Duplicate

Problem Statement

Quora uses a combination of machine learning algorithms and moderation to ensure high-quality content on the site. High question and answer quality has helped Quora distinguish itself from other Q&A sites on the web.

As the number of questions on Quora grows, there is an increasing likelihood that a new question may be a duplicate of an existing question. The text of the questions can vary significantly, but semantically might mean the same thing. We rely on human judgment to determine if 2 questions are considered to be duplicates and merge these questions together on Quora.

For this task, given Quora question text and topic data, determine if any 2 given pairs of questions are considered duplicates or not.

The following fields of raw data are given in JSON representing each question:

- question_key (string): Unique identifier for the question.
- question_text (string): Text of the question.
- context_topic (object): The primary topic of a question, if present. Null otherwise. The topic object will contain a name (string) and followers (integer) count.
- topics (array of objects): All topics on a question, including the primary topic. Each topic object will contain a name (string) and followers (integer) count.
- view_count (int): Number of views on the question.

After the raw data for each question, the list of known duplicate questions and non-duplicate questions among all the given questions are shared as pairs of questions with an integer representing whether that pair is a duplicate or not (1 for duplicate and 0 for non-duplicate). All other pairings of any 2 questions that are not in this list are unknown to be duplicates or not and the test set will come from this. Note that all questions referenced by the training and test sets will be present in the list of raw question data.

Input Format

- The first line will contain an integer $1 \leq Q \leq 60,000$, which represents the number of lines of questions in the training data to follow.
- The next Q lines will contain JSON encoded fields of raw question data.
- The next line will contain an integer $1 \leq D \leq 25,000$, which represent the number of lines of duplicate question pairs in the training data to follow.
- The next D lines will each contain 2 known duplicate questions identified by their unique question_key and an integer representing whether the pair are duplicates (1 for duplicate, 0 for non-duplicate) separated by spaces. (e.g. “abc def 1”)
- The next line will contain an integer $1 \leq N \leq 3,000$ for the number of lines of test question pair data to follow.
- The next N lines will each contain 2 test questions identified by their unique question_key separated by a space.

Output Format

- N lines of 2 question keys and an integer representing whether the 2 test questions are duplicates (1 for duplicate and 0 for non-duplicate). (e.g. “ghi jkl 0”)

Sample Input

```
3
{"view_count": 773, "question_text": "Which is the most intelligent alien or alien species in fiction?", "context_topic": {"followers": 130960, "name": "Science Fiction (genre)"}, "topics": [{"followers": 48, "name": "Science Fiction Books"}, {"followers": 130960, "name": "Science Fiction (genre)"}, {"followers": 1182, "name": "Extraterrestrial Intelligence"}, {"followers": 50056, "name": "Extraterrestrial Life"}, {"followers": 3883, "name": "Science Fiction Movies"}], "follow_count": 9, "question_key": "AAEAAJ/qtRMKkzXyA0tvjyz5tPRWgYizvOkCr9Z9CdJ4cood", "age": 413}
{"view_count": 3522, "question_text": "What is the best way to keep bookmarks?", "context_topic": {"followers": 513, "name": "Bookmarking"}, "topics": [{"followers": 1136, "name": "Pocket (app)"}, {"followers": 9, "name": "ReadItLater"}, {"followers": 1625, "name": "Pinboard"}, {"followers": 1275, "name": "Social Bookmarking"}, {"followers": 513, "name": "Bookmarking"}, {"followers": 5604, "name": "Delicious (web application)"}, {"followers": 4359, "name": "Instapaper"}, {"followers": 85, "name": "Web Bookmarks"}], "follow_count": 62, "question_key": "AAEAADJKxcVF6l23JZvf1Fz+QrKr35CTlMKayNnZebc8dQAY", "age": 1193}
{"view_count": 390, "question_text": "What is best for online bookmarks?", "context_topic": null, "topics": [{"followers": 1275, "name": "Social Bookmarking"}, {"followers": 285, "name": "Social Bookmarking Websites"}, {"followers": 513, "name": "Bookmarking"}], "follow_count": 4, "question_key": "AAEAAO3FKYrsnYH9uKAOnnXfYrGGTVFA3uzHz+Vltm5Ssii3", "age": 1211}
2
AAEAADJKxcVF6l23JZvf1Fz+QrKr35CTlMKayNnZebc8dQAY AAEAAO3FKYrsnYH9uKAOnnXfYrGGTVFA3uzHz+Vltm5Ssii3 1
AAEAADJKxcVF6l23JZvf1Fz+QrKr35CTlMKayNnZebc8dQAY AAEAAJ/qtRMKkzXyA0tvjyz5tPRWgYizvOkCr9Z9CdJ4cood 0
1
AAEAAJ/qtRMKkzXyA0tvjyz5tPRWgYizvOkCr9Z9CdJ4cood AAEAAO3FKYrsnYH9uKAOnnXfYrGGTVFA3uzHz+Vltm5Ssii3
```

Note: a more comprehensive sample input is available here (input, output). You can use this script to score your output for the sample dataset.

Sample Output

```
AAEAAJ/qtRMKkzXyA0tvjyz5tPRWgYizvOkCr9Z9CdJ4cood AAEAAO3FKYrsnYH9uKAOnnXfYrGGTVFA3uzHz+Vltm5Ssii3 0
```

Scoring

You will get 1 point for each pair of questions that you predicted correctly.

Your raw score is the sum of all points for all pairs of questions in the test case (N).

Your final score is $200 \cdot \frac{yourRawScore}{N}$.

Resource Limits

Your program is limited to 1024 MB of memory and must run in 60 seconds or less.

Notes

- This data set is not representative of all the data in Quora but intentionally sampled to make a good challenge to solve.