ALL CHALLENGES

Upvotes
Related Questions
Ontology
Wombats
Labeler
Duplicate
Answered
Interest
Views
Typeahead Search
Feed Optimizer
Answer Classifier
Python URI
Trend Analyzer

POWERED BY
HackerRank

Upvotes

At Quora, we have aggregate graphs that track the number of upvotes we get each day.

As we looked at patterns across windows of certain sizes, we thought about ways to track trends such as non-decreasing and non-increasing subranges as efficiently as possible.

For this problem, you are given N days of upvote count data, and a fixed window size K. For each window of K days, from left to right, find the number of non-decreasing subranges within the window minus the number of non-increasing subranges within the window.

A window of days is defined as contiguous range of days. Thus, there are exactly $N - K + 1$ windows where this metric needs to be computed. A non-decreasing subrange is defined as a contiguous range of indices $[a, b]$, [math]a, where each element is at least as large as the previous element. A non-increasing subrange is similarly defined, except each element is at least as large as the next. There are up to $K(K-1)/2$ of these respective subranges within a window, so the metric is bounded by $[-K(K-1)/2, K(K-1)/2]$.

Constraints
$1 \leq N \leq 100,000$ days
$1 \leq K \leq N$ days

Input Format
Line 1: Two integers, N and K

Line 2: N positive integers of upvote counts, each integer less than or equal to 10^9.

Output Format
Line 1..: $N - K + 1$ integers, one integer for each window's result on each line

Sample Input

```
1 5 3
2 1 2 3 1 1
```

Sample Output

```
1 3
2 0
3 -2
```

Explanation
For the first window of [1, 2, 3], there are 3 non-decreasing subranges and 0 non-increasing, so the answer is 3. For the second window of [2, 3, 1], there is 1 non-decreasing subrange and 1 non-increasing, so the answer is 0. For the third window of [3, 1, 1], there is 1 non-decreasing subrange and 3 non-increasing, so the answer is -2.

Live Submission (FAQ):

Source File: Choose File No file chosen

Enter your email address here

Submit and Test

Related Questions

For the purposes of this problem, suppose Quora has N questions, and question $i\ (1 \leq i \leq N)$ takes T_i time to read. There exists exactly one path from any question to another, and related questions form undirected pairs between themselves. In other words, the graph of related questions is a tree.

Each time Steve reads a question, he will see a list of related questions and navigate to one that he hasn't read yet at random. Steve will stop reading once there are no unread related questions.

Which question should we show first to Steve so that we minimize his total expected reading time? It is guaranteed that there is one unique question that is optimal.

Input Format
Line 1: A single integer, N
Line 2: N integers, T_i
Line 3...$N+1$: Each line contains two integers A, B indicating that question A and B are related

Output Format
Line 1: A single integer, X, the best question to show first.

Constraints
$1 \leq N \leq 10^5$
$1 \leq T_i \leq 10^6$

Sample Input

```
1 5
2 2 2 1 2 2
3 1 2
4 2 3
5 3 4
6 4 5
```

Sample Output

```
3
```

Live Submission (FAQ):

Source File: Choose File No file chosen

Enter your email address here

Submit and Test

Ontology

Quora has many questions on different topics, and a common product use-case for our @mention selectors and search service is to look-up questions under a certain topic as quickly as possible.

For this problem, imagine a simplified version of Quora where each question has only one topic associated with it. In turn, the topics form a simplified ontology where each topic has a list of children, and all topics are descendants of a single root topic.

Design a system that allows for fast searches of questions under topics. There are N topics, M questions, and K queries, given in this order. Each query has a desired topic as well as a desired string prefix. For each query, return the number of questions that fall under the queried topic and begin with the desired string. When considering topics, we want to include all descendants of the queried topic as well as the queried topic itself. In other words, each query searches over the subtree of the topic.

The topic ontology is given in the form of a flattened tree of topic names, where each topic may optionally have children. If a topic has children, they are listed after it within parentheses, and those topics may have children of their own, etc. See the sample for the exact input format. The tree is guaranteed to have a single root topic.

For ease of parsing, each topic name will be composed of English alphabetical characters, and each question and query text will be composed of English alphabetical characters, spaces, and question marks. Each question and query text will be well behaved: there will be no consecutive spaces or leading/trailing spaces. All queries, however, are case sensitive.

Constraints
For 100% of the test data, $1 \leq N, M, K \leq 10^5$ and the input file is smaller than 5MB
For 50% of the test data, $1 \leq N, M, K \leq 2 \cdot 10^4$ and the input file is smaller than 1MB

Input Format
Line 1: One integer N
Line 2: N topics arranged in a flat tree (see sample)
Line 3: One integer M
Line 4...M+3: Each line contains a topic name, followed by a colon and a space, and then the question text.
Line M+4: One integer K
Line M+5...M+K+4: Each line contains a topic name, followed by a space, and then the query text.

Output Format
Line 1...K: Line i should contain the answer for the ith query.

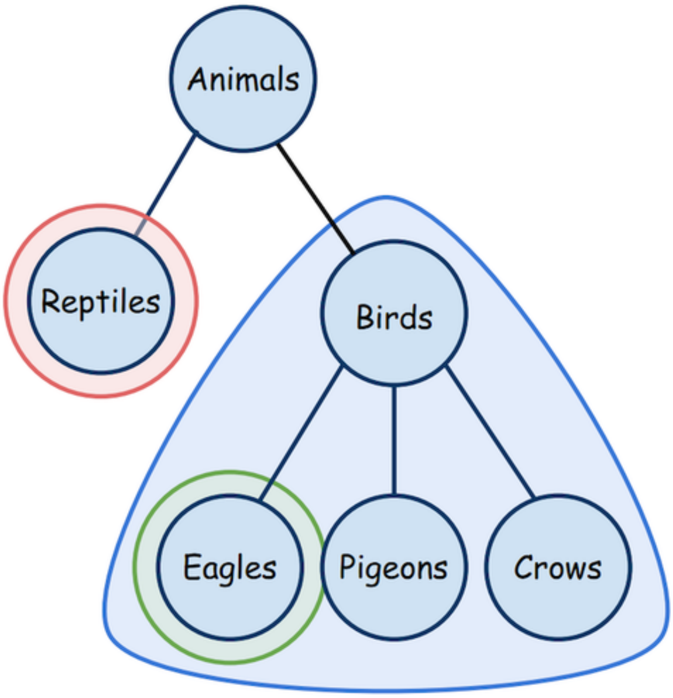
Sample Input

```
6
Animals ( Reptiles Birds ( Eagles Pigeons Crows ) )
5
Reptiles: Why are many reptiles green?
Birds: How do birds fly?
Eagles: How endangered are eagles?
Pigeons: Where in the world are pigeons most densely populated?
Eagles: Where do most eagles live?
4
Eagles How en
Birds Where
Reptiles Why do
Animals Wh
```

