CMPT 820

Project Proposal

Group Members
Fidelia Orji, fao583, fao583@mail.usask.ca
Amirabbas Jalali, amj301, amj301@mail.usask.ca
Seyedeh Mina Mousavifar, sem311, sem311@mail.usask.ca

Project Topic Title: Benching Machine Learning Methods for Expert Recommendation.

Data
Stackoverflow dataset

Project Topic Description / Abstract
Community question and answer (CQA) platforms, such as Stack Overflow, leverage the knowledge and expertise of users to provide answers to questions. Over time, these websites turn into repositories of knowledge. Knowledge acquisition and exchange are generally crucial, yet costly for both businesses and individuals, especially when the knowledge concerns various areas. CQA platforms offer an opportunity for sharing knowledge at a low cost, where community users, many of whom are domain experts, can potentially provide high-quality solutions to a given problem. In this project, we aim to recommend the experts who are most likely going to answer a given question. The project will have the following stages.

Data Collection: We will collect StackOverflow (SO) dataset. Stackoverflow is an open community where developers share knowledge and skills through posting questions and answering each other's questions. It has a "data-dump" where its complied datasets are kept so that people can download it for research purposes. The dataset has the following files: Posts, PostLinks, Tags, Users, Votes, Batches, and Comments. The Posts table consists of an attribute named PostTypeId which is 1 if the Post is a question and 2 if the Post is an answer to the question.

Exploratory Data Analysis (EDA): This stage will involve summarizing the files contained in the SO dataset using a visual method to discover patterns or anomalies which will guide us in the ~~other processes.~~

Feature Engineering: We will use data mining techniques to extract relevant features from the SO dataset. ~~We have to~~ determine features that will be relevant for recommending experts for questions. Then we will extract the features and pre-processing them to make them suitable for our machine learning methods. ~~For example, we might not need all the attributes in different files of the SO dataset.~~

Benchmarking Machine Learning Models: We aim to benchmark the following machine learning methods: *Naive Bayes, Support Vector Machine, Logistic Regression, Decision Tree, and Random Forest* for expert recommendation using StackOverflow dataset. We will asses the impact of feature engineering on the different methods and compare their performance base on their accuracy to determine the effectiveness of the methods.

Results Analysis: ~~We will present the results of our benchmarking machine learning methods for an expert recommendation and explore the circumstances in which these methods work.~~

Note: Link for StackOverflow data-dump - https://archive.org/details/stackexchange