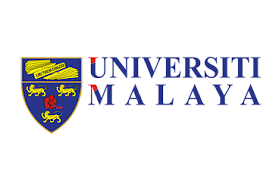
****

**Online Assignment week 7**

**Data Mining**

**WQD7005**

**Members:**

|  |  |
| --- | --- |
| **Name** | **Matric No.** |
| **Sidratul Muntaha** | **WQD180079/ 17199166** |
| **Nuraini Afiqah Rohanip** | **WQD180115/17198049** |

**Week 7 - Data Cleaning on Crawled Dataset**

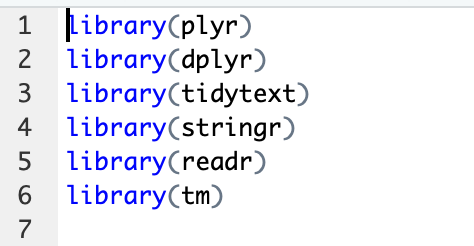
**Brief summary of the data**

There are several datasets that we have crawled in order to answer the objectives of our project namely reddit dataset, reddit comments dataset, and twitter dataset. All of these datasets are about consumer reviews on different laptop brands such as Apple, Dell, Asus, Lenovo, Acer, and HP.

**Cleaning Reddit dataset:**

**Step 1: Import library**

Import necessary libraries such as plyr, dplyr, tidytext, stringr, readr, tm and lubridate. Each library is chosen due to its applicable function to our dataset.



**Plyr**: Tools for splitting, applying and combining data.

**Dplyr**: Act as grammar for data manipulation. Consistent tool to work on objects in the data frame in memory and out of memory.

**Tidytext**: Text mining for word processing using tidy tools like dplyr, ggplot2 etc.

**Stringr**: String wrapper for common string operations. All function, argument names and positions are consistent, all functions deal with ‘NA’ values and zero length vectors in the same way and are easy to feed output of a function into another.

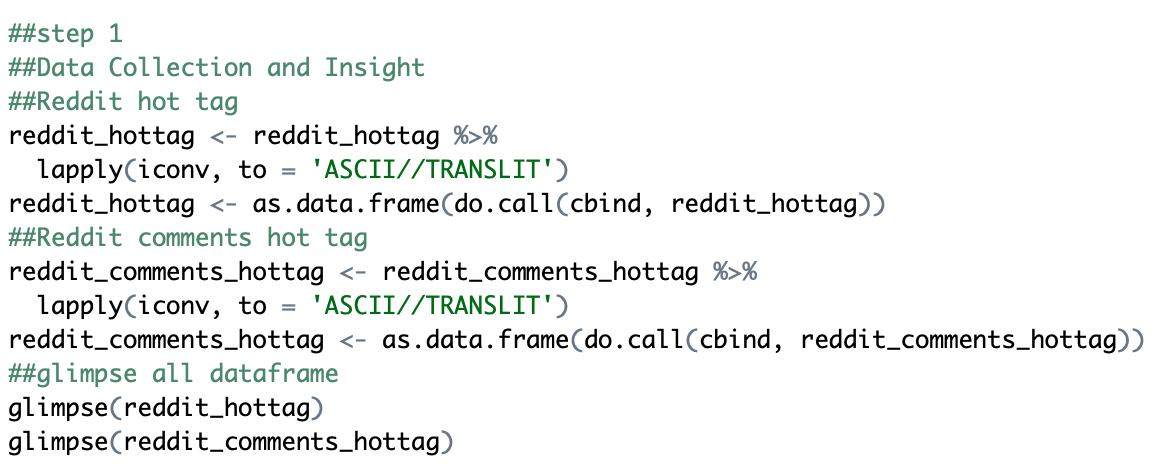
**Readr**: Provide a fast and friendly way to read rectangular data like csv.

**Tm**: Framework for text mining applications within R.

**Step 2: Import data into R**

Import the data into R by using read.csv() function and then glimpse() to take a look at the structure of the data.

The iconv function is used in order to convert the character vectors between encodings.