Sample Output

```
1
2
0
3
```

Explanation
The first query corresponds to the green area in the diagram, since it is looking for topics under Eagles, and the query string matches just one question: "How endangered are eagles?" The second query corresponds to the blue area in the diagram, which is the subtree of Birds, and matches two questions that begin with "Where". The third corresponds to the red area, which does not have any questions that begin with "Why do". The final query corresponds to the entire tree, since Animals is the root topic, and matches three questions.



Live Submission (FAQ):

Source File: Choose File No file chosen

Enter your email address here

Submit and Test

Wombats

One day at Quora, Jerry decides to start a stuffed animal collection on his desk. Fortunately, he comes across a massive tetrahedral pyramid of wombat stuffed animals, where the tetrahedron has side length N. Some of these stuffed animals, however, are cuter than others, so Jerry only wants to pick the best possible set of stuffed animals.

Given that each stuffed animal is a perfect sphere and has an integer cuteness value, help Jerry pick a (possibly empty) subset of the stuffed animals to pick up such that their sum of values is maximized. Note that the stuffed animals can have negative value, so picking up all the animals is not necessarily optimal. Jerry does not want to disorganize the arrangement, so he cannot take any stuffed animal without taking all of the (up to 3) stuffed animals above it.

See the sample input and diagram to see how the stuffed animals are organized as a tetrahedral pyramid and given as input.

Constraints
For 100% of the test data, $1 \leq N \leq 12$
For 40% of the test data $1 \leq N \leq 6$

Input Format
Line 1: One integer N
Line 2...N(N+1)/2+1: The formatted pyramid of cuteness values

Output Format
Line 1: One integer, the maximum sum of values achievable

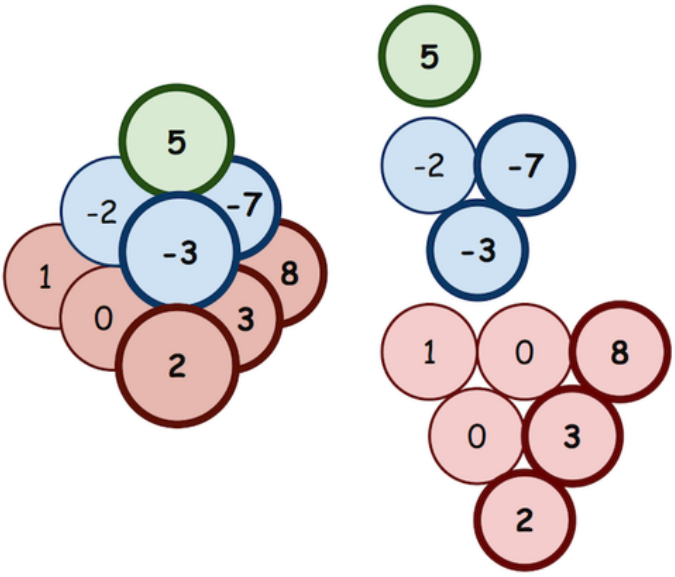
Sample Input

```
1 3
2 5
3 -2 -7
4 -3
5 1 0 8
6 0 3
7 2
```

Sample Output

```
8
```

Explanation
The optimal selection is shown in the diagram in bold. It is suboptimal to take 1 because that would require taking -2 as well, which would decrease the total. On the other hand, 8, 3, and 2 should all be taken because those outweigh -3 and -7.



Live Submission (FAQ):

Source File: Choose File No file chosen

Enter your email address here

Submit and Test

Labeler

When a user adds a question to Quora, we automatically suggest topics which might be relevant to the question. For example, a question about Dijkstra's algorithm would probably fit well under the topics "Algorithms" and "Graph Theory".

We have a pretty good machine learning system in place for generating these topics, but it could always be better. Your goal in this task is to design a question topic labeler.

The input data consists of a training dataset of T questions and an evaluation dataset of E questions. Your labeler should suggest 10 topics for each question in the evaluation set, ordered by relevance.

Input Format
The first line consists of two space-separated integers T and $E (1 \leq T \leq 20000, 1 \leq E \leq 1000)$.

The next $2T$ lines provide the training dataset. Each question in the training dataset is described by two lines:

One integer N $(1 \leq N \leq 25)$, followed by N positive integers (below 250) representing the topic IDs of the question in no particular order. On average, there are approximately 3 topics per question in the dataset.
One string (between 1 and 500 printable ASCII characters), the question text. The average question text length is 70 characters.

The next E lines provide the evaluation dataset. Each question in the evaluation dataset is described by one line (between 1 and 500 ASCII characters), the question text.

Output Format
There are T lines in total. The i-th line contains 10 space-separated integers, each representing the topic id of a suggestion for the i-th question in the evaluation dataset. To maximize your score, topic suggestions for a question should be ordered in descending order of relevance.

Sample Input

```
6 4
3 1 2 4
What is the meaning of life?
7 1 2 5 8 9 11 15
What is Quora?
2 14 178
What are the best Google calendar hacks?
3 117 93 125
Why does government of China not value the freedom of speech?
2 65 164
What is the best piece of design ever?
5 197 183 29 170 143
What was the last conversation you had with your father?
Are programming contests fun?
What is machine learning?
How do you code in C++?
Is it possible to sort in linear time?
```

Note: a more comprehensive sample input (with the same training data made available during evaluation) is available here (input , correct topics). You can use this script to score your output for the sample dataset.

