CS 6501 Text Mining:

An Question Recommendation System for Question Answer Community (Stackoverflow)

Presenter: Haoyu Chen Haoran Hou

What is Question Answering Community:

Community question answering (cQA) provides a platform for people with diverse background to share information and knowledge.

MedHelp

Corporation

MedHelp is an American private corporation that was founded in February 1994 and pioneered the field of consumer health information and communities on the Internet, before WebMD, Microsoft or Yahoo!. Wikipedia

stackoverflow

Quora

Website

Quora is a question-and-answer website where questions are created, answered, edited and organized by its community of users. The company was founded in June 2009, and the website was made available to the public on June 21, 2010. Wikipedia

Founded: June 2009

CEO: Adam D'Angelo

Founders: Adam D'Angelo, Charlie Cheever

StackExchange

YAHOO! ANSWERS

People need help!

"How do I take care of my pet potato?"

What we decided to work on:

There’s only one style of programming: stackoverflow oriented programming.

Exhibit A:
Result Ranking doesn't consider about the quality of answers.

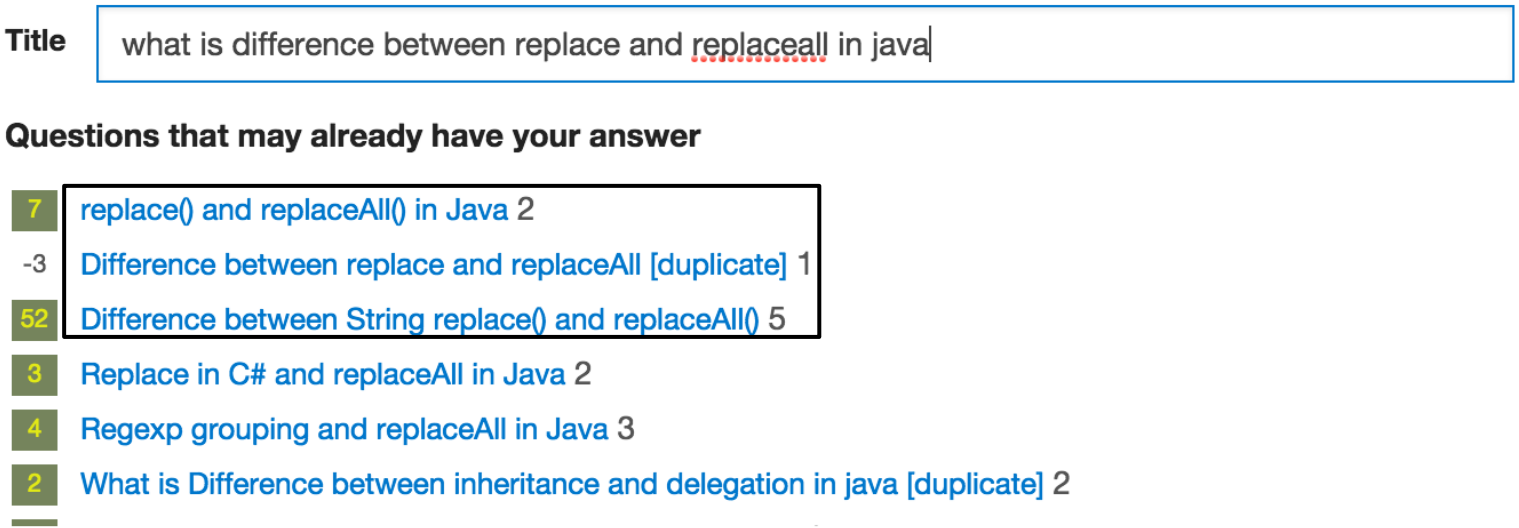


Exhibit B:
Result Ranking doesn't work well in some cases

Title: replace replaceall difference java

Questions that may already have your answer

0	I can't get replace / replaceAll to work 1
2	Java replace/replaceAll strange behavior 2
-3	java replace e replaceAll. ReplaceAll not working [closed] 2
7	replace() and replaceAll() in Java 2
-3	Difference between replace and replaceAll [duplicate] 1
58	Difference between String replace() and replaceAll() 6

What we aim to do:

- Find similar questions and list them in more reasonable order.
- Get answers in a faster and more convenient way.

