EXPLORING THE POWER OF HUGGING FACE DATASETS

UNDERSTANDING DATASETS' SIGNIFICANCE

HTTPS://GITHUB.COM/INSIGHTBUILDER

TRAINING A NLP MODEL : STEPS

1. DECIDE THE TASK THAT MODEL IS GOING TO ACCOMPLISH.
2. CHECK IF THERE IS EXISTING MODEL THAT CAN DO THIS TASK.
3. LEARN THE FEASIBILITY OF THE EXISTING MODEL TO BE FINE-TUNED.
4. IF NO MODEL IS AVAILABLE, CHECK FOR A MODEL THAT CAN PERFORM ALTERNATE TASK, WHICH CAN BE USED FOR TRANSFER LEARNING
5. LOOK FOR THE DATASET FOR TRAINING THE MODEL (FINE TUNING / TRANSFER LEARN)
6. IF DATASET UNAVAILABLE BUILD IT FROM SCRATCH AND FORMAT IT FOR TRAINING
7. DECIDE ON THE METRICS THAT MODEL WILL BE EVALUATED UPON
8. ONCE MODEL IS TRAINED DECIDE ON THE METRIC THAT MODEL WILL BE COMPARED WITH ANOTHER MODEL
9. DEPLOY THE MODEL INTO PRODUCTION, BY VARIOUS METHODS AVAILABLE
10. CONTINUE COLLECTING DATA FOR FURTHER IMPROVING THE MODEL

CHALLENGE SOLVED : DATASETS

1. SHARE DATASETS EASILY WHETHER IT IS TEXT, AUDIO OR VIDEO.
2. DATA INSIDE THE DATASET IS READY FOR THE MODEL TRAINING.
3. DATA IS HELD IN ARROW FORMAT WHICH PROCESSES LARGE DATASETS WITH ZERO-COPY READS. NO CONSTRAINT ON THE MEMORY.
4. CUSTOMISED DATASET CREATED BY YOU CAN BE UPLOADED TO HUGGINGFACE HUB, SO SHARING YOUR DATASET IS MATTER OF GIVING A NAME
5. PROCESSING THE DATA IS DONE USING METHODS AVAILABLE IN DATASET DICT INSTANCES
6. EACH DATASETS HAS CONFIGURATION, AND ASSOCIATED METRICS FOR EASY TRAINING
7. DATASETS ARE ALREADY SPLIT INTO TRAIN, VALIDATION AND TEST SETS IF REQUIRED
8. DATASETS CAN BE CONVERTED TO TF, PYTORCH,JAX FORMAT USING THE SET_FORMAT() METHOD

HTTPS://GITHUB.COM/INSIGHTBUILDER

CHALLENGE SOLVED : BIG DATASETS

- DATASETS USES THE ARROW FORMAT WHICH IS LANGUAGE AGNOSTIC
- LARGE DATASETS ARE CONVERTED TO STREAMING DATA USING BELOW CONSTRUCT

```
FROM DATASETS IMPORT LOAD_DATASET

IMAGENET = LOAD_DATASET("IMAGENET-1K",
SPLIT="TRAIN")  # DOWNLOADS THE FULL
DATASET
PRINT(IMAGENET[0])
```

```
FROM DATASETS IMPORT LOAD_DATASET

IMAGENET = LOAD_DATASET("IMAGENET-1K",
SPLIT="TRAIN", STREAMING=TRUE)  # WILL START
LOADING THE DATA WHEN ITERATED OVER

FOR EXAMPLE IN IMAGENET:
    PRINT(EXAMPLE)
    BREAK
```