Sample Output

```
1 2 5 3 7 9 11 10 14 15
2 9 29 3 117 197 1 183 178 15
15 197 170 143 8 5 1 7 2 14
1 8 14 164 125 15 2 65 164 3
```

Scoring
Your score for each question is determined as follows:

$$\frac{\sum_{i=0}^{9} \sqrt{10-i} \cdot (guess_i \in questionTopics)}{\sum_{i=0}^{min(|questionTopics|,10)-1} \sqrt{10-i}}$$

Your raw score is the sum of each question score.

$minScore$ = raw score for classifier that guesses 10 most frequent topics.

Your final score is $200 \cdot \frac{yourRawScore - minScore}{E - minScore}$.

Resource Limits
Your program is limited to 512 MB of memory and must run in 60 seconds or less.

Live Submission (FAQ):

Source File: [Choose File] No file chosen

[Enter your email address here]

[Submit and Test]

Duplicate

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 intellige
{"view_count": 3522, "question_text": "What is the best way to ke
{"view_count": 390, "question_text": "What is best for online boo
2
AAEAADJKxcVF6l23JZvf1Fz+QrKr35CTlMKayNnZebc8dQAY AAEAAO3FKYrsnYH9
AAEAADJKxcVF6l23JZvf1Fz+QrKr35CTlMKayNnZebc8dQAY AAEAAJ/qtRMKkzXy
1
AAEAAJ/qtRMKkzXyA0tvjyz5tPRWgYizvOkCr9Z9CdJ4cood AAEAAO3FKYrsnYH9
```

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.

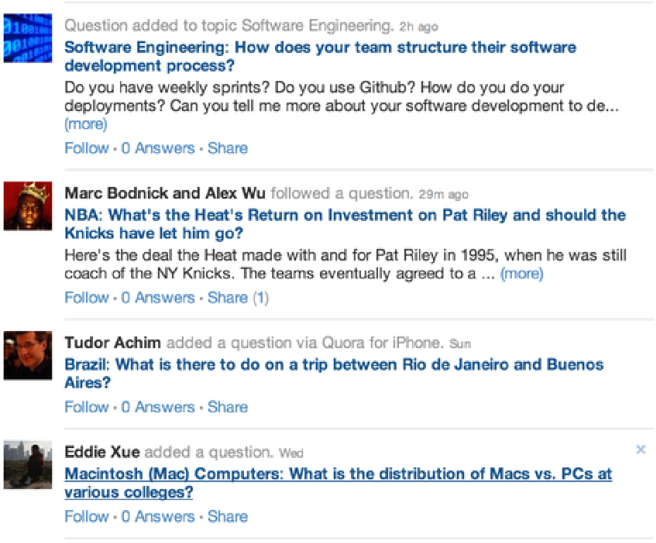
Live Submission (FAQ):

Source File: Choose File No file chosen

Enter your email address here

Submit and Test

Answered



Question added to topic Software Engineering. 2h ago

Software Engineering: How does your team structure their software development process?

Do you have weekly sprints? Do you use Github? How do you do your deployments? Can you tell me more about your software development to de... (more)

Marc Bodnick and Alex Wu followed a question. 29m ago

NBA: What's the Heat's Return on Investment on Pat Riley and should the Knicks have let him go?

Here's the deal the Heat made with and for Pat Riley in 1995, when he was still coach of the NY Knicks. The teams eventually agreed to a ... (more)

Tudor Achim added a question via Quora for iPhone. Sun

Brazil: What is there to do on a trip between Rio de Janeiro and Buenos Aires?

Eddie Xue added a question. Wed

Macintosh (Mac) Computers: What is the distribution of Macs vs. PCs at various colleges?

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 we get many questions every day, a challenge we have is to figure out good, interesting and meaningful questions from the bad. What questions are valid ones that can be answered? What questions attract reputable answers that then get upvoted? Can you tell which questions will likely get answers quickly, so that we can surface them in real-time to our users?

For this task, given Quora question text and topic data, predict whether a question gets an upvoted answer within 1 day.

Input Format
The first line contains N. N questions follow, each being a valid json object.
The following fields of raw data are given in json.
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.
anonymous (boolean): Whether the question was anonymous.
__ans__ (boolean): Whether the question got an up-voted answer within 1 day.
This is immediately followed by an integer T.

T questions follow, each being a valid json object.
The json contains all but one field *__ans__*.

Output Format

T rows of JSON encoded fields, with the *question_key* key containing the unique identifier given in the test data, and the predicted value keyed by *__ans__*.

Constraints
question_key is of ascii format.
question_text, name in topics and context_topic is of UTF-8 format.
0 <= followers <= 106
9000 <= N <= 45000
1000 <= T <= 5000

Training Data

Sample testcases can be downloaded here and used for offline training if desired.

Scoring

The answers are evaluated by accuracy.

