Lab Assignment 3.2: Create Pre-Trained BERT Vectors

Review the rubric before completing this assignment.

You will use the Yelp restaurant review data, which includes reviews for 1075 restaurants (the same dataset used in Lab Assignments 2 & 3).

For this assignment, you will complete the following tasks:

Section 1

Create a new dataframe with Base BERT vectors as features (refer to Lab 3.6).
Please ensure you:

- Investigate the token lengths for each review. Identify and list reviews that exceed Base BERT's token limits.
- For reviews longer than 512 tokens, truncate them.
- Your final dataframe should include columns for Cust_Rating, Datetime, Review (Original text), Restaurant, City, State, ZipCode, Business_Rating_Score, and Base BERT Vectors.
- Include a snippet (screenshot or a table) of the first 5 reviews with new BERT vectors from your dataframe in the assignment submission document.

Section 2

Create a new dataframe with Longformer vectors as features (refer to Lab 3.6).
Ensure your dataframe includes:

- Cust_Rating, Datetime, Review (Original text), Restaurant, City, State, ZipCode, Business_Rating_Score, and Longform Vectors.
- Include a snippet (screenshot or a table) of the first 5 reviews with new BERT vectors from your dataframe in the assignment submission document.

Section 3

Create a new dataframe with Base BERT vectors as features (refer to Lab 3.7) by splitting the long text into smaller chunks that fit within the token limits.
Make sure to:

- Include Cust_Rating, Datetime, Review (Original text), Restaurant, City, State, ZipCode, Business_Rating_Score, and BERT Vectors in your final dataframe.
- Include a snippet (screenshot or a table) of the first 5 reviews with new vectors from your dataframe in the assignment submission document.

Section 4

Create a document BERT vector embeddings for each restaurant (aggregate the reviews for each restaurant) and store them in a new dataframe (your final dataframe should contain 1075 rows). Each document includes all reviews related to a specific business (restaurant). You should:

- Start with the per-review vectors you created in Sections 1-3 using the pre-trained BERT models.
- Generate new embeddings by averaging all the per-review BERT embeddings for each business.
- Include a snippet (screenshot or a table) of the first 5 rows with new document vectors of your dataframe, along with the original metadata, in the assignment submission document.

Note: While not required, it is recommended that you save each section's dataframe as a .csv for later use in document classification tasks.

This assignment is due by 11:59 p.m. Pacific Time, the evening after the Module 8 class session.

Rubrics:

1. Token length review : Correct implementation of token lengths for each review and identify reviews that are longer than the 512 base bert token length limit.

2. Pre-trained base BERT as features : Python codes work correctly. The first five rows of vectors are provided in the submission file.

3. Pre-trained long former as Features: Python codes work correctly and produce correct results. The first five rows of vectors are provided in the submission file.

4. Based Bert features without truncating: Python codes work correctly and produce correct results. The first five rows of the vectors are provided in the submission file.

5. Bert document vectors : Python codes work correctly and produce correct results. The shape (e.g., row & columns) of the dataframe is correct. The first five rows of the new dataframe are provided in the submission file.

6. Code readability : Code is clean, understandable, and well-organized

7. Documentation: We can figure out what your program does just by reading your comments.