Please let me know what I can help.

Shiladitya suggested to remove the brand name from the VND\_ECOM\_DSC . I am working on it.

Is anything you’d like me to do?

Is anything I can help for scent analysis?

FYI: I created a folder “OpenAI + GenAI + LLM + GPT” to upload articles and links related to OpenAI + GenAI + LL

M + GPT. Enjoy the treading, and you are more than welcome to add more information.

Fine-Tuned OpanAI models - Confluence page

[Fine-Tuned OpenAI models - EIX - Enterprise Item Experience - Confluence (kroger.com)](https://confluence.kroger.com/confluence/pages/resumedraft.action?draftId=426281255&draftShareId=936203d3-cb34-4d65-90b1-24a0e5d5ba24&)

01.26.24

A big achievement for “KTD Release Announcement - Increased Flavor, Color & Scent on 579k Products leveraging AI / ML” and congratulate on your leadership to lead Finding Nemo Team! BTW, what KTD stands for? Is it related to ‘Kroger Taxonomy’?

## FYI: I added table of contents in [Autor leadshershp tpmation of normalized attribute extraction for new items.](https://confluence.kroger.com/confluence/display/EIS/Automation+of+normalized+attribute+extraction+for+new+items) The long link shows in ‘**Update Process diagram’.**

01.24.24

### The users needs to interact with config file often. Could you please review ‘JSON file basics’ and ‘Configuration (Config) file’? Shall we add the screenshots for updating config file and creating new config file?

### Hi Paul, I added a new session ‘‘JSON file basics’’ to help user understand the config file. Is it too technical? Shall I leave it as is? Or shall we save it as a fille and attach it to the document?

01.23.24

I’ll be released from this project. Today is my last day for this project.

It is a pleasure to work with you all for this challenging project. Thank all for help and support for my professional growth.

Do you have a list to add additional session(s) or adjustment for the document of ORE process handoff: [Automation of normalized attribute extraction for new items?](https://confluence.kroger.com/confluence/display/EIS/Automation+of+normalized+attribute+extraction+for+new+items)

If not, is anything I can help?

Do you know any Infosys project in Kroger which might use my skills?

01.22.24

Hi Paul, Thank you for updating the list of 'Attribute / Product Category schema as of 1/18/2024 in [Automation of normalized attribute extraction for new items.](https://confluence.kroger.com/confluence/display/EIS/Automation+of+normalized+attribute+extraction+for+new+items)

Just curiosity, how do you know the content will return “  
"content": "{'Color': 'YELLOW/ORANGE', 'Confidence': 90}" for sample\_input = 'marigold'?

* In your content you specified  what you want to return for input is 'Delphinium'. Based on this example, OpenAi knows what to return for input = marigold. Is it correct?
* Does confidence like a score? Is confidence higher better? If yes, why confidence for ‘OTHER’ is defined as 100?

1. FYI: In Azure studio chat, I used deployment = ‘acaopenai’ and content = ‘What is the color for this Floral item?: marigold. Separate multiple color values with a / and provide the answer and the confidence percentage in a JSON-like format. If no value is detected, return 'OTHER' with a confidence of 100. If the description contains a very rare color, report the more common variant of that color. Example: If the input is 'Delphinium', return {'Color': 'BLUE', 'Confidence': 90}’, I got result: {"Color": "YELLOW/ORANGE", "Confidence": 95}

Where can I find your code for [Extract Specific-Color data values for 'Floral'](https://jira.kroger.com/jira/browse/EISART-59799)?

It appears to me that the code ‘[ATTRIBUTE\_EXTRACTION’](https://adb-4812933386228410.10.azuredatabricks.net/?o=4812933386228410#notebook/3728363041126467/command/3728363041126468) is the consolidate code of ‘Configs for Normalized Attribution’ for Color/Flavor/Scent. Shall I put the name in confluence page:  [Automation of normalized attribute extraction for new items?](https://confluence.kroger.com/confluence/display/EIS/Automation+of+normalized+attribute+extraction+for+new+items)

Great job of the presentation at the review meeting. Please add 'The input file name can't contain space.' in your code ‘[ATTRIBUTE\_EXTRACTION’](https://adb-4812933386228410.10.azuredatabricks.net/?o=4812933386228410#notebook/3728363041126467/command/3728363041126468). If input file name has a space, there will be an issue to download output file.

Is anything I need to update the document? If not, is anything I can help today?

01.19.24

I use server name “11-eastus2-3534-idq-sqlmi.cc4c669dc8d1.database.windows.net” to connect SQL server via Azure Data Studio to access PIM tables. Can we access the PIM tables from Databricks?

If yes, could someone share the complete python code including:

* set connection between Server name and Databrciks
* Run a simple query “ SELECT  ITM\_ID, count(itm\_id)  from CURIDQ.PIMMART.CLS\_SPC\_ATB  group by itm\_id having count(\*) > 1” and save the result to a CSV file in databricks

I added a session “Notebooks in Azure Databricks” in the document. I’ll finalize the script name later. Could you please take a quick look new session?

# Below python code is from the article.

# [Query SQL Server with Azure Databricks](https://learn.microsoft.com/en-us/azure/databricks/connect/external-systems/sql-server)

remote\_table = (spark.read

.format("sqlserver")

.option("host", "hostName")

.option("port", "port") # optional, can use default port 1433 if omitted

.option("user", "username")

.option("password", "password")

.option("database", "databaseName")

.option("dbtable", "schemaName.tableName") # (if schemaName not provided, default to "dbo")

.load()

)

In my opinion, the example in the article is not complete. My blocks are:

1. I don’t use spark often. In order to run the code I think we need to import spark.connection to use spart.read.
2. Where to get the hostname? And port number?
3. After load the table, how to run the query?

My user case:

1. I extracted data from PIM table via Azure Data Studio and export the result to a CSV/excel file to my local folder.
2. Then upload the file to Databricks and run my application.
3. I’ll be nice if I can include the code to extract data from PIM tables in my application.

remote\_table = (spark.read

.format("11-eastus2-3534-idq-sqlmi.cc4c669dc8d1.database.windows.net")

.option("user", "iny2819")

.option("password", "Junchung@1")

.option("database", "CURIDQ")

.option("dbtable", " PIMMART.CLS\_SPC\_ATB” ) # (if schemaName not provided, default to "dbo")

.load()

)

01.18.24

Hi Mallik, I‘d like to clone your notebook: [Attribute\_Extraction\_Prod - Databricks (azuredatabricks.net)](https://adb-4812933386228410.10.azuredatabricks.net/?o=4812933386228410#notebook/1949532348594449/command/1949532348594457) to my folder and run the code. Since I don’t have modified permission of your folder and I am noy able to clone it. Could you please do something to allow me to clone your code?

1. Get access to Databricks

* Need to provide the instructions how to get access Databricks with end user.
* Had a phone conversation with Patel, Parav Kumar who is from ‘MX Core Data team regarding there profile to access MX Dataplatform.
* Per Patel the new profile to access MX Dataplatform is ‘jcs000-mx-insights-dev-8666**-domain name’** instead of jcs000-mx-insights-dev-8666 which is old APT for MX Dataplatform. Please refer to the document [MX Data Platform - MX Core - Confluence (kroger.com)](https://confluence.kroger.com/confluence/display/MC/MX+Data+Platform).   Per Patel, the new profile is domain-specific, and the **domain name can get from manager.**

1. **WIP: Tried to add a job in Databrick for** Code and Configs for Normalized Attribution
2. IM Si to review the confluence page:  [Automation of normalized attribute extraction for new items](https://confluence.kroger.com/confluence/display/EIS/Automation+of+normalized+attribute+extraction+for+new+items)
3. Updated confluence page:  [Automation of normalized attribute extraction for new items](https://confluence.kroger.com/confluence/display/EIS/Automation+of+normalized+attribute+extraction+for+new+items)

* Paul added **“Attribute / Product Category schema as of 1/18/2024’ and Sophia move them under the session of ‘**Configuration (Config) file’

**‘**

**Could you please share the domain name with us?**

I need to include how to get access to Databricks in my Confluence page for the new use who will have the same domain name as I have.

T**o access** MX Dataplatform, did you raise the APT for jcs000-mx-insights-dev-8666 or jcs000-mx-insights-dev-8666**-domain name?**

01.18.24

I’ll base on your code ‘‘NORMALIZED\_ATTRIBUTION\_WITH\_FILE\_UPLOAD\_011124.ipynb’’ to add one session ‘ How to run the code’ in "[Automation of normalized attribute extraction for new items](https://confluence.kroger.com/confluence/display/EIS/Automation+of+normalized+attribute+extraction+for+new+items)".

* Did the code have been finalized?
* Could you or someone please create a folder in the workspace in Databricks to save the code for end user to run the code?

I’ll try to create a job your script ‘NORMALIZED\_ATTRIBUTION\_WITH\_FILE\_UPLOAD\_011124.ipynb’ by using Dataflow in Databricks. If it works, we do not need a folder for end user to run the script.

01.17.24

Hi Shiladitya,

In your code ‘NORMALIZED\_ATTRIBUTION\_WITH\_FILE\_UPLOAD\_011124.ipynb’, I didn’t see ‘subs’ been used in def color\_match(item, color\_list\_org, subs,cutoff = 90).

I didn’t see ‘replacement’ been used in the code. Could you please check?

I understand. I am trying throw suggestions for my observations that ‘replacement’ and ‘subs’ have not been used in the code ‘NORMALIZED\_ATTRIBUTION\_WITH\_FILE\_UPLOAD\_011124.ipynb’.

* Observation 1: Per the code in PIM\_COLOR\_EXTRACTION\_floral\_121923.ipynb

pim\_dpt['EXTRACTED\_COLOR']  =  \

                pim\_dpt.MATCHING\_DESCRIPTION.str.upper().replace(replacements, regex = True).apply(lambda x:color\_match(x, color\_list\_org= color\_list))

* Observation 2: Per the code in Attribute\_Normalization\_color\_floral\_12192023**.**ipynb

def normalize\_colors(attr\_string, normalized\_list = normalized\_list ,delim\_read = '|', delim\_write = ';',rename = rename):

    attr\_list = attr\_string.split(delim\_read)

    attr\_list = [rename[i] if i in rename.keys() else i for i in attr\_list]

    attr\_list = [i for i in attr\_list if i in normalized\_list]

It might need to apply Observation 1 & Observation 2 to the code ‘NORMALIZED\_ATTRIBUTION\_WITH\_FILE\_UPLOAD\_011124.ipynb’.

01.16.24

Hi Vyas,

I like your idea of super generic and thank you for the recommendation. Regarding your recommendation for "The drop-down list would be 'Kitchen & Dining', 'Home Decor (excluding Rugs)", 'Home Decor (Rugs only)', 'Floral'’, the drop-down list is for departments for attribute ‘color’ which we did normalization in the past. The department will determine the normalization value list which is saved in the configure file and it is transparent to end user.

What your comments regarding the user cases from Vyas?

Use cases

* Update manual attribute for A gtin
* Update normalization values for a specific sub-com/category
* Etc

Thank you for adding the diagram of ‘Attribute Value Amplification’ to [Automation of normalized attribute extraction for new items](https://confluence.kroger.com/confluence/display/EIS/Automation+of+normalized+attribute+extraction+for+new+items). The diagram will benefit us and it makes the confluence page standout.

Hi Vyas, Thank you for the user cases and your recommendation for [the confluence page](https://confluence.kroger.com/confluence/display/EIS/Automation+of+normalized+attribute+extraction+for+new+items) and  I responded to your recommendation and will work on the user cases. ‘

Hi Shiladitya, I am a member of ‘MX Core Data (Public)’ team,  I raised a question to have the link or document to apply for the access of  'mx-core-data-stage-dbxws'. The response is “ You need to raise APT access for specific domain. Please check this document: [MX Data Platform - MX Core - Confluence (kroger.com)](https://confluence.kroger.com/confluence/display/MC/MX+Data+Platform)” Please read the document and let me know if it is a good idea to include the link in my confluence page? What could be for the specific domain?

Hi,

Thank you for the information. I have access to 'mx-core-data-stage-dbxws' and databricks workspaces. I am writing a document to guide the end user to access files we saved in Databricks. The end user would use the same domain I have. My EUID is INY2819. How do find my domain?

01.15.24

Please follow the instruction ‘Download output file‘ under the session ‘Output file’ in confluence [Automation of normalized attribute extraction for new items](https://confluence.kroger.com/confluence/display/EIS/Automation+of+normalized+attribute+extraction+for+new+items) if you are able to download the file ‘NormalIzed\_Color\_01102024\_Kitchen&Dining\_sample\_input.csv’ in Databrciks.

‘Could you please try t

01.12.24

Attribute is part of the name of config file.

Could you please take a look of PIM ATTRIBUTE EXTRACTION Flavor\_Scent\_Color.html? Is ‘Refining Results’ necessary?

01.11.24

The ITM\_CFIC\_FYT\_DPT\_NAM had been used to determine department for Color. Does it had been used for Flavor & Scent?

01.05.24

Hi Paul,

Regarding your comment of removing ‘COCOA, PEACOCK’ for [Re: Extract Color Data Values - Normalized Data (Floral)](https://nam12.safelinks.protection.outlook.com/?url=https%3A%2F%2Fjira.kroger.com%2Fjira%2Fbrowse%2FEISART-59357&data=05%7C02%7Csophia.yue%40kroger.com%7Ce692b3d0c8cf42af625b08dc0e26f297%7C8331e14a91344288bf5a5e2c8412f074%7C0%7C0%7C638400807640171557%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=9pIUsKTbIWdNMZco7W59P9fGUbW0V5wufkZqWshQR60%3D&reserved=0), I can change the code and recreate the file. I am wondering if we can skip manual validation part.

