Assignment 8.1

Problem Statement

Stack overflow tags prediction using Deep Sequence Modeling.

Data:

https://www.kaggle.com/datasets/stackoverflow/stackoverflow

Create a Kaggle notebook in this dataset to access Big Query data.

Task:

1. Using the bigquery-public-data.stackoverflow.stackoverflow_posts dataset, select first 10000 rows of id, title, and tags such that length of tags should be less than 20 and it should have python, r, c#, java, android, html, kotlin, c, and c++.
2. Preprocess the title, remove stopwords, hyperlinks, punctuations except # and +.
3. Convert tags string column into list column.
4. Apply multilabel binarizer on the tags.
5. Tokenize, and apply pad sequence on the title with maxlen as the longest title.
6. Create a sequence model using embedding with dimensions 50.
    a. Add LSTM, dropout and Batch normalization layers.
7. Display few title samples with actual and predicted tags.

Assignment 8.2

Problem Statement

Write a report on BERT and how to use it in python with code snippets in the document.

Note: Plagiarism will be checked.