```
Number correct classified / Total input size * 100%
```

The training and test set each will have approximately an equal number of each boolean type.

Your score will be based only on the hidden input. The sample input is only for your convenience.

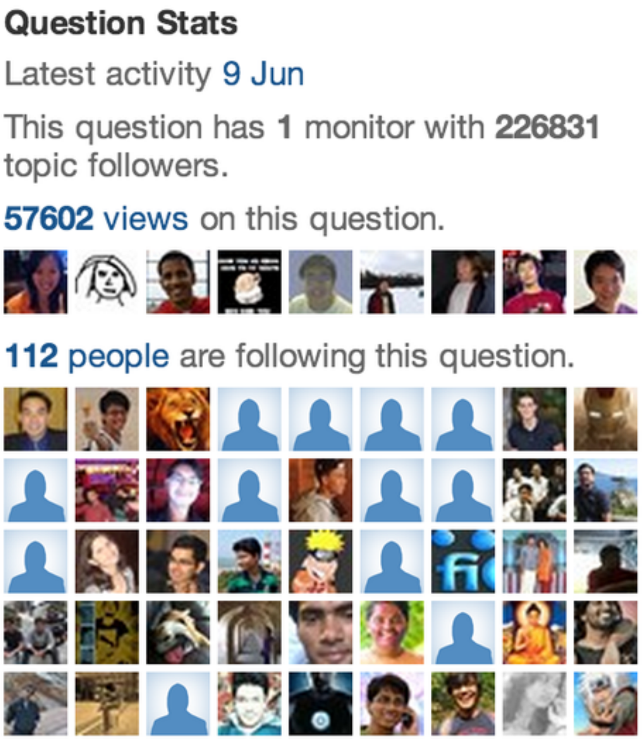
Live Submission (FAQ):

Source File: Choose File No file chosen

Enter your email address here

Submit and Test

Interest



Question Stats

Latest activity 9 Jun

This question has 1 monitor with 226831 topic followers.

57602 views on this question.

112 people are following this question.

Quora uses machine learning algorithms to try to generate interesting news feeds and digest emails for people.

Before a question gets an answer, we’d like to be able to make use of all the available information we have to be able to predict how interesting or relevant the question is to people. Ideally, we’d like to be able to tell this in real-time as soon as a few people have viewed it, by measuring the people following the question as a proxy of interest. Can you tell what questions will get the most followers?

For this task, given Quora question text and topic data for questions with 0 visible answers, predict the ratio of viewers to followers.

Input Format

The first line contains N. N questions follow, each being a valid json object.
The following fields of raw data are given in json.
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.
anonymous (boolean): Whether the question was anonymous.
__ans__ (float): The ratio of viewers to followers of the question.

This is immediately followed by an integer T.
T questions follow, each being a valid json object.
The json contains all but one field *__ans__*.

Output Format

T rows of JSON encoded fields, with the *question_key* key containing the unique identifier given in the test data, and the predicted value keyed by *__ans__*.

Constraints

question_key is of ascii format.
question_text, name in topics and context_topic is of UTF-8 format.
0 <= followers <= 106
9000 <= N <= 45000
1000 <= T <= 5000

Training Data

Sample testcases can be downloaded here and used for offline training if desired.

Scoring

Your solution is evaluated by a Root Mean Squared Logarithmic Error (RMSLE) metric. We then calculate a final score by how close it reaches a target score of 0.5, and scale that by 100%.

$$\frac{0.5}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log(x_i) - \log(y_i))^2}} \times 100\%$$

Your score will be based only on the hidden input. The sample input is only for your convenience.

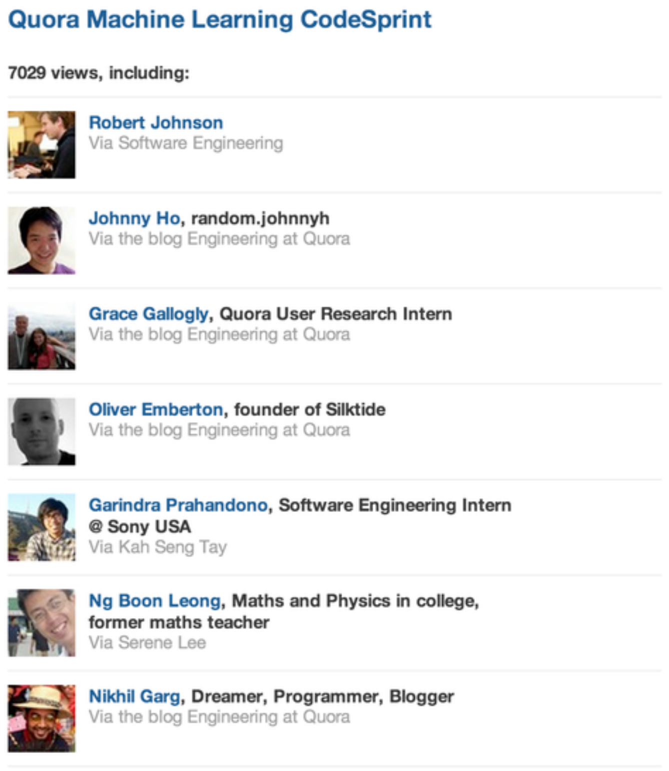
Live Submission (FAQ):

Source File: Choose File No file chosen

Enter your email address here

Submit and Test

Views



Quora Machine Learning CodeSprint

7029 views, including:

Robert Johnson
Via Software Engineering

Johnny Ho, random.johnnyh
Via the blog Engineering at Quora

Grace Gallogly, Quora User Research Intern
Via the blog Engineering at Quora

Oliver Emberton, founder of Silktide
Via the blog Engineering at Quora

Garindra Prahandono, Software Engineering Intern @ Sony USA
Via Kah Seng Tay

Ng Boon Leong, Maths and Physics in college, former maths teacher
Via Serene Lee

Nikhil Garg, Dreamer, Programmer, Blogger
Via the blog Engineering at Quora

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.

Not all of these questions are as interesting or appealing to people on the web. Can you tell what questions can organically attract the most viewers? What about questions that eventually become viral? Which questions are timeless and can sustain traffic?

For this task, given Quora question text, topic data, number of answers and number of people promoted to, predict the number of views per day in age of the question.

Input Format

The first line contains N. N questions follow, each being a valid json object.
The following fields of raw data are given in json.
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.
anonymous (boolean): Whether the question was anonymous.
num_answers (integer): The number of visible non-collapsed answers the question has.
promoted_to (integer): The number of people the question was promoted to.
__ans__ (float): The ratio of viewers to age of the question in days.

This is immediately followed by an integer T.
T questions follow, each being a valid json object.
