Qualcomm Problem Statement – Additional Details

I/O and constraints:

Since, this is an open domain problem you have to find/create your own dataset and formulize your approach in a python script. We will test your approach using our test data.

You'll have to create a python script which will take the following inputs:

1) Abstract csv with N abstracts
2) Full text csv with M full text articles

And your python script should return/create another similarity_matrix.csv which will have N rows and M columns where each cell (I,j) represents a similarity score between the abstract **I** and article **j**.

Each similarity score should be between [0,1] 1 being most similar and 0 being least similar.
You are free to use any mathematical/machine learning library.
Avoid using external pre-trained word embedding models like Word2Vec or GloVe.

Submission:

You will have to submit 3 files to be considered a valid submission:

1) Your Python Script
2) Requirements.txt which contains all the external libraries list which are required to run your script
3) README.txt which contains details about how to run your script and a brief description about your approach.

[Optional] You can also submit any Deep Learning/Machine Learning model you trained if any. If you trained any models, please provide the details about their training in README.txt.

Evaluation:

We will run your python script on our two test datasets:

1) Closed domain dataset. First dataset will contain summaries and articles from a single domain (like 5G, Neural Networks, IoT, etc.)
2) Cross domain dataset. Second dataset will contain summaries and articles across multiple domains.

We will evaluate your similarity matrices for above two datasets against our test matrices.
There will be 3 factors considered while evaluating the final results (in decreasing order of priority) :

1) Accuracy of the similarity measure
2) Novelty of approach
3) Inference time