Let the partial derivative equal to 0, $\frac{\partial \log P}{\partial \theta} = \frac{n}{\theta} - \frac{N-n}{1-\theta} = 0$.

We get $\theta = \frac{n}{N}$, so $p_t = s/|R|$.

Exercise 12.6

Consider making a language model from the following training text:

The martian has landed on the latin pop sensation ricky martin

a. Under a MLE-estimated unigram probability model, what are *P*(the) and *P*(martian)?

b. Under a MLE-estimated bigram model, what are *P*(sensation | pop) and *P*(pop | the)?

Solution

a. P(the) = 2/11, P(martian) = 1/11

b. *P*(sensation | pop) = 1, *P*(pop | the) = 0

Exercise 12.7

Suppose we have a collection that consists of the 4 documents given in the below table.

docID Document text

1 click go the shears boys click click click

2 click click

3 metal here

4 metal shears click here

Build a query likelihood language model for this document collection. Assume a mixture model between the documents and the collection, with both weighted at 0.5. Maximum likelihood estimation (mle) is used to estimate both as unigram models. Work out the model probabilities of the queries click, shears, and hence click shears for

each document, and use those probabilities to rank the documents returned by each query. Fill in these probabilities in the below table:

Query Doc 1 Doc 2 Doc 3 Doc 4

click

shears

click shears

What is the final ranking of the documents for the query click shears?

Solution

Language models

	click	go	the	shears	boys	metal	here
model1	1/2	1/8	1/8	1/8	1/8	0	0
model2	1	0	0	0	0	0	0
model3	0	0	0	0	0	1/2	1/2
model4	1/4	0	0	1/4	0	1/4	1/4

Collection model	7/16	1/16	1/16	2/16	1/16	2/16	2/16

model probabilities of the queries

query	doc1	doc2	doc3	doc4
click	0.4688	0.7188	0.2188	0.3438
shears	0.125	0.0625	0.0625	0.1875
click shears	0.0586	0.0449	0.0137	0.0645

Final ranking for the query click shears: doc4 > doc1 > doc2 > doc3