Thank you for updating ‘

|  |  |  |  |
| --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | |  | | --- | |  | | Thank you for updating [Investigate Azure OpenAI usage quotas](https://nam12.safelinks.protection.outlook.com/?url=https%3A%2F%2Fconfluence.kroger.com%2Fconfluence%2Fdisplay%2FEIS%2FInvestigate%2BAzure%2BOpenAI%2Busage%2Bquotas%3Fsrc%3Dmail%26src.mail.product%3Dconfluence-server%26src.mail.timestamp%3D1704480605457%26src.mail.notification%3Dcom.atlassian.confluence.plugins.confluence-notifications-batch-plugin%253Abatching-notification%26src.mail.recipient%3D8aa0e0198823b1590188d4f847020193%26src.mail.action%3Dview&data=05%7C02%7Csophia.yue%40kroger.com%7C3e18e5f216424c15c6a608dc0e1f2afe%7C8331e14a91344288bf5a5e2c8412f074%7C0%7C0%7C638400774234068839%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=8x4YbRHkzChf%2BKSTuDFWy3Dqf5BOY8m61UjbkVk9VDo%3D&reserved=0)  [Exploring OpenAI for Attribute Extraction](https://confluence.kroger.com/confluence/display/EIS/Exploring+OpenAI+for+Attribute+Extraction)  [Exploring OpenAI for Attribute Extraction - EIX - Enterprise Item Experience - Confluence (kroger.com)](https://confluence.kroger.com/confluence/display/EIS/Exploring+OpenAI+for+Attribute+Extraction) | |

# Enjoy the reading: [Fine-Tuning GPT-3 Using the OpenAI API and Python](https://www.datacamp.com/tutorial/fine-tuning-gpt-3-using-the-open-ai-api-and-python)

01.02-24

# Please review [Investigate Azure OpenAI usage quotas](https://confluence.kroger.com/confluence/display/EIS/Investigate+Azure+OpenAI+usage+quotas) and give me your feedback.

What is the hyper-specific refer to in hyper-specific FLAVOUR/attribute extraction?

12.22.23

Did you try to run the python code from [Deployments | Azure OpenAI Studio](https://nam12.safelinks.protection.outlook.com/?url=https%3A%2F%2Foai.azure.com%2Fportal%2Fd3d5be3f55654a9fb6019f0eaee99cc2%2Fdeployment%3Ftenantid%3D8331e14a-9134-4288-bf5a-5e2c8412f074&data=05%7C02%7Csophia.yue%40kroger.com%7C75f66caf45ae4e358dd908dc01958764%7C8331e14a91344288bf5a5e2c8412f074%7C0%7C0%7C638386988976056185%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=LC6cV81Epnu6RZAnZQVAL%2F88yj2oBAwF5vKoCSEPfvY%3D&reserved=0)?

Did you try to run the python code from [Deployments | Azure OpenAI Studio](https://nam12.safelinks.protection.outlook.com/?url=https%3A%2F%2Foai.azure.com%2Fportal%2Fd3d5be3f55654a9fb6019f0eaee99cc2%2Fdeployment%3Ftenantid%3D8331e14a-9134-4288-bf5a-5e2c8412f074&data=05%7C02%7Csophia.yue%40kroger.com%7C75f66caf45ae4e358dd908dc01958764%7C8331e14a91344288bf5a5e2c8412f074%7C0%7C0%7C638386988976056185%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=LC6cV81Epnu6RZAnZQVAL%2F88yj2oBAwF5vKoCSEPfvY%3D&reserved=0) successfully? I got "

SSLCertVerificationError.

12.27.23

I uploaded the file to Jira. You did a couple of manual validations. After Shiladitya and I reviewed your manual validation for rugs, we didn’t take your suggestion. You might chat with Shiladitya for a brief instruction for the color manual validation.

12.21.23

Could you please share your code in Confluence “[Exploring Exploring OpenAI for Attribute Extraction](https://confluence.kroger.com/confluence/display/EIS/Exploring+OpenAI+for+Attribute+Extraction)” under “**Learning to customize the output format”? Or you might upload the code to Confluence.**

**Hi Paul,**

For rugs, we combined 'ASSORTED' to ’MULTICOLOR’. For floral, shall we combine 'ASSORTED' and ‘ASSORTED COLORS’ to ’ MULTICOLOR’?

Would we combine 'HOT PINK' to ‘PINK’? no

Hi shila

Could you please write a brief instruction for the color manual validation and share it with Prateek and me?

Jira

* Discussed with Paul to finalize the colors to be discarded for floral are:

'ANTIQUE','APPLE','BARK','BERRY','BI-COLOR','BRONZE','CAYENNE','CHAMPAGNE',

'CHOCOLATE','CLAY','COFFEE','CONCRETE',

'COPPER','CORAL','CREAM','CRUSH','DARK','DEEP','FROST','HAPPY BIRTHDAY','HEIRLOOM',

'IRON', 'LIGHT','LOVE','MARIGOLD', 'MEDIUM','MELON','NATURAL','ORCHID','PEARL',

'PEONY','PINEAPPLE','SAGE','STEEL','STRIPED','URBAN','VALENTINE','VARIETY',

'WATERMELON','WOOD','ZINNIA'

* Discussed with Paul to finalize the colors to be combined for floral are:

'ASSORTED':'MULTICOLOR','ASSORTED COLORS':'MULTICOLOR', 'CLASSIC WHITE':'WHITE',

'MULTI-COLORED':'MULTICOLOR','MULTI COLOR':'MULTICOLOR',

'GREY':'GRAY','DARK GRAY':'GRAY','LIGHT GRAY':'GRAY',

'NAVY BLUE':'NAVY','DARK BLUE' : 'BLUE', 'LIGHT BLUE': 'BLUE',

'DARK BROWN' : 'BROWN', 'LIGHT BROWN' : 'BROWN',

'LIGHT PINK':'PINK', 'DARK PINK':'PINK', 'DARK RED': 'RED'. I.E.,

1. 'DARK RED' will be combined into 'RED'.
2. 'ASSORTED' and 'ASSORTED COLORS' will be combined into 'MULTICOLOR'.

* Applied Paul's feedback of normalization list for floral to create a file for validation.
* Uploaded validation file to ISART-59365 and assign it to Prateek.
* There are 408 samples in the validation file.

Please confirm that the colors in the list below would be discard:

['ANTIQUE','APPLE','BARK','BERRY','BI-COLOR','BRONZE','CAYENNE','CHAMPAGNE','CHOCOLATE','CLAY','COFFEE','CONCRETE',

'COPPER','CORAL','CREAM','CRUSH','DARK','DEEP','FROST','HAPPY IRTHDAY','HEIRLOOM','IRON','LIGHT','LOVE','MARIGOLD',

'MEDIUM','MELON','NATURAL','ORCHID','PEARL',','PEONY','PINEAPPLE','SAGE'','STEEL','STRIPED','URBAN','VALENTINE','VARIETY',

'WATERMELON','WOOD','ZINNIA']

12.20.23

Hi Paul,

Normalized\_COLORS\_FLORA\_122023.xlsx

Hi [FRIEMEL, PAUL A](https://jira.kroger.com/jira/secure/ViewProfile.jspa?name=PF75692),  Attached is the initial normalized Color list for Floral.

1. The file has two tabs.   
   Normalized\_COLORS\_FLORAL is the normalized Color list for Floral.
2. Normalized\_COLORS\_Combine combined the normalized Color list with discard column and rename colors for

* Floral: Column name with suffix ‘\_floral’
* Rugs: Column name with suffix ‘\_rug’
* home décor nonrugs: Column name with suffix ‘hd\_norug’
* Kitchen& dinning: Column name with suffix ‘\_kitdin’

The highlighted ones in the tab ‘Norm’ alized\_COLORS\_FLORAL are found to be ambiguous/error.

The tab Normalized\_COLORS\_Combine is for your reference. It helps me to determine which color need to be highlighted. Please go over the file and let us know your feedback or adjustment required.

12.19.23

I am working on the color – floral. There are minor changes from color – home décor. Is it a good idea to modularize the process with DPT\_NAM as a parameter for data extraction, normalization, creating validation file?

12.15.23

## [**The Differences Between OpenAI and Microsoft Azure OpenAI**](https://www.uscloud.com/blog/the-differences-between-openai-and-microsoft-azure-openai/#:~:text=Your%20data%20is%20your%20own,grow%20using%20advanced%20artificial%20intelligence.)

Hi ⁠M N, Mallik (NonEmp),

The ‘[COLOR\_NORMALIZED\_Home\_Decor\_Rugs\_12122023.csv’](https://adb-4812933386228410.10.azuredatabricks.net/files/tables/DATA_SCIENCE/COLOR/COLOR_NORMALIZED_Home_Decor_Rugs_12122023.csv) file is located at Databricks /dbfs/FileStore/tables/DATA\_SCIENCE/COLOR. Please create a power BI for it. Thanks.

“COLOR\_NORMALIZED\_Home\_Decor\_Rugs\_12122023.xlsx’ is the input for sampling. After Prateek completing the validation, I need to base on his findings to update the column ‘NORMALIZED\_COLOR’ for the xlsx file correspondingly and upload it to EISART -58990 (Extract … Rugs). Am I right?

12.14.23

My personal Azure account doesn’t have the Azure Subscription ID. To grant Azure OpenAi access, the Subscription ID is required. I do see the Subscription ID for my Azure Kroger account. Is it Ok I use the Kroger account for Quickstart: Get started generating text using Azure OpenAI Service?

I used Kroger Subscription ID and submitted the form ‘apply for access to Azure OpenAI’.

12.13.23

I am working on Lab 1.2. The cocd complained “KeyError: 'ELASTIC\_PROXY'”

load\_dotenv("env", override=True)

#if using the Elastic AI proxy, then generate the correct API key

if os.environ['ELASTIC\_PROXY'] == "True":

env

ELASTIC\_PROXY=True

12.12.23

The items in the list ['COPPER', 'STEEL','BRONZE', 'CHAMPAGNE','PEWTER', 'CHOCOLATE', 'CREAM']

are discarded from Kitchen& dinning. Shall I discard them from Color-rugs?

12.11.23

Hi Paul,

Thank you for the validation for Color – rugs normalization list.   Shall we omit 'LIGHT' normalization list? You are still the assignee for EISART-58990. GN!

[www.bing.com/chat](https://www.bing.com/chat)

[Yesterday 10:17 PM] Dua, Prateek (NonEmp)

The system is multimodal, meaning it can parse both images and text, whereas GPT-3.5 could only process text. This means GPT-4 can analyze the contents of an image and connect that information with a written question.

both words of questions and words of answers will be charged 'depeding length of answers/questions 'seperately\*\*\*\*

I got response from Paul. I am going to update the normalization list and run complete code of normalization. To generate the validation list shall I exclude marketplace?

12.8.23

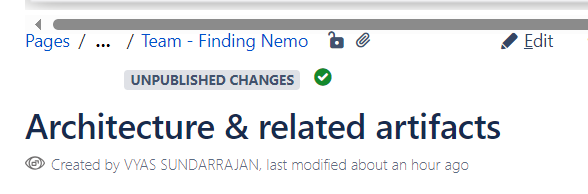
The CSV file didn't save colors. Attached please find the Normalized\_COLORS\_Rugs.xlsx.

The highlighted ones (52/100) are found to be ambiguous/error.  Please go over the file and let us know your feedback or adjustment required.

I am afraid you might need to redo “Normalized\_COLORS\_Rugs-paul-revised.csv”.

Hi ⁠Sundarrajan, Vyas,

Thank you for your section today and the section was informative. I browsed the chart in Confluence. Since the fonts are small, I used ‘Edit’ to browse the chart. I didn’t make any changes. However, you’ll see ‘UNPUBLISHED CHANGES’. Sorry about it and I’ll use CTRL+ to enlarge the fonts next time

C

Is Prateek  the original assignee of ticket 59095?

Hi Paul, I made a mistake to assign the ticket 59095 to you. Please ignore it.

My normalization file posted in Jira had excluded marketplace items.

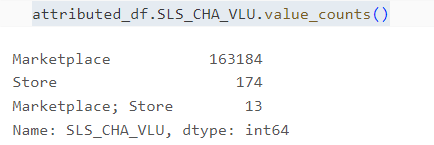
There are 187 non-marketplace items from the new run. The no is the same from yesterday.

Do we expect the number change?

I didn't do the filtering. I applied your code below and use valur\_counts() to 187 non-marketplace items for rugs.

mkt\_flg = pd.read\_csv('/dbfs/FileStore/tables/DATA\_SCIENCE/MARKETPLACE\_FLAG\_120723.csv')

attributed\_df = attributed\_df.merge(mkt\_flg, how = 'left', on = 'GTIN\_NO')



When I am waiting for Paul’s response, is anything I can do now?

12.6.23

Hi Dua, Prateek (NonEmp),

* Validation\_Data\_Color\_Rugs\_12062023.csv with 91 rows in /dbfs/FileStore/tables/DATA\_SCIENCE/COLOR/ is ready for validation.
* What is the ISART NO for color – rugs validation?
* Paul is reviewing the Rugs normalization. I’ll base on Paul’s suggest to changes the validation file.

12.5.23

There is no ‘ITM\_CFIC\_FYT\_DPT\_NAM == ‘Rugs’ in ‘CUSTOMER\_FACING\_ITEM\_DATA\_112823.zip'

Could you please check it?

Shall we use the same input for our Jira color assignment. Please review the SQL below and give me your feedback. I created a new column “COLOR\_VLU\_CD” which will make coding easier.

“After you review the SQL I’ll extract the data and upload to databricks.