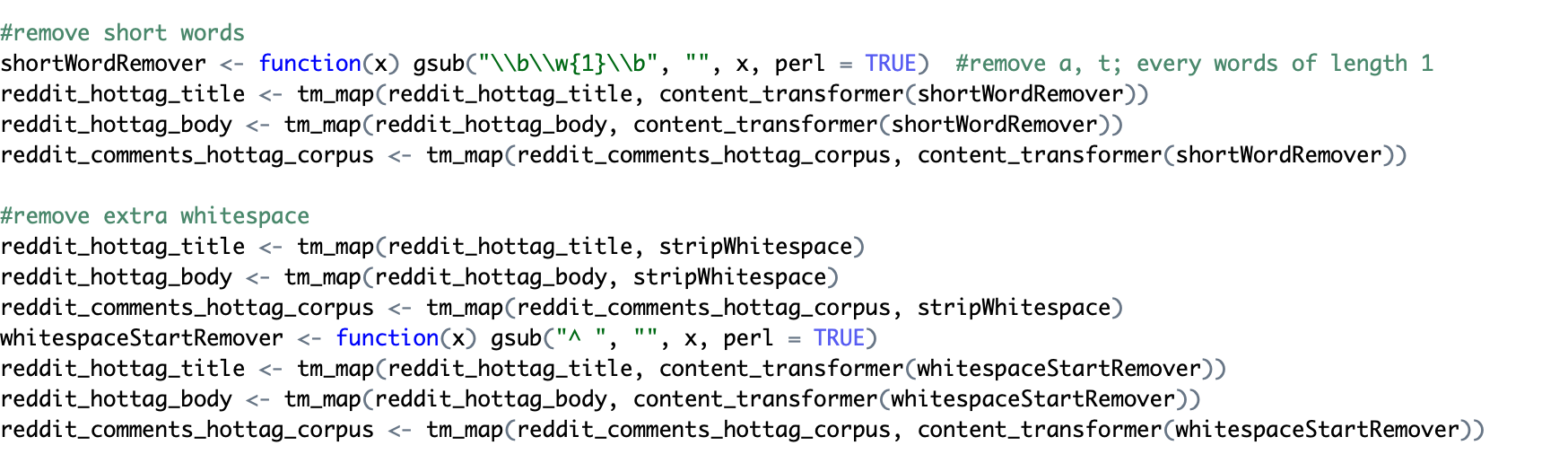


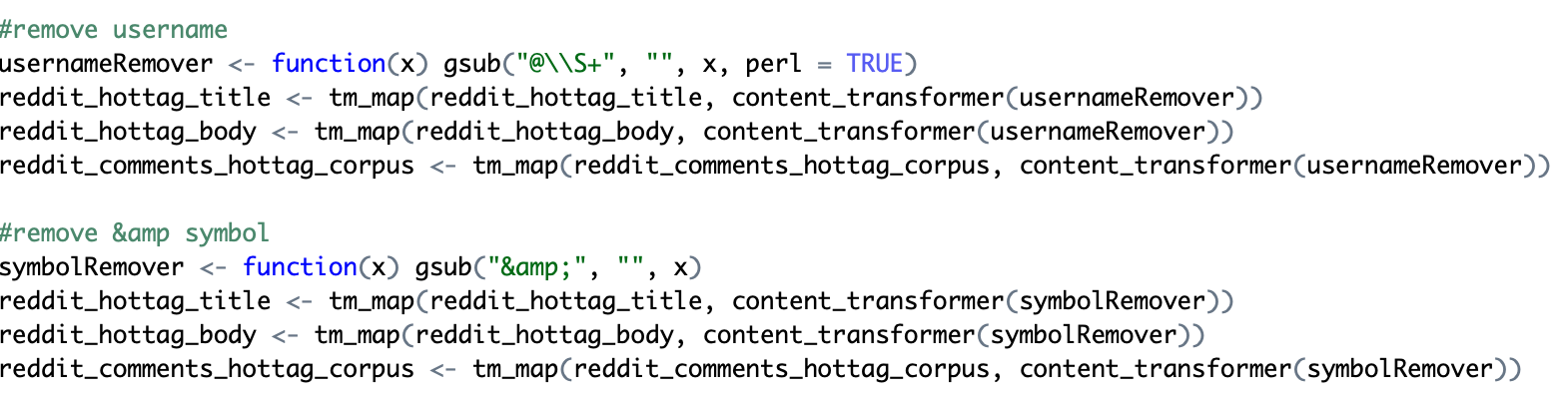
**Step 3: Removing unrequired data**

Removing all the contents that will not be required for further analysing. Like the short words, profane words, hashtag expressions, and irrelevant words and punctuations .

Removing Punctuations, profane words, whitespace, usernames, and symbols steps given below:







**Step 4: Clean the time column and save as csv file**

Use mutate() to append the new and clean columns to the dataframe. Drop the old columns using indexing. Glimpse() function is applied several times to make sure the data is now consisting of only clean columns.

Convert the date and time column into characters in order to format the column as year-month-day hour:minute:second using POSIXct. Then came the next part where we convert the time of the day into either morning, afternoon, evening, or night. This can be useful for later observations on the time of day laptop consumers frequently post on social media.

