Oduntan Oluwagbemiga Kolawole
Student Id: 28113306
Email: oko1n14@soton.ac.uk

Preprocessing of the Data: Data Cleaning

I use R Studio in preprocessing the unstructured data. Packages like "tidyr", "devtools", and DSR were installed in R studio for easy cleaning of the dataset. Syntax: ***install.packages(c("tidyr", "devtools"))***, ***devtools::install_github("garrettgman/DSR")***. Read the dataset into R Studio: ***microblog <- read.csv("~/Desktop/mongodb/microblog.csv")***. For every column and raw in the dataset, i check for every integer in the column that is not within 0-9 using: ***ind <- grep("[^0-9]", microblog$id)***. Where "***microblog$id***", represents the column i want to search for. I observed in the output that some rows have letters with the integers in $id.

I used the syntax: ***ind <- grep("[^0-9]", nameofdatabase$id)*** for every column in the dataset to output integers or strings that don't belong to the column and fill in missing values in every integer and strings in the column.

Installing and loading the rmongodb package

rmongod package have to be installed in R studio to be able to connect to Mongo. To install packages: ***install.packages("rmongodb")***. To load: ***library(rmongodb)***.

Connecting R studio to MongoDB

I first of all created a connection to the mongoldb installation: **mongo <- mongo.create()**. To know if its is connected: ***mongo.is.connected(mongo)***. It outputs ***"True"***.

Getting Databases and collections

To know the databases in Mongoldb connection: ***mongo.get.databases(mongo)***. To get the collections in a specific database: ***mongo.get.database.collections(mongo,"coursework")***. Where collection = "tweets", Database = "coursework".