* SQL to create “PIM\_Color\_Attributes\_120523.csv”

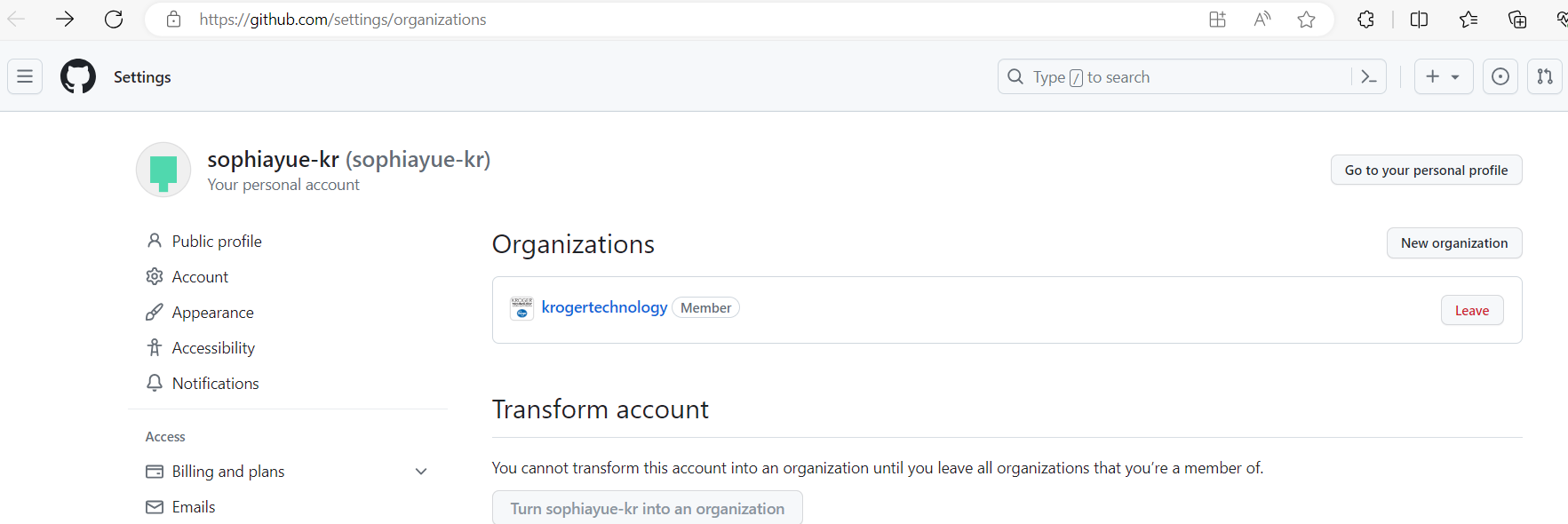
SELECT NAM\_IN\_KEY\_LNG\_NAM, ITM\_ID, ATB\_VLU AS 'COLOR\_VLU', ATB\_CD AS 'COLOR\_CD', COALESCE(ATB\_VLU, ATB\_CDS) as COLOR\_VLU\_CD

FROM CURIDQ.pimmart.CLS\_SPC\_ATB WHERE NAM\_IN\_KEY\_LNG\_NAM in ('Color', 'Other Color') order by ITM\_ID

* SQL to create “ITM\_ID\_GTIN\_Map\_120523.csv”

select GTIN\_NO, ITM\_ID FROM    CURIDQ.PIMMART.ITEM\_VIEW order by itm\_id

When I click the link you provided for Github Organization, I got the following screenshot. It looks like I am OK now. Am I right?



12.4.23

#### Are you the person to **invited you to join the @krogertechnology organization?**

#### **Jaden Zerbe**

Hi Team, we might use [Google Colab](https://colab.research.google.com/) in the training of “Fundamentals of AI and Language Models”.

You might set up a A Google account - to access Google Colab before the training.

11.24.23

Hi M N, Mallik (NonEmp), I  had updated <https://confluence.kroger.com/confluence/display/EIS/Finding+Nemo+-+Active+Work>. Please review my update and renumber Challenge, Assumption, and Methodology for me. It is for Attribute Normalization. Is it what you want me to do? Thanks!

Hi Chakraborty, Shiladitya (NonEmp),

* Please review the document in the link for Attribute Normalization as well. For your code ‘Attribute\_Normalization 2023-11-21’, it is case sensitive to create 'NORMALIZED\_COLOR' e.g, ‘BLUE’, ‘blue, ‘Blue’ are treaded as different attributes. Perhaps, the code might change ‘EXTRACTED\_COLOR’ and ‘normalized\_list’ to uppercase or lowercase and then checks if color in ‘EXTRACTED\_COLOR’ is in the normalized\_list.
* I reran your code and created “COLOR\_NORMALIZED\_HOUSEWARES\_11242023.csv” which contained ‘EXTRACTED\_COLOR’ and 'NORMALIZED\_COLOR'.

For some reason, the folder in GENERAL is empty and I was not able to save the file “COLOR\_NORMALIZED\_HOUSEWARES\_11242023.csv” in General. The file can be found in Databricks:'/dbfs/FileStore/tables/DATA\_SCIENCE/COLOR/'.

11.21.23

* Please refer to 'COLOR\_WAREHOUSE\_DPT\_112123.xlsx' which applied "matched = '|'.join(list(pd.Series(match\_list)))".
* The foloowing are examples in 'EXTRACTED\_COLOR'. They might need to tune.
  + RED|Red|Lollipop|rede
  + GREY|grey|Grey|Love|LOVE
* How do you filter out ‘Kitchen & Dining’ from 'COLOR\_WAREHOUSE\_DPT\_112123.xlsx'?

11.20.23

Could you please review the document I put in the code “PIM\_COLOR\_EXTRACTION\_111723.ipynb” and let me know if I can use it aa a template for “\_Flavor Identification\_ICECREAM.ipynb”?

I didn’t see ‘Kitchen & Dining’ or ‘Home Décor’ in the spreadsheet I created. I’ll investigate it. I used “PIM\_DATA\_W\_TAG\_DSC\_110323.zip“ as pim\_df. How did you create the file?

A: Those depts are from customer facting taxonomy, not KFT .

How many depts are from customer facting taxonomy? If I want to get color match for these dept, how do I do? Did you have a code which I can refer to?

A: We do not access to that taxonomy yet.  Kitchen & Dining is mostly included under 'Housewares' in KFT and Home Decor falls within Furn/ Home Décor

I’ll be OOO from Nov 23 to Dec 01. I’ll complete the document for “\_Flavor Identification\_ICECREAM.ipynb” before my vacation. Is anything else I can help before Nov 23?

11:19:23. Hi team,

* The “color\_match\_dpt\_dsc\_111723.xlsx” has color matched for all dpt.
* The ‘sumy’ summarized the ratio of other color, non other color, and run time for the matching process
* There are 95 DPT from colors\_DPT and there are 5 colors\_DPT either <OBSOLETE> or no observation in pim\_load. These DPT are excluded from matching process.
* The “PIM\_COLOR\_EXTRACTION\_111723.ipynb” is the code for the matching process. It used PIM\_Color\_Attributes\_111623 and ITM\_ID\_GTIN\_Map\_111623.csv as input. The length of color\_list is 6453. We might pass the items of color\_list from the code to Paul for normalization. I put detail comments in the very beginning.
* It took 20hh:5mm:9ss to run 90 DPT.

11.17.23

The pim\_dpt.MATCHING\_DESCRIPTION for dpt = <OBSOLETE> FOOD SERVICE

1300385 NaN

1300429 NaN

The job abended unless I remove “regex = True” from the following code. Is it ok to do so?

pim\_dpt.MATCHING\_DESCRIPTION.str.upper().replace(replacements, regex = True)

In your code “\_Flavor Identification\_ICECREAM.ipynb”, you based on the BASE FLAVOR CATEGORIES to construct Flavors\_combo. I am wondering how did you get the items for each category. E.g., How did you get the flavors for ‘NEW” category?

hs

All the categories are based on your analysis and knowledge m r

Where did you get the for you code “\_Flavor Igory?dentification\_ICECREAM.ipynb”

11.15.23

Hi team,

I uploaded the following articles in General under the folder “OpenAI + GenAI + LLM + GPT”:

Temperature and Top\_p in ChatGPT.docx

* This article is from Medium. It required a paid member to read a complete article. I download the article to word doc.
* The article is based on the model ‘gpt-3.5-turbo', and we are using "model": "Llama-2-70b-chat-hf" for LLaMA-text model.
* Shashank did a good job to explain Temperature and Top\_p. I added his explanation in appendix.

Fine-Tuning Large Large Models (LLMs).docx

* This article is from TDS. It required a paid member to read a complete article. I download the article to word doc.
* You might also read the article “Beginners Guide to Finetuning Large Language Models (LLMs).docx”.

ChatGPT API Using Python - Hanane D\_.pdf

11.13.23

You explained parameters of “temperature”, " and top\_p", and "max\_tokens" in LLaVa?

I am still not clear. Could you please elaborate them a little bit? Do we have document about them?

Hi team,

* ‘Flavor Match Process.docx’ is the document for rapidfuzz.process.extract**,**  and in-house function “flavor\_match”. Please browse/review of the document and let me know your feedback.
* The document is based on the code “\_Flavor Identification\_ICECREAM.ipynb” from Shiladitya. It’d be nice to have data-flow.
* I’ll see if I can use Visio or word to draw flowchart for data flow.

11.06.23

Is "NON\_IC\_FLAVOR\_MAPPING\_DEPT\_110623.xlsx what you are looking for rerun item-wise results?

Would you like me to use ‘PIM\_NON\_ICE\_CREAM\_FLV\_MAPPING\_110323.ipynb’ as a template to get the flavor for

h

11.03.23

* NON\_IC\_FLAVOR\_MAPPING\_DEPT\_110323.xlsx includes DEP\_DSC in [BABY, BAKERY,BEAUTY,BEEF,BEER,CANDY, CKY/CRKR/SNK]. Please review it and let me know your feedback.
* The code used “PIM\_DATA\_WITH\_FLAVOR\_ATTB\_102423.xlsx” and the most current match process.
* The performance is slow.

11.02.23-1

* I am able to use Falvor\_code and Flavor\_value and Flavor Other for the matching process.
* I got 13570/19978 ‘OTHER’ and you got 12692/19978 OTHER for “COFFEE SHOP”. I expected to get less OTHER than yours.
* I used ‘ITM\_ID\_GTIN\_Map\_100323.csv’ instead of using 'ITM\_NO\_GTIN\_100323.csv'.
* I am trying to investigate why I got more ‘OTHER’.
* Could you please upload 'ITM\_NO\_GTIN\_100323.csv'? I’ll see what could be if I use this file.

11.02.23-1

* The file 'ITM\_NO\_GTIN\_100323.csv' didn’t change the result. I ran your code with the change of file locations, I got 13731/19978 ‘OTHER’ in ‘Detected\_Flavor(s)’ which is the same result for combining Falvor\_code and Flavor\_value.
* After including ‘Flavor Other’, it reduced 161 ‘OTHER’ from ‘Detected\_Flavor(s)’.
* There is no missing value for Flavoe\_Code in PIM\_DATA\_WITH\_FLAVOR\_ATTB\_100323.csv with shape (41593, 17)
* There are many versions of “PIM\_DATA\_WITH\_FLAVOR\_mmddyy. All of them don’t have missing value for Flavoe\_Code except the PIM\_DATA\_WITH\_FLAVOR\_ATTB\_102423.xlsx with 5767/60340 missing Flavor\_Code.
* Is it a good idea to replace ‘PIM\_DATA\_WITH\_FLAVOR\_ATTB\_100323.csv’ with PIM\_DATA\_WITH\_FLAVOR\_ATTB\_102423.xlsx?
* Do we need to do ITM\_NO\_GTIN mapping for GTIN\_NO in 'PIM\_Data\_New\_50\_82Mn.zip’?
* I’ll try ‘Bakery’ Dept next and then try all the DPTs.

The ‘replacements’ and ‘Flavors\_combo’ in New Flavor Identification\_ICECREAM\_102723.ipynb are for ice cream. I’ll not use them .

11.01.23y all

* I didn’t see ‘Other Flavor’ been used for matching process. Perhaps “flavor\_list” needs to include ‘Other Flavor’
* For the following codes, the code get shape meeting criteria and pim\_load didn’t change. Is that what you want to do?
  + pim\_load[(pim\_load.COM\_CD.isin(flavors\_COM.COM\_CD.unique())) & (~pim\_load.Flavor\_Code.isin(['NONE']))].shape
  + pim\_load[~pim\_load.Flavor\_Code.isin(['NONE','OTHER','MISECLLANEOUS'])].shape

Below is the test code to combine ['Flavor\_Value'] and ['Flavor\_Code']. I got the error message: “**ValueError**: The truth value of a Series is ambiguous. Use a.empty, a.bool(), a.item(), a.any() or a.all().” Do you have similar code to check missing value? Do you know how to resolve the issue? Or if you have code to archive the purpose, please share?

pimdpt[ 'cmb\_flav\_val\_cd'] = pimdpt.apply( lambda x: x['Flavor\_Value'] if pimdpt['Flavor\_Value'] != np.NaN else pimdpt['Flavor\_Code'])

10.31.23

You want me to use the code “PIM\_NON\_ICE\_CREAM\_102423\_v2\_yue.ipynb” as a template to create files “FLAVOR\_MAPPING\_'+dpt+'\_DEPARTMENT.xlsx” for different DPT. Is it correct?

Thx. I got the message " flavors\_DPT\_df is not defined”? Please check your code where to create the DF?

Is ITM\_ID\_GTIN\_Map\_100323.csv the same as ITM\_NO\_GTIN\_100323.csv? if not, please upload ‘ITM\_NO\_GTIN\_100323.csv’or the most current version.

