BerkeleyX: CS100.1x Introduction to Big Data with Apache Spark

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?

- Yes
- No ✔

CHECK SHOW ANSWER

PART (2D) EXPLANATION (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. ✔

CHECK SHOW ANSWER

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. ✔

- The distribution of IDF values is very flat.

CHECK SHOW 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?

- Yes ✔
- No

CHECK SHOW 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?

- 0
- 0.1
- 0.2 ✔
- 0.5

- 0.85
- 1.0

CHECK SHOW 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?

- 0
- 0.1
- 0.2
- 0.5
- 0.85 ✔
- 1.0

CHECK SHOW ANSWER

SURVEY: LAB COMPLETION TIMES (3/3 points)

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

1 ✔

1

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

3 ✔

3

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

6 ✔

6

Please click "Check" to save your answers.

CHECK