Module 2 Quiz:
Data Collection

1. How do we measure annotation quality?

A. By calculating the accuracy of an annotator compared to ground-truth data
B. By calculating the overall agreement and reliability of a data set
C. By measuring inter-annotator agreement on a per-task basis to generate a confidence score for each training data label
D. By calculating similarity among data samples with annotations

Answer: ABC
Explanation: See lecture 2.4 slides

2. Why do we care about data quality?

A. High-quality data can be a major business asset, a unique source of competitive advantage.
B. Poor data quality can lower customer satisfaction.
C. Poor data quality can lower employee job satisfaction leading to excessive turnover and the resulting loss of key process knowledge.
D. Poor data quality can breed organizational mistrust and make it hard to mount efforts that lead to needed improvements.

Answer: ABCD
Explanation: See lecture 2.4 slides

3. How do we obtain high-quality data?

A. Prevention: Keep Bad Data Out of the Database/List
B. Detection: Proactively Look for Bad Data Already Entered
C. Repair: Let the Bad Data Find You and Then Fix Things
D. Allocating Resources: How Much for Prevention, Detection, and Repair

Answer: ABCD
Explanation: See lecture 2.4 slides

4. Which of the following data collection approaches likely to require the highest amount of time and effort:

A. Scraping
B. Download
C. API

Answer: A

4. Which of the following data collection approaches likely to require the highest amount of time and effort:

A. Scraping
B. Download
C. API

Answer: A

5. What data structure(s) does SQLite use for indices implementation?

A. Hashmap
B. Heap
C. B-Tree
D. Linked list

Answer: C

6. In SQLite, assuming that an index has already been created for a table column, what is the time complexity of searching for an item in that column?

A. O(n)
B. O(1)
C. O(n*log n)
D. O(log n)

Answer: D

7. Data stored in CSV (comma separated values) formats may not be easy to parse.

A. True
B. False

Answer: A

8. SQLite can easily store datasets larger than 200PTB

A. True
B. False

Answer: B