Mining developers’ questions

Mathias Birrer
Supervised by Pooja Rani
Developers have a lot of questions!
Where do they ask the questions?
How to collect these questions?
Web scraping

Web scraping is from a legal perspective often in a grey area and should be done carefully!
Rules for web scraping

- Use API if available
- Respect Terms of Service
- Limit numbers of requests
- Identify yourself
- Respect robots.txt
- Ask for permission or don’t do it if in doubt
How we started...
First source...
Stack Overflow API

```
/2.2/search/advanced?order=desc&sort=activity&site=stackoverflow
```
Stack Exchange Data Dump

<table>
<thead>
<tr>
<th>File Name</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>stackoverflow.com-Badges.7z</td>
<td>211.1M</td>
</tr>
<tr>
<td>stackoverflow.com-Comments.7z</td>
<td>3.8G</td>
</tr>
<tr>
<td>stackoverflow.com-PostHistory.7z</td>
<td>22.4G</td>
</tr>
<tr>
<td>stackoverflow.com-PostLinks.7z</td>
<td>74.1M</td>
</tr>
<tr>
<td>stackoverflow.com-Posts.7z</td>
<td>12.8G</td>
</tr>
<tr>
<td>stackoverflow.com-Tags.7z</td>
<td>758.2K</td>
</tr>
<tr>
<td>stackoverflow.com-Users.7z</td>
<td>410.0M</td>
</tr>
<tr>
<td>stackoverflow.com-Votes.7z</td>
<td>963.7M</td>
</tr>
</tbody>
</table>

https://archive.org/details/stackexchange
Stack Exchange Data Explorer

-- How many upvotes do I have for each tag?
-- how long before I get tag badges?

DECLARE @UserId int = ##UserId##

SELECT --TOP 20
    TagName,
    COUNT(*) AS UpVotes
FROM Tags
INNER JOIN PostTags ON PostTags.TagId = Tags.id
INNER JOIN Posts ON Posts.ParentId = PostTags.PostId
INNER JOIN Votes ON Votes.PostId = Posts.Id and VoteTypeId = 2
WHERE
    Posts.OwnerUserId = @UserId
GROUP BY TagName
ORDER BY UpVotes DESC
Next source...
Challenges

- No API
- Data not available publicly
- Hard to scrape
Is it legal to scrape Quora?
d. **Permitted uses.** If you operate a search engine, web crawler, bot, scraping tool, data mining tool, bulk downloading tool, `wget` utility, or similar data gathering or extraction tool, you may access the Quora Platform, subject to the following additional rules: i) you must use a descriptive user agent header; ii) you must follow `robots.txt` at all times; iii) your access must not adversely affect any aspect of the Quora Platform’s functioning; and iv) you must make it clear how to contact you, either in your user agent string, or on your website if you have one.
How to scrape Quora?

Browser Automation
- No API
- Handle JS
- Scriptable

Mimic humans
- Request limitation
- Navigation

Tools

21
How to manage **questions** from **heterogeneous sources**?
We need a **Question Management Tool!**

- Heterogeneous sources
- Data preprocessing
- Export for analysis
Data preprocessing

**Question**
- text
- tags
- answer
- source
- text_sanitized

**Answer**
- answer
- answer_sanitized

**Question**
- text
- tags
- answer
- source
- text_sanitized
- answer_sanitized
Technologies
Similarity Matcher

Fast sentence lookup based on sentence vectors.
Demo
Future work

- More sources
- Build corpus of high quality questions
- Derive possible classification of questions
- DSL for data preprocessing
Summary

- About developer’s questions
- Web scraping
- Mining Stack Overflow / Quora
- Question Management Tool
A prudent question is one-half of wisdom.

Francis Bacon