Assignment 1 Language Modelling

Prakash Kumar Uttam
SR -15247
https://github.com/pkuttam/ngram-assignment

1 Question - Perplexity Measure:

Tr-G = train on gutenberg dataset (80%)
Tr-B = train on brown dataset (80%)
Tr-GB = train on gutenberg and brown dataset (80%)*each*

Ts-G = test on gutenberg dataset (20%)
Ts-B = test on brown dataset (20%)

1.1 Add-K smoothing

—	uni-gram	bi-gram	trigram
Tr-B and **Ts-B**	399	291	Large
Tr-G and **Ts-G**	514	592	Large
Tr-GB and **Ts-B**	600	100	Large
Tr-GB and **Ts-G**	572	655	Large

1.2 kneser-Nay smoothing

—	uni-gram	bi-gram	trigram
Tr-B and **Ts-B**	-	63	97
Tr-G and **Ts-G**	-	98	160
Tr-GB and **Ts-B**	-	109	175
Tr-GB and **Ts-G**	-	75	110

2 Question - Sentence Generation :

2.1 add-K smoothing

Example- brown dataset
i have been a good deal of time . to be a “ great service ” . no permission to enter the university of chicago and all the way to the editor of the united states , and the other hand , the first time in the first two years

Example- Gutenberg dataset
i will not be afraid of the lord , and the lord . make thee a great deal of the house of the children of israel , and he said , ” i am sure i should have been a great many more . give you a great multitude ,

Example- Gutenberg dataset + brown dataset
i have not been able to go to the king of judah , and the lord , and he said , ” i am sure i should have been a great deal of the lord . been the case of the house of the children of israel , and i will

2.2 kneser-Nay smoothing

Example- brown dataset
i have been , and the other hand , the , country – and the “ the lord is my light and power company , and , in the first time in the world . to be a “ a ” . a good deal of the , disciplines that

Example- Gutenberg dataset
i will not be a great deal of the lord , and the lord . make thee a man of god , and he said , ” i am sure i should be the lord god of israel , and i will give you a great , and to the lord

Example- Gutenberg dataset + brown dataset
i have been a great deal of the lord , and the lord . not been in the land of egypt , and he said , ” i am sure i should be the lord god of israel , and i will not be a great many of the house of

3 observation

The best model is Gutenberg + Brown data-set trained on tri-gram with Kneser-Nay smoothing.