

Laboratory Experiment for Practical Exam AI & Application Lab (14B17CI671)

Names /
Roll NO:

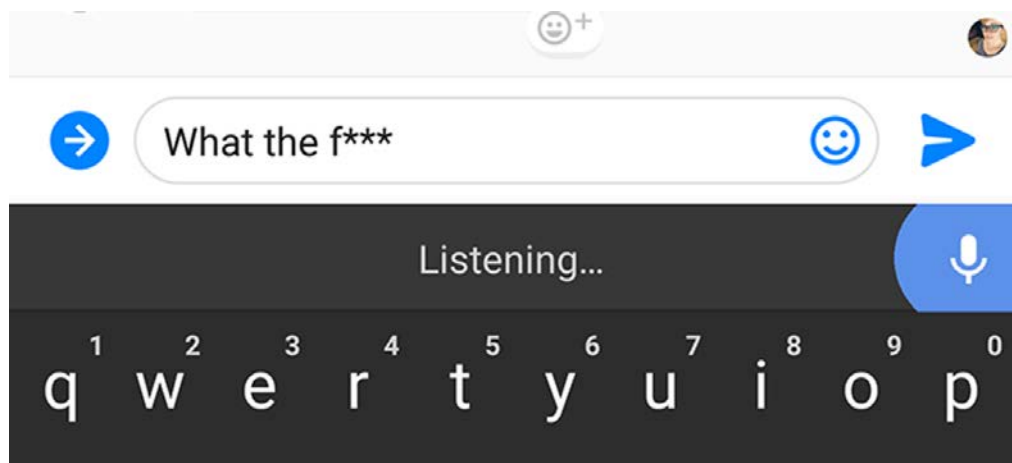
1. Yash Raj Pandey (191B295)
2. Samarth Dubey (191B304)
3. Aryman Tripathi (191B309)
Hrithik Bansal (191B298)

Dept: CSE

Expt. Title: **"Profanity Chat Filter"**

Date:
22/03/21

Description:



A Python script to detect language of text and filter out the OFFENSIVE Words.

Profanity Checker is a Python script that uses the Natural Language Toolkit (NLTK) library to solve the usual problem of toxicity in society.

The problem was divided into below three phases:

1. Detect the language of the text
2. Fetch offensive words for the language detected
3. Output the bad words in each line of the text

Language detection can be achieved by using the 'stopwords' function as provided by the Python's NLTK library. Datasets containing the most common bad words in each of the 4 languages - English, Spanish, French and German are attached as the CSV files. User can be asked to test out in any of the 4 languages mentioned above.

Approach/
Algorithms

- Importing the sys module of python
- Importing the time module
- Importing the NLTK Library for symbolic and statistical natural language processing for English written in the code

- Calculating the language probability based on the ratio probability of the stopwords
- We load the bad words for the specific language after the language detection
- Iterating over the words in each line we check for whether it is offensive or not
- Finally, printing over the line number containing the offensive word and the number of offensive words in the specific line

Test Procedures

- Importing all the libraries required for the code

```
import sys
import time
```

- Using stopwords from NLTK Library

```
try:
    from nltk import wordpunct_tokenize
    from nltk.corpus import stopwords
    print("nltk is installed. \n")
except ImportError:
    print("You need to install nltk (http://nltk.org/index.html)")
```

- Starting off with the calculate_language_ratios function, we compute per language included in nltk number of unique stopwords appearing in analyzed text

```
def calculate_languages_ratios(text):
    languages_ratios = {}
    tokens = wordpunct_tokenize(text)
    words = [word.lower() for word in tokens]

    # Compute per language included in nltk number of unique stopwords appearing in analyzed text
    for language in stopwords.fileids():
        stopwords_set = set(stopwords.words(language))
        words_set = set(words)
        common_elements = words_set.intersection(stopwords_set)

        languages_ratios[language] = len(common_elements) # language "score"

    return languages_ratios
```

- Moving on, we use the load_bad_words function to utilise the included offensive dataset for the specific language from the .csv files

```
def load_bad_words(language):
    if language.upper() in ['ENGLISH', 'FRENCH', 'SPANISH', 'GERMAN']:
        badwords_list=[]
        lang_file = open('datasets/'+language.lower()+'.csv', 'r')
        for word in lang_file:
            badwords_list.append(word.lower().strip('\n'))
    return badwords_list
```

