2022-05-10

Documentation of Machine Learning Software

- users of software documentation are not always software experts.
- scientest and engineers of varying expertise and depth of knowledge on data science and machine learning
- automated generation and adaptation of machine learning software documents for users with varying levels of expertise.
- Investigation of stack overflow Q&A and classifying the documentation related to Q/As within machine learning domain.
- Find trigger of problems and potential change requests to the documentation
- Build techniges for automatic documentation generation and extending to adoption, summarization, and explanation of software functionalities.

Introduction

- Documentation is any artifact which its purpose is to communicate information about the software system to which it belongs, to individuals involved in the production of that software.
- Problem: Existing hyper-technicality of software engineering causes users to do extra searching in other resources.
- develop expertise aware software daumentation for ML products
- Research Questions: 1) What type of problems do users face when using documentation of ML software?
  2) How does the expertise level of users impact their understandings of the documentation of ML Software?
  3) How does documentation evolve in relation to the SO questions?

Methodology

Research Question 1: "SO"

- Use ~~soft~~ stackoverflow platform to analyse problems and painpoints of ML software tools.
- Mining following questions: - what types of problems do users have with ML software, and in what areas?
  - which types of documentations are used and questioned?
  - Why thee documentations are referred to, and in which area?
  - How the documentation referrals happen?
- Gathered SO Q$As tagged as ML or popular ML Libraries. ie. Tensorflow or Pytorch
- Labeling found SO Q$A with mining questions mentioned above.
- Used two annotators for thematic analysis for identification by manual analysis
- the latter method helped in defining pithfalls of SO Q$A.

Research Question 2:

- Ultimately adapt software documentation for different user groups and information needs
- develop~~ing~~er understanding of software and domain (ML)

- Movashoutz et al :- The number of upvoted questions, number of up voted answers, number of accepted answers, number of down voted answers.
- consider # of posts tagged in relation to ML
- compare groups with different levels of expertise and reason out the extent of difference between their needs from software documentation.

Research Question 3:
- analyse the relationship between questions to the opened issues in the respective software repository and the commit messages while changing the documentation
- associate each question to a release of a software assuming that a valid question that is asked before Release$_i$ cant be responded in Release$_j$, where $j < i$
- Cosine similarity between the text of the SO question and issue in the repository
- for questions not matching issues, similarity of commit messages ~~o~~ were matched.

Summary and Future Work.

- understanding users behaviour in using ML documentation and understanding the process of evolution of the software documents in ML domain

For RQ1 →
- 500 questions with tensor flow tag and manually categorizing questions, 16.6% of question are related to documentation
  - 50 quesions are concerning parameter tuning, Model creation, and error/exception → 10.7%
  - Most questions were triggered as users werent able to replicate example in documentation and lack of description

22.6% ↘ 14.2% ↙ 24.8% ↘ 12.9%

- 60.7%. of documentations are related to official Tensorflow documentations, while others refer to third-party materials like tutorials, videos, books, scientific papers
- 72.8% hyperlinked to mentioned documentation and others used a screenshot or mentioned the name of the documentation.

What do developers Know about machine learning: a Study of ML discussions on Stack Overflow

- Machine learning can be used for bug prediction, and Software development effort estimation.
- Investigation to understand what educators should focus on, and how different online programming discussion communities can be more helpful.
- Some machine learning topics are significantly more discussed
- Latent Dirichlet Allocation(LDA) can suggest more appropriate tags that make machine learning tags more visible.
- The last point helps getting faster response. on sites like SO

Introduction

- Need to find specific gaps (in those interested) knowledge in machine learning
- Topic modeling approach for making machine learning questions more searchable through tagging system.

RQ1: what machine learning topics are discussed on SO?
RQ2: what exactly do the developers disuss about those machine learning topics?
RQ3: what are the characteristics of machine learning posts considering their popularity and efficiency?
RQ4: Do the developers tag machine learning posts correctly, and can we improve such tagging system with topic modeling?

Methodology

- Data collection:
  - Used java SAX to parse through dataset
  - found 28010 machine learning posts with machine learning tag.
  - entire analysis in this paper is based on SOTorrent

Topic Modeling using LDA:

- LDA is a method of topic modelling
- java based package for topic modelling; "Mallet"
- Requires set of documents and parameters α, β, and # of topics (K)
- Used results from data collections stage
- topic word matrix displayed top words
- document word matrix displag described topic weights for each input document.
- LDA dont have a label but are rather probabilistic distribution of words only

2022-05-11

Topic Categorization 2022-05-10

- LDA was used to identify the topics of machine learning post(28,010)
- Tried categorization with 20 topics and 50 topics (K = 20 & K = 50)
- Set α & β to 0.01, 0.25, 1.00 for each k value
- when $\alpha/\beta = 0.01$: - hard to identify leading topics
  - important topics remain hidden
  - many topics get associated with document when they arent related