**Stack Exchange**

**Metrics**

* Total Posts: 20 million
* Total Comments: 30 million
* Number of Categories: 6 → Technology, Culture/Recreation, Life/Arts, Science, Professional, and Business
* Number of Subcategories: 176
  + Technology: 73 Subcategories
  + Culture/Recreation: 46 Subcategories
  + Life/Arts: 25 Subcategories
  + Science: 22 Subcategories
  + Professional: 7 Subcategories
  + Business: 3 Subcategories
* Number of Tags: 62746 (59200 w/ >1 post, 878 w/ >8000 posts)

**User Interface**

* Topics can be filtered by:
  + Categories/Subcategories
  + Tags
  + Least-to-Greatest/Greatest-to-Least Upvoted items
  + Most recent to Least recent
  + Active
  + Bountied (offer reputation, a score connected to each user, to anyone who provides a good answer)
* Individual topics display:
  + The title, dates, and views at the top
  + The vote score of the post
  + Content of the post, along with authors and editors
  + Comments on the original post, along with vote scores, dates, and authors
  + Answers to the original post, along with their vote scores, dates, and authors/editors
  + Comments on each answer, along with vote scores, dates, and authors
* It doesn’t look particularly like the Stem-Away forums, which arrange their topics into finer categories in a grid. StackExchange arranges all the topics in a forum (like StackOverflow) in a list with multiple pages.

**Functionality / Reliability**

* Tags
  + Popular Tags: javascript, java, python, c+, php, android, html, jquery, c++, css, ios, mysql, sql, asp.net, r, node.js, arrays, c, ruby-on-rails, .net, json, objective-c, sql-server, swift, angularjs
  + Almost all of the posts on Stack Exchange have multiple tags that correlate with the information presented in the post
  + 46.9% of posts have a positive score, from which we can assume at least the tags of those posts are correct. We could scrape only these, because we’d still be left with nearly 10 million posts.
* Categories
  + Most posts are placed properly into their designated categories so viewers find posts related to what they are looking for
* Post Titles/Questions
  + Almost all of the Titles/questions on posts on Stack Exchange correlate with the information being presented
* Content
  + Popular content includes Syntax Errors, missing packages, invalid version errors, Runtime Errors, etc.
  + (add more)
* Comments
  + Most comments relate to the rest of the thread including the post’s title/question being presented
  + People comment about the content in the post or thank a certain user for helping out; for example, telling the user what they did wrong in their code and how to debug it
  + Comments tend to be related to the post but not directly answering it, or they are answers which aren’t fully-formed yet
* Answers
  + Answers are like more major comments
  + These tend to be directly related to the post, and contain more information than the original question

**Alternative forums**

* StackOverflow is a forum on StackExchange mostly devoted to technical/coding questions. This could exclude a lot of questions about bioinformatics or some very specific questions about ML, UX, or FS, which have their own forums.
* To resolve this, we could train on bioinformatics.stackexchange.com, but it has only ~3,400 posts. We could pick a broader category like biology.stackexchange.com and train on both forums, but this runs the risk of splitting our dataset across too many different forums.
* We could also train on the AI forums, ai.stackexchange.com, stats.stackexchange.com, and datascience.stackexchange.com, or the UX forum, ux.stackexchange.com, but this has the same problem as before.
* We chose the StackOverflow forum despite these issues because it is by far the biggest and has a wide range of topics covered which all relate to STEM.

**Conclusion**

* We think StackExchange and StackOverflow in particular are well-suited to our project. They have an enormous amount of posts, and the vote system allows us to scrape only the highest-quality posts as well to ensure a good training set. The content of the questions, comments, and answers is similar to the Stem-Away forums, as the subject matter is very similar, and despite the interface appearing different, the methods by which one can sort (tags, categories, and so on) are the same.