Proposal for Capstone 2 Project

Title: Using text data from Stack Exchange sites to identify keywords, topics and summaries

Problem: Stack Exchange and its associated websites provide a massive platform for sharing technical knowledge and collaboration on projects. Encompassing actively used sites such as Stack Overflow, Super Users and Ask Ubuntu, the Stack Exchange environment is a network for questions and answers across a wide range of topics. These forums serve as hubs for basic as well as complex queries providing a rich collection of technical text data. Such information often covers important topics and the latest trends in various industries. A way to summarize this text as well as identify topics in the data could help search engines sort user queries and provide quicker and more accurate results. It could also help other similar forums categorize future questions in an automated manner where users can more easily find solutions to their questions. In this project, we propose topic modeling and clustering techniques to analyze this text data and identify keywords related to each question.

Audience: Stack Exchange as well as similar platforms such as Quora, Reddit, LinkedIn groups, Facebook groups etc, could use this model or even adopt the model to data on their respective platforms to categorize text data. Since we are dealing with text data which is platform agnostic, this model and analysis can be extended to other sites and search engines such as Google and Bing. The analysis can also be applied to any sort of text data such as comments, reviews, item descriptions in platforms such as Yelp, Amazon, Twitter etc.

Data: The dataset is acquired from Kaggle and consists of questions from Stack Exchange users, consisting of 420668 rows. Each instance contains a column that has the question title and another column containing the body of the question which could consist of text as well as code. The text in the body is in HTML format which needs to be processed to extract only the meaningful text. There is also a target variable which contains the topics related to each document, serving as the target variable. The original kaggle competition is structured to predict these topics/labels. However, for our application, we can even ignore the target variable and conduct our own analysis on simply the title and the question treating the data as unstructured.

Analysis and Modeling: Clustering, topic modeling and summarization techniques can be used to achieve this task. Additionally, methods as Latent Dirichlet Allocation (LDA), bag of words and deep learning techniques such as Recurrent Neural Networks (RNN) could be used to help with the modeling.

Deliverables:

- Jupyter notebooks (code):
    - Data acquisition, cleaning and wrangling
    - Exploratory data analysis (EDA)
    - Text processing
    - Modeling and evaluation
- Project report
- Presentation slides