Building an Automatic Product Title Tagging Engine

Problem Statement

The Client is an E-commerce aggregator and collects the data on a range of products from multiple E-commerce portals and wants to aggregate the information. The business objective is to build a price comparison website and show the products with in the same category and provide the information on range of price points at which multiple retailers offer the same product.

As they are gathering the data from multiple websites, each individual retailer defines the category and subcategory for their products based on the various business parameters. So, while aggregating them there is a lot of inconsistency in labelling of the categories/subcategories. To fix this issue temporarily, they hired a 10-member support team to read and tag each product description and come up with a standard list of category and subcategory. However, as the volume of the product titles so huge, it takes 72 hours to tag the product category and subcategory and there were mistakes while labelling them.

They hired you as the Data scientist to solve this challenge and you are free to use manually tagged data as the training data and the goal is to come up with a scalable and consistent solution.

Data Set

You are provided with four csv files

1. Train.csv
2. Test.csv
3. Evaldata.csv
4. Predictions.csv

You need to build the models and test them and improve the model performance. You are required to check the performance of the model on the Evaldata and store the predictions in Predictions.csv and evaluate model performance through the shiny app provided .

Evaluations

The metric we are interested in this problem is Recall and Accuracy. You need to explain why both these metrics are important.

The benchmark accuracy and recall values are 80% and 80% respectively. You could use shiny application at the below URL to check model performance on Evaluation data set.

http://172.16.0.12:3838/

This will help you to tweak the algorithm parameters and/or implementing the necessary pre-processing steps. Please store the prediction results in the ''**predictions.csv**' and upload this file to check the model performance and do not submit the Evaldata.csv into the shiny app.

Submit your best predictions and relevant code and report through **piazza** by **6:00** PM. The folder name should be <*your full name*>.

You will be evaluated on the following tasks

1. **Data pre-processing**
2. **Visualizations**
3. **Tuning the parameters of the algorithms**
4. **Performance metrics (Accuracy and Recall)**
5. **Presentation skills**

Wish you good luck!