10.30.23

Please use the following code to ignore or prevent the SettingWithCopyWarning to be printed:

import warnings

from pandas.errors import SettingWithCopyWarning

warnings.simplefilter(action="ignore", category=SettingWithCopyWarning)

⁠Chakraborty, Shiladitya (NonEmp)

* ICE\_CREAM\_FLAVORS\_RMV\_VND\_BRN\_NAM102723\_v5.xlsx is the file to remove RMV\_VND\_BRN\_NAM to get flavor. It has the following tabs:
* The following tabs are results of removing VND\_BRN\_NAM
  + Sngle\_Flavors
  + 2\_Flavors
  + 3\_plus\_\_Flavors
  + OTHERs
* The following tabs are results of comparision: The suffix '\_x' is the one to remove VND\_BRN\_NAM. The suffix '\_y' is the one not to remove VND\_BRN\_NAM.
  + com 1 flav
  + cmp 2 flav
  + cmp 3\_plus flav
  + cmp OTHERs
  + cmp flavor all
* The code “New Flavor Identification\_ICECREAM\_RMV\_VND\_BRN\_NAM\_102723.ipynb “ used your code ‘RMV\_VND\_BRN\_NAM102723’ updated on oct.28 as a template.
* It used the file “ICE\_CREAM\_FLAVORS\_102723\_v5.xlsx “, you created on Oct 28, to do the flovor comparison.
* Please review the file and let me know your feedback.

10.27.23

⁠Chakraborty, Shiladitya (NonEmp)

In the code ‘New Flavor Identification\_ICECREAM\_102523.ipynb’, is any reason, you comment out process.extract for scorer = fuzz.token\_set\_ratio and fuzz.partial\_ratio?

* I am going to use “New Flavor Identification\_ICECREAM\_102523.ipynb” as a template VND\_ECOM\_DSC, TAG\_DESCRIPTION, and PRODUCT\_DESCRIPTION/MKT\_MSG and re-do flavor match.
* Do you have new version of New Flavor Identification\_ICECREAM\_102523.ipynb?

| Is any reason, you comment yoy wm | , |  |  |
| --- | --- | --- | --- |

You commented out

10.26.23

* The file ‘ICECREAM\_BRN\_NAM\_MCH\_102723.xlsx’ was created based on the request from Shiladitya ‘Please try to use 'VND\_BRN\_NAM' to  see if flavor match can be improved by removing the brand -specific words from the description before matching’ .
* The file had VND match score. I completed the VND matching score. We need to base on the score to determine removing the VND name from VND\_ECOM\_DSC and get the flavor.
* The code ‘ICE\_FLAVORS\_RMV\_VND\_BRN\_NAM\_101623.ipnnb’, ‘com\_code.py’ and the file ICECREAM\_BRN\_NAM\_MCH\_102723.xlsx’ are uploaded to General.
* Please take a look of the file and let me know your feedback.

ICE\_CREAM\_FLAVORS\_102423.xlsx is the input to create ICE\_CREAM\_FLAVORS\_102523.xlsx with flavor\_cnt. Please refer to the following tabs:

1)  IC Flavor & count = 1: flavor\_cnt= 1

2)  IC Flavor & count = 2: flavor\_cnt=2

3) IC Flavor & count > 2: flavor\_cnt>2

4) IC Flavor = OTHER : 'Flavor' = 'OTHER'

Data Cleaning

Create a new column ’ flavor\_mod ‘ from ‘Flavor’ with following modification and use flavor\_mod to do the flavor count.

* Replacing 'PEANUT BUTTER' with 'PEANUTBUTTER'
* Replacing ' AND' with ''.
* Replacing ' &' with ''

Is anything I can do/help for flavor analysis?

* Could you please upload ‘Multiflavor\_Maps.xlsThere is no ITM\_ID in 'ICE\_CREAM\_FLAVORS\_102423.xlsx’. After I get the most current GTIN\_NO and VND\_BRN\_NAM  from PIMMART.ITEM\_VIEW. How do I map them to ICE\_CREAM\_FLAVORS\_102423.xlsx’?
* Or you want to keep a separate file for the most current GTIN\_NO and VND\_BRN\_NAM  from PIMMART.ITEM\_VIEW.

x’ to General? thx

Which code would be the best reference/template for the new assignment?

10.25.23

I am trying to remove adj from ‘Flavor’ to count the number of flavor. For Flavor = ‘CLASSIC’, or Flavor = ‘HALF BAKED’, may I change it to ‘OTHER’. -> No

import pypyodbc as odbc

DRIVER\_NMAE = 'SQL SERVER'

SERVER\_NAME = '11-eastus2-3534-idq-sqlmi.cc4c669dc8d1.database.windows.net'

DATABASE\_NAME = 'CURIDQ'

connection\_string = f"""

DRIVER = {{{DRIVER\_NMAE}}};

SERVER = {SERVER\_NAME};

DATABASE= {DATABASE\_NAME} ;

ODBC;DSN=jg\_report;

Trusted\_Connection = yes;

"""

conn = odbc.connect(connection\_string)

print(conn)

I got the following message:

Error: ('IM002', '[IM002] [Microsoft][ODBC Driver Manager] Data source name not found and no default driver specified') Error: ('IM002', '[IM002] [Microsoft][ODBC Driver Manager] Data source name not found and no default driver specified')

10.24.23

Do you have a new version of ‘New Flavor Identification\_ICECREAM\_101223.ipynb’?

The file ‘CO\_CMD\_WITH\_FLAVOR\_ATTB\_102423.xlsx’ was created based on the request from Shiladitya,

1)  The unique COM\_CD for all items that have 'Flavor' and 'Other Flavor' available in CLS\_SP\_ATB

2) The total number of PIM items that fall within the above COMs

**CO\_CMD\_WITH\_FLAVOR\_ATTB\_102423.xlsx’**

* CO\_CMD\_WITH\_FLAVOR\_ATTB\_102423.xlsxwith the following tabs
  + ‘Flavor\_GTIN’ is the same as PIM\_DATA\_WITH\_FLAVOR\_ATTB\_102423.xlsx
  + ‘CO\_CMD\_Flavor\_GTIN\_Count’ with COM\_CD, GTIN\_count, no of flavor and list of flavors.
  + ‘SUMY NONE+OTH+MISC’: summary for CON\_CD containing COM\_CD with the flavor value/code containing ‘NONE, MISC ELLANEOUS, OTHER’ only.
* PIM\_DATA\_WITH\_FLAVOR\_ATTB\_102423.xlsx & CO\_CMD\_WITH\_FLAVOR\_ATTB\_102423.xlsx had been upload to General.

10.23.23

* PIM\_DATA\_WITH\_FLAVOR\_ATTB\_102323.xlsx is the file with ATB Flavor and GTIN information.
* There are 251 rows/60340 rows with missing COM\_CD and 32 rows/251 rows without missing Flavor value/code.
* Total number of COM\_CD is 403. 115, 336, 886 are three COM\_CD doesn’t have Flavor value/code.
* There are 5579 rows/60340 rows with missing Flavor value/code.
* There are 10402 rows/60340 rows with Flavor value/code = NONE.
* There are 4044 rows/60340 rows with Flavor value/code = MISCELLANEOUS.
* There are 3109 rows/60340 rows with Flavor value/code = OTHER.
* I’ll see if I can filter out COM\_CD with the flavor value/code containing ‘NONE, MISC ELLANEOUS, OTHER’ only.

,

There are 251 rows/60340 rows with missing COM 

10.20.23

Per my test, the fuzz.WRatio will get better score than fuzz.token\_set\_ratio. You might try fuzz.WRatio.

The function flavor\_matchx() is the revised of flavor\_match() with scorer as a variable.

Do you have a new version of ‘New Flavor Identification\_ICECREAM\_101223.ipynb’ which showed OTHER = 612? The code I have is from oct13.

* ICECREAM\_OTHER\_NEW\_DETECTED\_FLAVORS\_102023.xlsx is the revised version of ICECREAM\_OTHER\_NEW\_DETECTED\_FLAVORS\_101923.xlsx with the following modification.
  + Added 'PEPPER'] on Spices in Flavors\_combo
  + Spices = ['GINGER','CARDAMOM','CINNAMON','NUTMEG','TURMERIC','ANISE','ALLSPICE','SPICE', 'PEPPER']
  + After handle B&J and Blue Bunny,  apply flavor\_list = Flavors\_combo instead of flavor\_list = IC\_flavors\_new for scoring.
  + df\_oth\_res1: Get 'New\_Detected\_Flavor(s)'   with cutoff = 98 and 'New\_Detected\_Flavor(s)' != ‘OTHER’
  + df\_oth\_res2: Get 'New\_Detected\_Flavor(s)' with cutoff = 50  for  all the GTIN with 'New\_Detected\_Flavor(s)'  not been classified
  + The Icecream Flavor Extraction Synchup\_102023.ipynb is the code

Comments:

* There are 1066 rows from 1837 rows with ‘Detected\_Flavor(s)’ = ‘OTHER’ remains ‘OTHER’ in New\_Detected\_Flavor(s). 771 rows  are able to identify  New\_Detected\_Flavor(s). Shall we add  ‘COM\_DSC’ and ‘SUBCOM\_DSC’ for New\_Detected\_Flavor(s) ?
* Please check if it make sense for New\_Detected\_Flavor(s) != ‘OTHER’  with cutoff = 50’.
* New\_Detected\_Flavor(s) for Blue Bunny doesn't look great. Perhaps we can do some enhancement.
* Can we treat Blue Bell as Blue Bunny?

* ICECREAM\_OTHER\_NEW\_DETECTED\_FLAVORS\_102023.xlsx is the revised version of ICECREAM\_OTHER\_NEW\_DETECTED\_FLAVORS\_101923.xlsx with the following modification.
  + Added 'PEPPER'] on Spices in Flavors\_combo
  + Spices = ['GINGER','CARDAMOM','CINNAMON','NUTMEG','TURMERIC','ANISE','ALLSPICE','SPICE', 'PEPPER']
  + After handle B&J and Blue Bunny,  apply flavor\_list = Flavors\_combo instead of flavor\_list = IC\_flavors\_new for scoring.
  + df\_oth\_res1: Get 'New\_Detected\_Flavor(s)'   with cutoff = 98 and 'New\_Detected\_Flavor(s)' != ‘OTHER’
  + df\_oth\_res2: Get 'New\_Detected\_Flavor(s)' with cutoff = 50  for  all the GTIN with 'New\_Detected\_Flavor(s)'  not been classified
  + The Icecream Flavor Extraction Synchup\_102023.ipynb is the code

Comments:

* There are 1066 rows from 1837 rows with ‘Detected\_Flavor(s)’ = ‘OTHER’ remains ‘OTHER’ in New\_Detected\_Flavor(s). 771 rows  are able to identify  New\_Detected\_Flavor(s). Shall we add  ‘COM\_DSC’ and ‘SUBCOM\_DSC’ for New\_Detected\_Flavor(s) ?
* Please check if it make sense for New\_Detected\_Flavor(s) != ‘OTHER’  with cutoff = 50’.
* New\_Detected\_Flavor(s) for Blue Bunny doesn't look great. Perhaps we can do some enhancement.
* Can we treat Blue Bell as Blue Bunny?

‘ ‘bn

* ICECREAM\_OTHER\_NEW\_DETECTED\_FLAVORS\_102023.xlsx is the revised version of ICECREAM\_OTHER\_NEW\_DETECTED\_FLAVORS\_101923.xlsx with the following modification.
  + Added 'PEPPER'] on Spices in Flavors\_combo
  + Spices = ['GINGER','CARDAMOM','CINNAMON','NUTMEG','TURMERIC','ANISE','ALLSPICE','SPICE', 'PEPPER']
  + After handle B&J and Blue Bunny,  apply flavor\_list = Flavors\_combo instead of flavor\_list = IC\_flavors\_new for scoring.
  + df\_oth\_res1: Get 'New\_Detected\_Flavor(s)'   with cutoff = 98 and 'New\_Detected\_Flavor(s)' != ‘OTHER’
  + df\_oth\_res2: Get 'New\_Detected\_Flavor(s)' with cutoff = 50  for  all the GTIN with 'New\_Detected\_Flavor(s)'  not been classified
  + The Icecream Flavor Extraction Synchup\_102023.ipynb is the code

Comments:

* There are 1066 rows from 1837 rows with ‘Detected\_Flavor(s)’ = ‘OTHER’ remains ‘OTHER’ in New\_Detected\_Flavor(s). 771 rows  are able to identify  New\_Detected\_Flavor(s). Shall we add  ‘COM\_DSC’ and ‘SUBCOM\_DSC’ for New\_Detected\_Flavor(s) ?
* Please check if it make sense for New\_Detected\_Flavor(s) != ‘OTHER’  with cutoff = 50’.
* New\_Detected\_Flavor(s) for Blue Bunny doesn't look great. Perhaps we can do some enhancement.
* Can we treat Blue Bell as Blue Bunny?

Please check New\_Detected\_Flavor(s) for cutoff= 50

I expect ‘Dozen Yellow Roses’ would be classified as ‘Roses’ but CREAM PUFF,SOUR CREAM (cutoff= 50)

10.19.23

* CECREAM\_OTHER\_NEW\_DETECTED\_FLAVORS\_101923.xlsx is the file with New\_Detected\_Flavor(s) for the GTIN been classified as  'OTHER' in Detected\_Flavor(s) from ICECREAM\_OTHER\_FLAVORS\_101323.csv.
* The tab2 ‘New\_Detected\_Flavor’ with 1672/1837 as ‘OTHER’. 165 rows were able to identify the flavor.
* The tab2 New\_Detected\_Flavor\_Res applied extract\_bnj\_flavor to the ‘OTHER’ from tab2.
* The result from tab2 doesn’t look good. Please let me know if you have any idea to improve it.
* “Icecream Flavor Extraction Synchup\_101923.ipynb” is the code.

2

10.17.23

(Q). Can waffle be classified as a ice cream flavor?

(Q). At the meeting, what did 8000 refer to?

(A). Using  the mappings in  'Syndigo Taxo to GS1 Taxo.xlsx'

Around  82000 of the 5M items in  ''Syndigo\_Levels\_L1\_L2\_L3.csv'

can be mapped to unique GS1 Brick based on 'Level2'

#path\_file = path + 'Parts 1-2-3-4-5 Syndigo matches w taxonomy .xlsx'

* 'Parts 1-2-3-4-5 Syndigo matches w Syndigo Taxonomy.xlsx' is about the same as “'Parts 1-2-3-4-5 Syndigo matches w taxonomy .xlsx' “ which we received is the past.
* The file “syndiago Parts 1-2-3-4-5 matches w taxonomy classification\_101723.xlsx” with 164053 rows. The file concatenates Part 1 ~ Part 5 with level classification.

