**Problem Statement :** MCAT Deletion SEO Analysis Data Preparation

**Data Sources**

We have data from 3 sources:

1. IMPCAT- [impcat\_data\_june2020.csv](https://drive.google.com/file/d/12knOEvoc1OCB_fgci0D5_70CUeOjqbDb/view?usp=sharing)
2. PDP - [pdp\_data\_june2020.csv](https://drive.google.com/file/d/1jTP9muUKrh6Fps1cMeNnwbuR2YAHtG5l/view?usp=sharing)
3. GA - [ga\_data\_new\_sept20.csv](https://drive.google.com/file/d/1KkHn4lOfItGYLcGY-s3CxxdQKQ2JZfj1/view?usp=sharing)

**Data Preparation**

* Combine the three different files to form a single file, having entries from each file.
* Keep only the mentioned columns:

['Query','Impressions','Avg\_Position','GLCAT\_MCAT\_ID','GLCAT\_MCAT\_NAME','Total Unique Searches']

* For each **query in a MCAT**:
  + If only one entry for a query exists (in a MCAT) calculate and store the following as:

1. data[impression]=data[‘Impressions’]
2. data[unique\_searches]= data[‘Unique Searches’]
3. data[avg\_position]= data['Avg\_Position']
   * If more than one entry for a query exists (in a MCAT) calculate and store the following as:
4. data[impression]= Sum of Impressions
5. data[unique\_searches]= Sum of Unique Searches
6. data[avg\_position]= Average of 'Avg\_Position'

* This will ensure that there is only one entry of a particular query in a MCAT.
* Columns after this step:

mcat\_id, mcat\_name, query,impression, avg\_position, unique\_searches

* Save the results in a file (combined\_data.csv)

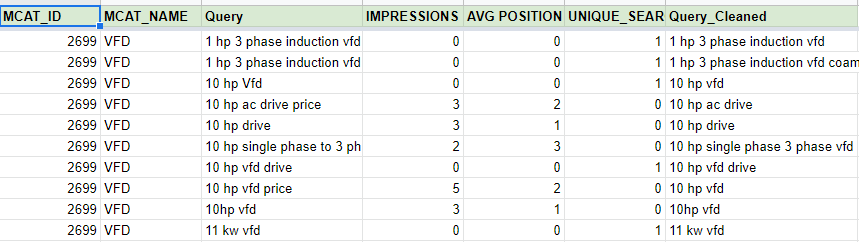
* For each query, we need to clean it by applying the following steps:
  + Conversion to lowercase
  + Keeping only alphabets and numeric entries
  + Removal of all english stop words
  + Removal of certain keywords mentioned below:

["manufacturer","manufacture","manufacturing","manufactured","price","wholesaler","wholesalers","wholesale","deal","dealer","deals","dealers","distribution","supplier","distributor","distributors","india","rate","cost","costs","near me","buy","buys","online","company", "exporter","exporters","good","topmost","business","trusted","finest","offer","organization","trader","organizations","traders","offers","indiamart","delhi","manufacturers","prices","suppliers","manufactures"]

* + Removal of location names provided in the pickle file ([location.pkl)](https://drive.google.com/file/d/14-n00fiKrUToXch7HZQ-HQOPoLeShdMl/view?usp=sharing)
  + Singularize the query
  + Replace one than one spaces with a single space
  + Trim the extra spaces (trailing and preceding)
* This will give us a clean\_query corresponding to each query in a MCAT.
* Columns after this step:

mcat\_id, mcat\_name, query, impression, avg\_position, unique\_searches, query\_cleaned

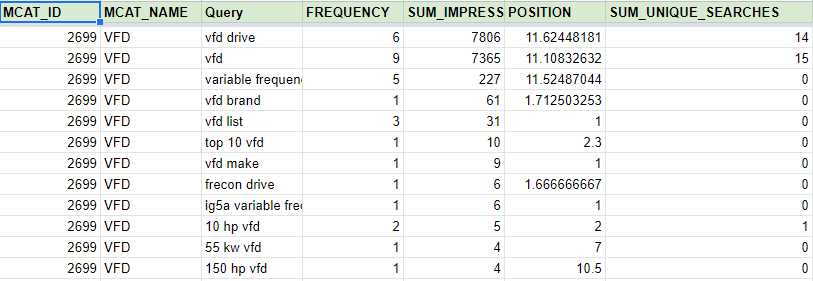
* Save the results in a file (cleaned\_queries.csv)



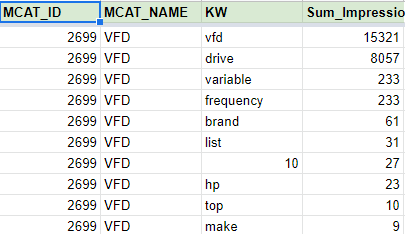
* Now, for each **clean\_query in a MCAT** (please note that many queries when cleaned might result in the same clean\_query), we will calculate and store the following:
  + data[‘Impressions’]= Sum of Impressions
  + data[‘unique\_search’]= Sum of Unique Searches
  + data[‘frequency’]= Count of the times ‘clean\_query’ appears in a MCAT
  + If the clean\_query exists as it is in the original data, we take the value of “avg\_position” for our clean\_query, else we take the Minimum of average\_position
  + For a MCAT, sort the values as per ‘Impressions’ of queries.(high to low)
* Columns after this step:

mcat\_id, mcat\_name, query (cleaned\_query) , frequency, sum\_impression, position, sum\_unique\_searches

* Save the results in a file (cleaned\_queries\_with\_freq.csv)



* For each MCAT:
  + Find all the uni-grams.
  + For each uni-gram, calculate the total sum\_impressions (summation of “sum\_impression” values for every entry where the uni-gram exists)
  + Return top-15 for every MCAT. (as per “sum\_impression”- high to low)
  + Save the results in a file (MCAT\_wise\_uni\_grams.csv)



* + Find all the bi-grams.
  + For each bi-gram, calculate the total sum\_impressions (summation of “sum\_impression” values for every entry where the bi-gram exists)
  + Return top-15 for every MCAT. (as per “sum\_impression”- high to low)
  + Save the results in a file (MCAT\_wise\_bi\_grams.csv)

