Close

Week 4 Practice Quiz

**9/10** points earned (90%)

Excellent!

Retake

[Next](https://www.coursera.org/learn/text-retrieval/exam/iq1A9/week-4-quiz)

Correct

1 / 1 points

1.

You are given a vocabulary composed of only three words: “text,” “mining," and “research.” Below are the probabilities of two of these three words given by a unigram language model:

|  |  |
| --- | --- |
| **Word** | **Probability** |
| text | 0.4 |
| mining | 0.2 |

What is the probability of generating the phrase “text mining research” using this unigram language model?



0



0.08



0.4



0.032

**Correct Response**

The probability of “research” is P(“research”) = 1 – ( P(“text”)+P(“mining”) ) = 1 – (0.4 + 0.2) = 0.4. The probability of generating the given phrase P(“text mining research”) = P(“text”) x P(“mining”) x P(“research”) = 0.4 x 0.2 x 0.4 = 0.032.

Incorrect

0 / 1 points

2.

You are given the query Q= “food safety” and two documents:

D1 = “food quality regulations”

D2 = “food safety measures”

Assume you are using the maximum likelihood estimator **without**smoothing to calculate the probabilities of words in documents (i.e., the estimated *p*(*w*|*D*) is the relative frequency of word *w* in the document *D*). Based on the unigram query likelihood model, which of the following choices is correct?



P(Q|D1) = 0 P(Q|D2) = 1/9



P(Q|D1) = 1/3 P(Q|D2) = 1/9

**Incorrect Response**



P(Q|D1) = 1/3 P(Q|D2) = 0



P(Q|D1) = 1/2 P(Q|D2) = 1/2

Correct

1 / 1 points

3.

Probability smoothing avoids assigning zero probabilities to unseen words in documents.



False



True

**Correct Response**

Correct

1 / 1 points

4.

Assume you are given two scoring functions:

*S*1(*Q*,*D*)=*P*(*Q*|*D*)

*S*2(*Q*,*D*)=log*P*(*Q*|*D*)

For the same query and corpus, *S*1 and *S*2 will give the same ranked list of documents.



True

**Correct Response**



False

Correct

1 / 1 points

5.

Assume you are using linear interpolation (Jelinek-Mercer) smoothing to estimate the probabilities of words in a certain document. What happens to the smoothed probability of the word when the parameter *λ* is **decreased**?



It does not change.



It becomes closer to the probability of the word in the collection language model.



It becomes closer to the maximum likelihood estimate of the probability derived from the document.

**Correct Response**

The smoothed probability can be thought of as a weighted average of the maximum likelihood estimate and the probability of the word in the collection. When *λ*decreases, the weight assigned to maximum likelihood estimate increases, thus having a higher effect on the smoothed probability value.

Correct

1 / 1 points

6.

Refer to the Rocchio feedback formula in the lectures. If you want to reduce the effect of the**relevant**documents in the updated query, which of the following should be done?



Reduce *β*

**Correct Response**

The weight assigned to the centroid of the relevant documents is directly proportional to *β*.



Increase *γ*



Increase *β*



Reduce *γ*

Correct

1 / 1 points

7.

Let *q* be the original query vector, *DR*={*P*1,...,*Pn*} be the set of positive document vectors, and *DN*={*N*1,...,*Nm*} be the set of negative document vectors. Let *q*1 be the expanded query vector after applying Rocchio on *DR* and *DN* with positive parameter values *α*, *β*, and *γ*. Let *q*2 be the expanded query vector after applying Rocchio on *DR* and *DN* with the same values for *α*, *β*, but *γ* being set to zero.

In which updated query do you expect stopwords to have higher weights?



*q*2

**Correct Response**



*q*1

Correct

1 / 1 points

8.

In the query likelihood model, why is smoothing necessary?



Without smoothing, if a term is neither in the query nor the document, then the likelihood becomes −∞.



Without smoothing, if a term in the document is not in the query, then the likelihood becomes −∞.



Without smoothing, if a term in the query is not in the document, then the likelihood becomes −∞.

**Correct Response**

Correct

1 / 1 points

9.

Which of the following is NOT correct about the unigram model?



The probability of generating the words A AND B is the product of the probability of generating A and the probability of generating B.



The probability of generating the word sequence "A" "B" "C" is the same as generating "C" "B" "A."



The probability of generating the word A OR B is the sum of the probability of generating A and the probability of generating B.

**Correct Response**

Correct

1 / 1 points

10.

Assume that *β*=1 is a good choice when performing relevance feedback using Rocchio's method. What is a reasonable value of *β* to use when relying on pseudo feedback?



Less than 1

**Correct Response**

When doing relevance feedback, the judgments are usually reliable since human assessors generate them after reading the queries and documents. However, in pseudo feedback, the top k documents retrieved by the system are “blindly” assumed to be relevant, which makes the judgments less reliable compared to relevance feedback. The reasonable choice is to lower the parameter *β*, which can be thought of as the degree of “confidence” in the documents being used as “positive” examples in feedback.



More than 1



1