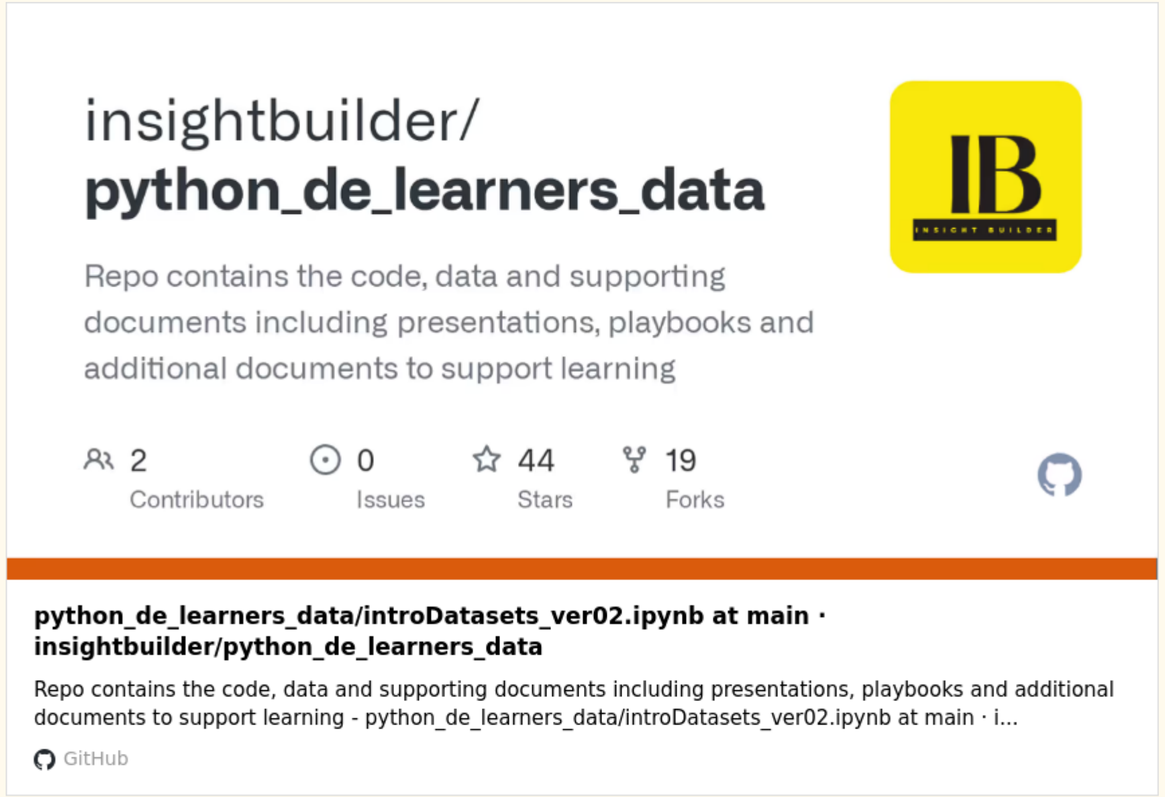
- DATASETS ARE CREATED USING DICTIONARIES. THEY CAN BE LOADED ENTIRELY OR PROGRESSIVELY FROM LOCAL OR REMOTE LOCATIONS
- DATASET SUPPORTS BOTH EAGER AND LAZY PROCESSING DEPENDING HOW IT IS INITIALIZED
- TO_ITERABLE_DATASET() METHOD CONVERTS DATASET TO ITERABLEDATASET

DATASET FEATURES : BACKBONE OF DATA

- FEATURES CONTAINS HIGH-LEVEL INFORMATION ABOUT EVERYTHING FROM THE COLUMN NAMES AND TYPES, TO THE CLASSLABEL
- FEATURE IS LIKE SCHEMA IN ANY DATABASE TABLE
- DIFFERENT FEATURE CAN BE ASSIGNED DEPENDING ON THE DATA TYPE
- MY_DATASET.FEATURES
- MY_DATASET['COL_1'].NUM_CLASSES
- MY_DATASET['COL_1'].NAMES
- MY_DATASET.INFO
- IMPORTANT METADATA IN THE DATASET IS
  - SPLIT - DESCRIPTION
  - CITATION - HOMEPAGE
  - LICENSE

METHODS TO PLAY WITH DATASETS

1. LOAD_DATASET : THERE ARE MULTIPLE WAYS, FROM DIFFERENT FILE FORMAT
2. VARIETY OF PROCESSING OPERATIONS POSSIBLE. FOLLOWING ARE THE MAIN
   - A. SORT, SHUFFLE, SELECT, SPLIT, SHARD
   - B. RENAME, REMOVE, FILTER, FLATTEN
   - C. MAP
   - D. SET-FORMAT
   - E. CONCATENATE
   - F. INTERLEAVE
   - G. SAVE & EXPORT



HTTPS://GITHUB.COM/INSIGHTBUILDER

THANKS FOR WATCHING

HTTPS://GITHUB.COM/INSIGHTBUILDER