Machine Learning Nanodegree
Capstone Proposal

Nujood Ahmed

January 9, 2019

Quora Duplicate Question Detection

(Project was inspired by Quora's Kaggle competition[1])

1. Domain Background

Quora is a place to gain and share knowledge—about anything. It's a platform to ask questions and connect with people who contribute unique insights and quality answers. This empowers people to learn from each other and to better understand the world.

Over 100 million people visit Quora every month, so it's no surprise that many people ask similarly worded questions. Multiple questions with the same intent can cause seekers to spend more time finding the best answer to their question, and make writers feel they need to answer multiple versions of the same question. Quora values canonical questions because they provide a better experience to active seekers and writers, and offer more value to both of these groups in the long term.

The main goal of this project is to build a high-quality knowledge base, it's important that we ensure each unique question exists on Quora only once. Writers shouldn't have to write the same answer to multiple versions of the same question, and readers should be able to find a single canonical page with the question they're looking for.[2]

An amazing research paper called: Detecting Misflagged Duplicate Questions in Community Question-Answering Archives[3] 'helped me to learn and gain more information about the subject as a whole and about the best practices and methodologies.

2. Problem Statment

most of the users don't check if their question has already been asked by someone else, which will lead to duplicate questions posts. This necessitates having a prediction system that can check if a question is duplicated. Before the user can post a question, the system will suggest an existing one.
So, the problem can stated as follow: predict whether a pair of questions are duplicates or not.

this will be tackled as a **Natural Language Processing** problem.

3. Datasets and Inputs

The dataset is provided by Quora on Kaggle page: https://www.kaggle.com/c/quora-question-pairs/data

The data is splitted into train and test set, the train set contain 404290 row and 6 columns while the test set contain 2345796 row and 3 columns. The train set contains the following fields:

Variable	Description	Type of Variable
id	the id of a training set question pair	Discrete Numerical
qid1	unique id of the first question in the question pair	Discrete Numerical
qid2	unique id of the second question in the question pair	Discrete Numerical
question1	full text of the first question	Text Variable
question2	full text of the second question	Text Variable
is_duplicate	the target variable, set to 1 if question1 and question2 have essentially the same meaning, and otherwise.	Categorical

For the **test set**, it includes the following fields: test_id, question1 and question2 only.
The target will be **is_duplicated**, and the other columns are the features/independent values. Any outliers or missing values will be detected and handled during the EDA phase.

4. Solution Statement

To tackle the problem described in section 2, I will be using Neural Networks to create the model since it's strong and popular in the field of NLP.

Before building the mode, I will do exploratory data analysis to understand the data. the analysis phase will include some practices like performing feature extraction and feature engineering and any other useful processes.

5. Benchmark Model

Before creating a complicated model I'll take a simple approach. for the benchmark model I'll create two predictors: percentage of words from question1 that appear in the question2 and vice-versa. Then I will use Random Forest to predict if the questions are duplicate.

6. Evaluation Metrics

The result will be evaluated using **log loss (or cross-entropy loss in binary classfication)**:

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{i,j}\log(p_{i,j})$$

7. Project Design

Before start training the model, I will first take a glimpse at the data see the shape of it, how it formatted. I will do some exploratory data analysis to detect outliers and impute missing values, and perform visualizations as well for better understanding of the data distribution. Then I will start doing my natural language processing and extract information such as character counts, sentence length, TF-IDF vector..etc. After that I will build the Network (try multiple networks with different hyper-parameters). Lastly, the model will be compared with the benchmark model.

8. References

[1]. https://www.kaggle.com/c/quora-question-pairs

[2]. quora. (2017). Semantic Question Matching with Deep Learning. [online] Available at: https://engineering.quora.com/Semantic-Question-Matching-with-Deep-Learning [Accessed 9 Jan. 2019].

[3]. Doris Hoogeveen, Andrew Bennett, Yitong Li, Karin M Verspoor and Timothy Baldwin. “Detecting Misflagged Duplicate Questions in Community Question-Answering Archives”, Australia, 2018.