COMP 472: Artificial Intelligence

n-gram Modeling

Solutions

1 Question

Assume that we are working with the Shloutan language. If you don't know Shloutan, don't worry; it is a simple language made of only 5 words: loola nikee aloka bibi vo.
You want to build a word language model for Shloutan. The training corpus that you use is the following:

"Loola nikee. Aloka bibi vo. Vo bibi loola. Loola nikee bibi vo. Vo. Vo. Aloka bibi loola. Loola aloka aloka. Loola loola. Nikee nikee nikee. Bibi vo. Bibi vo. Vo Vo. Nikee loola."

You can ignore case distinctions and sentence boundaries when answering the following questions.

(a) What is the value of P(vo | bibi)?

Solution

$$P(vo|bibi) = \frac{freq(\text{bibi vo})}{\sum_{w_i} freq(\text{bibi } w_i)} = \frac{4}{6}$$

(b) What is the value of P(bibi vo)?

Solution

$$P(bibi\ \ vo) = \frac{freq(bibivo)}{\sum_{w_j}\sum_{w_i} freq(w_j\ \ w_i)} = \frac{4}{32}$$

(c) Build a bigram language model based on this training corpus. Show the frequencies and the probabilities for each bigram.

Frequencies

Solution

	loola	nikee	aloka	bibi	vo
loola	3	3	1	0	0
nikee	1	2	1	2	0
aloka	1	0	1	2	0
bibi	2	0	0	0	4
vo	0	1	1	2	5

Conditional Probabilities

Solution

	loola	nikee	aloka	bibi	vo
loola	0.43	0.43	0.14	0	0
nikee	0.17	0.33	0.17	0.33	0
aloka	0.25	0	0.25	0.50	0
bibi	0.33	0	0	0	0.67
vo	0	0.11	0.11	0.22	0.56

(d) Smooth your bigram language model using "add 0.5". Show the frequencies and the probabilities for each bigram.

Frequencies

Solution

	loola	nikee	aloka	bibi	vo
loola	3.5	3.5	1.5	0.5	0.5
nikee	1.5	2.5	1.5	2.5	0.5
aloka	1.5	0.5	1.5	2.5	0.5
bibi	2.5	0.5	0.5	0.5	4.5
vo	0.5	1.5	1.5	2.5	5.5

Conditional Probabilities

Solution

	loola	nikee	aloka	bibi	vo
loola	0.37	0.37	0.16	0.05	0.05
nikee	0.18	0.29	0.18	0.29	0.06
aloka	0.23	0.08	0.23	0.38	0.08
bibi	0.29	0.06	0.06	0.06	0.53
vo	0.04	0.13	0.13	0.22	0.48

Using each language model from parts (c) and (d), which of the following 2 sentences is more probable. Show all your work.

sentence 1: Aloka vo nikee aloka.

Solution

Bigrams: (aloka vo), (vo nikee), (nikee aloka)
Model (c): $0 \times 0.11 \times 0.17 = 0$
Model (d): $0.08 \times 0.13 \times 0.18 \approx 0.0019$

sentence 2: Vo nikee nikee aloka.

Solution

Bigrams: (vo nikee), (nikee nikee), (nikee aloka)
Model (c): $0.11 \times 0.33 \times 0.17 \approx 0.0062$
Model (d): $0.13 \times 0.29 \times 0.18 \approx 0.0068$

Sentence 2 is more probable using either model.