NLP ASSIGNMENT 2 REPORT

Problem 1:

Part 1-

TASK 1: I tested the 2 class classifier using the file TestMotor.txt present in my TestFiles folder. This file contained information related to motorcycles and the model correctly predicted the class to be motorcycle with the following output:

Probability of rec.motorcycles = -736.7932577832585

Probability of rec.sport.baseball = -849.0339758961602

rec.motorcycles has highest probability

TASK 2 : Testcase 1- I tested the 20 class classifier using the file TestAth.txt present in my TestFiles folder . This file contained information related to atheism and the model correctly predicted the class to be alt.atheism with the following output:

Maximum prob is -2517.3851946007744

Maximum prob class is alt.atheism

Testcase 2 - I tested the 20 class classifier using the file TestMidPolitics.txt present in my TestFiles folder . This file contained information related to middleeast politics and the model correctly predicted the class to be talk.politics.mideast with the following output:

Maximum prob is -1081.5762746665932

Maximum prob class is talk.politics.mideast

Part 2-

TASK 1 : I tested the 2 class classifier using the file TestMotor.txt present in my TestFiles folder. This file contained information related to motorcycles and the model gave the following output for different values of k:

For k = 5

Probability of rec.motorcycles = -744.8454247317785

Probability of rec.sport.baseball = -820.5171284718913

rec.motorcycles has highest probability

For k = 10

Probability of rec.motorcycles = -1080.4623484805393

Probability of rec.sport.baseball = -1067.4311529134084

rec.sport.baseball has highest probability

For k = 100

Probability of rec.motorcycles = -1113.9792993627957

Probability of rec.sport.baseball = -1106.509786642353

rec.sport.baseball has highest probability

TASK 2:

I tested the 20 class classifier using the file TestAth.txt present in my TestFiles folder . This file contained information related to atheism and the model gave the following output for different values of k:

For k=5

Maximum prob is -2759.561337521286

Maximum prob class is alt.atheism

For k = 10

Maximum prob is -2895.049468566201

Maximum prob class is alt.atheism

For k=100

Maximum prob is -3316.1589944771313

Maximum prob class is soc.religion.christian

Part 3-

Difference between the 2 techniques above:

Add-1 smoothing assigns comparatively higher value to the class that is actually predicted by the model and which is also the correct class for the document.

While Add-k smoothing assigns the probability value which is lesser than that assigned by the Add-1 smoothing to the correct class.

Which one should produce better results and happens on increasing values of k?

Add-1 smoothing should produce better results than Add-k smoothing. We can also observe this above for the 2 -class classifier as well as the 20-class classifier:

For 2- class: I took the file which actually should be classified to class motorcycle and on k=1 and k=5 it is correctly classified while for k=10 and k-100 it is wrongly classified to baseball class.

Same for 20-class classifier the file gets wrongly classified to class Christianity for k=100.

On increasing values of k the document gets classified to possibly overlapping class in 20-class classifier like class atheism has a slight overlap with the class Christianity.

Problem 2:

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

TASK 1: For class rec.sport.baseball

1. Unigram
   writes would article year game one dont
   log(probability of the sentence) = -15.506434723436277

2. Bigram
   i think that the game in the game
   log probablity of the sentence = -9.093419873907768
3. Trigram
   in article aprcscornelledu tedwardcscornelledu edward ted fischer writes in
   log probablity of the sentence = -1.9734806596694643

TASK 2: For class rec.motorcycles

1. Unigram
   writes article bike dod like one get

log(probability of the sentence) = -15.06548056215115

2. Bigram
   i was a few weeks i was a
   log probablity of sentence = -8.594653756099255
3. Trigram
   in article newsdukeedu infanteacpubdukeedu andrew infante writes well as
   log probablity of sentence = -2.824112407429834

Part 2,3,4: In part 4 I am using Good-Turing Smoothing technique

TASK 1: For class rec.sport.baseball ,

I am using the input sentence as:

Hello, let us play baseball today, as i love sports

1. **Unigram**
   log(probability of input sentence) = -25.94624372993785
   perplexity = 1750.977100015434

   Using Good Turing baseball unigram, log(prob) of sentence = -26.937394817036648

2. **Bigram**
   The log probability of the sentence = -37.31819633285345
   perplexity = 5392.866050373287

   Using Good Turing baseball bigram , log(prob) of sentence = -21.621023387817754

3. **Trigram**
   Sentence probablity = -40.572558767609216
   perplexity = 11409.217956147444

   Using Good Turing baseball trigram , log(prob) of sentence = -2.5946292292766233

TASK 2: For class rec.motorcycles ,

I am using the input sentence as:

Hey, let us go driving the bike on the highway today it is a good day today

1. **Unigram**
   log(probability of input sentence) = -33.21296766033492
   perplexity = 1045.5882554466934

   Using Good Turing motorcycle unigram,
   log(prob) of sentence = -32.613576733401246

2. **Bigram**
   The log probability of the sentence = -53.761929898976824
   perplexity = 1453.6721316089993

   Using Good Turing motorcycle bigram ,
   log(prob) of sentence = -40.67395404908519

3. **Trigram**
   Sentence probablity = -63.79781055311344
   perplexity = 5659.947273984326

   Using Good Turing motorcycle trigram ,
   log(prob) of sentence = -29.991944324705823

While doing Good-Turing Smoothing to overcome the drawback of getting a N(c+1) value to be 0 , I used the Katz equation and observed a threshold i.e the value of k for the 2 classes for the different model . If the count value is greater than the threshold then the revised count is not calculated for those values and the old count is only taken. The words for which the counts is less than the threshold the revised counts are calculated using the Katz equation.

ASSUMPTIONS:

1. The sentence given as input for Problem 2 is not Null i.e empty sentence.
2. Also, for trigram the sentence given as input is of length atleast 2 after preprocessing.