**CS794-CS892 Final Year Project Synopsis**

|  |  |
| --- | --- |
| **Team Number** | 1 |
| **Group Members** | Sudipta Halder(10400116059) , Smita Bandopadhyay(10400116078) |
| **Project Mentor** | Amit Kumar Das |
| **Project Title** | **Quora Question Pairs**  Can you identify question pairs that have the same intent? |

**Objective**

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. so main aim of project is that predicting whether pair of questions are similar or not. This could be useful to instantly provide answers to questions that have already been answered. Credits: Kaggle

**Scope**

* This dataset contains various possible question pairs asked in Quora and whether they are duplicate or not.
* Real world/Business Objectives and Constraints:
* The cost of a mis-classification can be very high.
* You would want a probability of a pair of questions to be duplicates so that you can choose any threshold of choice.
* No strict latency concerns i.e. if a new question is posted in Quora, it can take upto few minutes to find whether it is duplicate with any question or not.
* Interpretability is partially important.
* At first, we will do some basic feature extraction like word\_common, word\_share, frequency\_qid etc. which follows text pre processing by removing html tags, punctuations, performing stemming, expanding contractions etc.
* Then we need to do some advanced feature extraction like fuzz\_ratio, fuzz\_partial\_ratio, token\_sort\_ratio, first\_word\_equivalent, longest\_common\_substring\_ratio etc.
* In addition to these basic and advanced features we will create a bag of words model using NLP and therefore we will have near about 350+ features all total.
* Based on these 350+ features(xi) and is\_duplicate\_value (0/1) (yi) we will build a binary classification model.
* We need to build several models and check with the help of performance metric which model is giving us better accuracy.

**Plan**

|  |  |
| --- | --- |
| **Task Description** | **Planned completion date** |
| Data consolidation and finalise project plan | August 15, 2019 |
| Data cleaning and exploration | September 21, 2019 |
| Feature extraction (basic & advanced) | November 20, 2019 |
| Creating Bag of Words model | December 5, 2019 |
| Development of initial data analysis and prediction model | January 31, 2020 |
| Completion and finalisation of data visualization and prediction model | March 30, 2020 |
| Finalise project report and presentation | April 20, 2020 |

**References**

Site for Quora Question PairsDatabase source: <https://www.kaggle.com/c/quora-question-pairs/data>

|  |  |
| --- | --- |
| **Mentor’s Endorsement**  **Signature and Date** |  |