You should complete Lab 3 before answering these quiz questions.

PART (2D) TOKENS WITH THE SMALLEST IDF

(1/1 point)

In part (2d), do you think the 11 terms (tokens) are useful for entity resolution?

Top of Form

 Yes No No - correct

Bottom of Form

**EXPLANATION**

Answer the next quiz question for the explanation.

CHECKYOUR ANSWER **HIDE ANSWER**

PART (2D) EXPLANATION

(1/1 point)

In part (2d), why do you think the terms are useful or not useful for entity resolution?

Top of Form

 These terms are useful for entity resolution because they describe distinguishing tokens in product descriptions These terms not useful for entity resolution because they are generic terms for marketing, prices, and product categories. These terms not useful for entity resolution because they are generic terms for marketing, prices, and product categories. - correct

Bottom of Form

**EXPLANATION**

For this question, the answer is the explanation - the terms are too generic to be useful in entity resolution.

CHECKYOUR ANSWER **HIDE ANSWER**

PART (2E) IDF HISTOGRAM

(1/1 point)

Using the plot in (2e), what conclusions can you draw from the distribution of weights?

Top of Form

 The distribution of IDF values is very dense. You cannot draw any conclusions from the histogram. There is a long tail of rare words in the corpus - these have large IDF values. There is a long tail of rare words in the corpus - these have large IDF values. - correct The distribution of IDF values is very flat.

Bottom of Form

**EXPLANATION**

There are gaps between IDF values because IDF is a function of a discrete variable, i.e., a document count.

CHECKYOUR ANSWER **HIDE ANSWER**

PART (3E) PERFORM A GOLD STANDARD EVALUATION

(1/1 point)

In part (3e) you used the "gold standard" data to answer the following questions:

\* How many true duplicate pairs are there in the small data sets?

\* What is the average similarity score for true duplicates?

\* What about for non-duplicates?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Based on the answers to the questions in part (3e), is cosine similarity doing a good job, qualitatively speaking, of identifying duplicates?

Top of Form

 Yes Yes - correct No

Bottom of Form

**EXPLANATION**

Cosine similarity looks useful, because duplicates on average are 250X more similar than non-duplicates. As long as variance isn't too high, that's a good signal.

CHECKYOUR ANSWER **HIDE ANSWER**

PART (5C) LINE PLOTS - PART 1

(1/1 point)

Using the plots in (5c), what is the optimal threshold value to maximize the F-measure?

Top of Form

 0 0.1 0.2 0.2 - correct 0.5 0.85 1.0

Bottom of Form

**EXPLANATION**

F-measure is maximized with the threshold equal to ~0.2, so that is the optimal threshold if we value precision and recall equally.

CHECKYOUR ANSWER **HIDE ANSWER**

PART (5C) LINE PLOTS - PART 2

(1/1 point)

If false-positives are considered much worse than false-negatives, how does that change your answer?

Top of Form

 0 0.1 0.2 0.5 0.5 - correct 0.85  1.0

Bottom of Form

**EXPLANATION**

If we wanted to really avoid false positives, that means we want higher precision at the cost of lower recall, in which case ~0.5 offers the best trade-off. If we didn't care at all about recall, ~0.85 has peak precision, and would be the best choice.

CHECKYOUR ANSWER **HIDE ANSWER**

SURVEY: LAB COMPLETION TIMES

(3/3 points)

How long did Lab ONE take you to complete (in hours - decimals are OK)?

 correct

1

Loading

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

How long did Lab TWO take you to complete (in hours - decimals are OK)?

 correct

3

Loading

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

How long did Lab THREE take you to complete (in hours - decimals are OK)?

 correct

30

Loading

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Please click "Check" to save your answers.