About stackoverflow

- No need for sentiment analysis
- Few duplicated questions
- Provide tags
- Ordered Answer: Voting
- Full data provided

New query
->Best existing post with most similar query
->Return best answer

Our thoughts on improvement:

- query-answer matching: After finding similar existing queries, compute the similarity between the new query and the best answer
- Adding tag matching along with query matching
- Find the reasonable ‘return-best-answer’ strategy

query-answer matching

Query: difference replace replaceall java

Question title

Difference between String replace() and replaceAll()

Question content

59

15

I am asking this question out of curiosity.

What's the difference between java.lang.String 's `replace()` and `replaceAll()` methods, other than later uses regex. For simple substitutions like, replace `.` with `/` , is there any difference?

java string replace

share edit

edited Jul 14 '14 at 7:30
user2864740
27.6k • 4 • 22 • 58

asked May 31 '12 at 5:24
Jijoy
2,132 • 7 • 23 • 41

Best answer

47

In `java.lang.String` , the `replace` method either takes a pair of char's or a pair of `CharSequence` 's (of which String is a subclass, so it'll happily take a pair of String's). The `replace` method will replace all occurrences of a char or `CharSequence` . On the other hand, both `String` arguments to `replaceFirst` and `replaceAll` are regular expressions (regex). Using the wrong function can lead to subtle bugs.

share edit

edited May 31 '12 at 5:33

answered May 31 '12 at 5:28
emilan
2,796 • 4 • 18 • 29

⟹ Only compute new query and existing query

Adding tag matching

Compute the similarity between existing queries, as well as their tags

e.g.

new query: *difference replace replaceall java*

existing query: *difference between string replace() and replaceall()*

tags: java string replace

Find answer:

More votes -> acceptance

Favor vote more than acceptance

Return even if there’s no (good) answer: comments

Let's start from Solr

Solr is the popular, blazing-fast, open source enterprise search platform built on Apache Lucene

--- The Headline on Solr Official Website

Key Facts on *Stackoverflow data*

Open -- Under CC BY-SA 3.0(ShareAlike and Attribution)
API -- E.g. Search Users, Answer, Questions
Updation -- every Monday
Size -- 8 million questions (28G)