- Coming on to the detect_language function, we calculate the ratios of the language detected from the calculate_language_ratios function and then maxing out on the ratio for a specific language

```
def detect_language(text):
    ratios = calculate_languages_ratios(text)
    most Rated language = max(ratios, key=ratios.get)
    return most Rated language
```

- Now, we take input from the user on the whereabouts of the file. We input the whole text file as a paragraph and then read it line by line

```
print ("Language can be in the form of: english, french, german and spanish.")
print ('\n')
testlang = input ("Language to test: ")
print('\n')
filename = 'test-files/'+testlang.lower()+'.txt'
file = open(filename, 'r')
text = ''
line_count=1
for i in file:
    text+=str(line_count)+'| '+i
    line_count+=1
time.sleep(2)
```

- Once we get the input, we read the file and show the user his written paragraph with line number and then move on to show the language detected. Continuing we move on to check for offensive words in the file.

```

print ('-----Input Text-----')
print ('\n')
print (text)
print ('\n')
print ('-----Text Read-----')
print ('\n')
language = detect_language(text)
print ('\n')
time.sleep(1)
print ('-----')
print ('Language Detected: ', language.upper())
print ('-----')
print ('\n')
time.sleep(1)
print ('-----')
print ('Checking for offensive words in '+language.upper()+'.')
print ('-----')
print ('\n')
badwords = load_bad_words(language)
badwords = set(badwords)
text_list = text.split('\n')

```

- Iterating over the words in each line, we remove the unnecessary punctuations and convert all the words into lowercase so as make it easy to read.

```

for sentence in text_list:
    line_number = str(text_list.index(sentence)+1)
    for key in [',', '.', '!', '?', '"', '"', '!', '!', ':', ';', '(', ')', '[', ']', '{', '}']:
        sentence = sentence.replace(key, '')
    abuses=[i for i in sentence.lower().split() if i in badwords]

```

- Lastly, checking for abusive words in the specific language we go through each and every line to identify and print the offensive word

```

if abuses == []:
    continue
else:
    time.sleep(0.5)
    print (str(len(abuses))+' Offensive Words found at line number: '+line_number)
    x_words=''
    for i in abuses:
        x_words+=i+' '
    print ('Offensive Words: '+x_words[:-2])
    print ('\n')

```

Results:

- Running the code for “English”, we run the script in Terminal and get the offensive words as the output on our screen

```
C:\Users\yashp\Desktop\AIA Project>python "Profanity Filter.py"
nltk is installed.
Language can be in the form of: english, french, german and spanish.
Language to test: english

-----Input Text-----
1 There is a passage that I got memorized, Ezekiel 25:17. "The path of the righteous man is harassed on all sides
2 and the inequality of the selfish and the tyranny of evil men. Blessed is he who, in the name of charity
3 and goodwill, shepherds the weak through the valley of darkness, because he is truly the guardian of his brother
4 and the search engine for lost children. And I'm going to tear you down with great vengeance and furious anger
5 those who try to rob you of your brothers. And they will know that I am the Lord when I put my revenge
6 over you. "Now...I've been saying that shit for years, and if you ever heard it, that meant your ass.
7 I'm dead right now, I never thought much about what it meant. I just thought it was a cold-blooded thing
8 I'll tell a son of a bitch before he blew a hat on his ass. But I saw some shit this morning made me think
9 twice. Look, now I'm thinking: maybe it means you're the bad man. And I am the just man. And he
10 9mm here...he is the shepherd who protects my ass right in the valley of darkness. Or it could mean
11 you are the righteous man and the shepherd, and the world is evil and selfish. And I would like
12 that. But that shit is not the truth. The truth is that you are the weak and I am the tyranny of evil men.
13 But I'm trying, Ringo. I'm trying to be the pastor.
14
15 This man, Ringo, he is an asshole.
16 Sometimes, I really think of dumping him and fucking his dirty ass.
17 He is a bloody jerk.

-----Text Read-----

Language Detected: ENGLISH

Checking for offensive words in ENGLISH.

2 Offensive Words found at line number: 6
Offensive Words: shit, ass

3 Offensive Words found at line number: 8
Offensive Words: bitch, ass, shit

4 Offensive Words found at line number: 10
Offensive Words: ass

4 Offensive Words found at line number: 12
Offensive Words: shit

4 Offensive Words found at line number: 15
Offensive Words: asshole

2 Offensive Words found at line number: 16
Offensive Words: fucking, ass
```