Could you please forward Deevraj’s meeting invitation to me? For some reason, I don’t have it?

Did you show me the meeting invitation from the Deevraj’s at the team meeting? That is what I refer to. I think the meeting is 1:PM (EST)

10.13.23

I tried to run your code ‘‘Flavor Code Attribution\_IceCream\_100623.ipynb’’. The file ‘PIM\_Data\_New\_50\_82Mn.csv’ doesn’t have ‘ITM\_LVL\_DSC’. Perhaps, ‘PIM\_Data\_New\_50\_82Mn.zip’ has ‘ITM\_LVL\_DSC’. Could you please upload ‘PIM\_Data\_New\_50\_82Mn.zip’ to General? thx

[3:09 PM] Yue, Sophia (NonEmp)

Thank you for joining my performance review meeting and nice comments about me.

Could you please upload ‘icecream\_additional\_ITEM\_VIEW.csv’ to General? thx

10.12.23-1

* There are 3444 rows from the file ‘ICECREAM\_FLAVORS\_101123\_comma\_delim.csv‘ with flavor = ‘OTHER’
* I used ‘4.99, 5.99, Cookies, CSTN, frozen, gelato, Fudge bar, MILKSHAKE, sherbet, Sorbet, TOY, Yogurt’ as keyword to scan VND\_ECOM\_DSC and identify rows can be removed.
* I identified ‘gelato, sherbet, Sorbet’ are not ice cream.
* From the remaining rows, there are 1389 rows with ‘ice cre’ in VND\_ECOM\_DSC and 1392 row without ‘ice cre’ in VND\_ECOM\_DSC which can be removed.
* There are 1399 rows need do further investigation if we can identify flavor other than ‘OTHER’.
* Please refer to ‘ICECREAM\_FLAVORS\_101123\_comma\_delim\_101223.xlsx’

10.12.23-0 can

The following code is from your code ‘Flavor Code Attribution\_IceCream\_100623.ipynb’

pim\_df = pd.read\_csv(path + 'PIM\_Data\_New\_50\_82Mn.zip')

pim\_load = pd.read\_csv(pth + 'PIM\_DATA\_WITH\_FLAVOR\_ATTB\_100323.csv')

pim\_load = pim\_load.merge(pim\_df, how = 'right', on = list(pim\_df.columns))

'PIM\_DATA\_WITH\_FLAVOR\_ATTB\_100323.csv' has combined PIM data mode

Where can I get ‘ice\_cream\_items\_org.csv’ and ‘PIM\_Data\_New\_50\_82Mn.zip’?

Do you mean ice\_cream\_items\_org.csv’ is the same as ice\_cream\_items.csv’

* PIM\_DATA\_WITH\_SCENT\_ATTB\_101123.xlsx is the file of to combine SCENT from CLS\_SPC\_ATB with GTIN information in PIM\_Data\_New\_50\_82Mn.csv & ITM\_DSC\_ATTRIBUTES\_NEW\_parquet.gzip
* Used ITM\_ID\_GTIN\_Map\_101123.csv to map GTIN\_NO to TIM\_ID.
* The ITM\_ID (794780727, 794786276, 794791871,794793081,794794421) in PIM\_DATA\_WITH\_SCENT\_ATTB\_101123.xlsx can’t find matching GTIN\_NO. Perhaps, we need to refresh PIM\_Data\_New\_50\_82Mn.csv & ITM\_DSC\_ATTRIBUTES\_NEW\_parquet.gzip
* After merging, the type of GTIN\_NO changed from int64 to float64. It was caused by missing GTIN\_NO. I manually changed the format of TIM\_ID in the xlsx file.
* The code PIM\_DATA\_WITH\_ATTB\_101123 is flexible to handle different attributes, e.g., scent and flavor.

10.11.23

If you use fuzz for scoring, you might add the parameter ‘processor=utils.default\_process’ in fuzz.token\_set\_ratio to improve score. E.g.,

score = fuzz.token\_set\_ratio(pim\_match.PIMMART\_Matching\_Text.iloc[pim\_ix],j, processor=utils.default\_process)

“

My pleasure. Indeed, I did sahre “The match score did improve by added parameter ‘processor=utils.default\_process’ to fuzz.token\_set\_ratio.” with team in Jun.

10.10.23

For my code, some of input our files use GTIN and some use GTIN\_NO, which one would be our preference for output file?

For the ouytpurt GTIN\_NO and GTIN )

Please review Scent\_CLS\_SPC\_ATB.xlsx in general. It has 170 rows. Do you want me to add PIM\_DATA like I did for 'PIM\_DATA\_WITH\_FLAVOR\_ATTB\_100323.csv'?

Is anything I can do/help for ice cream flavor analysis?

* Please review the file ice\_cream\_items\_pim\_100923.xlsx in General. I got the required information from ITM\_DSC\_ATTRIBUTES\_NEW\_parquet.gzip and PIM\_Data\_New\_50\_82Mn.csv.
* There are 5737/12386 in ice\_cream\_items.csv with GTIN\_NO match with PIM\_Data.
* There are 4 Unnamed column from ice\_cream\_items.csv. The 'Unnamed: 12'(5 not missing), 'Unnamed: 13' (2 not missing),,'Unnamed: 14'(all missing), 'Unnamed: 15'(all missing). 10.09.23
* Python treats (null) as non-missing

PIM\_DATA\_WITH\_FLAVOR\_ATTB\_100923.xlsx is the file to add the 'Other Flavor'  field from CLS\_SPC\_ATB into 'PIM\_DATA\_WITH\_FLAVOR\_ATTB\_100323.csv'.fuzzy

have flavor fro

10.06.23

* There are 4972 rows in **CLS\_SPC\_ATB WHERE** NAM\_IN\_KEY\_LNG\_NAM = 'Other Flavor'
* All the 'Flavor\_Code' are missing and there are 188 rows where 'Flavor\_Value' with values.
* When you build ‘PIM\_DATA\_WITH\_FLAVOR\_ATTB\_100323.csv', you might need to consider NAM\_IN\_KEY\_LNG\_NAM = 'Other Flavor' as well.

Where can I find PIM\_Data\_New\_50\_82Mn.zip? I can find PIM\_Data\_New\_50\_82Mn.csv but not PIM\_Data\_New\_50\_82Mn.zip in Databricks.

10.05.23

How do you build ‘PIM\_DATA\_WITH\_FLAVOR\_ATTB\_100323.csv'? Did you write a code to build it?

[Yesterday 6:00 PM] Chakraborty, Shiladitya (NonEmp)

1. SELECT ITM\_ID, ATB\_CD AS 'Flavor\_Code', ATB\_VLU AS 'Flavor\_Value'

FROM CURIDQ.pimmart.CLS\_SPC\_ATB

WHERE NAM\_IN\_KEY\_LNG\_NAM = 'Flavor(s)'

1. then join with 'PIM\_Data\_New\_50\_82Mn.zip'  using ITM\_ID

Flavor Distribution Counts

10.04.23

What is the input for GROCERY\_COMs\_for\_Flavor\_Attribution.csv?

'PIM\_Data\_New\_50\_82Mn.zip'

Is the 'PIM\_Data\_New\_50\_82Mn.zip'  the input for PIM\_DATA\_WITH\_FLAVOR\_ATTB\_100323.csv? Where did you get Flavor\_Value and Flavor\_ode?

he same file and same filtering  -

PIM\_DATA\_WITH\_FLAVOR\_ATTB\_100323.csv

Article from Confluence

* [**RNN Traffic**](https://confluence.kroger.com/confluence/display/SCIEN/RNN+Traffic)
* [**RNN Model**](https://confluence.kroger.com/confluence/display/8451SE/RNN+Model)

Article from website

* [Introduction to Recurrent Neural Network](https://www.geeksforgeeks.org/introduction-to-recurrent-neural-network/)

10.03,23

Hi, For my YE Evaluation, I asked Deevraj to be my Evaluation Manager. Would you mind being the Reviewer for me?

I think it is another way around. How about I put your name as reviewer, if the system complains, I’ll put Deevraj as my reviewer?

hithe reviewer has to be someone more senior like someone's Deevraj reports to.

10.03.23

I uploaded [ACDS Introduction.pdf](https://confluence.kroger.com/confluence/display/8451MERCH/3CM+Tools%2C+Libraries%2C+and+Sources+Overview?preview=%2F392431509%2F392442815%2FACDS+Introduction.pdf) which is informative from Confluence to General -UOM. Hopefully, we can get access to ACDS in a near future.

# [ACDS Certification - Process Overview](https://confluence.kroger.com/confluence/display/8451EC/ACDS+Certification+-+Process+Overview) from [krogerpedia](https://krogerpedia.kroger.com/asset/39567e66-dd5e-4aa7-90d6-539c317443f8?type=dataset2) is also a good reference.

FYI: I am heading Infosys Phoenix office to get some stationary, e.g., print paper.

Hi Deevraj,

Is it ok you would be my review manager for YE evaluation? Please read my emailed at Infosys. Thx.

Chakraborty, Shiladitya (NonEmp)f? have comments regarding the observations from Mallik’s PP for GTIN – 250810000008. Per < 0 then haps we can check the observations with descriptions in the dictionary.

1. Could we ask Graig why scn\_unt\_qy,cly\_dol\_am, rtl\_dol\_am, and  net\_spend\_amt having Value < 0 in  acds\_transactions?
2. Please read my email for the details.

What is the product /item name for gtin\_no = 94011 in acds\_transactions. There is a scenario that gtin\_uom\_am = 1.31 and cly\_dol\_am = -0.9. I assume if cly\_dol\_am < 0 then gtin\_uom\_am < 0.

9.28.23

For your’ file ‘[Syndigo\_Level1\_GS1\_Segment\_Mapping\_092823.csv](https://kproductivity.sharepoint.com/:x:/r/teams/MerchandisingDataPlatform599/Shared%20Documents/General/Syndigo_Level1_GS1_Segment_Mapping_092823.csv?d=wb73c522204b8447ba4f93e91ca776929&csf=1&web=1&e=ptRsEL)’, did you use ‘Syndigo\_Final\_ALL.csv’ for Syndigo Level 1? Which file did you get ‘Corresponding GS1 Segments’

Is the file ’GPC as of November 2022 v20221212 EN.xlsx’ created by our team?

Chakraborty, Shiladitya (NonEmp)

Some of ‘SegmentTitle/Corresponding GS1 Segments’ in GPC mapped more than one L1 in Syndigo.

E.g., the ['Food/Beverage/Tobacco'] from ‘SegmentTitle/Corresponding GS1 Segments’ in GPC mapped 3 L1 in Syndigo:

* Food / Beverages
* Beer / Wine / Spirits
* Tobacco Products

We might need to consider/incorporate the following for the mapping:

* ‘ClassTitle – second level of GPC
* ‘FamilyTitle’ - second level of GPC. F
* ‘BrickTitle’ – the lowest level of GPC.

E.G., for row no 25778 in [GPC as of November 2022 v20221212 EN.xlsx](https://kproductivity.sharepoint.com/:x:/r/teams/DataanalysisCoreTaxonomy/Shared%20Documents/General/May/GPC%20as%20of%20November%202022%20v20221212%20EN.xlsx?d=w20e52de4c3e7451c9ae7fa57743653fa&csf=1&web=1&e=AWkQwy), the ‘SegmentTitle/Corresponding GS1 Segments’ is ['Food/Beverage/Tobacco'] .

* The ‘ClassTitle’ Alcoholic Beverages (Includes De-Alcoholised Variants)
* The ‘FamilyTitle’ is Beverages
* The ‘BrickTitle’ is Beer

9.28.23-1

Is anything I can help for the analysis of GIS data? ‘KWIKEE’ which has GTIN and GPC

I was searching GPC from Confluence and I found the article [Kwikee](https://confluence.kroger.com/confluence/display/DSE/Kwikee) which has GTIN and following GPC items in Kwikee database. The article has Python code to retrieve data from Kwikee. The article was published by  [DAVID BELAIS](https://confluence.kroger.com/confluence/display/~DB55644), on  [May 31, 2017](https://confluence.kroger.com/confluence/pages/diffpagesbyversion.action?pageId=63971930&selectedPageVersions=52&selectedPageVersions=53). David looks like not working for Kroger anymore. I am not sure Kwikee database is still around and how to apply for access.

* **gpc\_attributes\_assigned** (datetime)
* **gpc\_brick\_id** (str): Brick is the lowest-level of Global Product Classification (GPC). The brick ID is an 8-digit string composed of integers, representing a unique ID for the classification. For more information, see: <http://www.gs1.org/gpc>.
* **gpc\_brick\_name** (str): Brick is the lowest-level of Global Product Classification (GPC). For more information, see: <http://www.gs1.org/gpc>.
* **gpc\_class\_id** (str): Class is the second-lowest-level of Global Product Classification (GPC). Class ID is an 8-digit string composed of integers, representing a unique ID for the The classification. For more information, see: <http://www.gs1.org/gpc>.
* **gpc\_class\_name** (str): Class is the second-lowest-level of Global Product Classification (GPC). For more information, see: <http://www.gs1.org/gpc>.
* **gpc\_family\_id** (str): Family is the third-lowest-level of Global Product Classification (GPC). The family ID is an 8-digit string composed of integers, representing a unique ID for the classification. For more information, see: <http://www.gs1.org/gpc>.
* **gpc\_family\_name** (str): Family is the third-lowest-level of Global Product Classification (GPC). For more information, see: <http://www.gs1.org/gpc>.
* **gpc\_segment\_id** (str): Segment is the highest level of Global Product Classification (GPC). The segment ID is an 8-digit string composed of integers, representing a unique ID for the classification. For more information, see: <http://www.gs1.org/gpc>.
* **gpc\_segment\_name** (str): Segment is the highest level of Global Product Classification (GPC). For more information, see: <http://www.gs1.org/gpc>.

ind

9.27.23

There are 3 access roles for Productboard in IAM. Which one shall I apply?

