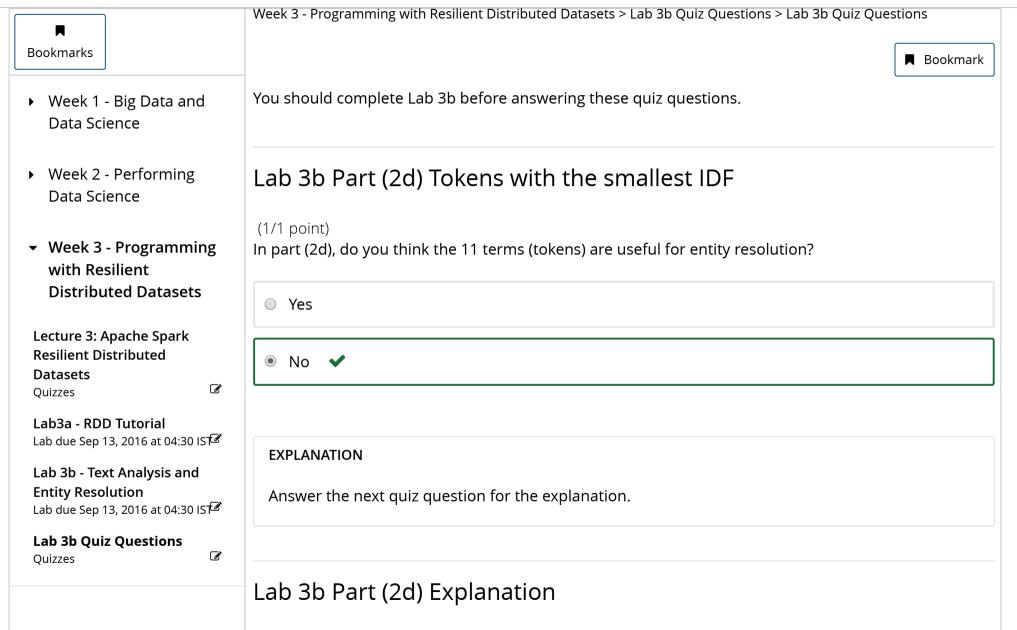


BerkeleyX: CS110x Big Data Analysis with Apache Spark



(1/1 point)

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

- 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.

EXPLANATION

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

Lab 3b Part (2e) IDF Histogram

(1/1 point)

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

- 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.	V
--	---	----------

The distribution of IDF values is very flat.

EXPLANATION

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

Lab 3b 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?

● Yes ✔			
○ No			
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.			
Lab 3b 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?			
○ 0			
O 0.1			
O.5			

0.85	
1.0	

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.

Lab 3b 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?











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.



© edX Inc. All rights reserved except where noted. EdX, Open edX and the edX and Open EdX logos are registered trademarks or trademarks of edX Inc.





8/31/2016











