**Using Machine Learning to Recommend Similar Questions on Stack Overflow**

**== Motivation ==**

In this project you will explore information retrieval on Stack Overflow. Information retrieval is an important task on question answering forums such as Stack Overflow and Quora. Question retrieval is helpful in the following two cases:

(1) **Question search**: search inside Stack Overflow

(2) **Similar question recommendation**: recommend its most similar questions. This task is especially important when the current question is not answered yet, it can help the question asker find answers from the similar questions.

Information retrieval on Stack Overflow is challenge. The challenge mainly lies in the difficulty in finding *semantically* related questions. For the software development domain, in particular, part of the challenge comes from the programming logic encoded in the questions and answers. Simple keywords matching is not enough for understanding such logic. For example, the following questions contain almost the same keyword tokens, however, they contain different meanings:

**Question 1**: [How to check whether Java Plugins are installed or not in a browser?](https://stackoverflow.com/questions/5631506/)

**Question 2**: [How to check if Java is installed on System (\*not\* in browser)?](https://stackoverflow.com/questions/14366608/)

On the other hand, the following questions are duplicated, yet they contain very different word tokens:

**Question 3:** [How do i force a number with unknown digits behind the decimal mark](https://stackoverflow.com/questions/39357584/)

**Question 4:** How to merge two ints into a double in java

How to help Stack Overflow users questions that are *semantically* related? In this project, you will explore this research question.

**== Learning to Retrieve Semantically Related Questions using Question Links ==**

Existing work shows that machine learning is effective in identifying semantically related question and answer pair. In this paper [1], we find that we could use Stack Overflow question links as the training labels for learning to rank Stack Overflow questions, because these links are very similar to the relevance judgment.

**== Question Link Dataset ==**

The LinkSO dataset: <https://sites.google.com/view/linkso>

This dataset contains Stack Overflow questions in three programming languages: Java, JavaScript and Python.

Download the dataset and unpack it. Each folder contains the following files:

1. The questions and answers: language\_qid2all.txt:
   1. Each line contains 4 components (split by “\t”), which are the question id, title, question body, and the top-2 voted answers.
2. The linked (and unlinked questions): language\_cosidf.txt
   1. Each line contains 4 components (split by “\t”): q1, q2, TFIDF, is\_linked (whether q1 and q2 are linked)
   2. Each q1 contains 30 q2s
3. The training/development/testing folds: language\_train/dev/test\_qid.txt

**== Project Ideas ==**

Your project will be focusing on exploring a retrieval algorithm using the dataset LinkSO. It can be any of the following directions, or combining any two directions:

1. Traditional retrieval methods:
   1. Explore one retrieval function you learned in class on the LinkSO dataset, e.g., BM25, KL-divergence method, the translation model. For this direction, the minimum requirement is to implement something non-trivial, e.g., you need to implement something at least as hard as [the translation model.](http://maroo.cs.umass.edu/getpdf.php?id=811)
2. Ensemble methods of learning to rank:
   1. Explore RankLib with a list of features you define, empirically compare the performance using different features / different learning to rank methods. You must propose features and show the reason why you define those features.
3. Neural ranking methods:
   1. Explore how to use the MatchZoo toolkit. You can report an empirical study on parameter tuning with the MatchZoo tool.
4. Open-ended research question with bonus in the final score: if you want to challenge yourself and publish a paper, propose a ranking algorithm where your goal is to outperform existing state of the art on the LinkSO dataset. Talk to me (Xueqing) if you are interested in this direction. You don’t have to actually outperform SOTA to receive bonus, but the report needs to show your thoughts on the problem.

**== Related work ==**

[LinkSO: A Dataset for Learning to Retrieval Stack Overflow Questions](http://xliu93.web.illinois.edu/pdf/nl4se18.pdf)

1. Learning to rank (ensemble method):

Prof. Zhai’s course: <https://www.coursera.org/lecture/text-retrieval/lesson-6-1-learning-to-rank-part-1-mFYTD>

Ranklib: <https://www.lemurproject.org/ranklib.php>

1. Neural IR:

[Learning deep structured semantic models for web search using clickthrough data](https://www.microsoft.com/en-us/research/publication/learning-deep-structured-semantic-models-for-web-search-using-clickthrough-data/)

[aNMM: Ranking Short Answer Texts with Attention-Based Neural Matching Model](https://arxiv.org/abs/1801.01641)

1. The MatchZoo tool:

<https://github.com/faneshion/MatchZoo>