An NLP Tip

Javad.Pourmostafa

December 2019

1 EM-Algorithm for PLSA Topic Model

Consider the following tiny document: One fly flies, two flies fly.

After Lemmatization: **One fly fly, two fly fly**. So let us use this version of the text below.

Consider ϕ matrix from the latest M-step:

	topic1	***topic2***	***topic3***
fly	0.1	0.8	0.2
one	0.4	0.1	0.3
two	0.5	0.1	0.5

Table 1: Phi Matrix

And θ column for the document:

	Document
Topic 1	0.2
Topic 2	0.7
Topic 3	0.1

Table 2: Theta Matrix

1) Compute posterior topic probabilities of E-step for the word fly.
2) Compute n_{wt} count for the word fly and topic 2. (Assume there are no other documents in the corpus).

1.1 Solution:

$$\frac{0.8 \times 0.7}{(0.1 \times 0.2) + (0.8 \times 0.7) + (0.2 \times 0.1)} = \frac{0.56}{0.6} = 0.9333 \tag{1}$$

$$n_{wt} = 4 \times 0.9333 = 3.7333 \tag{2}$$