The Python Program I Used to Calculate Relative Frequency of Letters in Bangla Text

The program is named “bang4.py” that as input takes the pdf files of novels “Lalshalu (1948)” by Syed Waliullah and “Nondito Noroke (1970)” by Humayun Ahmed which are the 2 renowned novels of Bangla literature (these are also uploaded on the Google Classroom with the same name). It uses the font Nikosh to display Bangla letters on the program generated bar chart (The font file named “nikosh.ttf” is uploaded in Google Classroom). The program outputs the generated image containing the bar chart (uploaded as ProgramOutputFig.png in Google Classroom) that is produced according to calculation in the same folder and a TSV (uploaded as Bn_out.tsv in Google Classroom) file that can also be used for creating bar charts.

Required Libraries

Running this python program requires python-3 preinstalled and installing the python libraries numpy, matplotlib and pdftotext. Of the libraries numpy and matplotlib that are used for numerical calculation and plotting respectively can easily be installed using the command

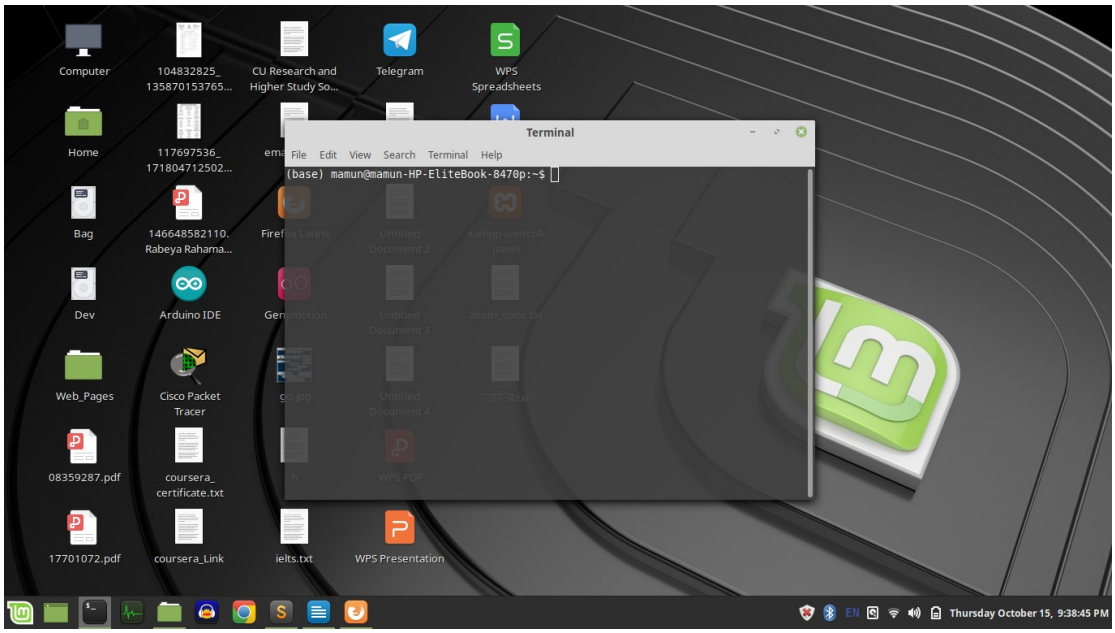
```
pip install numpy, matplotlib
```

in the command line interface (CLI; CMD for Windows, Terminal for Linux and Mac). Installing the library pdftotext is quite a bit tricky. The instructions given in this web-page where operating system specific guidelines and CLI command(s) are provided is to be followed for installation of the last required library.

