Media Product Classification

Problem Statement

Indix hosts the world's largest collection of programmatically accessible structured product information in the cloud. The products in our database belong to 25 verticals and that translates approximately to 6000 sub-categories. Every product that we carry in our database gets stamped with information about the "category" it belongs to. This problem of classifying a product into a particular category is very important to serve various use cases – like, helping search, performing product matching, providing category specific insights, and so on.

The problem of stamping every product in our catalog into a category is a Multi Label Hierarchical Classification problem. Your challenge is to build a micro version of this classifier where you will predict 4 classes.

Data Description

The datasets are available at the following locations

- Training data : https://drive.google.com/open?id=1tKuZW05YpMnsLGYTwVJ10rICe8SxTD7B
- Test data : https://drive.google.com/open?id=1VBwFrY_Rj5MPsnhK0VODcTzl4SVZN8BZ

Training data has the following fields and contains **300000** records

1. **Id** - unique id for the record
2. **additionalAttributes** - Product attribute related to a particular product. These are key, value pairs that can be found in tabular format as product information for most products in e-commerce websites.

   An example of additionalAttributes


```
{"ASIN": " B000JJRY9M",
"Amazon Bestsellers Rank": "  in DVD & Blu-ray (See Top 100 in DVD & Blu-ray)",
"Average Customer Review": " Be the first to review this item",
"Classification": "  Exempt",
"DVD Release Date": " 26 Feb. 2007",
"Format": " AC-3, Colour, Dolby, DVD-Video, PAL",
"Language": " English",
"Number of discs": " 1",
"Region": " Region 2 (This DVD may not be viewable outside Europe. Read more about DVD formats.)",
"Run Time": " 287 minutes",
"Studio": " Hip-O Records"}
```

3. **label** - The class to which a product belongs. Values belong to the finite set ('books','music', 'videos','rest')

The problem statement is to classify the products into any one of the buckets

(i) Books
(ii) Music
(iii) Videos
(iv) Rest - A default class for products which doesn't belong to (i),(ii) or (iii) category.

Evaluation Criteria

- You should submit a file `submissions.csv" which contains the test set with an additional column called label which. The evaluation metric would be accuracy score on the test set that would be provided. There are **60000** records in the test set all of which would need a prediction.
- The submission should also contain a Jupyter Notebook or Python/Scala/Julia Script with the model(s) actually built. It should contain in detail the following
    - All the Exploratory data analysis done on the dataset.
    - Preprocessing logic and rationale for them if any.
    - Model Training code.
    - Model Prediction code.
- A Readme that can help us replicate the submission score.

Questions

For any questions or doubts , please reach out to any of us below

Krishna - krishnasangeeth@indix.com
Sriram - sriram@indix.com