -64.986719
The trigram perplexity of the given sentence in baseball class is 43843.829788
*******
```

4. **Advance smoothing of a given sentence:**

The technique used : Good Turing smoothing

```
Enter the sentence They argue with each other and plan rides together.
the unigram prob. of class motorcyc -28.86921042570865
the unigram prob. of class baseball -31.2944502117113
the bigram prob. of class motorcyc -5.201970915114129
the bigram prob. of class baseball -5.291804256860793
the trigram prob. of class motorcyc -3.4640533534837252
the trigram prob. of class baseball -3.4600248651071976
```