Question 1 **Max. score: 100.00**

Category and subcategory of articles

Category and subcategory of articles

You are given PDF files representing articles or news from various industries such as entertainment, finance, education, etc.

Task

You are required to predict the industry and the sub-category that a news/PDF/article belongs to.

For Example, the industry is Finance and the sub-category is Banking

Note:- We may try to reproduce your results so make sure to write a code that is reproduceable i.e use a fix seed values while using random functions.

Dataset description

The dataset folder contains the following :

- **train(folder):** 31779 .txt files
- **test(folder):** 7945 .txt files
- **train.csv:** 31779 x 3
- **test.csv:** 7945 x 1
- **sample_submission.csv:** 5 x 3

The columns provided in the dataset are as follows:

Column name	**Description**
File_name	Represents a unique name of a file
Industry	Represents the category of an article
Sub-category	Represents the sub-category of an article

Evaluation metric

```
score_industry =  100*metrics.f1_score(actual["Industry"] ,predicted["Industry"], average='micro')
score_sub_category = 100*metrics.f1_score(actual["Sub-category"] ,predicted["Sub-category"], average='micro')
score =  ( score_industry + score_sub_category )/2
```

Result submission guidelines

- The index is **File_name** and the targets are the **Industry** and **Sub-category** column.
- The submission file must be submitted in **.csv** format only.
- The size of this submission file must be 7945 x 3.

Note: Ensure that your submission file contains the following:

- Correct index values as per the test file
- Correct names of columns as provided in the **sample_submission.csv** file

Download dataset