Machine Learning Engineer Nanodegree

Capstone Proposal

Brett Clubb
February 18, 2018

Proposal

Domain Background

Natural Language Processing (NLP) is an area of active research and development in which a computer analyzes a set of text for the purpose of achieving human-like language processing for a range of tasks or applications. A common goal of NLP systems is to represent the true meaning and intent of an inquiry. [7]

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. [1]

I chose this project to better understand Natural Language Processing and ways it can be utilized. I am involved in another project where Customers are interacting with a simple chat bot and asking simple IT Support related questions. I hope to apply some of the concepts learned from this project to programmatically understand which questions are being asked most often to identify problem areas and perhaps add more specialized bot capabilities to better assist Customers.

Problem Statement

This project aims to solve a binary classification problem to predict whether or not a provided pair of questions have the same or similar meaning through the use of natural language processing. By determining whether or not question pairs are duplicates, we can improve the experience in finding high quality answers for Quora writers, seekers, and reader by providing a single page for questions with a similar intent. [1]

Datasets and Inputs

The Quora dataset provides a set of question pairs and indicates whether or not the questions are requesting the same set of information. The data is provided by Quora via Kaggle's competition website.

Data Fields (Inputs):

- id - the id of a training set question pair
- qid1, qid2 - unique ids of each question (only available in train.csv)

- question1, question2 - the full text of each question
- is_duplicate - the target variable, set to 1 if question1 and question2 have essentially the same meaning, and 0 otherwise.

The dataset is split into two files, one for training and one for testing. [2]

Training

- Question Pairs: 404,290
- Questions: 537,933
- Ratio of Duplicate Pairs: 36.92%

> **Sample**
>
> | id | qid1 | qid2 | question1 | question2 | is_duplicate |
> |---|---|---|---|---|---|
> | 0 | 1 | 2 | "What is the step by step guide to invest in share market in india?" | "What is the step by step guide to invest in share market?" | 0 |
> | 1 | 3 | 4 | "What is the story of Kohinoor (Koh-i-Noor) Diamond?" | "What would happen if the Indian government stole the Kohinoor (Koh-i-Noor) diamond back?" | 0 |
> | 2 | 5 | 6 | "How can I increase the speed of my internet connection while using a VPN?" | "How can Internet speed be increased by hacking through DNS?" | 0 |

Test

- Question Pairs: 2,345,796

> **Sample**
>
> | test_id | question1 | question2 |
> |---|---|---|
> | 0 | "How does the Surface Pro himself 4 compare with iPad Pro?" | "Why did Microsoft choose core m3 and not core i3 home Surface Pro 4?" |
> | 1 | "Should I have a hair transplant at age 24? How much would it cost?" | "How much cost does hair transplant require?" |
> | 2 | "What but is the best way to send money from China to the US?" | "What you send money to China?" |

Solution Statement

The solution will predict whether or not a question pair is considered a duplicate by setting the target variable, is_duplicate, to 1 if question1 and question2 have the same meaning and 0 otherwise. To tackle this problem, I will first use the TF-IDF (Term Frequency, Inverse Document Frequency) to determine the importance of words in the questions based on how frequently those words appear across all provided questions. [5] I will then extract features from the dataset to train a binary classifier to differentiate between duplicate and non-duplicate question pairs.

Benchmark Model

Quora has noted that they currently use a Random Forest model to identify questions. [1] Because the goal of the competition is to improve on the current implementation, I will utilize a Random Forest Classifier from sklearn to fit a Raondom Forest model to use as a benchmark.

Evaluation Metrics

As defined by Quora in the Kaggle competition, I will evaluate the model on the log loss between the predicted values and the ground truth. For each ID in the test set, there should be a prediction on the probability that the questions are duplicates (a number between 0 and 1). [3] The goal of this model will be to minimize this value, since log loss increases as the predicted probability diverges from the actual label. [4]

Log Loss is defined as:

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

where N is the number of samples or instances, M is the number of possible labels, $y_{ij}y_{ij}$ is a binary indicator of whether or not label j is the correct classification for instance i, and $p_{ij}p_{ij}$ is the model probability of assigning label j to instance i. A perfect classifier would have a Log Loss of precisely zero. Less ideal classifiers have progressively larger values of Log Loss. [6]

Project Design

I will start by doing some initial inspection of the data to better understand the dataset. From there, I will begin an exploratory data analysis to summarize the data, using visualizations where they make sense.

After gathering these details, I should have a good enough understanding of the dataset to move forward. I will begin with using TF-IDF to provide some meaning to the words. This will weigh the words based on how common they are across all questions. Extremely common words will carry less importance since they are shared and are not good indicators of a question's intent.

I will then begin to train and select a model. Because this is a binary classification problem, there are several methods available (Decision Trees, Random Forests, Bayesian Networks, Support Vector Machines). I will attempt to implement at least 3 of these, then use cross-validation to determine which model performs best and fine-tune its parameters.

Finally, I will evaluate the model trained using the LogLoss function to determine how well the model performs.

References

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

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

[3] https://www.kaggle.com/c/quora-question-pairs#evaluation

[4] http://wiki.fast.ai/index.php/Log_Loss

[5] https://stevenloria.com/tf-idf/

[6] http://www.exegetic.biz/blog/2015/12/making-sense-logarithmic-loss/

[7] Liddy, E.D. 2001. Natural Language Processing. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc.