DLP - Week-1 Graded Assignment-1

1) Choose all the correct statements about the dataset at the following URL 1 point

"https://huggingface.co/datasets/ai4bharat/naamapadam"

↳ 11 configs

Note: We strongly recommend using the appropriate functions in the datasets module to obtain the dataset's information, rather than accessing that information directly from the hub.

- [x] The dataset contains 11 configs(subsets)
- [ ] The dataset contains 22 configs(subsets)
- [x] The Hindi subset of the dataset contains 985787 **training** samples
- [ ] The Tamil subset of the dataset contains a **total** of 985787 samples.
- [x] The number of classes (labels,tags) is 7

2) **Common data** 1 point

Download the Tamil sub-dataset from the following URL https://huggingface. co/datasets/ai4bharat/naamapadam, and store in a variable named "ds".

Find the location of the cache directory (one way to determine the cache directory location is by using a method from the Dataset class; please refer to the documentation). Which of the following files are in the cached directory corresponding to the downloaded dataset?

- [x] naamapadam-train.arrow
- [x] naamapadam-test.arrow
- [x] naamapadam-validation.arrow
- [ ] naamapadam-all.arrow
- [x] dataset_info.json

since there are 3 types of setting in the dataset → train, test, validation

but also, if you check the folder a json file is created that stores the info. regarding the dataset.

3) What is the disk space size of the downloaded dataset in MB? → check NB

180

Yes, the answer is correct.
Score: 1

Accepted Answers:

(Type: Range) 179,181

1 point

4) Enter the number of training samples in the dataset. → check NB

497882

5) Create a new column named "num tokens". Compute the number of tokens (words) in each sample and store the results in the newly created column. Reassign the modified dataset to the same variable "ds". How many tokens (in millions) are there in the entire dataset?

- [ ] 4
- [ ] 5
- [x] 6
- [ ] 7

around 6 millions

→ modified dataset has the new column with total no. of words.

↳ this makes the disk usage as 360 MB

6) The statement that the modified dataset increased the disk space requirement about two fold is 1 point

- [ ] True
- [x] False

No, the answer is incorrect.
Score: 0

Accepted Answers:

True

7) Concatenate all the samples across the splits in the following order: [train:test:validation]. Currently, each sample contains a list of words (tokens). Create a sentence by joining the individual words (tokens) in each sample using a single white space as a delimiter and store the resulting sample in a new column named "text". Create a new dataset by removing the columns "ner tags" and "tokens". Store the new dataset in the same variable "ds". Enter the total number of samples in "ds"

501435

Yes, the answer is correct.
Score: 1

Accepted Answers:

(Type: Numeric) 501435

1 point

8) Each modification to the dataset introduces new cache file in the cache directory, thereby increasing the disk space. The statement is 1 point

- [ ] True
- [x] False

No, the answer is incorrect.
Score: 0

Accepted Answers:

Yes, the cache files keeps on increasing after every iteration.

For eg.-pycache-

9) Filter the dataset so that all the samples in the dataset should have at least six tokens (any symbol separated by a white space is considered a token). Enter the number of samples in the dataset after filtering. Enter the exact number. If the answer is 123456, enter it as 123456.

370495

Yes, the answer is correct.

Score: 1

Accepted Answers:

(Type: Numeric) 370495

1 point

10) Download all the splits of Tamil sub-dataset "inltkh.ta" of https://huggingface.co/datasets/ai4bharat/indic_glue .

Filter the dataset so that each sample contains at least six words (separated by a single white space).Then Interleave the resultant dataset with the above filtered dataset of naamapadam. Take 80% of samples from naamapadam and 20% from indic glue. Enter the number of samples after interleaving the datasets.

Note: Set the value of the seed argument to 42.

32354

Notebook problems.