AffinityAnswers Assignment

SANJAY.N|DATE : 04-12-2022

Problem Statement

Imagine there is a file full of Twitter tweets by various users and you are provided a set of words that indicates racial slurs. Write a program that can indicate the degree of profanity for each sentence in the file.

In [6]:
```
import re
from nltk.tokenize import word_tokenize
from nltk.stem.porter import *
stemmer = PorterStemmer()

def remove_pattern(input_txt, pattern):
  r = re.findall(pattern, input_txt)
  for i in r:
    input_txt = re.sub(i, '', input_txt)

  return input_txt

t_file = open('affine.txt',encoding="utf8")                    # Opening file

racial_slur = {"Nigger","wasabi","white"}                        # list of racial slurs words


lines = t_file.readlines()
s =0
for line in lines:
    line = line.lower()
    line = remove_pattern(line,"@[\w]*")            #we need to remove Twitter handles if files contains that.

    line = re.sub(r"[^a-zA-Z0-9]"," ",line)         # Remove special characters, numbers, punctuation
    line = word_tokenize(line)

    #degree_of_profanity of that line
    degree_of_profanity = sum(1 for t in line if t in racial_slur)/ len(line)
    print("line",s,":",degree_of_profanity)
    s +=1
```

```
line 0 : 0.0
line 1 : 0.0
line 2 : 0.07692307692307693
line 3 : 0.0
line 4 : 0.04
```

2. Which is an interesting data set you discovered recently?

Currently I'm working under my prof on Research topic Music Therapy. where we took data from students who loves to listen music and who don't then we planned to observe there results and other extra curricular activities. It was an intresting topic to me.

3. Why do we need a database? We can store everything in a file, no?

you are correct we can store everything in a file but while we will be dealing with large amount of data, searcing or retrieval will be very difficult. i.e while i was working in microsoft i was required to optimize database query so i make use of index to store and get the data which makes retrieval very easy and O(1) other wise it would be O(n). it makes files inefficient in dealing with large data.

lot of security issues will arises and there will be always vulnerable to attacks.

4. How well versed are you on the Unix command line?

I'm preety decent with that. but still there is lot to learn.

In [ ]:
```

```