Datathon@IndoML 2024 Phase 2

Organized by shubhadip_nag - Current server time: Oct. 26, 2024, 7:53 a.m. UTC

▶ Current

Final (Phase 2)

Sept. 20, 2024, noon UTC

End

Competition Ends

Oct. 28, 2024, 7 p.m. UTC

Learn the Details | Phases | Participate | Results | Forums

Note: This is only for Phase 2 of the competition.

Please read the rules here before joining Phase 2. You can find it in the left sub-tab - "Terms and Conditions".

Task: Attribute-Value Prediction From E-Commerce Product Descriptions

In recent years, e-commerce has grown tremendously, with major online retailers offering billions of products and shipping millions of packages daily. However, given the sheer volume of offerings, sellers find it extremely difficult to fill in extensive sets of product attributes, resulting in incomplete product profiles. E-commerce platforms, on the other hand, depend on such structured metadata, typically in the form of attribute-value pairs, for a deeper understanding of the products, and for facilitating critical downstream applications, such as search, product recommendation, question answering; as well as, for providing an enhanced customer experience.

Predicting attribute-value pairs from unstructured product descriptions is, therefore, a fundamental challenge for worldwide e-commerce catalogs such as Amazon, Walmart, and Alibaba. In this Datathon, your task would be to develop a model that automatically predicts attribute-value pairs for a given product description.

Phase 1:

Phase 1 is already over. One can directly participate in Phase 2

Phase 2:

Along with a short product description obtained from receipts/invoices, you are provided with the retailer and price information. Please however note that the "retailer" information is anonymized. Hence, either the value may not be an actual retailer or the retailer may not even be associated with the product. You will then be required to predict the values for a total of 4 attributes related to the product: Super Group, Group, Module, and Brand. Please note that the values of these attributes may or may not appear in the product title. Also, unseen values for all the attributes may appear in the hidden test data.

Datathon @ IndoML 2024

TLDR: Important Dates and Steps

- 🗓️ Phase 2 Starting Date: 20th September
- 🗓️ Phase 2 Deadline: 25th October 11:59 PM IST
- 🏆 Win prizes worth INR 1.1 Lakhs
- 👉 Register: https://forms.gle/cVzA1cwyzuvXtzh17
- 🚢 Join Discord Channel: https://discord.com/invite/V2W7gY8DRa
- 🚀 Tutorial 1 (Classification-Based): https://tinyurl.com/ymem4sxk
- 🚀 Tutorial 2 (Generative-Based) & Submission Guidelines: https://youtu.be/V5-KZNYaAEY?si=kkXmHIOeFyeSWMTJ
- 🚀 First AMA Session: https://youtu.be/BWrDWUpj8j8

Phase 2 / Final Phase

Data Format

The data will be distributed in JSONL format with one example per line (see http://jsonlines.org for more details).

Each line will have an ID associated with it which will be useful for submitting predictions.

The data files have 2 types of extensions. One is '*.features' and the other one is '*.labels'

The '*.features' files contain the input features of data points to be used to train/test the model. Whereas '*.labels' files contain a set of labels corresponding to the data points. Both files have a field called 'indoml_id' that will be used to map the '.features' file to the '.labels' file.

The `'.features'` file will contain the following fields:

`indoml_id: "id_num"`

`id_num: id of that data point`

`description: Short Product Description`

`retailer: Anonymized retailer names (May not be actual retailers. May not even be connected to the product)`

`price: Price of the product`

Example:

```
{"indoml_id": 4897, "description": "cad 5 pk creme egg 197 g", "retailer": "greenzen", "price": "1.99"}
```

- This example was taken from the `'train.features'` file

The `'.labels'` file will contain the following fields:

indoml_id: "id_num" which is associated with the corresponding `'.features'` file.

`id_num: id of that data point`

`supergroup, group, module, brand: These are the attributes to be predicted.`

Example:

`{"indoml_id": 4897, "supergroup": "biscuits & confectionery & snacks", "group": "chocolate novelties", "module": "chocolate novelties", "brand": "cadbury creme egg"}`

- This example was taken from the `'train.labels'` file.

Files

You will only be provided with a training dataset consisting of two files "train.features" and "train.labels". You may choose to train your models on the entire data. You may also choose to split the data into train and validation splits, train your models on the train split, and keep checking the model's performance on an unseen validation split. **The test data will be kept hidden for now and will be released 2 weeks before the deadline**.

Answer Submission Instructions

Submit the answers file in a `ZIP file` containing your predictions (`The prediction file format should be: test*.predict` , where '*' is any string. **The best is to name your file as** `'test_teamname.predict'` . Each line in the file should be a dictionary and should adhere to the following format:

{"indoml_id": 4897, "supergroup": "biscuits & confectionery & snacks", "group": "chocolate novelties", "module": "chocolate novelties", "brand": "cadbury creme egg"}

The order of the `'indoml_id'` **must be preserved.**

Please note that the value of "indoml_id" is an integer. Don't submit it as a string.

There is no restriction on the zip file name. It can be named anything, such as `'submission_1.zip'.`

Contact Us:

- Competition Website
- Discord
- Use this link to join the Discord channel if you have not already: Discord Invite

---

Phase 1 / Development Phase

Phase 1 is already over. Please go to this link to get the details of Phase 1.

#	Username	Score
1	AmanSinha	0.4764
2	codecrafter	0.4432
3	chaithra	0.4345