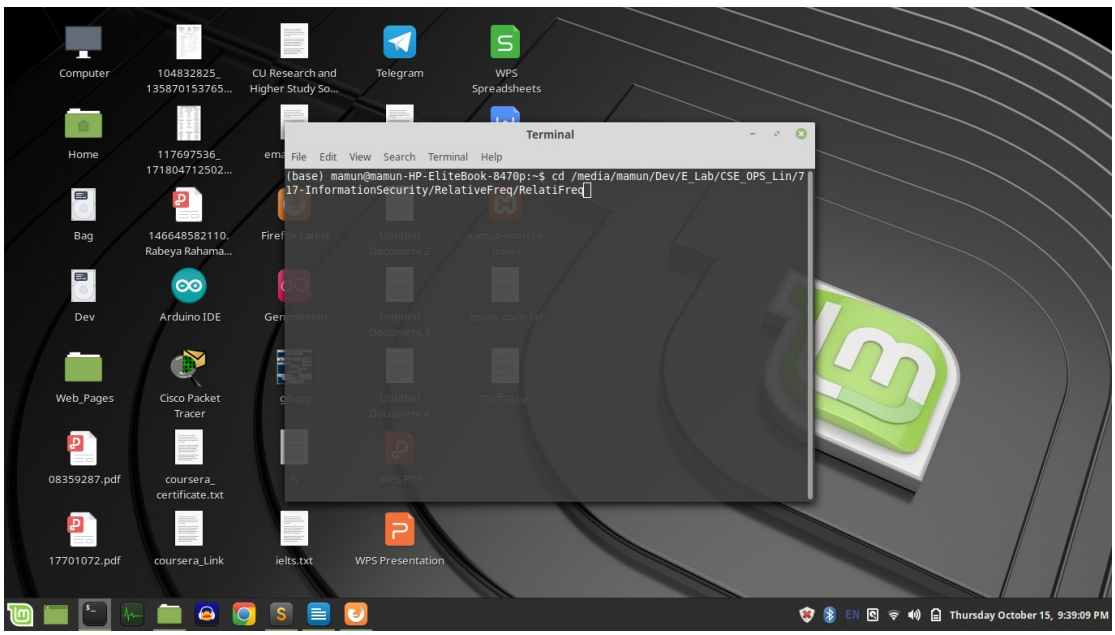
Running the Program

This Python program can be easily run from the command line interface (CLI; CMD for Windows, Terminal for Linux and Mac) as below.

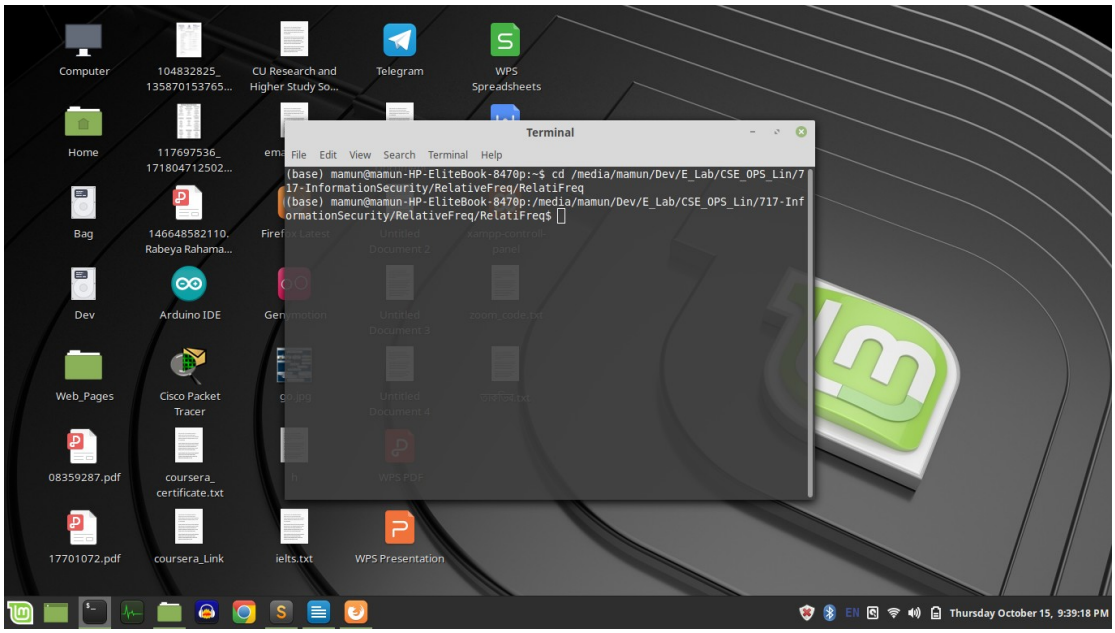
1. At first the command line interface is to be opened.



