19BCE245-Aayush Shah

Practical 7 – Virtual Lab of Language Models

1) N-Grams

Aim of the practical: The practical's goal is to determine the likelihood of a sentence using the chance of a word sequence happening in it. The Markov assumption states that the likelihood of a word appearing in a phrase is determined by the probability of the word appearing immediately before it. Bigram or Markov models are examples of this type of model. This is the formula for it:

$$P(W_n | W_{n-1}) = P(W_{n-1}, W_n)/P(W_{n-1})$$

Objective of this practical: To learn to calculate bigrams from a given corpus and calculate probability of a sentence.

N grams

Corpus 1

N-Grams

Corpus A

Select Corpus

(eos) Can I sit near you (eos) You can sit (eos) Sit near him (eos) I can sit you (eos)

Find Bigram Probabilities

	(eos)	I	you	him	can	near	sit
(eos)	0	.2	.2	0	0.2	0	0.2
I	0	0	0	0	.5	0	0.5
you	0.66	0	0	0	0.33	0	0
him	1	0	0	0	0	0	0
can	0	0.33	0	0	0	0	0.66
near	0	0	0.5	.5	0	0	0
sit	.25	0	.25	0	0	0.5	0

Submit

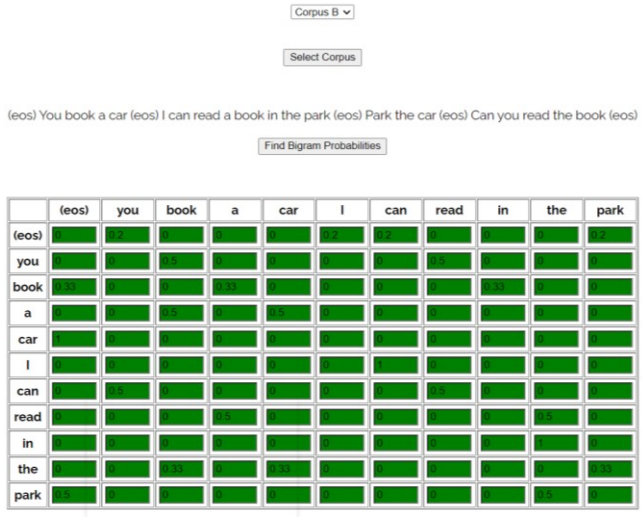
Find probabilities of the following sentences:

Sentence	Probability
I sit you EOS	0.0825
Can you sit near I EOS	0
I can sit EOS	0.0825
You sit EOS	0

Submit

Wrong Answer

Corpus 2



	(eos)	you	book	a	car	I	can	read	in	the	park
(eos)	0	0.2	0	0	0	0.2	0.2	0	0	0	0.2
you	0	0	0.5	0	0	0	0	0.5	0	0	0
book	0.33	0	0	0.33	0	0	0	0	0.33	0	0
a	0	0	0.5	0	0.5	0	0	0	0	0	0
car	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	0	0
can	0	0.5	0	0	0	0	0	0.5	0	0	0
read	0	0	0	0.5	0	0	0	0	0	0.5	0
in	0	0	0	0	0	0	0	0	0	1	0
the	0	0	0.33	0	0.33	0	0	0	0	0	0.33
park	0.5	0	0	0	0	0	0	0	0	0.5	0

Find probabilities of the following sentences:

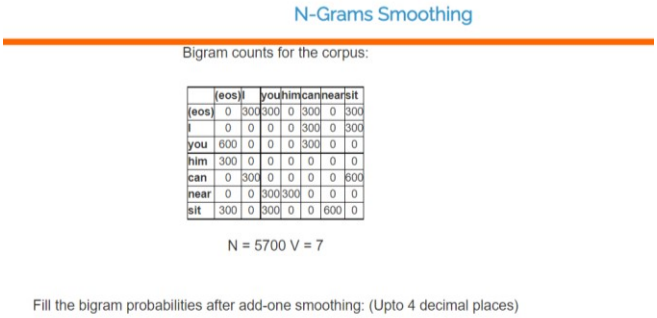
Sentence	Probability
I read EOS	0
Can you park car EOS	0
I book park EOS	0
You park in the park EOS	0

Submit

2) N Gram Smoothing

Corpus A

Question :



	(eos)	I	you	him	can	near	sit
(eos)	0	300	300	0	300	0	300
I	0	0	0	0	300	0	300
you	600	0	0	0	300	0	0
him	300	0	0	0	0	0	0
can	0	300	0	0	0	0	600
near	0	0	300	300	0	0	0
sit	300	0	300	0	0	600	0

Output :

Fill the bigram probabilities after add-one smoothing: (Upto 4 decimal places)

	(eos)	I	you	him	can	near	sit
(eos)	0.0002	0.0527	0.0527	0.0002	0.0527	0.0002	0.0527
I	0.0002	0.0002	0.0002	0.0002	0.0527	0.0002	0.0527
you	0.1053	0.0002	0.0002	0.0002	0.0527	0.0002	0.0002
him	0.0527	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
can	0.0002	0.0527	0.0002	0.0002	0.0002	0.0002	0.1053
near	0.0002	0.0002	0.0527	0.0527	0.0002	0.0002	0.0002
sit	0.0527	0.0002	0.0527	0.0002	0.0002	0.1053	0.0002

Conclusion: We learned how to use N-grams smoothing on a corpus's N-gram (bigram counts) table to identify re-evaluated probabilities that are not zero-probabilities.