FILES PROVIDED:

1. cross_validate_bow.py
2. cross_validate_tfidf.py
3. data_set_200.csv
4. data_set_200_creation.py
5. data_set_all_not_shuffled.csv
6. data_set_all_not_shuffled.py
7. shuffled_data.csv
8. shuffled_data_creation.py

WHAT EACH FILE DOES –

1. **data_set_200_creation.py**: Creates data set of 200 docs from each book with 150 sentences each while processing the data (Data preparation and preprocessing).
   Output: data_set_200.csv

2. **data_set_all_not_shuffled.py:** Creates data set of multiple docs from each book with 150 sentences each while processing the data (Data preparation and preprocessing).
   Output: data_set_all_not_shuffled.csv

3. **shuffled_data_creation.py:** Creates data set of multiple docs from each book with 150 sentences each while processing the data. It also shuffles the words of each sentence while creating the sentence column (Data preparation and preprocessing).
4. **Output: shuffled_data.csv**

Steps:

1. **Start with creating the data sets using any of the three scripts:**
   data_set_200_creation.py (default data set used)
   data_set_all_not_shuffled.py
   shuffled_data_creation.py

2. **Run** cross_validate_bow.py to get results for Cross validation with Bag of words.
3. **Run** cross_validate_tfidf.py to get results for Cross validation with TF_IDF.