1. Productboard Contributor
2. Productboard Maker
3. Productboard Viewer

9.26.23

1. The file ‘acds\_trans\_gtin\_agg\_col\_val\_092623.xlsx’ in General have a list of unique values for all columns in ‘acds\_transactions.xlsx’ and 'gtin\_agg.xls'
2. There are 21 GTIN in in ‘acds\_transactions.xlsx’ and can’t find match the GTIN with ‘[Active\_GTINs\_chunked.zip](https://kproductivity.sharepoint.com/:u:/r/teams/DataanalysisCoreTaxonomy/Shared%20Documents/General/Active_GTINs_chunked.zip?csf=1&web=1&e=G7mcBS)’, or ‘PIM\_Data\_New\_50\_82Mn.csv’, or 'Syndigo\_Final\_ALL.csv'

The zip file has 5 files. What shall I change in “pd.read\_csv("Active\_GTINs\_chunked.zip", compression="zip") “ to read the individual file, e.g. Active\_GTINs\_1.csv ?

Do you have an example to read the individual file from the winzip file

09.25.23

1. There are 10 CSV files in "/dbfs/FileStore/tables/DATA\_SCIENCE/" started with “GS1”. The notebook “GS1 files in Databricks.ipynb” has information about them.
2. The code “FuzzyMapping\_Initial.ipynb” used 'GS1\_FoodBev\_2022.csv' with shape = (135,600, 19) to do the fuzzy prediction.

09.24.23

* Based on the sheet  'MLP\_SMOTE\_test.xlsx', the file “Level1\_MLP\_SMOTE\_Optimal\_Thresholds\_092423.csv” was created to get the accuracy by using threshold values [0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 0.96, 0.97, 0.98,0.99].
* Accuracy is the proportion of “TRUE” samples to all samples - only considering results where score >= threshold.
* The file and the code and the code “Level1\_MLP\_SMOTE\_Optimal\_Thresholds.ipynb” had been uploaded into General.

09.22.23

Please take a look of the file “Level1\_MLP\_SMOTE\_Optimal\_Thresholds\_092223.csv”and the code “Level1\_MLP\_SMOTE\_Optimal\_Thresholds.ipynb” in General and let me know if it is what you want me to do.

I created df1 by filtering Score >= thereshold from the input

I ran “clust\_itm\_atb.dtypes['ClusterLabel'] “ in Databricks and I got “Out[13]: dtype('O')”

09.20.23

synd\_ALL = pd.read\_csv(DBFR + 'Syndigo\_Final\_ALL.csv', dtype='unicode')

have mixed types. Specify dtype option on import or set low\_memory=False.

9.19.23

Hi

* At one of meetings with Samiran, you showed us the notebook of Cluster\_Attribute\_Distance\_081423.ipynb.
* The code based on “Non-Null Attributes” to create a new column “'AttributeVector” with all the value =0 or 1 in a list.
* The 'AttributeVector' would create an array ‘attr\_vects’ which would be used to build model.
* After clustering, it’ll be nice to map the number to description.
* I didn’t follow well for the code at the meeting. Per my observation, Samiran got confused as well.

The code Cluster\_Attribute\_Distance\_hdb\_kmeans\_091923.ipynb is cloned from Cluster\_Attribute\_Distance\_081423.ipynb and made the following modifications

1. Add more comments
2. After clustering, mapped the clusters which were number to description.
3. Modified the function “cluster\_attribute\_overlap “ to calculate the overlap
4. Apply pyplot which has moose over function to plot.

9.15.22

1. The ML\_Classifier\_Comparision\_091523.xlsx is the modified version of ML\_Classifier\_Comparision\_091423.xlsx with a new tab ‘LR\_MLPSMOTE weighted difference’.
2. For each class, the weighted difference = the sum of ( accuracy difference \* GTIN count)/ (sum of GTN count)
3. The GTIN count is from syndigo\_gtin\_level\_257K\_081523.csv
4. drop Level 1 if one of level 2 doesn't have a predicted value
5. The code ‘LR\_MLPSMOTE average weighted difference\_091523.ipynb’ applied groupby to remove the whole level1 if one of level 2 doesn't have a predicted value generate the weighted difference.
6. The weighted difference didn’t change too much from previous run.

The code ‘LR\_MLPSMOTE weighted difference\_091423.ipynb’ applied groupby to generate the weighted difference

9.14.22

1. The ML\_Classifier\_Comparision\_091423v1.xlsx is the modified version of ML\_Classifier\_Comparision\_091423.xlsx with a new tab ‘LR\_MLPSMOTE weighted difference’.
2. For each class, the weighted difference = the sum of ( accuracy difference \* GTIN count)/ (sum of GTN count)
3. The code ‘LR\_MLPSMOTE weighted difference\_091423.ipynb’ applied groupby to generate the weighted difference.

The spreadsheet looks like already have GTIN count in col K, T, and AC. However, the number are different for LR, MLP\_ SMOTE and MLP.

Do you know about [**sklearn.metrics.silhouette\_score? We might use it to compare models.**](http://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html)

For each L1 if there are 3 L2

Get GTIN counts for each L2, e.g., n1, n2, n3

Get the difference of Median Accuracy for test between LR and MLP + SMOTE, e.g, d1, d2,d3

Get the weight difference by di\* ni

1TN count \* CNT c,

Would we compare Median Accuracy for test to determine

9.13.22

“unlabel\_L2\_MLP\_predict\_091323.ipynb” is the coed to use MLP to predict the un-label data. mlp\_unlabel\_scores\_091323.zip is the result. The shape of the file is (3,799,519, 25).

Shall I use Syndigo data and apply MLP model for each level1 and use the model to apply to un-label data? Yes

Would ‘Level2’ be the target variable for un-label data for this exercise?

09.12.23.1o

The LR and SMOTE will not be able to handle level1 with only one level2. MLP is able to builds model for it. Does it make sense to build the MLP model for it?

Shila: There is no need to predict level2 in that case, since level1 uniquely determines level2

Typically, the macro-averaged F1 score is the most suit able measure for multiclass classification. We are trying to build LP, MLP, MLP\_SMOTE for each level1. Shall we compare macro-averaged F1 for each model and pick up the winner and apply the winner for label and un-label data?

I used the following code to calculate the measures. Accuracy is very close to f1\_score\_micro. Shall I use f1\_score\_macro to pick up the winner?

accuracy = accuracy\_score(y\_test,preds)

f1\_score\_micro = f1\_score(y\_test, preds, average='micro')

f1\_score\_weighted = f1\_score(y\_test, preds, average="weighted")

f1\_score\_macro = f1\_score(y\_test, preds, average='macro')

Shia: The first experience. n can be accuracy. if the accuracy is the same, then the macro F1 can be used to  designate the  winner.

I’ll assume MLP is the preference. If there is a tie among LR, MLP and SMOTE, the winner would be MLP. Am I correct?

09.11.23.2

Shila,

1. There is 32+ level1 classifications. I applied MLP + SMOTE to build 32+ models. How do I apply them to un-label data?
2. Shall we have un-label input with level1 classification and based on the level1 classification to pick up the right models to do the level2 classification for un-label data?

All the 21 gtin\_no in ‘acds\_transactions.xlsx’ are not in any one of the CSV file in ‘Active\_GTINs\_chunked.zip’.

F the CSV file

09.11.23.1

1. “CLASS REPORT SMOTE.txt” is the SMOTE result based on the max count of Level 1 to determine the sampling\_strategy
2. It is required to filter out the rows with target count <= 2 and determine the value k\_neighbors. It had been handled before split.
3. After split, it is required to do the filter out process as well. It is WIP for it.
4. Need to filter out the data if the size of class 1 is too small which will cause error in train\_test\_split. It is WIP for it.

9.8.23

Could you please elaborate “modify 'desired\_levelcount'  for each Level 1” ?

SMOTE error

9.7.23

I tried SMOTE for L1\_l2 combination to see if we can improve the performance for W2vec

I used 0.3 population of Syndigo data to test and I got “M **emoryError**: Unable to allocate 3.08 GiB for an array with shape (2554740, 162) and data type float64“. After Applying SMOTE oversample, the size of y is 2,554,740

Perhaps, I’ll skip SMOTE and try RandomForestClassifier or XGBClassifier to replace LogicRegression to see the performance.

Shila\_’1

Did you try SMOTE for our project?

I’ll try SMOTE and SMOTE + RandomUnderSampler for L1\_l2 combination to see if we can improve the performance. Shall I focus on W2vec only for the test?

Shila\_2

The following code are in L1\_l2 model build:

A\_train, A\_test, B\_train, B\_test = train\_test\_split(syndigo\_mapped.GTIN.tolist(), syndigo\_mapped['l1\_l2'].tolist(), test\_size=test\_size, random\_state=42, stratify= y)

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size= test\_size, random\_state=42, stratify= y)

My code

When I do oversample fit, for X-Train, y\_train. I also need to apply oversample fit for A\_Train and B\_train. Am I right?

oversample = SMOTE(k\_neighbors =1)

X\_train, y\_train = oversample.fit\_resample(X\_train, y\_train)

A\_train, B\_train = oversample.fit\_resample(A\_train, B\_train)

9.5.23

Hi, Whenever you are available, Could you please walk through your cede “Cluster\_Attribute\_Distance\_081423.ipynb”with me? I am confused about ‘ast’ and TSNE.

Sure we can schedule some time tomorrow.

Meanwhile, ast allows us to read a python list from a file properly ( e.g.  '[ 1,2,3] '  should be interpreted as a list [1,2,3] instead of a string.

TSNE is a method to project high - dimensional data to 2D or  3D for ease of visualization. [sklearn.manifold.TSNE — scikit-learn 1.3.0 documentation](https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html#sklearn.manifold.TSNE)

If you can cover how  to interpret the picture from  TSNE, it’d be much better.

To Shila

I am looking for the file “SuperClusterResults\_081723.csv’ you presented on 08-25-23 and the code to create the file.

* I use the following code to load function of 'elapse\_time' for the local run

path\_code = 'C:\\users\\iny2819\\kroger\\Code\\'

f\_com\_code = path\_code + "com\_code.py"

exec(compile(open(f\_com\_code , "rb").read(), f\_com\_code, 'exec' ))

* In Databricks, I saved ‘com\_code.py’ in my workspace. What path\_code would be in Databricks?

Do you have anything you’d like me to do? If not, I’ll try to run PyCaret for L1\_l2 comb.

8.31.23 The following files are created by excluding non-item SUBCOMs from un-label data and including ‘stratify= y’ when I did train\_test\_split. (code: imbed\_L1\_L2\_Syndigo Mapping MachineLearning\_090123.ipynb)

* L1\_L2\_syndigo\_pred\_cls\_rpt\_size0.3\_Stratify\_083123.xlsx
* L1\_L2\_syndigo\_pred\_tst\_size0.3\_Stratify\_083123.xlsx
* l1\_l2\_unlabel\_pred\_map\_W2vec\_Stratify\_083123

When I built the model, I didn’t exclude non-item SUBCOMs and I didn’t specify ‘stratify= y’ when I did train\_test\_split. I didn’t exclude non-item SUBCOMs from the un-label data for prediction.

Shall I re-build the model to include these I missed?

I am going to re-build the model to specify ‘stratify= y’ and exclude non-item SUBCOMs.

In order to run through ‘stratify= y’ in train\_test\_split, I need to exclude rows in y with count < 2.

After re-build the model, I’ll exclude non-item SUBCOMs from un-label data for the prediction.

Stratify\_

when I did train\_test\_split

by excluding rows in y with count < 2 and wr

to build the ML model and predict the un-label data. I’ll exclude non-item SUBCOMs exclude rows from syndigo\_mapped abS ‘y’ with count < 2 and specify ‘stratify= y’ and when I did train\_test\_split to

8.30.23t specify nd

To team

L1\_L2\_label\_accuracy\_unlbel\_validation\_C100\_W2vec\_tsize0.3\_083023.csv is the file with L1\_L2 accuracy for label test result with threshold = [0.5, 0.6, 0.7, 0.75, 0.8, 0.85] and corresponding validation required/not required for un-label data when I did the

* l1\_l2\_unlabel\_predition. red\_map\_W2vec\_083023.csv is the file with W2vec prediction for unable data PIM\_Data\_New\_50\_82Mn.csv. The file is loading nw. The Accuracy in the file should be Probability.
* I used pickle file for prediction. It doesn’t take too long to run the prediction. However, there was memory challenge, and I need to divide the %M data into two part to run.

dc

Howe

8.29.23

ShilaDitya

I'll try to predict it use W2vec. Could you please browse the tab ‘Consolidate Metrics’ in L1\_L2\_syndigo\_pred\_cls\_rpt\_size0.3\_082923.xlsx and let me know what the C value you’d like me to run? C = 100

What is 'ITM\_ID\_y' in PIM\_Data\_New\_50\_82Mn.csv for? Is it a derived variable? There are 4,823,143 rows missing value and the rest are 1. There .9490 are missing.

⁠Hi M N, Mallik (NonEmp), I didn’t run un-label data for W2vec yet. Is PIM\_Data\_New\_50\_82Mn.csv for the un-label data? Do you have code to predict un-label data?

The code “ imbed\_L1\_L2\_Syndigo Mapping MachineLearning\_082923.ipynb” to able to build models for C=0.1, 1, 10, 100 and test\_size = 0.3. I ran imbed = ‘TFIDF’ & “W2vec’ locally to create the following files for all the C values.