Link:http://data.stackexchange.com/help

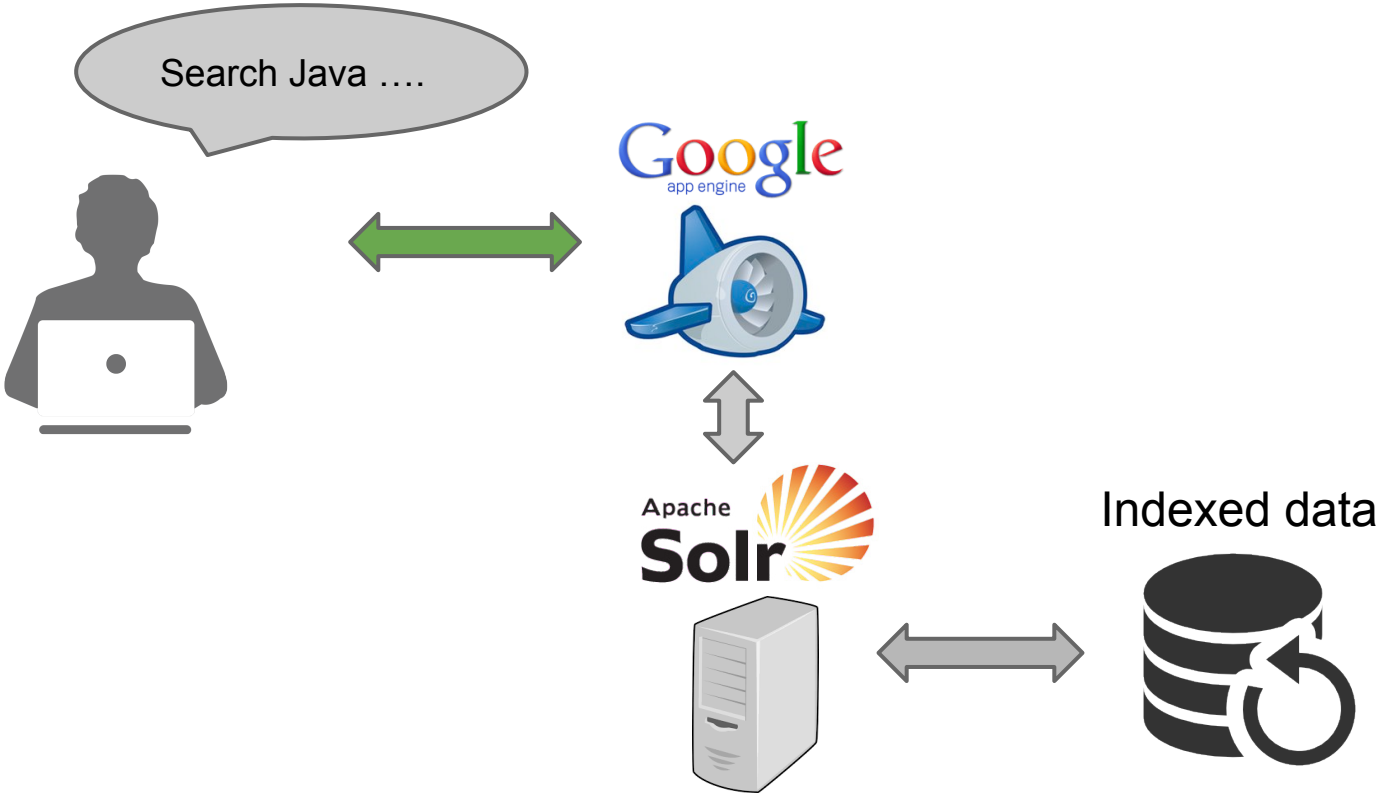
Preprocessing *Stackoverflow data*

Select Useful features -- Tags, QuestionsID, Titles

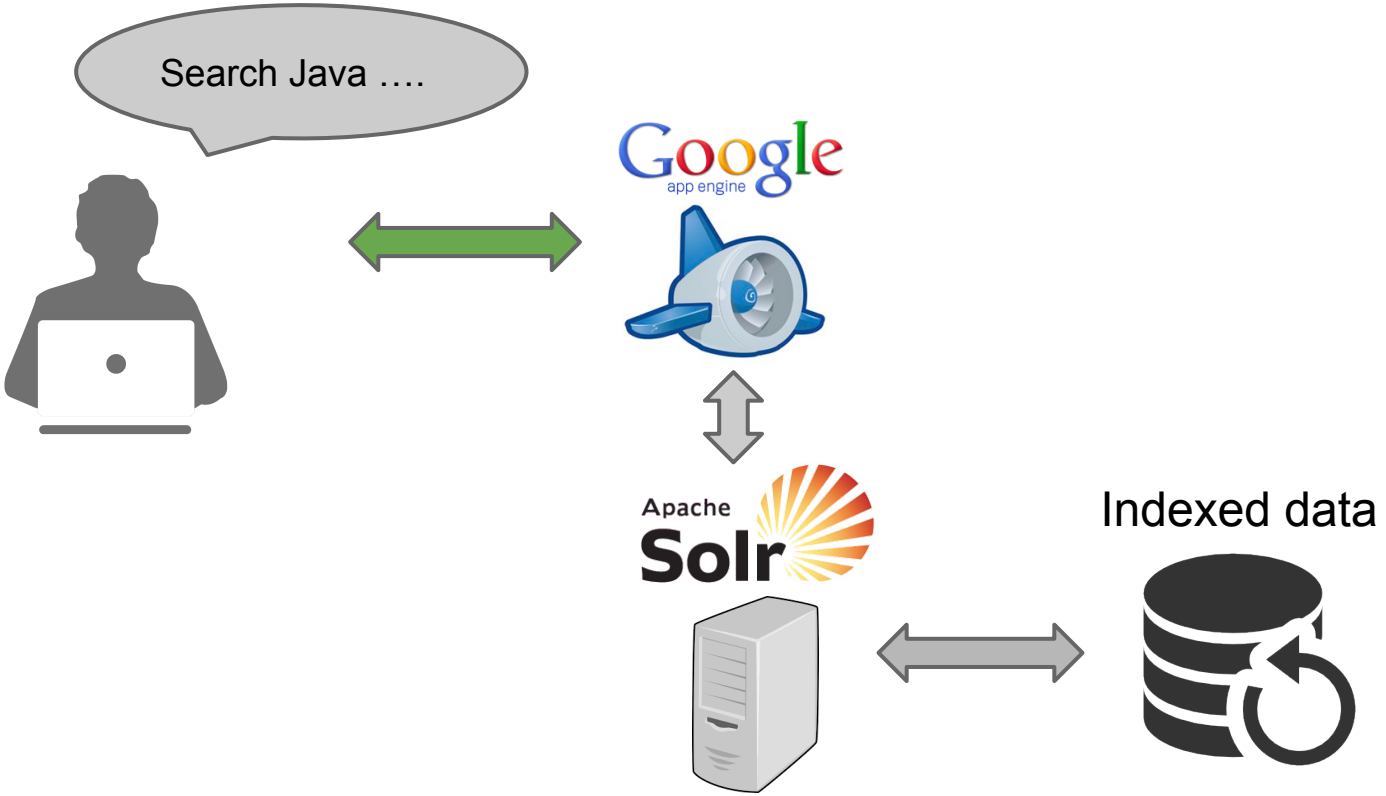
Convert it into Solr input format

Result: 28G -> 1.6G

Search Flow Chart



Search Flow Chart



Solr similarity algorithm:

$$\text{score}(q,d) = \text{coord}(q,d) \cdot \text{queryNorm}(q) \cdot \sum_{t \text{ in } q} \left(\text{tf}(t \text{ in } d) \cdot \text{idf}(t)^2 \cdot t.\text{getBoost}() \cdot \text{norm}(t,d) \right)$$

