

# Topic Models [ Solution by Karthikeyan.S ]

LATEST SUBMISSION GRADE

100%

1.Question 1

How many parameters does PLSA topic model have?

Let us denote the vocabulary size by  $|W|$ , the number of documents by  $|D|$ , the length of the corpus by  $|N|$ , and the number of topics by  $|T|$ .

☐

$$|T| \cdot |N| \cdot |T| \cdot |N|$$

☒

$$|T| \cdot |W| \cdot |T| \cdot |W| + |T| \cdot |D| \cdot |T| \cdot |D|$$

☐

$$|W| \cdot |D| \cdot |W| \cdot |D|$$

Correct

2.Question 2

Which assumptions are made in PLSA topic model?

☒

Bag of words assumption

☐

Topic distributions are sparse and diverse

☒

Conditional independence:  $p(w \mid t, d) = p(w \mid t)p(w \mid t, d) = p(w \mid t)$

☐

Distributions of words in topics have Dirichlet prior

☐

Conditional independence:  $p(t \mid w, d) = p(t \mid d)p(t \mid w, d) = p(t \mid d)$

Correct

3.Question 3

Let's see how EM-algorithm for PLSA works.

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

Before building a topic model, one would usually apply lemmatization and obtain the following: **One fly fly, two fly fly**. So let us use this version of the text below.

Consider  $\Phi$  matrix from the latest M-step:

	topic 1	topic 2	topic 3
<b>fly</b>	0.1	0.8	0.2
<b>one</b>	0.4	0.1	0.3
<b>two</b>	0.5	0.1	0.5

And  $\Theta$  column for the document:

	document
<b>topic 1</b>	0.2
<b>topic 2</b>	0.7
<b>topic 3</b>	0.1

1) Compute posterior topic probabilities of E-step for the word **fly**.

2) Compute  $n_{\{wt\}} n_{wt}$  count for the word **fly** and **topic 2**. (Assume there are no other documents in the corpus).

Enter  $n_{\{wt\}} n_{wt}$  value with 2 digits after the decimal point.

If you have difficulties with this question, get back to the last in-video question in the corresponding video. There is a full explanation of the solution there.

**Correct**

#### 4.Question 4

Imagine you are analysing news flow for a company. You want to know what topics are being mentioned when people discuss the company, and how they change over time.

For each news article there are several modalities that you want to use: English text, time, author and category. Your final goal is to track, how topics change over time.

Which additive regularizers would you add to your topic model?

☐

Visualizable

☒

Multimodal

☐

Dynamic

☐

Multilingual

☐

Hierarchical

**Correct**