NLP Project: Document similarity (Quora question pair similarity)

Overview

Quora is a dynamic platform where users can ask questions and receive answers from the community. One of the challenges is identifying whether two different questions are semantically similar (duplicate) or not. This project aims to classify pairs of questions as duplicates or non-duplicates, contributing to the quality of content and user experience on Quora.

Objective

Classify pairs of questions to determine whether they are duplicates, improving content relevance and minimizing redundancy on Quora.

Submission Guidelines

- Code Repository: Share a GitHub repository containing your code.
- Final Report: Submit a PDF report detailing your approach, methodology, results, and conclusions.
- Presentation: Prepare a set of slides summarizing your project for a presentation.

Guidelines for presentation

- Provide a brief overview of the project, mention the dataset and key challenges
- Describe your approach to data preprocessing, feature engineering, and model selection. Summarize your modeling techniques and justify your choices.
- Present performance metrics (e.g., accuracy, F1-score) with relevant comparisons.
- Discuss how your solution can be scaled to larger datasets
- Discuss a potential cloud deployment plan
- Reflect on what worked well and what could be improved