The json contains all but one field *__ans__*.

Output Format

T rows of JSON encoded fields, with the *question_key* key containing the unique identifier given in the test data, and the predicted value keyed by *__ans__*.

Constraints

question_key is of ascii format.
question_text, name in topics and context_topic is of UTF-8 format.

0 <= followers <= 106
9000 <= N <= 45000
1000 <= T <= 5000

Training Data
Sample testcases can be downloaded here and used for offline training if desired.

Scoring

Your solution is evaluated by a Root Mean Squared Logarithmic Error (RMSLE) metric. We then calculate a final score by how close it reaches a target score of 0.5, and scale that by 100%.

$$\frac{0.5}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log(x_i) - \log(y_i))^2}} \times 100\%$$

Your score will be based only on the hidden input. The sample input is only for your convenience.

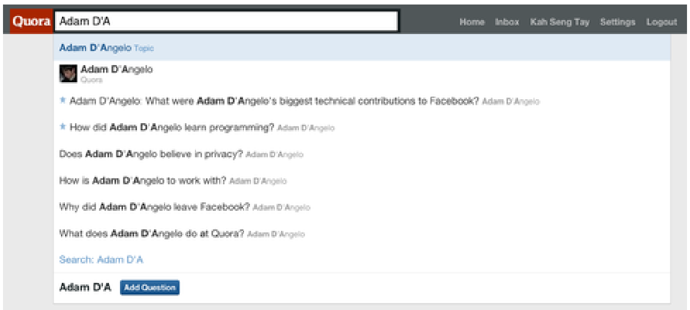
Live Submission (FAQ):

Source File: Choose File No file chosen

Enter your email address here

Submit and Test

Typeahead Search



Adam D'Angelo Topic

Adam D'Angelo
Quora

Adam D'Angelo: What were **Adam D'Angelo**'s biggest technical contributions to Facebook? Adam D'Angelo

How did **Adam D'Angelo** learn programming? Adam D'Angelo

Does **Adam D'Angelo** believe in privacy? Adam D'Angelo

How is **Adam D'Angelo** to work with? Adam D'Angelo

Why did **Adam D'Angelo** leave Facebook? Adam D'Angelo

What does **Adam D'Angelo** do at Quora? Adam D'Angelo

Search: Adam D'A

Adam D'A Add Question

The search bar at the top of every page on Quora allows you to search the most up-to-date people, topics and questions on our site.

We want to quickly return the most relevant results that match the search query entered into the input text field. Every time a new user, question or topic is added (or old ones deleted), subsequent queries must reflect those changes immediately. We currently use a fast in-memory service to handle this.

Input comes into the service as the following commands:
ADD <type> <id> <score> <data string that can contain spaces>
Adds the following <type> of item (user | topic | question | board) with the unique <id> string and <score> float, corresponding to the <data string> that would be used to match queries.
DEL <id>
Deletes the item specified by unique identifier <id>.
QUERY <number of results> <query string that can contain spaces>
Queries for the specified integer number of results (up to 20) that match a given query string. For a data item to be considered a matching result to a query, each token in the query must be found in the data string as a case-insensitive prefix to any token in the data string. The results are ranked in descending order of score, and we only take the top few results as specified. When there is a tie in the score, newer items (more recently added) should be

ranked higher. If there are no results, just output the empty string on that line.

WQUERY <number of results> <number of boosts> (<type>:<boost>)* (<id>:<boost>)* <query string that can contain spaces>

Same as QUERY, except that instead of using the raw input score specified by ADD for the particular item, the scores are weighted by the optional number of boosting factors. The boosts are floats that should be multiplied to the raw score, and affect either whole types, or specific items with the given <id>s. If there are multiple boosts applicable, they each are multiplied commutatively to the raw score.

Your task will be to write an equivalent service as a standalone program, with input files that correspond to the queries and updates to the data, and expected output files that correspond to the results obtained for each query.

Input format (read from STDIN):

Your program will be given an integer N on the first line of stdin, followed by N lines of the form:

<command> <command data>

The input commands are: ADD | DEL | QUERY | WQUERY
Types are: user | topic | question | board
Ids are alphanumeric strings without spaces or punctuation and will not include the same strings used for types.
Data strings can be any ASCII character, but are delimited by spaces or tabs. We will not be using anything special unprintable characters or \r and \n in the data string.

Command data for each command is as specified above. For example:

