Merchant Classification

Problem Statement

Merchant categorisation is an important task for Fintech businesses which ensures that a payment carried out at say McDonalds is properly categorised as a food related transaction.

One of the major challenges with this problem is the lack of data. We've collected ranked google search and map results for **around 7000** mercha nt names using each merchant name as a search query to augment the training data. Each merchant is labelled under one of the **15** categories such as travel, housing, entertainment, etc.

The goal of this task is to build a system that can accurately classify previously unseen merchant's name into the right category.

Evaluation Criteria

The task will be evaluated using **accuracy** as a metric. The baseline accuracy for this problem is **currently at 75%.** We are looking for a model that performs better than benchmark accuracy.

Hint: Feel free to try any novel approaches for feature engineering apart from sparse features.

Data

Files description

- *labels.csv - category labels for all merchant names*
- *search.json - google search results for all merchant names*
- *maps.json - google maps results for all merchant names*

Labels.csv:

For each merchant name, this includes its corresponding category label

- *merchant_name* - name of the merchant
- *category* - category label of the merchant

search.json

For each merchant name, this includes structured data for

- *organic results* - main search results, may include reviews, title, snippets, etc.
- *local results* - for some requests, may include reviews, title, type, address, etc
- *knowledge graph* - for some requests, google includes a knowledge graph block
- *ad results*

maps.json

For each merchant name, this includes structured data for maps local results and maps place results

- *local results* - this may contain title, reviews, rating, gps_coordinates, data_id, etc.
- *place results* - this may contain title, description, type, reviews, address, etc.

Submission File Format -

ipython notebook

Confidentiality -

This problem statement, along with sample data is extremely confidential for Jupiter and should not be shared with anyone. Same should not be used for any purpose whatsoever, except the purpose of the ML assignment by Jupiter.