PyStack: A Python Toolkit for Analyzing Stack Exchange Sites

Jiankai Sun · Mengxue Zhang

Updated: November 8, 2017

Abstract PyStack is an open source software project for analyzing Stack Exchange sites in Python. It aims to provide a simple and easy tool to pre-process Stack Exchange sites data dump, and perform research on various topics such as question routing in Community Question Answering services (CQAs). Particular focus is made on the software usability and interoperability with other tools. PyStack can transform the raw XML file to CSV and pickle file, which can be read easily by other tools. Additionally, PyStack supports fast implementations of existing algorithms of qeustion difficulty and user expertise estimation, and question routing in CQAs, which can be used as baselines for new algorithms. The toolkit is available at: https://github.com/zhenv5/PyStack

Keywords Python · Stack Exchange Sites · Toolkit

1 Introduction

Community question answering services (CQAs) such as Stack Overflow and Stack Exchange sites are examples of crowdsourcing platforms, with their usage being examples of an important type of computer supported cooperative work in practice. The usage of such CQAs has seen a dramatic increase in recent years, in both the frequency of questions posted and general user activity. This, in turn, has given rise to several interesting problems ranging from expertise estimation to question difficulty estimation, and from automated question routing to incentive mechanism design on such collaborative websites [1], [2], [3], [4].

Jiankai Sun, and Mengxue Zhang
Department of Computer Science and Engineering
The Ohio State University
E-mail: sun.1306@osu.edu, zhang.8689@osu.edu

Stack Exchange, Inc provides an anonymized dump of all user-contributed content on the Stack Exchange network [1]. Each site is formatted as a separate archive consisting of XML files zipped via 7-zip using bzip2 compression. Each site archive includes Posts, Users, Votes, Comments, PostHistory and PostLinks.

References

1. L. et al., "Estimating domain-specific user expertise for answer retrieval in community question-answering platforms," ADCS'2016.
2. S. et al., "Question/answer matching for cqa system via combining lexical and sequential information," AAAI'2015.
3. F. et al., "Community-based question answering via heterogeneous social network learning," in *AAAI'2016*.
4. R. et al., "Beyond questioning and answering: Teens' learning experiences and benefits of social q&a services," CSCW'2017.

[1] https://archive.org/details/stackexchange