document contains more query's term the higher

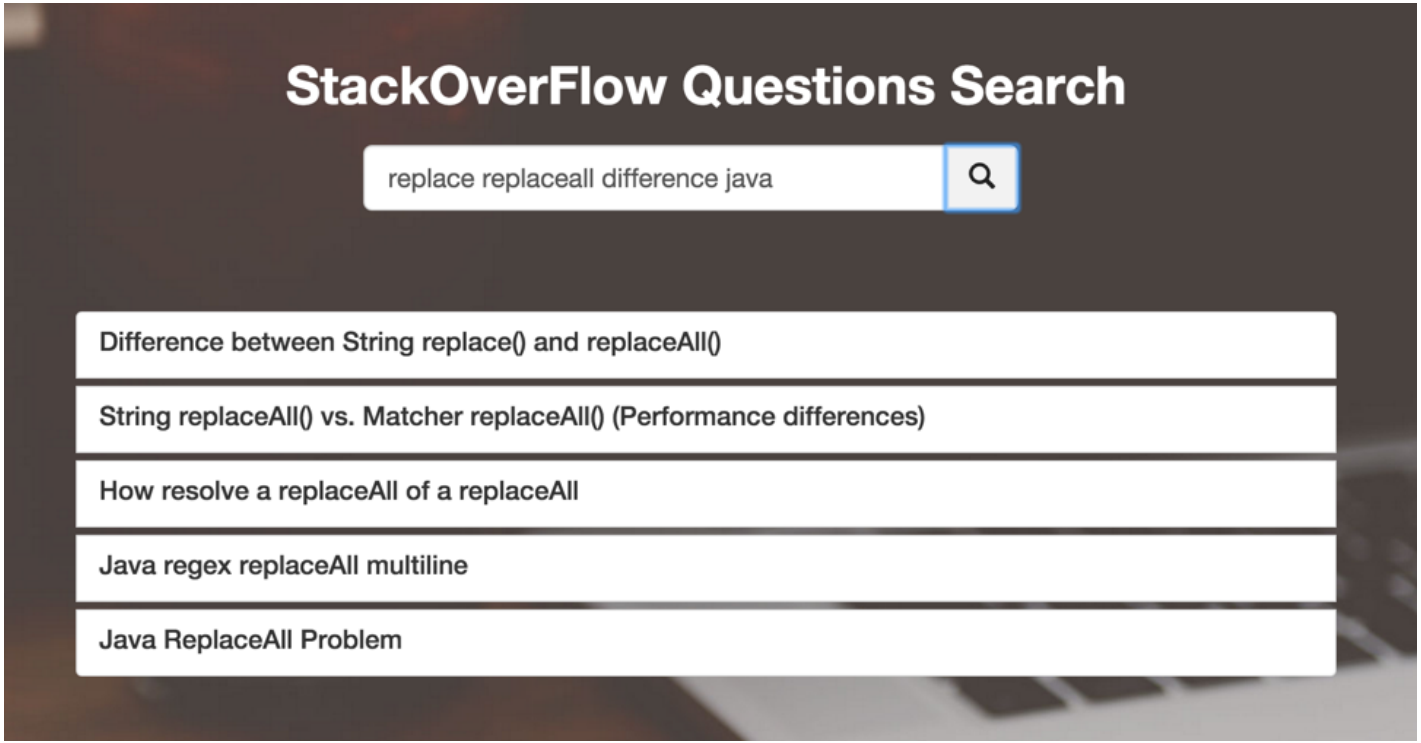
make scores between queries comparable

Normalize document with boost

$$\frac{1}{q.\text{getBoost}()^2 \cdot \sum_{t \text{ in } q} \left(\text{idf}(t) \cdot t.\text{getBoost}() \right)^2}^{1/2}$$

$$doc.\text{getBoost}() \cdot \text{lengthNorm} \cdot \prod_{\text{field } f \text{ in } d \text{ named as } t} f.\text{getBoost}()$$

Let’s Demo Our Tools!



Let’s Demo Our Tools!

Features:

- Auto change detection
- Answer overview - (More responsive than StackOverflow version)

