LOAD DATASETS

Why did you convert the xlsx to csv and then read in csv instead of reading-in the xlsx files?

ANSWER: Simply because its much faster to read csv than xlsx.

PRE-PROCESSING THE DATA – Seems clear

CLUSTERING UNSUPERVISED ANALYSIS

I think I understand that this section is creating the DTM on the unsorted data

ANSWER: I thought you wanted to classify unsorted data.

K-MEANS CLUSTERING & K-MEANS CLUSTERING PLOT & HIERARCHICAL CLUSTER

I believe these sections computes clustering, distances and then multidimensional scaling on the unsorted data. I don’t understand why these computations are performed on the unsorted data.

ANSWER: Same answer as previous one.

MODEL PERFORMANCE

Again, I am interpreting in the belief that this is the unsorted data, so I don’t understand how the prediction come from 'predicted' = kmeansResult$cluster.

When I run this code: #ggplot(data = result.agg, aes(x=actual, y=predicted, size = counter)) + geom\_point(). I get the error that “result.agg” not found. I assume this should be “result” not “result.agg”. Unless of course there was some function applied to “result” to obtain “result.agg.”

When I change “result.agg”to “result” then I get the error, “object 'counter' not found.” I searched the counter and cannot find “counter” mentioned elsewhere.

ANSWER: Here, you will get final data frame from predictions of unsorted data and clustered to groups. For example, in the future you have to load unsorted data, do preprocessing and run K-Means algorithm, where it will give final data frame with 2 columns actual(actual text) and predicted(clustered).

SUPERVISED ANALYSIS & PRE-PROCESS & SPARSE & SPLIT DATA INTO TRAINING AND TESTING

These analyses are performed on the sorted data and seems straightforward.

CART MODEL

When I run this code:

# Build a CART model

outputCART <- rpart(clusters ~ ., data = trainSparse, type = "class")

prp(outputCART)

I get this error:

Error in rpart(clusters ~ ., data = trainSparse, type = "class"): Argument type not matched

Traceback: 1. rpart(clusters ~ ., data = trainSparse, type = "class")

2. stop(gettextf("Argument %s not matched", names(extraArgs)[indx ==

. 0L]), domain = NA)

ANSWER : Ignore CART Model. I ran several algorithms to find which one fits our data. Instead you can use Random Forest.

I had to stop here.