```
15
ADD user u1 1.0 Adam D’Angelo
ADD user u2 1.0 Adam Black
ADD topic t1 0.8 Adam D’Angelo
ADD question q1 0.5 What does Adam D’Angelo do at Quora?
ADD question q2 0.5 How did Adam D’Angelo learn programming?
QUERY 10 Adam
QUERY 10 Adam D’A
QUERY 10 Adam Cheever
QUERY 10 LEARN how
QUERY 1 lear H
QUERY 0 lea
WQUERY 10 0 Adam D’A
WQUERY 2 1 topic:9.99 Adam D’A
DEL u2
QUERY 2 Adam
```

Output format (write to STDOUT):

For each QUERY and WQUERY command, you should output the following line:

<sorted result ids>

Where each line contains the <id>s in descending score order, up to the required number of results, as specified above according to the QUERY and WQUERY commands. If there are no matches, output the empty line. For example:

```
u2 u1 t1 q2 q1
u1 t1 q2 q1

q2
q2

u1 t1 q2 q1
t1 u1
u1 t1
```

Constraints:

0 < N < 100000
0 < number of ADDs < 40000
0 < number of DELs < 10000

0 < number of QUERYs < 20000
0 < number of WQUERYs < 1000
0 < number of boosts < 25
0.0 < each score < 100.0
0 < data string length < 100 chars

You should aim to have your algorithm be fast enough to solve our largest test inputs in under 5 seconds, or be as close to that as possible.

Notes
In this problem, we are presenting a representative simplified version of our service for the purposes of the contest.

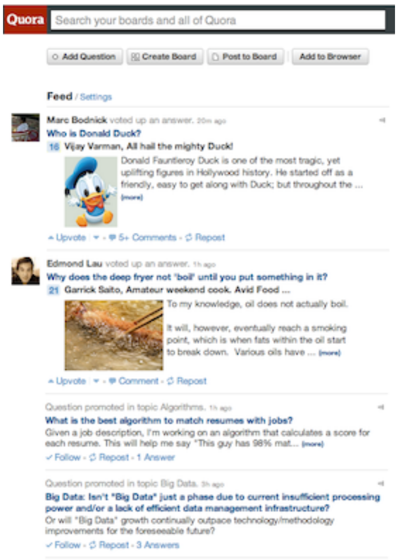
Live Submission ([FAQ](#)):

Source File: Choose File No file chosen

Enter your email address here

Submit and Test

Feed Optimizer



Marc Bodnick voted up an answer. 20m ago
Who is Donald Duck?
16 Vijay Varman, All hail the mighty Duck!
Donald Fauntleroy Duck is one of the most tragic, yet uplifting figures in Hollywood history. He started off as a friendly, easy to get along with Duck; but throughout the ... (more)
Upvote · 5+ Comments · Repost

Edmond Lau voted up an answer. 1h ago
Why does the deep fryer not "boil" until you put something in it?
21 Garrick Saito, Amateur weekend cook. Avid Food ...
To my knowledge, oil does not actually boil.

It will, however, eventually reach a smoking point, which is when fats within the oil start to break down. Various oils have ... (more)
Upvote · Comment · Repost

Question promoted in topic Algorithms. 1h ago
What is the best algorithm to match resumes with jobs?
Given a job description, I'm working on an algorithm that calculates a score for each resume. This will help me say "This guy has 98% mat... (more)
✓ Follow · Repost · 1 Answer

Question promoted in topic Big Data. 3h ago
Big Data: Isn't "Big Data" just a phase due to current insufficient processing power and/or a lack of efficient data management infrastructure?
Or will "Big Data" growth continually outpace technology/methodology improvements for the foreseeable future?
✓ Follow · Repost · 3 Answers

Quora shows a customized feed of recent stories on a user's home page. Stories in Quora refer to activities that happen on the site, for example, when a user posts a note, adds a question, or upvotes an answer. We score each story based on its type and other characteristics, and this score represents the value we think a story brings to the user. We want to be able to generate quickly the feed of best and most recent stories for the user every time they reload their home page.

Your task will be to design the algorithm that picks the stories to display in the feed.

You are given a list of stories, each having a time of publication, a score and a certain height in pixels that it takes to display the story. Given the total number of pixels in the browser available for displaying the feed, you want to maximize the sum of scores for the stories that you can display in the feed at each time the user reloads their home page. We only want to consider recent stories, so only stories that were published in a recent time window from the time of reload should be considered. You do not have to use up all the pixels in the browser.

Input format (read from STDIN):

The first line of input will be 3 positive integers: N the number of events, W the time window representing the window of recent stories, and H the height of the browser in pixels.

There will be N lines following that, each beginning with ‘S’ if it is a story event, or ‘R’ if it is a reload event. A story event will have 3 positive integers: the time of publication, the score and the height in pixels of that story. A reload event will have 1 positive integer: the time of reload.

The events will always be in chronological order, and no two events will happen at the same time.

For example:

```
9 10 100
S 11 50 30
R 12
S 13 40 20
S 14 45 40
R 15
R 16
S 18 45 20
R 21
R 22
```

Output format (write to STDOUT):
For each reload event given in the input, you are to output a line of integers. First, the maximum score of stories you can show in the feed. This should be followed by the number of stories picked and the id number for each story picked, in increasing order. Stories are given an id starting from 1 in the order of their time of publication.
If two sets of stories give the same score, choose the set with fewer stories. If there is still a tie, choose the set which has the lexicographically smaller set of ids, e.g. choose [1, 2, 5] over [1, 3, 4].

For example:

```
50 1 1
135 3 1 2 3
135 3 1 2 3
140 3 1 3 4
130 3 2 3 4
```

Explanation:
There are 4 stories (with ids 1 to 4) and 5 reload events. At the first reload, there is only one story with score of 50 available for display. The next two reloads, we have 3 stories that take up 90 of the 100 pixels available, for a total score of 135. When we reload at time 21, there are 4 stories available for choosing, but only 3 will fit into the browser height. The best set is [1, 3, 4] for a total score of 140. At the last reload event, we no longer consider story 1 when choosing stories because it is more than 10 time units old. The best set is thus [2, 3, 4].

Constraints
All times are positive integers up to 1,000,000,000.
All scores are positive integers up to 1,000,000.
All heights are positive integers.
0 < N <= 10,000
0 < H <= 2,000
0 < W <= 2,000

You should aim to have your algorithm be fast enough to solve our largest test inputs in under 5 seconds, or be as close to that as possible.

Notes
For this contest, we are considering a smaller-scale version of our feed optimization problem with simplified constraints. (The actual feed on our site uses a different algorithm.)

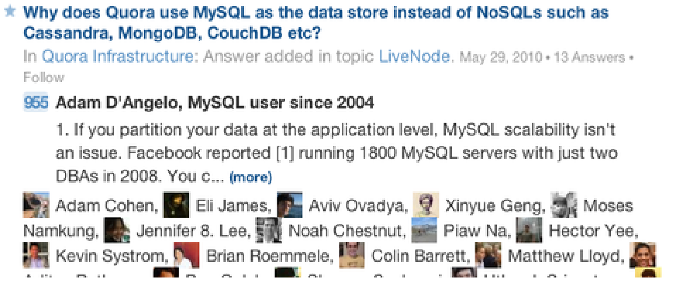
Live Submission (FAQ):

Source File: Choose File No file chosen

Enter your email address here

Submit and Test

Answer Classifier



Why does Quora use MySQL as the data store instead of NoSQLs such as Cassandra, MongoDB, CouchDB etc?

In Quora Infrastructure: Answer added in topic LiveNode. May 29, 2010 • 13 Answers • Follow

955 **Adam D'Angelo, MySQL user since 2004**

1. If you partition your data at the application level, MySQL scalability isn't an issue. Facebook reported [1] running 1800 MySQL servers with just two DBAs in 2008. You c... (more)

Adam Cohen, Eli James, Aviv Ovadya, Xinyue Geng, Moses Namkung, Jennifer 8. Lee, Noah Chestnut, Piaw Na, Hector Yee, Kevin Systrom, Brian Roemmele, Colin Barrett, Matthew Lloyd,

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

Your task will be to devise a classifier that is able to tell good answers from bad answers, as well as humans can. A good answer is denoted by a +1 in our system, and a bad answer is denoted by a -1.

Input format (read from STDIN):

The first line contains N, M. N = Number of training data records, M = number of parameters. Followed by N lines containing records of training data. Then one integer q, q = number of records to be classified, followed by q lines of query data

Training data corresponds to the following format:
`<answer-identifier> <+1 or -1> (<feature-index>:<feature-value>)*`

Query data corresponds to the following format:
`<answer-identifier> (<feature-index>:<feature-value>)*`

The answer identifier is an alphanumeric string of no more than 10 chars. Each identifier is guaranteed unique. All feature values are doubles.
0 < M < 100
0 < N < 50,000
0 < q < 5,000

For example:

