|  |
| --- |
| **Data Dictionary**   * I took racial slur words from the link below:   <https://ora.ox.ac.uk/objects/uuid:9232010f-72c5-44d7-aaa6-9e7896e13ea8/download_file?file_format=application%2Fpdf&safe_filename=Racial%2Bslurs%2Bin%2Bdictionaries&type_of_work=Conference+item>   * I took twitter public tweets. |

**Api\_keys and Access\_tokens**

* First I created configuration file(config.ini) as it contains api\_keys and access\_tokens. So that we can use that file in the code withour exposing the keys and tokens.
* Then using the configparser library, accessed the config.ini file and its information.
* Then did authentication using tweepy package and accessed the public tweets.
* Checked the tweets, time, user of each tweet and kept all tweets into a dataframe “tweets.csv”.

**Data Cleaning:**

* Accessed the tweets.csv file and took only tweets from it.
* Using join, I appended all the tweets into paragraph.
* In the main function: I make data into lowercase, removed numbers and links of https.
* Then did lemmatizaton but did not work properly. So I did parts of speech tagging.
* Did stopwords removal using STOPWORDS package from wordcloud.

**Calculating the degree of profanity:**

* Created separate column, so that it took numbers and then counted number of slur words are there in total.
* Calculated total number of sentences
* I calculated degree of profanity for each sentence using “total number of slur words/total number of sentences” and got degree as 0.2222222222222222
* And kept the entire data cleaning and profanity calculating in main function.