Project Proposal: Analysing Stack Exchange Forums

Valeria Chernenko
valeriia.chernenko@tum.de

Kerstin Gottelt
ke.gottelt@in.tum.de

Oleksandr Shchur
shchur@in.tum.de

Tizian Sarre
tizian.sarre@gmx.de

1. DATA SET

The data set [1] we are intending to work on contains information about user activity on Stack Exchange forums. Stack Exchange [2] is a network of Q&A communities. Each website within the network is focused on a different topic such as Software Engineering, Linux or Mathematics. Each board has its own data set. Every data set contains information about posts (a.k.a. questions), tags, answers, user profiles, comments and votes.

2. MOTIVATION

Our first idea is to develop an algorithm to find highly similar posts. One possible application for it is checking new posts for uniqueness. It often happens that inexperienced users ask questions that have already been answered. In this case pointing directly to the answer will save the user's time and will reduce the amount of redundant data on the forums. The current implementation of this feature on Stack Exchange lists possible duplicates based only on the title of the question. Our approach will take into account not only the title, but also the body and tags of the post. This will further help users to avoid posting duplicate questions.

A second possible application of our technique is to provide links to similar questions for existing ones that are not yet marked as accepted. Sometimes the questions could not be resolved by other users due to a lack of details or bad formulation. If there are other similar questions that already have a solution, pointing to them could help finding a correct answer to the original question. This will also allow users to find better results on Stack Exchange via external search engines.

Another potential way to apply our solution is to suggest tags for new posts that have few or none at all. For instance, on StackOverflow currently more than 5 out of 9 million posts have no tags assigned to them. Resolving this issue will make it easier for experts to find questions they might be able to answer. This will allow the askers to get faster responses, make navigation easier for experienced users and hopefully increase the rate at which questions are being answered. As of now, the Stack Exchange platform suggests tags only if there are hyperlinks in the question's body. A question containing no hyperlinks will get no tag recommendations from the system.

3. IMPLEMENTATION DETAILS

As our first step we intend to apply different sorts of preprocessing techniques on the data, such as stripping HTML markup, stop-word elimination and stemming. Then, the main challenge will be finding a meaningful representation for the actual posts. It is going to be an iterative process and we will refine the methodology as we get feedback from our models. The initial idea is to use bag-of-words representation augmented by other features like urls in the body of the posts or tags. Next, it will be necessary to come up with an appropriate similarity measure and a corresponding locality-sensitive hashing (LSH) technique for it's efficient computation. At this point we are going to have an algorithm (an executable or a script) that given any input post performs the nearest-neighbors search within our data set based on the custom metric discussed above.

The result from the last step can then be used to detect the duplicates as well as do tag recommendation. Finally, we will evaluate the performance of our method by comparison with the existing Stack Exchange implementation. This can be done by brute-force, i.e. by writing some very similar questions ourselves and seeing how they are handled in both cases.

4. TIMELINE

To keep track of our work progress, we set up following timeline:

- **Week 1 (- 29/11):**

  Convert raw data to a more usable format. Preprocess the converted data.

- **Week 2 - 4 (30/11 - 20/12) :**

  Design and optimize the similarity measure with the corresponding LSH technique.

- **Midway goal:**

  Implementation of nearest-neighbor search for posts in the dataset.

- **Week 5 (4/1 - 10/1) :**

  Application of the neares-neighbor search on the Stack Exchange datasets in order to perform some simple duplicate detection and tag recommendation.

- **Week 6 (11/1 - 17/1) :**

Development of a performance metric and comparison with the existing implementation of Stack Exchange.

- **Week 7 - 8 (18/1 - 31/1) :**

  Application of the algorithm to larger datasets (i.e. Stack Exchange forums with more posts). Measurement of scalability.

5. REFERENCES

[1] Internet Archive. Stack exchange data dump. `https://archive.org/details/stackexchange`.

[2] Stack Exchange. About stack exchange. `http://stackexchange.com/about`.