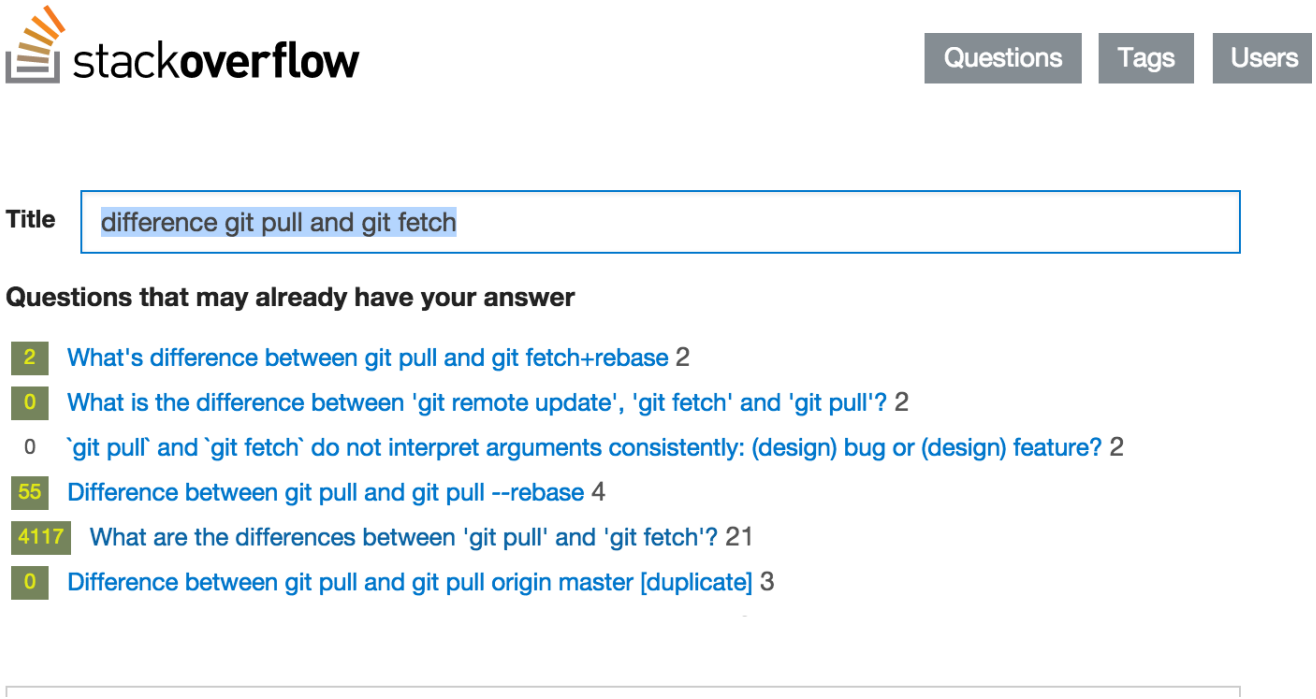
Difference:

- Search not just for title, but also tags.
- Show answer with the largest votes

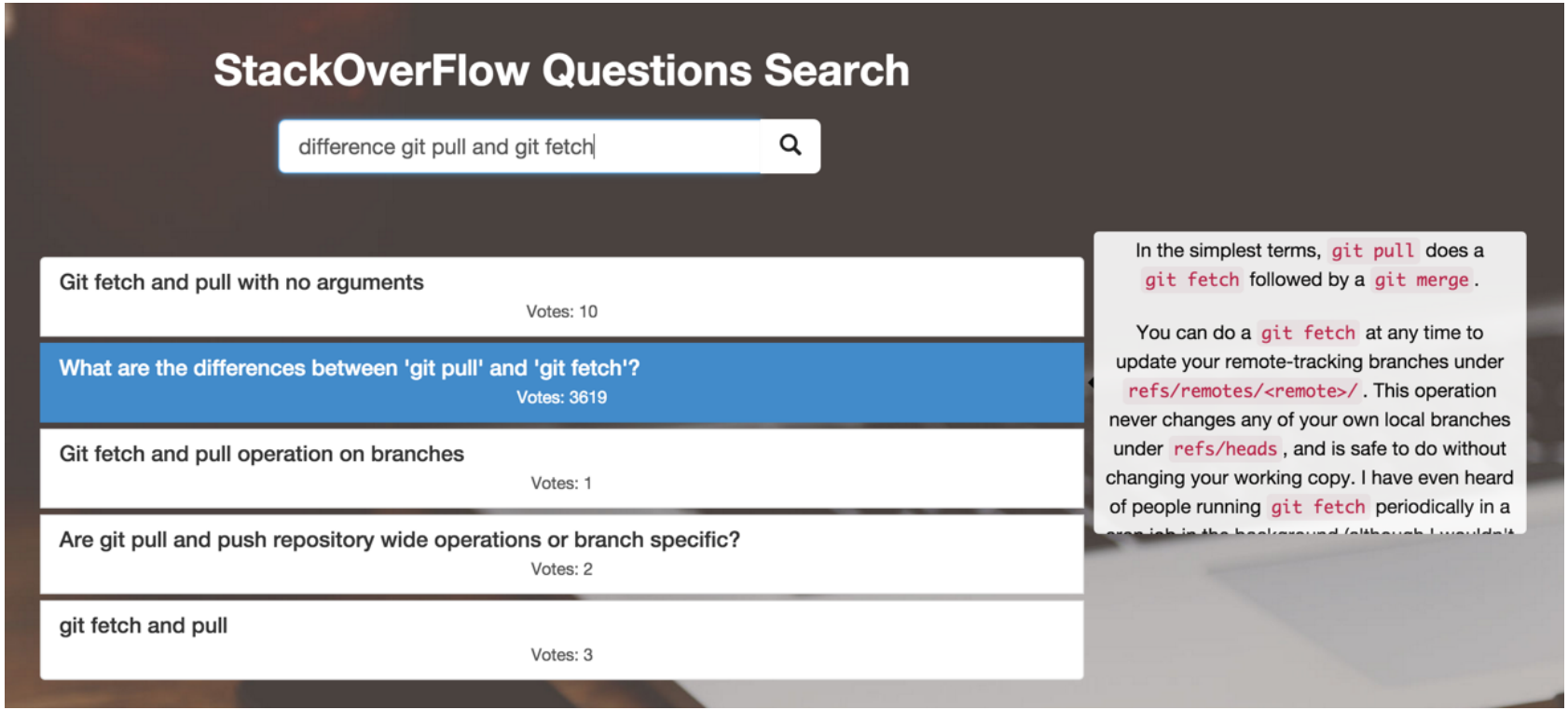
Testing Questions:

