StackOverflow Data

Download: [R Data] [CSV]

StackOverflow.com is an online question-and-answer site for programmers. Started in fall 2008, its rich feature set brought rapid popularity: users can ask and answer questions, collaboratively tag and edit questions, vote on the quality of answers, and post comments on individual questions and answers.

The owners of StackOverflow.com have released a "data-dump" of their website. From this, I compiled a dataset of answers posted between February 18, 2009 and June 7, 2009.

Within the zip file is a Matlab .mat file and an R .Rdata file. The matrix it contains has the following attributes:

- qid: Unique question id
- i: User id of questioner
- qs: Score of the question
- qt: Time of the question (in epoch time)
- tags: a comma-separated list of the tags associated with the question. Examples of tags are ``html", ``R", ``mysql", ``python", and so on; often between two and six tags are used on each question.
- qvc: Number of views of this question (at the time of the datadump)
- qac: Number of answers for this question (at the time of the datadump)
- aid: Unique answer id
- j: User id of answerer
- as: Score of the answer
- at: Time of the answer

As this is a dataset with a row for every *answer*, note that qid values will be repeated, once for every answer to that question.

The dataset is sparse, with about 236,789 answers from 15,098 unique users, with the majority of users answering less than 50 questions. There is also a SQLite database with far more information than in this dataset if you're interesting in having more data to work with.

Potential Project Ideas:

Predicting Answerers

Given the deluge of questions arriving every day, some users may want a filter that provides only questions they might be interested in based on their past activity. One approach is to build a classifier for each user that predicts whether they will answer each question and serve only those with a high probability of being answered. For very active users, this seems reasonable; for users with little activity, some creativity may be required.

Predicting Time until Answer

One advantage of large scale collaborative sites like Wikipedia and StackOverflow is the speed at which new information accumulates. Some questions on StackOverflow seemed to get answered immediately while others sit around for a while. As you are writing your question, it would be cool to have an idea of how long it might take to get answered. How well can you predict the time until a question gets answered? This will likely depend on the subject matter of the question, since some programming communities are more active than others.

Expert Finding Algorithms

The StackOverflow community looks to a points-based system to gauge users' expertise. These points are earned for good questions, good answers, and so on. Other algorithms exist for finding important nodes in a network (e.g. PageRank). You can also query the datadump for the reputation points for all the users and see if your measure is correlated with the points system. What advantages or disadvantages does your measure have?

Contact

Feel free to email Chris DuBois (duboisc 'at' ics 'dot' uci 'dot' edu).

Helpful Links

- Latest data dump (torrent)
- Python script for converting xml data dump files to sqlite database