* tender\_dictionary.txt is the file which contains embeddings for 424 tenders.
* tender\_array.txt file contains all the tenders reference number.
* output\_data.xlsx file converts tender\_dictionary.txt into excel file.
* merged\_file.xlsx file contains 424 tenders with all the columns from Tenders\_WA\_descriptionExtracted.xlsx file.
* merged\_file.xlsx contains 1064 tenders , some of the reference have multiple tenders with different UNSPSC code, hence the cluster isn't efficient. Taking the unique tenders to create cluster.
* unique\_references.xlsx contains unique tender reference
* some of the tenders are missing in the "Tenders WA\_descriptionExtracted" sheet we have total "424" tenders given by Andre's embeddings but only 416 tenders was found in Tenders WA\_descriptionExtracted excel(26146) records , in original tender we have 26476 recorsd available.

for now i am going ahead with 416 records.

* So far i have tried DBSCAN and K-means clustering method.

so where eps = 0.1 there lots of points are found as outliers

and where eps increases to 0.2 lots of points are found as cluster 0 few points are in cluster 1.

when eps = 0.3 outliers decreases how ever most of the points fall in cluster 0 and no other clusters are made. looking at this i feel like DBSCAN might not be a good fit for my data.

* I have also tried Kmeans clustering method , to find optimal k values we used elbow method for sum of square distances , Silhouette Score and silhouette analysis to find k.   
  After analyzing every thing so far k = 10 and 11 looks better than other clusters, although we could see  few negative silhouette scores, this configuration seems more reasonable.
* We could also try using 7 or 8.
* updated\_excel\_file.xlsx file contains tenders with cluster group.
* cluster\_file.xlsx contains clusters with few related columns.

Upon individually checking the cluster i can see that the cluster looks good, it formed around some common words.  
Using TF\_IDF to avoid such words are highly recommended.