- Running the code for “French”

```
C:\Users\yashp\Desktop\AIA Project>python "Profanity Filter.py"
nltk is installed.
Language can be in the form of: english, french, german and spanish.
Language to test: french

-----Input Text-----
1 Il y a un passage que j'ai mémorisé, Ézéchiel 25:17. "Le chemin de l'homme juste est harcelé de tous les côtés
2 par les inégalités de l'égoïsme et la tyrannie des hommes mauvais. Béni soit celui qui, au nom de la charité
3 et la bonne volonté, les berges, les faibles à travers la vallée de l'obscurité, parce qu'il est vraiment le gardien de son frère
4 ceux qui essaient d'empoisonner et de détruire mes frères, et ils savent que je suis le Seigneur quand je mets ma vengeance
5 sur vous. "Maintenant, je ne pense pas à ce que cela signifiait. Je pensais juste que c'était une chose de sang-froid
6 que vous m'avez dit. Je pensais à un chapeau sur le cul. Mais j'ai vu de la merde ce matin et ça m'a fait penser
7 deux fois. Regardez, maintenant je pense, peut-être que cela signifie que vous êtes le méchant. Et je suis l'homme juste. Et M.
8 Telle est la pensée qui traverse mon cul juste dans la vallée de l'obscurité. Ou cela pourrait signifier
9 cette. Mais cette merde, c'est parce que le monde est mauvais et méchant. Et je voudrais
10 Ruler j'essaye de l'être. La vérité est que vous êtes les faibles et je suis la tyrannie des hommes méchants.
11
12 Cet homme, Ringo, il est un conard.
13 Un jour, je pense vraiment à le jeter et à baiser son sale cul.
14 Il est un chétif rampant.

-----Text Read-----

Language Detected: FRENCH

Checking for offensive words in FRENCH.

2 Offensive Words found at line number: 6
Offensive Words: merde, cul

3 Offensive Words found at line number: 8
Offensive Words: pute, cul, merde

4 Offensive Words found at line number: 10
Offensive Words: cul

4 Offensive Words found at line number: 12
Offensive Words: merde

4 Offensive Words found at line number: 15
Offensive Words: connard

2 Offensive Words found at line number: 16
Offensive Words: baiser, cul
```

- Running the code for “German”

```
C:\Users\yashp\Desktop\AIA Project>python "Profanity Filter.py"
nltk is installed.
Language can be in the form of: english, french, german and spanish.
Language to test: german

-----Input Text-----
1 Es gibt eine Passage, die ich auswendig gelernt habe, Hesekiel 25:17. "Der Weg des Gerechten ist von allen Seiten belästigt
2 und die Ungleichheit der Selbstsuche und die Tyrannei der Bösen. Segen sei dem, der im Namen der Nächstenliebe
3 und der Gutmütigkeit, die Schwachen durch das Tal der Finsternis, weil er wirklich der Hüter seiner Brüder
4 ist, die versuchen, sie zu vergiften und zu zerstören. Und sie wissen, dass ich der Herr bin, wenn ich mich räche.
5 Aber ich habe ihnen schon seit Jahren gesagt, und wenn sie es jemals gehört haben, hat das deinen Arsch bedeutet.
6 Ich bin jetzt wieder daran zu denken, dass es vielleicht bedeutet, dass du der Böse bist, und ich der Gerechte.
7 Und ich habe heute Morgen gesehen, dass du ein Scheißkerl bist, der dich selbst zum Nachdenken gebracht hat.
8 Ich habe heute Morgen gesehen, dass du ein Scheißkerl bist, der dich selbst zum Nachdenken gebracht hat.
9 Ich habe heute Morgen gesehen, dass du ein Scheißkerl bist, der dich selbst zum Nachdenken gebracht hat.
10 Ich habe heute Morgen gesehen, dass du ein Scheißkerl bist, der dich selbst zum Nachdenken gebracht hat.
11 Ich habe heute Morgen gesehen, dass du ein Scheißkerl bist, der dich selbst zum Nachdenken gebracht hat.
12 Ich habe heute Morgen gesehen, dass du ein Scheißkerl bist, der dich selbst zum Nachdenken gebracht hat.
13 Ich habe heute Morgen gesehen, dass du ein Scheißkerl bist, der dich selbst zum Nachdenken gebracht hat.
14 Ich habe heute Morgen gesehen, dass du ein Scheißkerl bist, der dich selbst zum Nachdenken gebracht hat.
15 Ich habe heute Morgen gesehen, dass du ein Scheißkerl bist, der dich selbst zum Nachdenken gebracht hat.
16 Ich habe heute Morgen gesehen, dass du ein Scheißkerl bist, der dich selbst zum Nachdenken gebracht hat.

-----Text Read-----

Language Detected: GERMAN

Checking for offensive words in GERMAN.

4 Offensive Words found at line number: 6
Offensive Words: arsch

2 Offensive Words found at line number: 8
Offensive Words: hurensohn, arsch

4 Offensive Words found at line number: 10
Offensive Words: arsch

4 Offensive Words found at line number: 12
Offensive Words: scheißkerl

4 Offensive Words found at line number: 15
Offensive Words: arschloch

2 Offensive Words found at line number: 16
Offensive Words: arsch, ficken
```