* L1\_L2\_syndigo\_pred\_cls\_rpt\_size0.3\_082923.xlsx
  + The run time for TFIDF & W2vec which can be found in the tab of ‘Consolidate Metrics’
* L1\_L2\_syndigo\_pred\_tst\_size0.3\_082923.xlsx

8.28.23,

I tried to use 0.3 population of Syndigo data to run W2vec, The following code looks like take forever to run. Am I missing something?

X = np.array(list(syndigo\_mapped.ITEM\_SUBCOM\_text.apply(lambda x: get\_item\_vector(x.split(' ')))))

To colin

I am from Infosys Core Taxonomy team under Deevraj Nair. Could you please grant me access to <https://app.mural.co/t/test5624/m/test5624/1681952645036/e4f1e597bf820cb3798fc97ed70a4714251ab0d6?sender=ue4891c98590017f95ae325928.25.23>? Thanks, Sophia

8.27.23

⁠Hi M N, Mallik (NonEmp) & ⁠V Ray, Shashank (NonEmp),

Per my conversation with Shiladitya on 08.24.23, “Let's run the L1\_L2 LR model on  training set (70% of 257K)  with  a few different values of 'C"  -  say 0.1, 1, 10, 100, and obtain and compare metrics on test set (remaining 30%). We can do it this week, and share the code + results to offshore - they can pick it up when they are back on Monday and redo it with word2vec”

The imbed\_L1\_L2\_Syndigo Mapping MachineLearning.ipynb is code to build models for C=0.1, 1, 10, 100. I ran imbed = ‘TFIDF’ locally to create the following files. You might modify the code for ‘word2vec’.

* L1\_L2\_syndigo\_pred\_tst\_size0.03\_082527.xlsx – predict file for C = 0.1, 1, 10, 100, an
* L1\_L2\_syndigo\_pred\_cls\_rpt\_size0.03\_082527.xlsx – classification report for C0.1, 1, 10, 100

Pickle files

Below are pickle files for different C values. You might upload the pickle files to predict unlabeled data.

* L1\_l2\_LR\_tfidf\_tst\_siz0.3\_C0.1.pkl
* L1\_l2\_LR\_tfidf\_tst\_siz0.3\_C1.pkl
* L1\_l2\_LR\_tfidf\_tst\_siz0.3\_C10.pkl
* L1\_l2\_LR\_tfidf\_tst\_siz0.3\_C100.pkl

The files and code had uploaded to General.

Hi

<https://app.mural.co/t/test5624/m/test5624/1681952645036/e4f1e597bf820cb3798fc97ed70a4714251ab0d6?sender=ue4891c98590017f95ae32592>8.25.23

What is the link for? It is from “Core Taxonomy” meeting yesterday. I got page not found message.

Shila: actually you can just log in using kroger email and create a new account to access it

It is not working for me. See the message below. How do I ask for access? Shall I ask Johnston, Colin

`to grant me the access?

The signed in user 'sophia.yue@kroger.com' is blocked because they are not a direct member of a group with access, nor had access directly assigned by an administrator. Please contact your administrator to assign access to this application.

8.24.23

I referred to the example in the article ‘[Deep dive into multi-label classification’](https://towardsdatascience.com/journey-to-the-center-of-multi-label-classification-384c40229bff), and I was trying to use LabelPowerset(LogisticRegression()) for multi label classification for Syndigo data and I got “**MemoryError**: Unable to allocate 67.6 GiB for an array with shape (181359, 50000) and data type float64”. It happened in Databricks as well. Do you know how to fix the issue. Please refer to the output “preds, preds\_lrtf, probs, preds\_train, probs\_train, lr\_tf = proc\_predict()” in the code “L1\_L2\_Syndigo Mapping ML Multi Label.ipynb” in General.

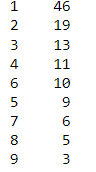
8.23.23

* To get one f1\_score for Multi-Class Classification, set the average parameter.
  + V
  + f1\_score(y\_true, y\_pred, average='macro')
  + from sklearn.metrics import f1\_score
* To get f1\_score for every class, set the average parameter to None
  + f1\_score(y\_true, y\_pred, average=None) - get every f1\_score every class

ref: https://www.baeldung.com/cs/multi-class-f1-score

The code ‘L1\_L2\_Syndigo Mapping MachineLearning.ipynb ‘, ‘com\_code.py ‘ and the file ‘L1\_L2\_cls\_rpt\_Test\_Metrics.xlsx’ had been uploded.

8.22.23



For L1 & L2 combination, there are 46 L1\_L2 with value\_counts =1,  19 L1\_L2 with value\_counts =2.

When we  build a model, shall we throw away L1\_L2 if value\_counts  < n .  There  are  122  L1\_L2 with value\_*counts < 10*

8.21,23

I created a ‘Articles’ which is a new folder in General. I posted my collections of articles in the folder for your reference. The articles include:

* Handle imbalance data
* [**Multi-Label Classification with Scikit-MultiLearn**](https://www.section.io/engineering-education/multi-label-classification-with-scikit-multilearn/#:~:text=September%2024%2C%202021&text=Multi%2Dlabel%20classification%20allows%20us,outputs%20for%20a%20given%20prediction.)  - We might use some algorithms from this article.
* [**A Gentle Introduction to Threshold-Moving for Imbalanced Classification**](https://machinelearningmastery.com/threshold-moving-for-imbalanced-classification/)

Our classification is “Hierarchical Multi-Label Classification”. However, I don’t find too much information.

8.21.23

“Syndigo Mapping MachineLearning\_def.ipynb” is the modified version of “Syndigo Mapping MachineLearning.ipynb” with the following modifications:

* Modularize the functionality to extract data, prediction and save results.
* Create a dictionary from Syndigo date with Level1 and files created. It can run through all the Level1 and save file in a loop.
* The code had been tested and uploaded. It only for TFIDF and it can be easy to change to W2vec

8.16.23

So far Shiladitya and I only have the following files for test classification test result. Could one of you provide us with a list of complete train, test result files and location?

* LivestockPetSupplies.csv
* HomeVenueDecoration.xlsx
* GardeningOutdoors.csv
* ToysGamesHobbies.csv
* KitchenBathroom.xlsx
* F&B (1).xlsx

Office Supplies, Home & Venue Decoration

8.15.23

The ‘accuracy\_cnt\_normd\_median.xlsx’ created from ‘l2\_accuracy\_prob\_score.ipynb’ is a part of information required by “Accuracy and Probability scores.xlsx”. The file has count, normalized count for labelled data ‘syndigo\_gtin\_level\_257K\_081523.csv’ , and count, normalized count, and median for labelled data ‘ Syndigo\_Levels\_L1\_L2\_L3.zip’.

The file and code had uploaded.

There are two columns ‘y\_test’, ‘soln’ in GardeningOutdoors.csv. Which column shall I used to geiilelet median probability, and median accuracy?

There are two columns ‘Actual’, ‘Predict’ in ‘KitchenBathroom.xlsx’. Which column shall I used to get median probability, and median accuracy?

I am creating a file to merge ‘Syndigo\_GTIN\_Mapping\_060723.csv’ with ‘Syndiago Parts 1-2-3-4-5 matches w taxonomy classification.csv ‘ derived from ‘Parts 1-2-3-4-5 Syndigo matches w taxonomy .xlsx’. I am struggling to give a file name for the merger file. The merged file has 257K. The merged file only has GTIN, Level 1 ~ Level 9. Is ‘syndigo\_gtin\_level\_257K\_081523’ good for the merger file? Or you might suggest a better one. yes

The shape of Syndigo\_Levels\_L1\_L2\_L3.csv is (5,211,106, 22).    Are the predictions in syndigo\_Levels\_L1\_L2\_L3.csv  and all the files you  upload on last Friday  based on 25syndigo\_0K Syndigo data or 95 K Syndigo data?

8.14.23

* Would 'Level1', 'Level2', 'L2\_scores’ in Syndigo\_Levels\_L1\_L2\_L3.zip be used to calculate unlabeled median probability?
* I unzip the zip file and read the file with shape = (1,048,575, 22). Does the size look OK to you?
* Do you have an example of python code to read the zip file directly?

df\_unlabel.czip file and olumns

Index(['VND\_ECOM\_DSC', 'Level1', 'L1\_Scores', 'Level2', 'L2\_Scores', 'GTIN',

'PMY\_DPT\_CD', 'PMY\_DPT\_DSC', 'REC\_DPT\_CD', 'REC\_DPT\_DSC', 'DPT\_CD',

'DPT\_DSC', 'COM\_CD', 'COM\_DSC', 'SUBCOM\_CD', 'SUBCOM\_DSC', 'ITM\_ID',

'L1\_Percent', 'L2\_Percent', 'L2\_Flag', 'Score', 'Level3'],

dtype='object')

8.11.23

Can you take a look or run ‘Level 1 Classification’? below are my observations

1. The synd\_rest = pd.read\_csv(DBFO + 'Left\_Out\_1\_1Mn\_Full\_data.csv'). All the ‘Level1’ are null. After execute cmd 5 and cmd 6, the syndigo\_classifier is empty. I don’t think pim\_gtin\_mapped = synd\_rest in cmd 5 is necessary.
2. It looks like that the code only use VND\_ECOM\_DSC + DPT\_DSC to build the model.
3. The code doesn’t do train-split for train model and no validation for the model.

Code

1. In cmd 5,

pim\_gtin\_mapped = synd\_rest

SUBCOMs = pim\_gtin\_mapped[pim\_gtin\_mapped['Level 1'].notnull()]['SUBCOM\_CD'].unique() #

remove\_subcoms = [97000, 37120, 92004, 37122, 97004, 97001, 97010, 96800, 97008, 96400, 97003, 34260, 97002, 96401, 97006, 34529, 96900, 73531, 87633, 97009, 34265, 34155, 74202]

SUBCOMs = [int(i) for i in SUBCOMs if int(i) not in remove\_subcoms]

print(len(SUBCOMs))

1. in cmd 6

syndigo\_classifier = pim\_gtin\_mapped.loc[pim\_gtin\_mapped['SUBCOM\_CD'].isin(list(SUBCOMs))]

print(len(pim\_gtin\_mapped.loc[pim\_gtin\_mapped['SUBCOM\_CD'].isin(list(SUBCOMs))]))

pim\_gtin\_mapped = synd\_rest

synd\_rest = pd.read\_csv(DBFO + 'Left\_Out\_1\_1Mn\_Full\_data.csv') if you can get the same result of cmd 22

offshore my run (1) my run (2)

248441 255947 0

1112596 3553830 0

241735 248754 0

549363 2795970 0

8.8.23

Thank you for walking through the code Syndigo\_Mapping\_ML\_060723.ipynb with me. I am trying to use your code to create the input file to fit different models and find the best model for Level 1 classification. Below are questions/observations for you from your code:

1. Step0: cmd 40 to cmd 96 were for information purpose and not for the preparation to create input for modeling. Am I right?
2. Step1: cmd 97 to cmd 121 used genism to vectorize VND\_ECOM\_DSC + SUBCOM\_DSC and the vectorized results will be the features vectors for model.

* I didn’t see the code to concat syndigo\_concat to syndigo\_mapped.
* I didn’t see the code to create a subset S1 to exclude the rows from syndigo\_mapped with SUBCOM\_DSC in pimmart and not in syndago\_mapped.

1. Step2: cmd 121 to cmd 141 used TFIDF to vectorize VND\_ECOM\_DSC + DPT\_DSC and the vectorized results will be the features vectors for model.

* I didn’t see the code to create a subset S2 to exclude the rows in syndigo\_mapped from S1 or rows from syndigo\_mapped with DEP\_DSC in pimmart and not in syndigo\_mapped.

### Step3: cmd 121 to cmd 141 used to train L2 classifier for a given L1

* I didn’t see Level 1 in sndigo\_mapped been updated by the predict results from Step1 and step2

8.6.23

* The Syndigo\_L2\_Optimal\_Thresholds.xlsx is the file with adjusted threshold.
* Syndigo\_L2\_Optimal\_Thresholds.ipynb is the code to generate the TRUE FALSE percentage before/after adjustment for the following files:

1. KitchenBathroom (1).csv
2. HomeVenueDecoration (1).csv
3. GardeningOutdoors.csv
4. ToysGamesHobbies.csv
5. LivestockPetSupplies.csv

* The code and file had been uploaded.
* Syndigo\_Level\_Optimal\_Thresholds.ipynb is the code from Shiladitya.

8.4.23

Would the concept to separate input into ' VND\_ECOM\_DSC + SUBCOM\_DSC'   and VND\_ECOM\_DSC + DPT\_DSC apply to RF? No, in RF only VND\_ECOM\_DSC is compared between items to find a match.

Q: Do we train RF2 classifier for a given RF1?

A: They are independent since there is nothing to train really unlike an ML model

2)  We are using different X inputs, depending on the data:

      a) For items that have SUBCOMs common with the labeled data from SYndigo ( ~250K items that you compiled a few days ago):

                 VND\_ECOM\_DSC + SUBCOM\_DSC

      b) For items that have SUBCOMS outside of the 250K labeled data, but have DPT\_DSC common with that data, we use:

                  VND\_ECOM\_DSC + DPT\_DSC

Thank you for your explanations.

If we have 32 unique Level1

1. For feature variable = ‘VND\_ECOM\_DSC + SUBCOM\_DSC’, we’ll have one model for Level 1 and 32 models for Level 2.
2. For feature variable = ‘VND\_ECOM\_DSC + DPT\_DSC’, we’ll have one model for Level 1 and 32 models for Level 2.

In this case, we’ll have 64 models for Level2 giving Level 1. Am I right?

8.3.23

Your presentation for tomorrow only covers clustering and not covers Syndigo level prediction. Is it correct? Yes

Can clustering be replaced by Syndigo level prediction?

Level 2 Classification is based on Level 1 Classification. Shall we use the predict level 1 in Syndigo to train level 2?

