SOP - Product classification using FastText:

Important Links:

Install R- https://cran.r-project.org/bin/windows/base/
Install R studio - https://www.rstudio.com/products/rstudio/download/
Install cygwin - http://www.cygwin.com/install.html

Objective:- To map products in most relevant mcat(brand, non brand or PMCAT), based on following parameters(textual inputs):

- Product Name(PC_ITEM_NAME).
- Specification response.
- Product description(only bulleted descriptions).

Defining the SubCAT at 3 levels of taxonomy.

1. PMCAT(except super PMCAT).
2. Brand MCAT.
3. Non Brand MCAT.

Steps:-

1.) **Extracting products data using SQL query.** The extracted data contains the fields like: PRIME_MCAT_ID,PRIME_MCAT_NAME PRIME_MCAT_IS_GENERIC,FK_MCAT_TYPE_ID,GOOD PMCAT,OTHER_MCATS, FK_GLCAT_MCAT_ID, PC_ITEM_ID, PC_ITEM_NAME,PC_ITEM_IMG_ORIGINAL PC_ITEM_GLUSR_USR_ID, PC_ITEM_GLCAT_MCAT_ID_LIST, PC_ITEM_GLCAT_MCAT_NAME_LIST, PC_ITEM_STATUS_APPROVAL PC_ITEM_DESC_SMALL,FK_IM_SPEC_MASTER_ID, FK_IM_SPEC_MASTER_DESC FK_IM_SPEC_OPTIONS_ID, FK_IM_SPEC_OPTIONS_DESC,Response MCAT PC_ITEM_ATTRIBUTE_MOD_DATE,Current MCAT.

Pls use the following **SQL query** to get above data.

2.) **Data processing in Excel.**
   a.) Since data extracted from query contains to many fields for our analysis, we proceed with required fields like: PRIME_MCAT_ID,PRIME_MCAT_NAME PRIME_MCAT_IS_GENERIC,FK_MCAT_TYPE_ID,GOOD PMCAT,PC_ITEM_ID,PC_ITEM_NAME,PC_ITEM_DESC_SMALL,FK_IM_SPEC_OPTIONS_DESC.

b.) Also our data contained multiple rows for same product having different specification options. So we constructed a single row per product thus combining all specification options to get combined data against unique PC_ITEM_ID.

i.) Concat all Options_Desc to get filled responses against unique Item_Id.

ii.) For PMCAT training data- Take PMCAT and PMCAT flag against each MCAT_ID. Consider MCAT having flag 2 as good pmcat, and if its child we consider its parent with flag '2' as good PMCAT.

iii.) Consider only filtered good PMCAT data for training.

iv.) To get training data for child_mcats just untick MCAT_FLAG '2' from Good_PMCAT_data for training the model on child mcats(2nd level).

3.) **Data Cleaning and processing in R:**

- After getting excel file with unique PC_ITEM_IDs we need to process the file in an R-script to clean the data and create the corpus for training and testing files. Here we have divided training and testing dataset into 10 folds to reduce bias in the model.
- **PC_ITEM_NAME -** Only change to lower case and remove the special character.
- **PRODUCT_DESCRIPTION-** Keep only bulleted data from the description and ignore flowery contents.
- **IM_SPEC_OPTION_DESCRIPTION-** After concatenating entire options, remove options like: *50Kg, 50 Kilogram* etc.
- Then contact all three column to separate training and testing files.
- Use this **R-Script** to perform above steps.

****K-Fold validation:-** It provides a robust estimate of the performance of a model on unseen data. It does this by splitting the training dataset into k subsets and takes turns training models on all subsets except one which is held out, and evaluating model performance on the held out validation dataset. The process is repeated until all subsets are given an opportunity to be the held out validation set. The performance measure is then averaged across all models that are created.

4.) **Developing model in Cygwin with k-fold:-**

a.) Create a list in vi editors with contains numeric values from 1 to 10. Use command: *vi file* and insert values. We can also create a file list containing different file names so that we can train the entire set of data in one go.

- To create a list as per our choice use below command:

*find /cygdrive/c/Users/Ashutosh/Desktop/corpusrar/ -name *.txt -printf "%f\n" > list*

b.) Train the model using following command:
while read p; do fasttext/fasttext.exe supervised -input /cygdrive/e/KFold/Wheel\ Loader/Training\ Files_brand/$p.txt -output /cygdrive/e/KFold/Wheel\ Loader/Training\ Files_brand/bin/$p -lr 0.8 -minn 4 -epoch 25 -wordNgrams 1 -lrUpdateRate 100 -thread 4 -loss hs; done < file

c.) Testing command to get output data:
while read p ; do fastText/fasttext.exe predict-prob /cygdrive/e/KFold/Wheel\ Loader/Training\ Files_brand/bin/$p.bin /cygdrive/e/KFold/Wheel\ Loader/Testing_pmcat/$p.txt 2 > /cygdrive/e/KFold/Wheel\ Loader/out_brand/$p.out ;done < file

d.) After getting output data, we combine the output result(.out) file and testing data-set to get final output. Use following commands to get final output(combined data).
while read p ; do paste /cygdrive/e/KFold/Wheel\ Loader/Testing_pmcat/$p.txt /cygdrive/e/KFold/Wheel\ Loader/out_brand/$p.out |column -s $'\t' -t > /cygdrive/e/KFold/Wheel\ Loader/out_brand/$p.final ; done < file

e.) The combined data is a (.final) file. To convert (**.final**) files to (**.csv**)--> Open command prompt and run *"ren *.final *.csv"* in the final files directory.

5.) **Processing combined output file in R to get output in readable format.**
We have created another r-script which will convert all the output files into readable format in one go.
Use following **R-script** to get output data in readable format.
This script will create **K** output files(.csv) or (.xlsx) for K different folds , which can be directly used for further analysis.

Same Procedure to be followed for BL classification without using K-fold cross validation:

1. We will train the model on
    a. Entire PMCAT products data
    b. Entire Brand products data

c. Entire Child Products data

2. Testing to be done on BL data of same MCAT/PMCAT. Same command to be used w/o kfold.