- Running the code for “Spanish”

```

C:\Users\yashp\Desktop\AI4 Project>python "Profanity Filter.py"
Itk is installed.
Language can be in the form of: english, french, german and spanish.
Language to test: spanish

-----Input Text-----
1 Hay un pasaje que me han memorizado, Ezequiel 25:17, "El camino del hombre justo es acosado por todos lados
por las desigualdades de los egoístas y la tiranía de los hombres malvados. Bienaventurado el que, en nombre de la caridad
y buena voluntad, pastorea a los débiles a través del valle de las tinieblas, porque así es verdaderamente el ciudadano de su hermano
y el buscador de niños perdidos. Y voy a devolverte con gran venganza y enojo fútilos
aquellos que habían de enseñarte a destruir a mis hermanos. Y sabrás que yo soy el Señor cuando puse mi venganza
sobre ti. Ahora... me estoy diciendo mierda por dentro. Y si alguna vez la escuchaste, eso significaba un fracaso.
estar muerto ahora mismo. nunca pensó mucho sobre lo que significaba. Solo pensó que era una cosa de zanja fría
digo, mira, ahora estoy pensando: tal vez significa que eres el hombre malo. y yo soy el hombre justo. Y el Sr.
"me acordé... Ahí es el pastor que pastorea al hombre justo en el valle de la oscuridad. O podrá significar
que eres el hombre justo y yo el pastor, y el mundo es malo y egoísta y me gustará.
ese. Pero era mierda no es la verdad. La verdad es que usted es el débil y yo soy la tiranía de los hombres malvados.
Pero lo estoy intentando, Ringo. Estoy tratando de ser el pastor.
Este hombre, Ringo, es un imbécil.
En algún momento, realmente pienso en tirarlo y follar su sucio culo.
Así es un maldito idiota.

-----Text Read-----

Language Detected: SPANISH

-----Checking for offensive words in SPANISH-----
1 Offensive Words found at line number: 6
Offensive Words: mierda
2 Offensive Words found at line number: 8
Offensive Words: puta, mierda
1 Offensive Words found at line number: 12
Offensive Words: mierda
1 Offensive Words found at line number: 15
Offensive Words: imbécil
2 Offensive Words found at line number: 16
Offensive Words: follar, culo
2 Offensive Words found at line number: 17
Offensive Words: maldito, idiota

```

Conclusion/ Remarks

Completing this Project, Profanity Checker was a lot overwhelming for us (Yash, Samarth and Aryman).

- Learned over NLTK Library and usage in language detection
- Learned the usage of stopwords and the removal of articles despite the meaning remaining the same
- Added bad-words CSV datasets for 4 languages so as the data set can be used while reading the input
- Facility for user to give in the info for which language he'd like to be tested by our code

Evaluation/ Marks