WORD SIMILARITY – AMAZON REVIEWS DATASET

Word Similarity: using cosine function:

Clear the Ram and import the dataset:

```
> setwd("C:/Users/hp/Desktop")
> amazon=read.csv("amazon.csv")
```

Data Cleaning and Re-organizing:

```
> library(tm)
Loading required package: NLP
> regex_func=function(x){
+   return (gsub('[^a-z ]','',x))
+ }
> custom_stopwords=c()
> common_stopwords=stopwords()
> all_stop_words=c(custom_stopwords,common_stopwords)
> docs=as.character(amazon$reviewText)
> docs=VCorpus(VectorSource(docs))
> docs=tm_map(docs,content_transformer(tolower))
> docs=tm_map(docs,content_transformer(regex_func))
> docs=tm_map(docs,removeWords,all_stop_words)
> docs=tm_map(docs,stripWhitespace)
> dtm=DocumentTermMatrix(docs)
> df_dtm=as.data.frame(as.matrix(dtm))
```

User defined Function to display similarity between words:

```
> library(LSAfun)
Loading required package: lsa
Loading required package: SnowballC
Loading required package: rgl

> similarity_between_words=function(word,df_dtm){
+   result=data.frame()
+   for (column in colnames(df_dtm)){
+     if (word!=column)
+     {
+       x=cosine(df_dtm[,column],df_dtm[,word])
+       temp=data.frame(word,column,x)
+       result=rbind(result,temp)
+     }
+   }
+   colnames(result)=c("Word","Column Name","Cosine_Value")
+   filtered_result=result%>%arrange(-Cosine_Value)%>%head(10)
+   return(filtered_result)
+ }

> similarity_between_words("nook",df_dtm)
```

WORD SIMILARITY – AMAZON REVIEWS DATASET

Words having highest similarity with the word “nook”:

```
   Word Column Name Cosine_Value
1  nook         can    0.7174085
2  nook      kindle    0.6755382
3  nook       books    0.6664996
4  nook        like    0.6178033
5  nook        book    0.6071614
6  nook        also    0.5954705
7  nook        read    0.5868381
8  nook        will    0.5741792
9  nook        look    0.5628406
10 nook     reading    0.5609512
```