Exploring What Makes a High Quality Stack Overflow Question

Motivation

We would like to know what makes a Stack Overflow post "high quality". With Stack Overflow being widely utilized by students and professionals alike, we believe it could be beneficial to know of any specific characteristics that will lead to questions being answered.

Goal

Our goal to find whether or not there are key characterstics that lead to a highly ranked Stack Overflow post. If theses characteristics exist, we would like to be able to define them in a way that could be utilized by any question poster. If we can't find any characterstics, we would like to be able to definitively say that they don't exist.

Proposed Method

We would like to do some natural language processing on the titles, bodies, and possibly tags of the Stack Overflow questions in the "60k Stack Overflow Questions with Quality Rating" dataset. This dataset divides the questions into one of three groups: "HQ: High-quality posts without a single edit", "LQ_EDIT: Low-quality posts with a negative score and multiple community edits. However, they still remain open after those changes.", and "LQ_CLOSE: Low-quality posts that were closed by they community without a single edit." We would like to specifically see what the shared characteristics are between the HQ ranked questions.

Evaluation Expected

We will do some comparisons between the bodies of the questions to see if there are any patterns there. We'll also do the same with the titles. From there we would also like to see if either the titles or the bodies are a better predictor of a high quality post. We will utilize a plethora of models for evaluating the Stack Overflow dataset via PyCaret along with other modeling types, such as Neural Networks via PyTorch.

Division of Roles

- Teancum Price:
    - Gathering data and determining what models to use for classification.
- Reagan Hill:
    - Evaluating models we train and test on data.
- Dillon Johnson:
    - Testing and training the models.