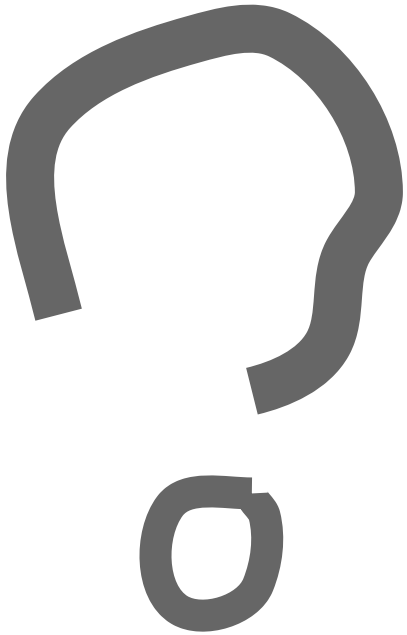
- Replace

Demo 1


stackoverflow
Questions
Tags
Users
Title
difference git pull and git fetch
Questions that may already have your answer
2 What's difference between git pull and git fetch+rebase 2
0 What is the difference between 'git remote update', 'git fetch' and 'git pull'? 2
0 `git pull` and `git fetch` do not interpret arguments consistently: (design) bug or (design) feature? 2
55 Difference between git pull and git pull --rebase 4
4117 What are the differences between 'git pull' and 'git fetch'? 21
0 Difference between git pull and git pull origin master [duplicate] 3

Demo 1


StackOverFlow Questions Search
difference git pull and git fetch
Git fetch and pull with no arguments
Votes: 10
What are the differences between 'git pull' and 'git fetch'?
Votes: 3619
Git fetch and pull operation on branches
Votes: 1
Are git pull and push repository wide operations or branch specific?
Votes: 2
git fetch and pull
Votes: 3
In the simplest terms, git pull does a git fetch followed by a git merge.
You can do a git fetch at any time to update your remote-tracking branches under refs/remotes/<remote>/. This operation never changes any of your own local branches under refs/heads, and is safe to do without changing your working copy. I have even heard of people running git fetch periodically in a

Future steps

- Distribute different weight to question title and tags
- Dig more information provided by comments
- Recommend tag using MoreLikeThis feature