```
5 23
2LuzC +1 1:2101216030446 2:1.807711 3:1 4:4.262680 5:4.488636 6:8
LmnUc +1 1:99548723068 2:3.032810 3:1 4:2.772589 5:2.708050 6:0.0
ZINTz -1 1:3030695193589 2:1.741764 3:1 4:2.708050 5:4.248495 6:0
gX60q +1 1:2086220371355 2:1.774193 3:1 4:3.258097 5:3.784190 6:0
5HG4U -1 1:352013287143 2:1.689824 3:1 4:0.000000 5:0.693147 6:0.
2
PdxMK 1:340674897225 2:1.744152 3:1 4:5.023881 5:7.042286 6:0.000
ehZ0a 1:2090062840058 2:1.939101 3:1 4:3.258097 5:2.995732 6:75.0
```

This data is completely anonymized and extracted from real production data, and thus will not include the raw form of the
answers. We, however, have extracted as many features as we think are useful, and you can decide which features make sense to be included in your final algorithm. The actual labeling of a good answer and bad answer is done

organically on our site, through human moderators.

Output format (write to STDOUT):
For each query, you should output q lines to stdout, representing the decision made by your classifier, whether each answer is good or not:

```
<answer-identifier> <+1 or -1>
```

For example:

```
PdxMK +1
ehZ0a -1
```

You are given a relative large sample input dataset offline with its corresponding output to finetune your program with your ML libraries. It can be downloaded here: http://qsf.cf.quoracdn.net/Quora...

Scoring
Only one very large test dataset will be given for this problem online as input to your program for scoring. This input data set will not be revealed to you.

Output for every classification is awarded points separately. The score for this problem will be the sum of points for each correct classification. To prevent naive solution credit (outputting all +1s, for example), points are awarded only after X correct classifications, where X is number of +1 answers or -1 answers (whichever is greater).

Timing
Your program should complete in minutes. Try to achieve as high an accuracy as possible with this constraint in mind.

Live Submission (FAQ):

Source File: Choose File No file chosen

Enter your email address here

Submit and Test

Python URI

In Python, write a class or module with a bunch of functions for manipulating a URI. For this exercise, pretend that the urllib, urllib2, and urlparse modules don't exist. You can use other standard Python modules, such as re, for this. The focus of the class or module you write should be around usage on the web, so you'll want to have things that make it easier to update or append a querystring var, get the scheme for a URI, etc., and you may want to include ways to figure out the domain for a URL (british-site.co.uk , us-site.com , etc.)

We're looking for correctness (you'll probably want to read the relevant RFCs; make sure you handle edge cases), and elegance of your API (does it let you do the things you commonly want to do with URIs in a really straightforward way?,) as well as coding style. If you don't know Python already, then this is also an exercise in learning new things quickly and well. Your code should be well-commented and documented and conform to the guidelines in the PEP 8 Style Guide for Python Code. Include some instructions and examples of usage in your documentation. You may also want to write unit tests.

EMAIL SUBMISSION
Send your solutions to challenges@quora.com

Trend Analyzer

Note: This is an open-ended problem geared towards our Data Scientist role. Data scientists at Quora work with data at all levels of our product, allowing our team to make informed decisions about which features to pursue and which areas need improvement. Being an information repository in itself, efficient and accurate use of data is central to everything we do at Quora.

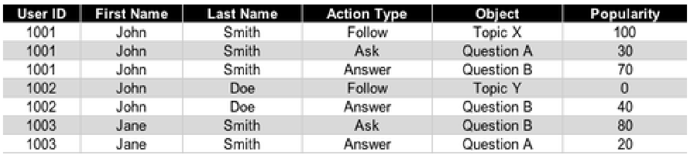
At Quora, we deal with a lot of data. As the number of users on our site grows, so does the volume of content and the number of interactions that go on daily.

We want to be able to determine various social and intellectual topics that are trending over time. However, we take privacy very seriously, and so we want to do this in a way that does not put any individual’s personal information at risk.

Your task is to design and implement a trend analysis engine. This engine should be able to take as input a table of raw data and present an interface for trend analysis. You are free to design this interface and choose the data analysis features that you wish to support. It is important that the engine only exposes to the user data that does not have any personal identifiers and protects sensitive attributes from being revealed. We provide some links to resources about how to think about this in the Resources section.

Scoring
Your submission will be evaluated according to the following criteria:
Code Quality and Design (25%):
Is the code organized well and readable?
Were there good design decisions made and documented?
Privacy (25%):
Is there adequate preservation of privacy?
Is there a reasonable way to trade off between protecting personal information and getting sufficiently high quality of data for analysis?
Analytics (25%):
What analytical metrics do you support?
How do you detect or compute trends in the data?
Visualization (25%):
Are there compelling visual aids for presenting data?
Are there ways to interactively explore the data?

A Primer on Privacy Protection
Suppose there is a table with the following raw data:

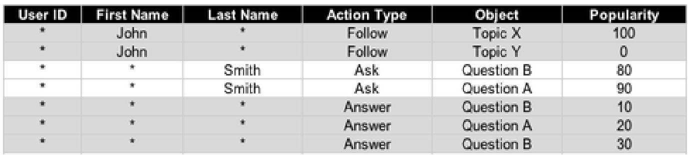


User ID	First Name	Last Name	Action Type	Object	Popularity
1001	John	Smith	Follow	Topic X	100
1001	John	Smith	Ask	Question A	30
1001	John	Smith	Answer	Question B	70
1002	John	Doe	Follow	Topic Y	0
1002	John	Doe	Answer	Question B	40
1003	Jane	Smith	Ask	Question B	80
1003	Jane	Smith	Answer	Question A	20

Table 1: Raw data with personal information exposed

