








Original data source:

https://www.kaggle.com/stefanoleone992/rotten-tomatoes-movies-and-critic-reviews-dataset?select=rotten_tomatoes_critic_reviews.csv

You do not need to download this; I already preprocessed the data.

Scripts:

-  0-Data Preparation.ipynb
-  1-CNN-W2V.ipynb
-  2-BERT.ipynb
-  3-CNN&Bert.ipynb
-  4-CNN-BertWordEmbeddings.ipynb
-  5-inferSent.ipynb
-  6-BOW Logistic Regression.ipynb

All those files are best viewed inside Google Colab



Scripts 0 will generate the preprocessed data for the other 6 scripts.

Each script is a model following the same order in the report.

To run 1-CNN-W2V.ipynb and 3-CNN&Bert.ipynb, you need to first create a file directory 'w2c/models' in your working directory

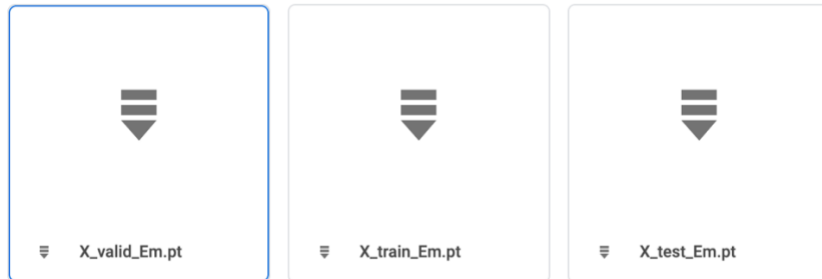
Data:

..

-  InferSent-master
-  movie_review_RT50K.csv

“movie-review-RT50K.csv” is the preprocessed data for all the scripts, load it inside your working directory

For “5-inferSent.ipynb”, you need to load all the content inside “inferSent-master” to your working directory, since it is a sentence encoder pre-trained. You also need to download the three already generated embeddings shown below and add into your working directory from <https://drive.google.com/drive/folders/1Ayt9xogFqAyMBsdQMhkM5m0BwFi6N3pw?usp=sharing>



Otherwise, you would have to follow the step 2-4 inside the scripts to download Glove and encode the sentence, which takes 3 hours.

Note: most of the time inside the scripts, I load data form Google Drive, you might to change the data path to your working directory. The instructions will be inside each script.