Readme

Akhter Al Amin

December 9, 2021

1 Readme.md

To run this the code, user only need to know about three files:

1. wimp_corpus.zip: Please unzip this folder and place it in the same file directory of notebook. The annotated word importance score has been extracted from this folder.

2. switchboard_word_alignments.tar.gz: Please extract the transcripts folder named 'swb_ms98_transcriptions' from this file. We will read the transcriptions from this folder.

3. Data.csv: This file contains contextualized word embedding of 11,000 words and their corresponding annotated weights. Number of columns each row have is 769 wherein we have 768 features that is a composed version of 3 dimensional BERT-generated embedding. This file has been created by a separate code segment that has been rescinded from this final version of the project to keep notebook understandable and easily executable.

Note that, to run these codes, you have to keep the above files in the same file directory of notebook: 'Word_Importance_Model_Final.ipynb' .

In each 'Performance Measure' code segment, since we are analyzing a total 44 documents, you might observe the progress like this: 1/44, 2/44.

While running this code you might encounter warning in several output logs, please ignore those. Here is the code segments where you may get such warning: "Importing Libraries", "Performance Measure" within this section "Incorporating POS with Word Embedding in generating Word Importance ", "Context-aware Term Weight Determination Using BERT".

Please run the code sequentially. For some analysis, it might take some time, specially in the word embedding segment.