# Week 3 Practice Quiz

**10/10** points earned (100%)

Excellent!

Retake

[Course Home](https://www.coursera.org/learn/text-mining/home/welcome)

Correct

1 / 1 points

1. You are given two unigram language models, θ1 and θ2, as defined in the table below:

|  |  |  |
| --- | --- | --- |
| **w** | **P(w|θ1)** | **P(w|θ2)** |
| the | 0.4 | 0.05 |
| of | 0.4 | 0.05 |
| technology | 0.1 | 0.5 |
| machine | 0.1 | 0.4 |

Suppose we are using a mixture model for document clustering based on the two given unigram language models, θ1 and θ2, such that P(θ1)=0.3 and P(θ2)=0.7. To generate a document, first, one of the two language models is chosen according to P(θi), and then all the words in the document are generated based on the chosen language model.

The probability of generating a document composed only of the one word “technology” using the given mixture model is P(“technology”)=

1. 0.7
2. 0.58
3. **0.38**

**Correct Response**

0.1 \* 0.3 + 0.5 \* 0.7

1. 0.3

Correct

1 / 1 points

2. Assume the same given as in Question 1. What is the probability of generating a document composed only of the phrase “the technology”, i.e., P(“the technology”)?

1. 0.1444
2. **0.0295**

**Correct Response**

0.4 \* 0.1 \* 0.3 + 0.05 \* 0.5 \* 0.7

1. 0.0589
2. 0.3

Correct

1 / 1 points

3. In mixture model, why do different components tend to assign high probability on different words?

1. Because during training, when different components assign high probability to the same model, the training restarts
2. **Because it gives a higher overall likelihood**

**Correct Response**

If they all assign the high probability to the same words, then it generally indicates that the chosen number of components is too many.

1. Because the model was initialized with components with high probability assigned to different words

Correct

1 / 1 points

4. Why it is good to have the "background" component? Check all that apply.

* To improve model likelihood

**Correct Response**

* **To prevent overfitting**

**Correct Response**

By applying human knowledge that words in background components does not form any useful topic, adding such background components improves the robustness of the model

* **To better filter topic words into other components**

**Correct Response**

As all non-topic words such as stop words goes into the background component

Correct

1 / 1 points

5. What type of words are usually assigned with high probability in the background component?

1. "car", "cat", "catch"
2. "computer", "information", "data"
3. **"the", "he", "she", "is"**

**Correct Response**

those are words with very high frequency (stop words)

Correct

1 / 1 points

6. Which of the following about the EM algorithm is false?

1. It always increase the likelihood.
2. **The result of the EM algorithm does not depend on the initialization.**

**Correct Response**

EM algorithm has different results for different initialization

1. It is generally considered a fast algorithm for optimizing likelihood.
2. It can be trapped into a local optimal solution.

Correct

1 / 1 points

7. In EM, what does the E-step do?

1. Given the predicted values of unseen data, maximizes the joint likelihood
2. **Predicts values of unseen (hidden) variables**

**Correct Response**

Correct

1 / 1 points

8. Which of the following generative descriptions is not TRUE about PLSA?

1. To generate a word, a topic is drawn from the document's topic weight distribution, and a word is drawn according to the topic's word distribution.
2. **To generate a topic assignment for a word, a coin is tossed to decide if the topic is from the background topic or not, and the probability of the background is a constant specified by the user.**

**Correct Response**

the probability of background component weight is learnt not fixed.

1. To generate a document, a distribution of topic weights (multinomial distribution) is assumed, which is considered part of the model.

Correct

1 / 1 points

9. In PLSA, which of the following variables are part of the model? Check all that apply.

* *z*, topic assignment

**Correct Response**

* ***λ*, background percentage**

**Correct Response**

* ***θ*, topic's word distribution**

**Correct Response**

Correct

1 / 1 points

10. True or false? Let *θ*1,...,*θk* be the k unigram language model's output by PLSA. Then, for a specific word w, the following relation always holds: ∑*ki*=1*P*(*w*|*θi*)=1.

1. **False**

**Correct Response**

∑*wP*(*w*|*θi*)=1

1. True