Task 2 Description

For task 2, I performed the following actions. First of all, I loaded training, validation, test and dialogsum_clustered csv's and converted them into dataframe. Then, I checked whether we had any missing data, and we did not. Then I merged train, test and validation csv's with dialogum_clustered csv to have clusters in each. After I made sure that we are using only the first ones of summary or dialogue in the test csv. I used TF-IDF (Term frequency-inverse document frequency) to convert data into numerical features and see the importance of words bigrams in the context. I tried using alternatives such as Word2Vec, however it gave a lower weighted F1 score.

Further, after trying dozens of algorithms, I preferred to use SGD (Stochastic Gradient Descent) classifier with adjusted regularization and with cross-validation evaluation, because it was giving the best weighted F1 score and was running the fastest. Perhaps there are some other classifiers that would give a better score (I'd like to see the score of Quadratic Linear Regression as it could be a good one for this problem), but after trying many of them, if they were running too long (over minutes) I did not consider them. So, after fitting the model on the train, I got the cross-validated F1 scores, weighted F1 score for validation, and after I got the weighted F1 score on test set.