Readme

Instructions:

- You can run train any model using ipynb file as it is.
- If you want to only test the model then use the evaluation script. change the only model name in the evaluation.ipynb file.

Dataset Creation Files:

- Translation using *mbart-large-50-one-to-many-translation model*. (https://huggingface.co/facebook/mbart-large-50-one-to-many-mmt)
  File Name: Translation_code/*squad_traslated_using_mbart.ipynb*
- Translation using *googletrans*
  File Name: Translation_code/*tydiqa_traslated_using_googletrans.ipynb*
- Extracted telugu and bengali translations from ai4bharat/IndicQuestionGeneration:
  File Name: Translation_code/*squad-context-only.ipynb*
- Extracted question translation from ai4bharat/IndicQuestionGeneration:
  File Name: Translation_code/*squad-que-translate.ipynb*

Datasets:

1) tydiqa
2) squad (only english)
3) squad_mbart_translated: 8139 questions of the squad dataset are translated into telugu and bengali language using the mbart model.
   (https://huggingface.co/datasets/Gautam9595/Squad_Translated)
4) squad_ben_tel_context: Keep only those question-answer pairs in which the answer is in context.
   (https://huggingface.co/datasets/krinal214/squad_ben_tel_context)
5) squad_ben_tel_que_only: Used completely translated dataset of squad into bengali and telugu, then replace the translated questions with original questions in the original dataset.
   (https://huggingface.co/datasets/krinal214/squad_ben_tel_que_only)
6) tydiqa_ben_tel: 20k questions are translated into bengali and telugu each from different languages of the tydiqa dataset using googletrans.
   (https://huggingface.co/datasets/krinal214/tydiqa_ben_tel)

Base Case:

Filename: bert-all-tydiqa-only.ipynb
Dataset: tydiqa
Additional Dataset: None
Best Model: *bert-base-multilingual-cased*
Model: *krinal214/bert-all*
(link: https://huggingface.co/krinal214/bert-all)

Additional Cases:

1)

Filename: bert-all-tydiqa+squad.ipynb
Dataset: tydiqa
Additional Dataset: squad (english only - original)
Model: *krinal214/augmented*
(link: https://huggingface.co/krinal214/augmented)

2)

Filename: bert-all-tydiqa+tydiqa_que-translation.ipynb
Dataset: tydiqa
Additional Dataset: tydiqa_que-translation (20k questions of tydiqa are translated into both bengali and telugu using googletrans)
Model: *krinal214/bert-all-translated*
(link: https://huggingface.co/krinal214/bert-all-translated)

3)

Filename: bert-all-tydiqa+squad_translated_using_mbart.ipynb
Dataset: tydiqa
Additional Dataset: Translated Squad (Only questions are translated into telugu and bengali using mbart)
Model: *krinal214/augmented_Squad_Translated*
(link: https://huggingface.co/krinal214/augmented_Squad_Translated)

4)

Filename: bert-all-tydiqa+squad+squad_que.ipynb
Dataset: tydiqa
Additional Dataset: Squad (english only - original) + Translated Squad (Only questions are translated into telugu and bengali)
Best Model: *krinal214/bert-all-squad_que_translated*
(link: https://huggingface.co/krinal214/bert-all-squad_que_translated)

5)

Filename: bert-all-tydiqa+squad+squad_context.ipynb
Dataset: tydiqa
Additional Dataset: Squad (english only - original) + Translated Squad (Used translated data of squad from hugging face)
Best Model: *krinal214/bert-all-squad_ben_tel_context*
(link: https://huggingface.co/krinal214/bert-all-squad_ben_tel_context)

6)

Filename: bert-all-tydiqa+squad+squad_que+squad_context.ipynb
Dataset: tydiqa
Additional Dataset: Squad (english only - original) + Translated Squad (Only questions are translated into telugu and bengali) +Translated Squad (Used translated data of squad from hugging face)
Best Model: *krinal214/bert-all-squad_all_translated*
(link: https://huggingface.co/krinal214/bert-all-squad_all_translated)