# Week 4 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. Which of the following is NOT a use of text clustering?

1. Grouping similar words together
2. **Grouping similar pictures together**

**Correct Response**

This is generally considered as image processing or computer vision

1. Grouping similar websites together
2. Grouping similar documents together

Correct

1 / 1 points

2. Suppose we are performing clustering on a collection of documents using a mixture model as discussed in the lecture **Text Clustering: Generative Probabilistic Models (Part 3)**. Then, if we add more documents to the collection such that no new words are added to the vocabulary, the number of parameters to be estimated by the EM algorithm, i.e., P(θi) and P(w|θi), will:

Note: Do **not** count the probabilities associated with the hidden variables (i.e., those estimated in the E-step) as parameters.

1. Increase
2. **Stay the same**

**Correct Response**

because there is no more new words nor new components.

1. Decrease

Correct

1 / 1 points

3. The following table shows the **similarity** values between a set of emails as well as a binary label associated with each email indicating whether it is spam (label=1) or ham (label=0).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **D1** | **D2** | **D3** | **D4** | **D5** | **D6** | **Label** |
| **D1** | 100.0 | 0.1 | 0.5 | 0.8 | 0.82 | 0.85 | 1 |
| **D2** | 0.1 | 1000.0 | 0.85 | 0.05 | 0.12 | 0.7 | 0 |
| **D3** | 0.5 | 0.85 | 10000.0 | 0.1 | 0.1 | 0.6 | 0 |
| **D4** | 0.8 | 0.05 | 0.5 | 100000.0 | 0.9 | 0.1 | 1 |
| **D5** | 0.82 | 0.12 | 0.1 | 0.9 | 1000000.0 | 0.3 | 1 |
| **D6** | 0.85 | 0.7 | 0.6 | 0.1 | 0.3 | 1.0 | ? |

Suppose we use {D1,D2,D3,D4,D5} as our training dataset and use the k-Nearest Neighbor classifier to predict the label of email D6. If k=1, then the prediction of the classifier for D6 is:

1. **1**

**Correct Response**

D1 is most closed to D6

1. There is a tie and thus 0 or 1.
2. 0

Correct

1 / 1 points

4. Assume the same setup as in Question 3. If k = 2, then the prediction would be:

1. 0
2. 1
3. **There is a tie and thus 0 or 1.**

**Correct Response**

the next closed data is D2

Correct

1 / 1 points

5. Which of the following is TRUE about the mixture model?

1. **Words of the document are drawn from a mixture of topics where the mixing weight depends on different documents.**

**Correct Response**

1. Topics are a mixture of words where the mixing weight depends not only on the topics but also the documents.

Correct

1 / 1 points

6. Which of the following is NOT true about the maximal likelihood of a set of documents?

1. If we have a document "w1 w2 ... wn" changed into "wn ... w2 w1", the maximal likelihood does not change.
2. If we exchange every word "A" and "B", the maximal likelihood does not change.
3. **If we have every document doubles (a document "w1 w2 ... wn" becomes "w1 w1 w2 w2 ... wn wn"), then the maximal likelihood does not change.**

**Correct Response**

say the original likelihood is P, then after the change, it becomes *P*2

Correct

1 / 1 points

7. If we have a large collection of documents to train PLSA with, what is the best way to initialize the model?

1. Initialize each topic as a distribution with probability 1 on a random single word but zero everywhere else and documents' topic weight to be 1 on a random topic but 0 everywhere else
2. Randomly initialize
3. **Train PLSA on a small subset collection of documents and use the model to initialize, and for other documents randomly initialize the documents' topic weights**

**Correct Response**

Using a small set of data to train PLSA for initialization is the best choice for a large dataset

Correct

1 / 1 points

8. Which of the following is correct about K-means and PLSA?

1. **Both algorithms require the user to specify the number of clusters/topics.**

**Correct Response**

the number of clusters/topics is given by user

1. Only the results of PLSA depend on the way it was initialized.
2. Both of them have a clear objective function.
3. Only K-means is an iterative algorithm.

Correct

1 / 1 points

9. What is the disadvantage of using a model-based clustering algorithm?

1. It's much slower to train.
2. The performance is much worse than other methods.
3. **It is difficult to substitute a different similarity measure.**

**Correct Response**

Correct

1 / 1 points

10. What is the difference between direct and indirect evaluation for a clustering algorithm? Check all that apply.

* **Indirect evaluation requires a user specified application to test with.**

**Correct Response**

* **Direct evaluation requires a human annotated gold standard cluster.**

**Correct Response**

* Direct evaluation is better than indirect evaluation.

**Correct Response**

there is no clear preference between the two