Udacity MLE Nanodegree - Capstone Proposal

Zhaoyi Hou (Joey)

Background

As a critical component of the developer community, Stack Overflow detailedly documented countless communication between developers, in particular, the question-answering activities. Not only do the "QnA" solve a critical problem the developer had, but it also paved the way for all the future developers to save their time on debugging and, therefore, to make a bigger impact with their code. From this perspective, making sure that the questions asked on Stack Overflow have high quality becomes a significant matter: if there were too many low-quality questions, developers would be overwhelmed and find it difficult to navigate a clear error case that they have. If there were a computer software or model that can comprehend the quality of a given question before it was posted, the low-quality questions would be reduced.

In this field, multiple researchers have made contributions in the past. For example, a group of researchers in Arizona State University have conducted a research back in 2017[1]. In this research, they built a machine learning with linguistic-based features and achieved around 60% accuracy on a 4-label prediction task. Based on this example, I am positive that this particular task is deliverable using machine learning techniques.

Problem Statement

In this project, the expected input data is the Stack Overflow question text and the output is the quality of the question, with three possible answers: "HQ"(high quality), "LQ_EDIT" (low quality but remains open after edits), "LQ_CLOSE" (directly closed without edits). In other words, this is a classification task with a multi-label target. In a broader sense, the solution to this particular problem can be applied to other learning communities (e.g. Piazza) or question answering communities (e.g. Reddit) to improve the overall quality of the discussion.

[1] Assessing Question Quality using NLP (2017)

Dataset

- The original data source: [60k Stack Overflow Questions with Quality Rating]
- Data
  - Size, Class distribution, etc.

	Size	Class Distribution
Train	45,000	1:1:1
Test	15,000	1:1:1
All	60,000	1:1:1

  - Attributes

Attribute	Id	Title	Body	Tags	CreationDate	Y
Type	Int	String	String (HTML Text)	String	String	String

  - Sample Data

```
train.head(2)
```

	Id	Title	Body	Tags	CreationDate	Y
0	34552656	Java: Repeat Task Every Random Seconds	<p>I'm already familiar with repeating tasks e...	<java> <repeat>	2016-01-01 00:21:59	LQ_CLOSE
1	34553034	Why are Java Optionals immutable?	<p>I'd like to understand why Java 8 Optionals...	<java> <optional>	2016-01-01 02:03:20	HQ

  - Data Summary

    As we can see from the basic analysis, the data I found is clean enough to be directly handed to the feature engineering part. Also, since the class (label) distribution is balanced across three classes, it would be appropriate to use simple accuracy for evaluation.

Solution Statement

The proposed solution to this problem is to train a machine learning model using the questions with human classification (the training set). The expected output should be a functional model that takes in a question text and predict whether or not the question is of high quality. In order to be successful in this prediction, my target accuracy for this classification is 85%.

Benchmark Model

Simple linear classification based on the length (the number of words) of the question.

Evaluation Metrics

Since the problem is a multi-label classification problem, I will use the exact match ratio (subset accuracy) as my evaluation criterion. More details can be found: sklearn.metrics.accuracy_score — scikit-learn 0.24.2 documentation

Project Design

My proposed solution to this problem consists of the following steps:

- Step 1: Data cleaning, preprocessing, and train-validation-test splitting
- Step 2: Feature Engineering
- Step 3: Build two different models
    - Model #1: Text only, LSTM network
    - Model #2: Feature only, traditional machine learning model
- Step 4: Hyperparameter tuning for both models
- Step 5: Model evaluations

Work Cited

1. Kristopher J. Kopp, Amy M. Johnson, Scott A. Crossley, and Danielle S. McNamara, Assessing Question Quality using NLP (2017), 18th International Conference on Artificial Intelligence in Education (pp. 523-527). Wuhan, China