UnitedHopes Technologies
Technopark - 695581
Trivandrum, Kerala

Media Product Classification

OVERVIEW

The task is multi-class classification for media products into 4 categories - movies, books, music and rest

DATA DESCRIPTION

Training data has the following fields and contains 603201 records

1. storeId - a unique number for identifying a website

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. breadcrumbs - breadcrumb captured at the page. Breadcrumbs typically appear horizontally across the top of a Web page, often below title bars or headers.

An example of breadcrumb

subjects > travel > world regions > europe > european nations > france

4. label - The class to which a product belongs. Values belong to the finite set ('books','music',

'videos','rest')

It is possible that for some products only one among (2) or (3) might be available. 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

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 442041 records in the test set all of which would need a prediction.

The submission should be in Python only with proper comments explaining the code written. It should contain in detail the following

- All the Exploratory data analysis done and data visualization with matplotlib, plotly, d3.js, tableau etc. 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.

The train and test files are in google drive.

Train

https://drive.google.com/open?id=18p_pwAcL50lGVqM86GeNTwOzn-g-VQHl

Evaluation

https://drive.google.com/open?id=1TgcblBDsxIe6JPaLY7vaiaVgBX8_OkcP