THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Group: 5
Group Members: Abhradeep Das (G38747350)
Gourab Mukherjee (G32241729)
Richik Ghosh (G31267506)
Shreya Sahay (G36642286)
Final Project: Natural Language Processing, Fall 2024
Date: 10/29/2024

Project Proposal

Problem Statement:

The problem involves identifying duplicate question pairs on Quora, where questions with the same intent but different phrasing create redundancy. This redundancy makes it harder for users to find relevant answers, impacting engagement quality. Reducing duplicates would improve user experience by streamlining access to canonical answers, benefiting both question seekers and contributors. This task has direct implications for Quora's content organization, search optimization, and user satisfaction.

Dataset Selection:

We will use the Quora Question Pairs dataset, which was made available as part of a Kaggle competition in 2017. This dataset comprises approximately 404,290 question pairs, each labeled to indicate whether the questions are duplicates, with additional test data to mitigate overfitting. This dataset is ideal for machine learning due to its size, accurate labels, and relevance to this specific problem.

NLP Methods:

For identifying duplicate questions, we plan to apply several NLP techniques and machine learning models, with a focus on feature engineering and classification:
1. ***Pre-processing and Feature Extraction:*** Techniques including tokenization, stemming, and stopword removal will be used to process and prepare the text.
2. ***Fuzzy Matching and Similarity Measures:*** Metrics among common word ratio, longest common subsequence, and Jaccard similarity will help us quantify the similarity between question pairs.
3. *Classical Machine Learning Models* will be applied and evaluated for their effectiveness in identifying duplicate vs. non-duplicate pairs. If feasible, we may experiment with deep learning methods to handle more nuanced question variations.

Packages and Tools:

Key Python libraries for efficient data processing, feature engineering, model training, and result visualization in our project will include:
- ***NLTK*** and ***spaCy*** for text preprocessing,
- ***scikit-learn*** for model building,
- ***fuzzywuzzy*** for string similarity calculations,
- ***Pandas*** for data manipulation
- ***Matplotlib/Seaborn*** for data visualization.

Performance Evaluation and Metrics:

To evaluate model performance, we will focus mostly on:
- **Log-Loss:** Suitable for classification tasks and provides insights on model calibration.
- **Confusion Matrix:** Offers a clear breakdown of model accuracy, highlighting true positives, false positives, true negatives, and false negatives.

We will also use visualizations of the output classes' distribution to ensure a balanced and generalized model.

Project Timeline:

- **Week 1**: Initial data exploration, preprocessing, and visualization of dataset features.
- **Week 2**: Implement feature engineering, select models, and conduct initial training.
- **Week 3**: Refine feature engineering and optimize model selection through evaluations.
- **Week 4**: Finalize the model, conduct thorough evaluations, and document findings for submission.

GitHub: https://github.com/abhradeepd/NLP-Final-Project-Group-5
Dataset: https://www.kaggle.com/c/quora-question-pairs