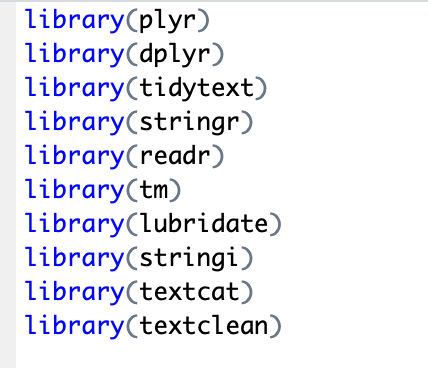
**Run is.na()** to see whether there are any missing values from the reddit dataset. Since there are none, we proceed to the final step which is saving the clean dataset as a csv file.



**Cleaning Twitter data set:**

**Step 1: Importing the library.**

The following plyr, dplyr, tidytext, textcat libraries were called. Some additional libraries were imported while cleaning the twitter dataset.



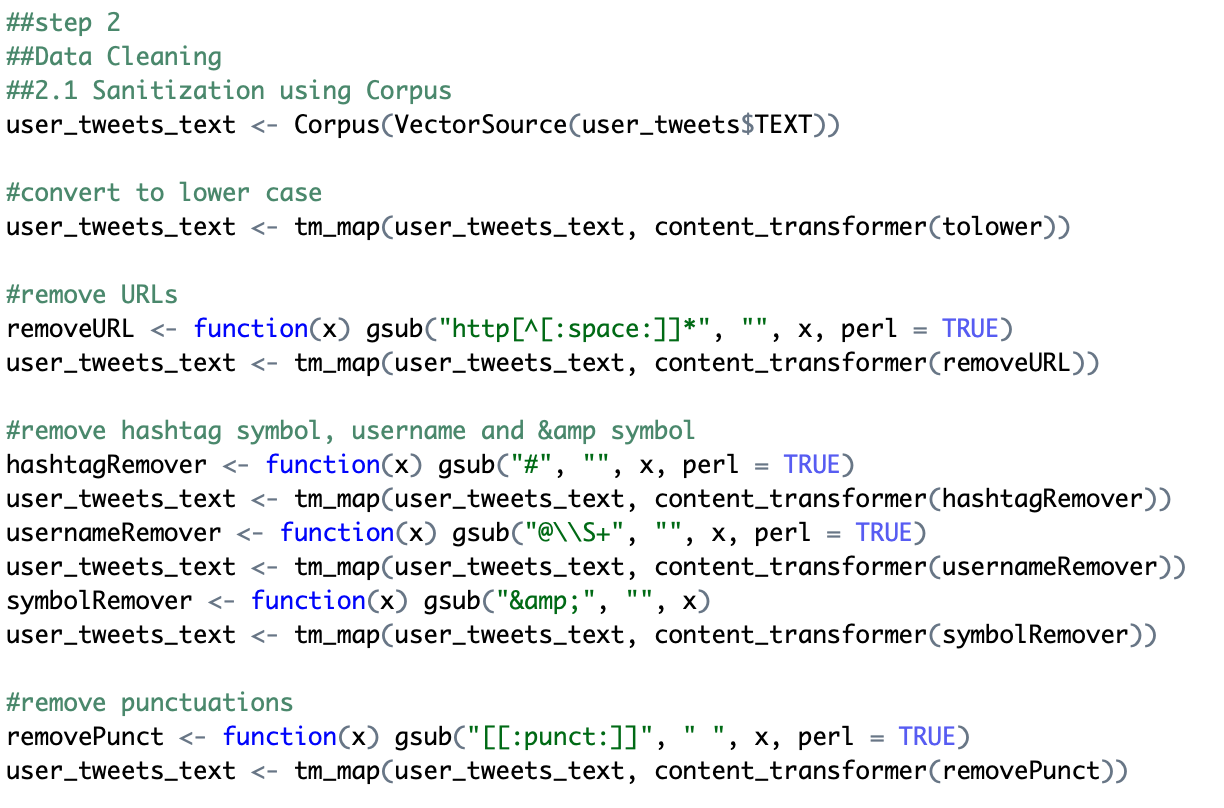
**Step 2: Import the data into R**

After importing the libraries the twitter dataset was called and converted to required type as before. The difference in twitter dataset than what we did for reddit dataset is there is an addition of applying iconv to convert the characters for latin words too, as opposed to only ASCII characters.



**Step 3: Cleaning the text column**

The corpus library is used here. **Corpus** is a collection of text document over which text mining outlines to derive inferences. In tm R package, there are functions to create corpus from files, vectors etc. Along with transforming all texts to lower the urls,symbols and punctuations are erased from the data.



**Step 4: Handle missing values and save as csv file**

Contrary to the Reddit dataset where there are no missing values, twitter dataset has several empty cells in the text column. We replace the empty cells with NA and then remove all rows with missing values. Lastly, the clean set of dataframe is saved as a csv file.