Syndigo -> level1 prediction modeling –> Syndigo with level1 prediction -> select Syndigo with specific level 1 prediction -> level2 prediction -> Syndigo level2 prediction -> -> select Syndigo with specific level 2 prediction -> level 3 prediction

8.2.23

The code didn’t train “Level 2 Classification” but “Train L2 classifier for a given L1”. Is it correct?

Regarding “Train L2 classifier for a given L1”, if we have 32 Level 1 Classification, do we need to build 32 models to train L2 classifier for a given L1? Yes

From cmd 29 to cmd 37, I didn’t see ‘syndigo\_mapped’ concat to ‘syndigo\_concat, for syndigo part 3 ~ Part 5”. I expect ‘syndigo\_mapped’ will concat to ‘syndigo\_concat’ for doing modeling. Is it correct?

syndigo\_mapped = pd.read\_csv('/dbfs/FileStore/tables/DATA\_SCIENCE/Syndigo\_GTIN\_Mapping\_060723.csv')

In Syndigo\_Mapping\_ML\_060723.ipynb cmd 110, is ‘get\_item\_vector’ required for the following code ?

#Target Vectors

level1\_id\_map = dict(zip(syndigo\_mapped['Level1'].fillna('Unclassified').unique(), range(syndigo\_mapped['Level 1'].fillna('Unclassified').nunique())))

y\_ = syndigo\_mapped['Level 1'].fillna('Unclassified').map(level1\_id\_map)

8.1.23

Where is the code you create the profile? The Cluster\_ATttribute\_Profiling\_2023-07-12 12\_17\_52.ipynb  appears not the code to me.

You made suggestions at today’s meeting, I am wondering if I can help Malik for something.

7.31.23

1. The file ‘Parts 1-2-3-4-5 Syndigo matches w taxonomy .xlsx’ with 164,053 unique GTIN, no duplicate GTIN, had been save to Databricks /dbfs/FileStore/tables/DATA\_SCIENCE/Syndiago\_Parts\_1\_2\_3\_4\_5\_matches\_w\_taxonomy.csv
2. ‘Syndiago Parts 1-2-3-4-5 matches w taxonomy classification.csv with Level 1 ~ Level 9 had been upload to Databricks and General

There are kwas created from

1. Wrote a code to read the 'Syndigo\_2\_3\_4\_5\_072823.xlsx' to process taxonomy and create L1 ~ L9 classification
2. The output file “Parts 2-3-4-5 Syndigo matches w taxonomy classification.xlsx” and the code “syndigo\_2\_3\_4\_5\_with\_level\_072823.ipynb” had been uploaded.

7.28.23

Could find\_topics() in bertopic be used to answer the question from Samiran, use our current playground to "match a new Product"?

Are you family with the process/data flow of how Mallik to get L1 and L2 prediction?

A

The following links are informative. We might try to play around the parameters to improve the performance. find\_topics can bertopic can predice new data.

Hyperparameter tuning

- <https://grabngoinfo.com/hyperparameter-tuning-for-bertopic-model-in-python/>

- https://www.youtube.com/watch?v=PbX4nHEOe0o&list=PLVppujud2yJpx5r8GFeJ81fyek8dEDMX-&index=6

- <https://colab.research.google.com/drive/1SXAz1qtoB1Wou5dfBr8WSYvKg91ExZwe?usp=sharing>

Topic Modeling with Deep Learning Using Python BERTopic

* <https://grabngoinfo.com/topic-modeling-with-deep-learning-using-python-bertopic>

Step 9: Topic Model Predictions On New Data

* https://www.youtube.com/watch?v=PbX4nHEOe0o&list=PLVppujud2yJpx5r8GFeJ81fyek8dEDMX-&index=5
* <https://colab.research.google.com/drive/1jhP3-UV0DCCh0JvFYv_8Pwsz1PQ-TkFo?usp=sharing#scrollTo=qYnTlARzz2Xa>



7.27.23

pimview = pd.read\_parquet('/dbfs/FileStore/tables/DATA\_SCIENCE/ITM\_DSC\_ATTRIBUTES\_NEW\_parquet.gzip') # (4937400, 34)

### Add correct ITM\_ID and KFT level codes

pim\_ITM\_ID\_CODES = pd.read\_csv('/dbfs/FileStore/tables/DATA\_SCIENCE/PIM\_CORRECT\_ITM\_ID\_GTIN\_\_LEVEL\_CODES\_061423.csv')

pimview = pimview.merge(pim\_ITM\_ID\_CODES, on = 'GTIN\_NO', how = 'left')

7.26.23

1. Please find the file pimmart\_gtin\_cnt\_subcom\_not\_in\_sandago.xlsx which has count of GTIN for each subcom\_dsc. The file had been uploaded to general
2. The input file is SUBCOM\_NOTCOVERED\_SYNDIGO.xlsx (sheet2) and ITM\_DSC\_ATTRIBUTES\_NEW\_parquet.gzip
3. There are 1942 rows in SUBCOM\_NOTCOVERED\_SYNDIGO.xlsx.
4. The number of unique SUBCOM\_DSC is 1768.
5. The number of unique SUBCOM\_CD is 1892.

7.24.23

Each item is assigned to different levels of KFT hierarchy, as described in the tree diagram : PMY\_DPT\_DSC,  REC\_DPT\_DSC, etc

has context menu

7.19.23

1. Wrote a code to read the ‘Parts 3-4-5 Syndigo matches w taxonomy.xlsx’ to process taxonomy and create L1, L2 & L3 classification
2. The output file “Parts 3-4-5 Syndigo matches w taxonomy classification.xlsx” and the code “process Parts 3-4-5 Syndigo matches w taxonomy.ipynb” had been uploaded.

get l1, l2.... values from the file.

7.17.23

1. Please take a look of the file Beer\_Wine\_Spirits\_ML\_RF\_validation.xlsx which I wrote a code to simulate manual validation.
2. I didn’t keep the color code in the excel sheet.
3. The file has 82580 rows and 79653 rows had been validated.

7.12.23

Hi Mallik,

Could you please upload the file with col names ‘Final Level1’,’Final Level2’, and ‘Final Level3’ … in the tab ‘SOLUCTION(1)’ . It’d be a good reference for my manual validation. Thanks.

7.12.23

Hi Mallik,

Could you please upload the file ‘level1 and level2 coverage\_crosstab.xlsx’? Thanks.

Shashank provided me with 'Syndigo 113k Levels.csv', There are 113,898 rows from the file and there are 67,839 unique GTIN. I expect there are 95K unique GTIN. Do you know What is missing?

. I created 'Syndigo 113k Levels.xlsx' from the csv file.

I tried to match the GTIN in ‘Beer\_Wine\_Spirits\_ML and RF.xlsx’ with 'Syndigo\_GTIN\_Mapping\_060723.xlsx'. The mismatch rate is 99.88%. Is it normal?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| file name/Description | DF name | Total Records | Unique GTIN | Example |
| Syndigo\_GTIN\_Mapping\_060723.xlsx | df\_syndiago | 113898 | 95032 |  |
| Beer\_Wine\_Spirits\_ML and RF.xlsx | df\_ml\_rf | 82580 | 78912 |  |
| total GTIN in both files |  |  | 93 | [11034410055, 11034410079, 11034420054, 11034420078, 11034430053, 12354071995] |
| total GTIN in df\_ml\_rf and not in df\_syndiago |  |  | 78819 | [222, 284, 765, 2004, 2011, 2035, 2042, 2066, 2073, 2080] |
| Percentage of GTIN in df\_ml\_rf and not in df\_syndiago = 99.88% |  |  | 0.998821472 |  |

Shiladaity

The attribute no for Cheese Product Type in Attributes\_Index\_with\_importance\_0629.csv is y. it should be 896. Could you please correct it?

7.10.23

TC

1. Wrote a code to simulate the process of manual validation for Beer\_Wine\_Spirits\_ML and RF.xlsx.
2. The code would identify which rows required mannual validation.
3. Complete beer and spirits.

TP

Continue to work on the code.

7.6.23

V Ray, Shashank (NonEmp)

1. Do you want me validate ‘Toys\_Games\_Hobbies\_RF. xlsx’?
2. Per my understanding, the ‘Toys\_Games\_Hobbies\_L2.xlsx’ I manually validated is not updated one. The General still shows the version I updated. Am I supposedly to get new file from you and manually re-validated it?

V Ray, Shashank (NonEmp)

The ‘Toys\_Games\_Hobbies\_RF. xlsx’ is not L2 file. Do you want me ignore the ‘Toys\_Games\_Hobbies\_L2.xlsx’ and do the validation for The ‘Toys\_Games\_Hobbies\_RF. xlsx’?

V Ray, Shashank (NonEmp)

1. I used ‘subcom’ validated misclassified sequentially for Livestock&PetSupplies\_L2.xlsx.
2. I validated ‘subcom’ up to 7206 with vendor started FM. I also reviewed the ones identified by you with highlighted as yellow.
3. Color definition

* Grey color: Misclassified ones identified by me
* Blue: Might be Misclassified or SUBCOM\_DSC is incorrect
* Green: Be identify as misclassified but might be not

I upload ‘Livestock&PetSupplies\_L2.xlsx’ with my update. Please continue to validate L2 file manually and I’ll continue the validation after you.

6.30.23

TC

1. Manually validated SUBCOMs are misclassified for Livestock&PetSupplies\_L2.xlsx
2. Wrote a Python code to use LDA topic modeling to group the cluster name in cluster\_sumary\_0627.xlsx to 10 groups and uploaded the result Cluster\_Summaries\_group\_063023.xlsx

TP

1. Continue to validate work SUBCOMs are misclassified for Livestock&PetSupplies\_L2.xlsx and other L2 files
2. Continue to deep dive attribute files

6.27.23

1. Wrote a python code dim\_itm\_oc\_atb\_col\_val.ipynb to create the dim\_itm\_oc\_atb\_col\_val.xlsx which combined the Attributes\_Index\_Readable\_Headers\_060923.csv (Attributes\_Index), created from CURIDQ.PIMMART.DIM\_ATB\_INF\_LKP (DIM), and CURIDQ.PIMMAR.ITM\_OC\_ATB (ITM)
2. "Attributes Changed Since Last Merge" is corresponding to column name "ATB\_CHG\_SIN\_LST\_MER\_TX" in ITM\_OC\_ATB.
3. Upload the code and file

TP

1. Streamline the attribute names in attribute\_index for ITM\_OC\_ATB

6.27.23

TC

1. Deep-dive the relationship between Attributes\_Index\_Readable\_Headers\_060923.csv (Attributes\_Index\_), created from CURIDQ.PIMMART.DIM\_ATB\_INF\_LKP (DIM), and CURIDQ.PIMMAR.ITM\_OC\_ATB (ITM)
2. Wrote itm\_oc\_atb\_deep\_dive.ipynb to get count value distribution for deep-dive
3. Created ITM\_OC\_ATB deep-dive.xlsx for the deep-dive results for ITM\_OC\_ATB

TP

1. Match attribute name in Attributes\_Index\_Readable\_Headers\_060923.csv with ITM\_OC\_ATB

RB/Challenges

1. Match attribute name in Attributes\_Index\_Readable\_Headers\_060923.csv with ITM\_OC\_ATB

6.23.23

TC

1. Deep-dive the relationship between ATB\_NAM in CURIDQ.PIMMART.DIM\_ATB\_INF\_LKP (ATB) and NAM\_IN\_KEY\_LNG\_NAM in CURIDQ.PIMMART.CLS\_SPC\_ATB (CLS)
2. Create CLS\_SPC\_ATB\_Deep\_dive.xlsx which include the research results for CLS\_SPC\_ATB and queries. It’ll serve ‘one-stop’ excel for CLS\_SPC\_ATB

TP

1. Deep-dive the relationship between CURIDQ.PIMMART.DIM\_ATB\_INF\_LKP and CURIDQ.PIMMART .ITM\_OC\_ATB
2. Streamline the attribute name in CURIDQ.PIMMART.DIM\_ATB\_INF\_LKP – might use LDA topic modeling for streamline.

RB/Challenges

1. It would be challenging to streamline the attribute name in CURIDQ.PIMMART.DIM\_ATB\_INF\_LKP

6.21.23

TC

1. Used Azure Studio to create a CSV file ‘CLS\_SPC\_ATB\_062123.csv’ by extracting data from ‘CURIDQ.PIMMART.CLS\_SPC\_ATB’ and dedup ‘itm\_id’ from the file. The file has 4,634,069 rows.
2. Without deduping, the count of the table ‘CURIDQ.PIMMART.CLS\_SPC\_ATB’is 22,131,836
3. Please refer to my email for the deduping.

TP

1. Deep dive more tables specified in ATB\_TBL\_NAM/’Table in PIM-Mart’ in Attributes\_Index\_Readable\_Headers\_060923.csv
2. Find relationship between attribute number and itm\_id

RB

1. The Attributes\_Index\_Readable\_Headers\_060923.csv only specified the col name in ‘Table in PIM-Mart’. Without the value of col name, it is not easy to do the mapping.
2. FuzzyMapping\_Initial

download dats from

6.14.23

TCLearned to use ‘workspace’ and ‘data’, download/upload data/file and ran python code in DataBricks via the help of Shiladitya

Reviewed BERtopic from the code PIMM

1. Will apply BERtopic for the ‘attribution definition’ for clustering from ‘Attributes\_Index\_Readable\_Headers\_060923.csv'

RB

1. ART\_BERTopic\_Meat\_0612.ipynb written by Shiladitya

TPHaving connection issue and not able to connect to SQL server Azure Studio

6.20.23

TC

1. Reviewed document of RapidFuzz and code of FuzyMapping Initial.ipynb
2. The match score did improve by added parameter processor=utils.default\_process) to fuzz.token\_set\_ratio.
3. Be able to access Confluence.

TP

1. Will apply BERtopic for the ‘attribution definition’ for clustering from ‘Attributes\_Index\_Readable\_Headers\_060923.csv'

RB

1. Having connection issue and not able to connect to SQL server Azure Studio