# Week 5 Quiz

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

Quiz passed!

[Back to Week 5](https://www.coursera.org/learn/text-mining/home/week/5)

Correct  
1 / 1 points

1. Assume that documents are being classified into two categories, c1 and c2, such that a document can belong to more than one category. The table below shows the prediction of a classifier, denoted by “y” or “n", in addition to the true label (ground truth) represented by a “+” or “-”, where a correct prediction is either y (+) or n (-).

|  |  |  |
| --- | --- | --- |
|  | **c1** | **c2** |
| **D1** | y(+) | y(+) |
| **D2** | n(-) | y(+) |
| **D3** | n(+) | n(-) |
| **D4** | y(-) | y(+) |
| **D5** | n(+) | n(-) |

Let P(ci) and R(ci) denote the precision and recall associated with category ci, respectively. The precision and recall of c1 and c2 are:

1. **P(c1) = 1/2 R(c1) = 1/3 P(c2) = 1 R(c2) = 1**
2. P(c1) = 1/3 R(c1) = 1/2 P(c2) = 1 R(c2) = 1
3. P(c1) = 1/2 R(c1) = 1/2 P(c2) = 1 R(c2) = 1
4. P(c1) = 1/2 R(c1) = 1/2 P(c2) = 1/2 R(c2) = ½

Correct  
1 / 1 points

2. Given the same data as in Question 1, the classification accuracy of the classifier is:

1. 3/10
2. **7/10**
3. 8/10
4. 9/10

Correct  
1 / 1 points

3. Given the same data as in Question 1, what is the recall of the classifier using **micro-averaging** (i.e., by pooling all decisions together)?

1. 1
2. 4/5
3. **2/3**
4. 5/6

Correct  
1 / 1 points

4. Suppose we are performing document clustering on a collection of N documents using a mixture model as discussed in the lecture **Text Clustering: Generative Probabilistic Models (Part 3)**. Let the number of clusters be K and the vocabulary size be M. What is the number of parameters that the EM algorithm tries to estimate? Consider each P(θi) or P(w|θi) as a separate parameter.

1. MK
2. MNK
3. **K+MK**
4. KN+MK

Correct  
1 / 1 points

5. Which one of the following statements is **not** an opinion?

1. PLSA always performs similarly to LDA.
2. PLSA is the best method for a topic mining task.
3. **PLSA is a mixture model.**

Correct  
1 / 1 points

6. True or false? Word unigrams are the best performing features for sentiment classification.

1. **False**
2. True

Correct  
1 / 1 points

7. True or false? Suppose we are using logistic regression for binary classification (i.e., k=2) where the number of features is M. Then, the number of parameters to be estimated is M+1.

1. False
2. **True**

Correct  
1 / 1 points

8. True or false? Assume we are using word n-grams as features to perform sentiment classification. Then, higher values of n will usually be **less** prone to overfitting (i.e., for higher values of n, the difference between training and testing accuracies will be smaller).

1. True
2. **False**

Correct  
1 / 1 points

9. Why is accuracy sometimes not good for classification evaluation? Check all that apply.

* **Some decisions are more serious than others.**
* **For imbalanced dataset, high accuracy does not imply good performance.**
* Computation of accuracy is difficult.

Correct  
1 / 1 points

10. If you want to put more emphasis on precision than recall, how should you adjust the value of*β*?

1. **Choose a low value of *β***
2. Choose a high value of *β*