Natural Language Processing

Kaustav Vats | 2016048 | Assignment-2

Multinomial Naive Bayes

Performed Add-K smoothing on 2 class and 20 class of 20 newspaper dataset.
Implemented Multinomial Naive Bayes from scratch.

Accuracy for both tasks	K = 1	K = 5	K = 10	K = 100
2 Class Models	0.9916	0.98	0.98	0.868
20 Class Models	0.82166	0.7805	0.7394	0.5472

For 2 class model we can see that there's not much change in accuracy on increasing K value. For K = 100 there's an exception for 2 class model where accuracy nearly decrease by 10%.
For class 2 I took 70:30 Data split.

For 20 class model, it's clear that the accuracy is decreasing by a significant amount. Probable reason for this decrease is that the actual probabilities are getting reduced by large amount on increasing K value. Which makes the probability of words incorrect thus resulting in lower accuracy.
For 20 Class I took 91:9 Data split for Training and testing of each class.

Add-1 should produce better results. Each word probability is reduced by a small amount, and by Add-1 smoothing we have also avoided zero probability of any words.

Language Modelling

Q1 - Sentence Generation- [Baseball, Motorcycles] [UniGram, BiGram, TriGram]

UniGram sentence Generation-

Class	Sentence	Log10 Probability
Baseball	the to a and of in	-4.796
Motorcycles	the a to i and of	-4.980

BiGram Sentence Generation-

Class	Sentence	Log10 Probability
Baseball	i think that the game in	-13.272
Motorcycles	i was a few weeks i	-15.074

TriGram sentence Generation-

Class	Sentence	Log10 Probability
Baseball	i do not know what it	-13.203
Motorcycles	if you want to go left	-13.029

Q2, 3 - Predicting probability and perplexity of input sentence

Sentence - *How about a Geeky temporary tatoo?* [Taken from motorcycles class]

N-Gram \ Class	Baseball Class [Log Probability]	Motorcycles Class [Log Probability]
Uni Gram	-22.812	-20.547
Bi Gram	-21.956	-20.399
Tri Gram	-20.361	-18.386

N-Gram \ Class	Baseball Class [Perplexity]	Motorcycles Class [Perplexity]
Uni Gram	6339.512	2658.084
Bi Gram	4563.887	2511.472
Tri Gram	2475.206	1159.792

Q4 - Predicting Probability using Good Turing Smoothing

Sentence - *the sale of the Orioles to anyone is likely* [Taken from Motorcycles Class]

N-Gram \ Class	Baseball Class [Log Probability]	Motorcycles Class [Log Probability]
Uni Gram	-23.123	-19.998
Bi Gram	4.112	7.071
Tri Gram	23.279	18.313

Assumptions

1. Sentence must only contain words.
2. There's no need to add count for frequency of last word in BiGram and TriGram.
3. We can take at least 1 frequency for each N-Gram to avoid log undefined and division by zero.
4. Good Turing is only required for words with count less than 5.