Lambton College | In Toronto

Natural Language Processing

(AML 2204)

PROJECT PROPOSAL

Quora Duplicate Question Classifier

Team Members

Anto Francis (C0825095)
Omer Volkan (C0831373)
Rupesh Chandran (C0826779)
Sachin Sreekumar (C0825096)

Submitted to
Prof. Harriet Huang

Our Motivation

Quora is a question-answering platform that lets users ask questions and get answers on them. It connects users having the same problems and allows them to share knowledge with the public. This allows people to get knowledge quickly which makes their life easy. The most common issue which has been faced by Quora users is question duplication. Sometimes users ask similar questions which have been answered before which results in question duplication. This makes the writers feel like they have to answer the same questions multiple times which reduces quora's experience.

Our aim is to resolve this problem by applying advanced NLP techniques to classify whether questions are repeated or not. This allows users to find good-quality answers easily.

Source of Data

We are taking the dataset from Kaggle which has more than 800 question IDs and question pairs. We will be using different mechanisms to identify whether questions are similar or not. We will be analyzing common words, first and last words, and fuzz ratios for our analysis. We will also be using cosine similarity and Euclidean distance to see the similarity.

Finally, we will be using SVC and Random Forest Classifiers to classify duplicate questions.

Team Member Responsibilities

S. No	Module	Team Member
1	Data Cleaning ● Case conversion ● Handling symbols ● Handling URLs ● Removing HTML tags ● lemmatizing	Anto Francis
2	Feature Extraction ● Letter count ● Word count ● Fuzz ratio calculation	Rupesh Chandra
3	Finding distance between vectors ● Cosine similarity ● Euclidean vectors	Sachin Sreekumar
4	Creating classifications models ● SVC ● Random Forest Classifier ● Accuracy score calculation	Omer Volkan Guney

References

- *Quora question pairs*. Kaggle. (n.d.). Retrieved March 22, 2022, from https://www.kaggle.com/c/quora-question-pairs/data