DAY 01 (10-04-2023)

**TASKS**

1. Data set - in kaggle datasets related to research papers - search

2. Analyse the datasets and understand

3. Training datasets and test dataset and validation datasets

UNDERSTANDING DATASET

**Files**

* **train** - the full text of the training set's publications in JSON format, broken into sections with section titles
* **test** - the full text of the test set's publications in JSON format, broken into sections with section titles
* **train.csv** - labels and metadata for the training set
* **sample\_submission.csv** - a sample submission file in the correct format

**Columns**

* Id - publication id - note that there are multiple rows for some training documents, indicating multiple mentioned datasets
* pub\_title - title of the publication (a small number of publications have the same title)
* dataset\_title - the title of the dataset that is mentioned within the publication
* dataset\_label - a portion of the text that indicates the dataset
* cleaned\_label - the dataset\_label, as passed through the clean\_text function from the [Evaluation page](https://www.kaggle.com/c/coleridgeinitiative-show-us-the-data/overview/evaluation)
* Publications are provided in JSON format in the folder named train
* The **train.csv** file provides a summary of the training data
* A percentage of the public test set publications are drawn from the training set.
* not all datasets have been identified in train.

so these unidentified datasets have been used as a portion of the public test labels. These should serve as guides for the difficult task of labeling the private test set.

* Train and train.csv file will act as private test data.
* Both train and train.csv can be used as training data set for developing ML model.
* Train folder will help to develop more complex NLP model.
* Train.csv is more structured and compact representation of the data. which can be easy to work with and process.
* Nothing wrong with the selection of both, selection of training set wont affect the accuracy of the model.it depend on person.
* The objective of the competition is to identify the mention of datasets within scientific publications
* Your predictions will be short quotation from the publications that appear to note a dataset, that is dataset\_label column in train.csv
* Predictions that more accurately match the precise words used to identify the dataset within the publication will score higher.
* Predictions should be cleaned using the clean\_text function, that is cleaned\_label column in train.csv