Tuesday Oct 1, 2024

In-class Handout
COSC 426A NLP
Prof. Forrest Davis

The following is to help you ground your understanding of ngram computations, namely MLE and smoothing.[1] We will be working with a very small corpus (ignoring start and end special tokens). Assume we observe the following corpus:

a a a b

a b b a

c a a a

First, let's calculate the unigram and bigram MLE probabilities. Again, we are assuming that our vocabulary is all three "words", with **no** <unk>, BOS, and EOS. Unigram MLE probabilities are:

	MLE prob
p(a)	8/12
p(b)	3/12
p(c)	1/12

and the Bigram MLE probabilities are:

	MLE prob
p(a|a)	4/6
p(b|a)	2/6
p(c|a)	0
p(a|b)	1/2
p(b|b)	1/2
p(c|b)	0
p(a|c)	1
p(b|c)	0

[1] This is inspired by a great handout of Prof. Kauchak from Pomona CS

p(c|c)	0

For the following, we are interested in modifying our bigram probabilities. Why?

First, let's try **add-K smoothing** with k = 2.

	equation	MLE prob
p(a|a)	(4+2)/(6+2*3)	6/12
p(b|a)	(2+2)/(6+2*3)	4/12
p(c|a)	(0+2)/(6+2*3)	2/12
p(a|b)	(1+2)/(2+2*3)	3/8
p(b|b)	(1+2)/(2+2*3)	3/8
p(c|b)	(0+2)/(2+2*3)	2/8
p(a|c)	(1+2)/(1+2*3)	3/7
p(b|c)	(0+2)/(1+2*3)	2/7
p(c|c)	(0+2)/(1+2*3)	2/7

Next, let's try interpolation smoothing with λ=0.8.

	equation	MLE prob
p(a|a)	(8/10)*(4/6)+(2/10)*(8/12)	80/120
p(b|a)	(8/10)*(2/6)+(2/10)*(3/12)	38/120
p(c|a)	(8/10)*(0)+(2/10)*(1/12)	2/120
p(a|b)	(8/10)*(1/2)+(2/10)*(8/12)	64/120
p(b|b)	(8/10)*(1/2)+(2/10)*(3/12)	54/120
p(c|b)	(8/10)*(0)+(2/10)*(1/12)	2/120
p(a|c)	(8/10)*(1)+(2/10)*(8/12)	112/120
p(b|c)	(8/10)*(0)+(2/10)*(3/12)	6/120
p(c|c)	(8/10)*(0)+(2/10)*(1/12)	2/120

I’ve filled in cells where you should really focus your attention. They highlight the purpose of smoothing and key contrast between approaches. If you understand them, you are well on your way to understanding smoothing.