The first field (User ID), or the second and third field if used together (First and Last Name), will uniquely identify the person in this data set. If you are trying to determine aggregate big picture trends of Popularity from this data, there is no need to deal with data of this resolution and granularity.

Instead, if you knew the fields which are identifiers and quasi-identifiers, you could produce the following table after an anonymization process.



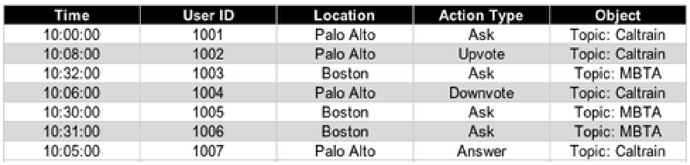
User ID	First Name	Last Name	Action Type	Object	Popularity
*	John	*	Follow	Topic X	100
*	John	*	Follow	Topic Y	0
*	*	Smith	Ask	Question B	80
*	*	Smith	Ask	Question A	90
*	*	*	Answer	Question B	10
*	*	*	Answer	Question A	20
*	*	*	Answer	Question B	30

Table 2: Raw data after anonymization process

Grouped in the following way, you now have a table whose rows are placed into groups where you can no longer distinguish an individual record from the table. And you can still observe that asking of questions is yielding a high popularity score compared to answering questions, while following of topics has a mixed response.

A Primer on Time-Series Data

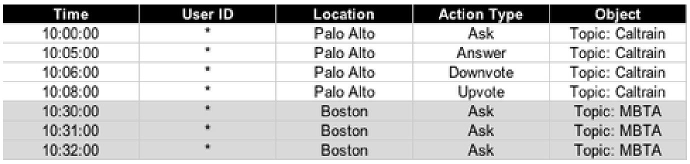
As another example, here is a different data set that contains time-series data.



Time	User ID	Location	Action Type	Object
10:00:00	1001	Palo Alto	Ask	Topic: Caltrain
10:08:00	1002	Palo Alto	Upvote	Topic: Caltrain
10:32:00	1003	Boston	Ask	Topic: MBTA
10:06:00	1004	Palo Alto	Downvote	Topic: Caltrain
10:30:00	1005	Boston	Ask	Topic: MBTA
10:31:00	1006	Boston	Ask	Topic: MBTA
10:05:00	1007	Palo Alto	Answer	Topic: Caltrain

Table 3: Raw data with no apparent patterns or trends

At first glance, it is difficult to detect any patterns or trends in the data. But if you strip out personal info and sort by Location and then by Time, a few things become clear.



Time	User ID	Location	Action Type	Object
10:00:00	*	Palo Alto	Ask	Topic: Caltrain
10:05:00	*	Palo Alto	Answer	Topic: Caltrain
10:06:00	*	Palo Alto	Downvote	Topic: Caltrain
10:08:00	*	Palo Alto	Upvote	Topic: Caltrain
10:30:00	*	Boston	Ask	Topic: MBTA
10:31:00	*	Boston	Ask	Topic: MBTA
10:32:00	*	Boston	Ask	Topic: MBTA

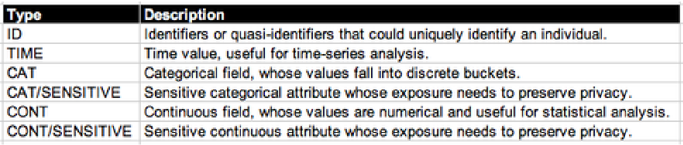
Table 4: Raw data after sorting exposes different patterns and trends

First, there is a high correlation between the Location of the user and the Object being acted upon. In this case, it happens to be that the users seem to be discussing the public transport systems in their location (MBTA for Boston and Caltrain for Palo Alto). It is also clear that a question was asked in about the Caltrain, and after a short wait, there was a flurry of activity around that topic, with an answer followed by 2 votes. Whereas in Boston, it was just a spike in questions getting asked about the MBTA.

Input Format

Your engine will be tested by us on several data sets, with both synthetic and real data. Your engine should take as input a path to a comma-delimited text file. The first line will be the field names, and the second line will be the field type. The remaining rows will be the raw data.

The table below lists the field types we want you to support. If you have ideas for other types, feel free to include support for them and provide us an input file for us to evaluate them.



Type	Description
ID	Identifiers or quasi-identifiers that could uniquely identify an individual.
TIME	Time value, useful for time-series analysis.
CAT	Categorical field, whose values fall into discrete buckets.
CAT/SENSITIVE	Sensitive categorical attribute whose exposure needs to preserve privacy.
CONT	Continuous field, whose values are numerical and useful for statistical analysis.
CONT/SENSITIVE	Sensitive continuous attribute whose exposure needs to preserve privacy.

Table 5: Field types to be specified in the input data set

The above 2 sample data sets can be described by the following schema.

Example 1
User ID, First Name, Last Name, Action Type, Object, Popularity
ID, ID, ID, CAT, CAT/SENSITIVE, CONT

Example 2
Time, User ID, Location, Action Type, Object
TIME, ID, CAT, CAT, CAT

Submission Format

Your submission should be the URL to a standalone repository that we can clone, build and run with minimal effort.

You have a week for this problem and you may ask clarification questions by emailing challenges@quora.com . We will try our best to respond to your questions but given the open-ended nature of this problem, you may also use your best judgement to make decisions and let us know how you dealt with these by documenting them.

Resources

We give links below to 3 papers that we believe are excellent starting points to think about preserving privacy in microdata. You may extend these ideas with other state-of-the-art in privacy and data research.

In your submission, you may also use any third party software libraries that are helpful to you, provided you cite them properly in your comments.

[1] L. Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570
http://epic.org/privacy/reidenti...

[2] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In Proc. 22nd Intnl. Conf. Data Engg. (ICDE), page 24, 2006.
http://www.cs.cornell.edu/~vmuth...

[3] N. Li, T. Li, and S. Venkatasubramanian. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. IEEE International Conference on Data Engineering (ICDE), 2007
http://www.cs.purdue.edu/homes/n...

Sample Data
You may download test datasets as a tarball from:
http://qsf.cf.quoracdn.net/Quora...

The original data was sourced from http://opendata.socrata.com/ , but we've touched it up a little and added the relevant field types to make them applicable to this problem. For each dataset in the zip file, we provide both the original raw data and the cleaned annotated version, as well as the script used to do the cleaning.

EMAIL SUBMISSION

Send your solutions to challenges@quora.com