2. Then the current directory should be changed to the directory where the program resides as below.



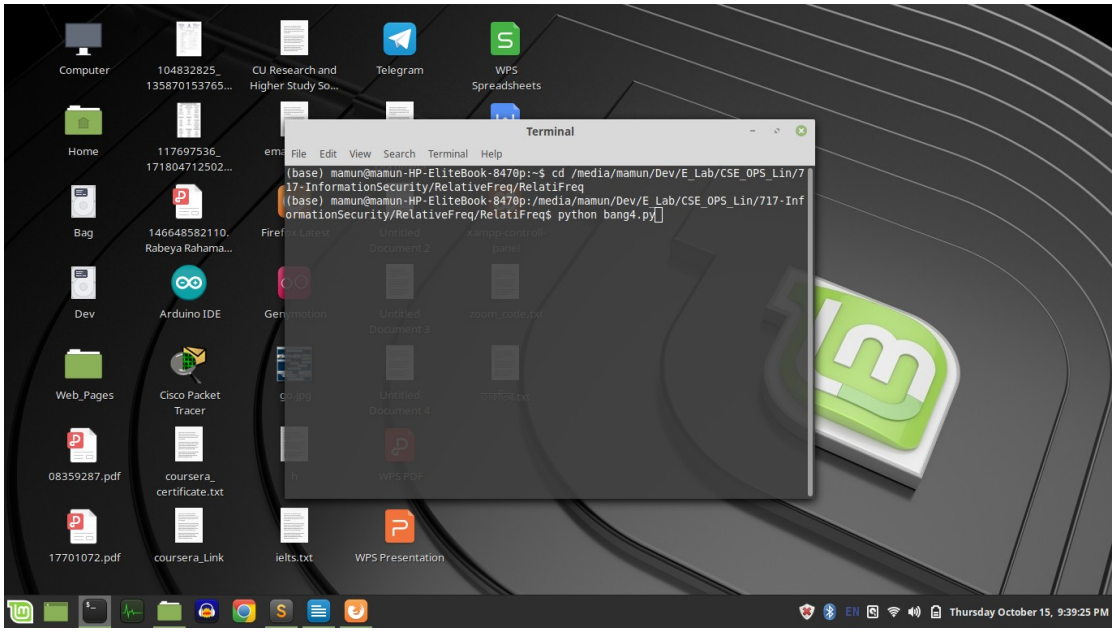
3. When the required directory containing the program is accessed as below,



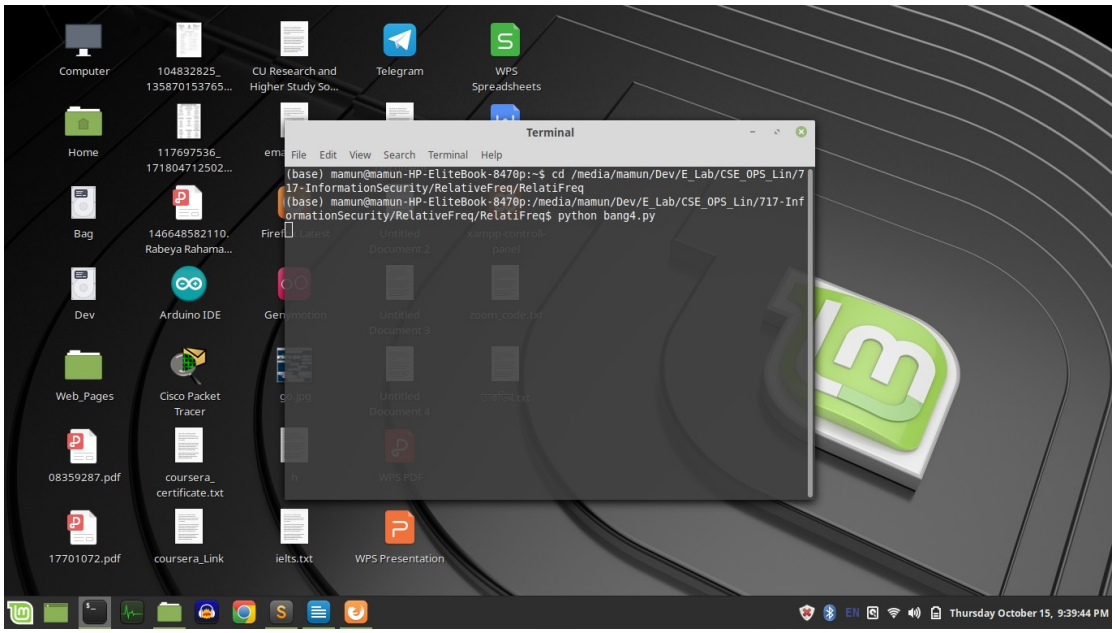
the command

python bang4.py

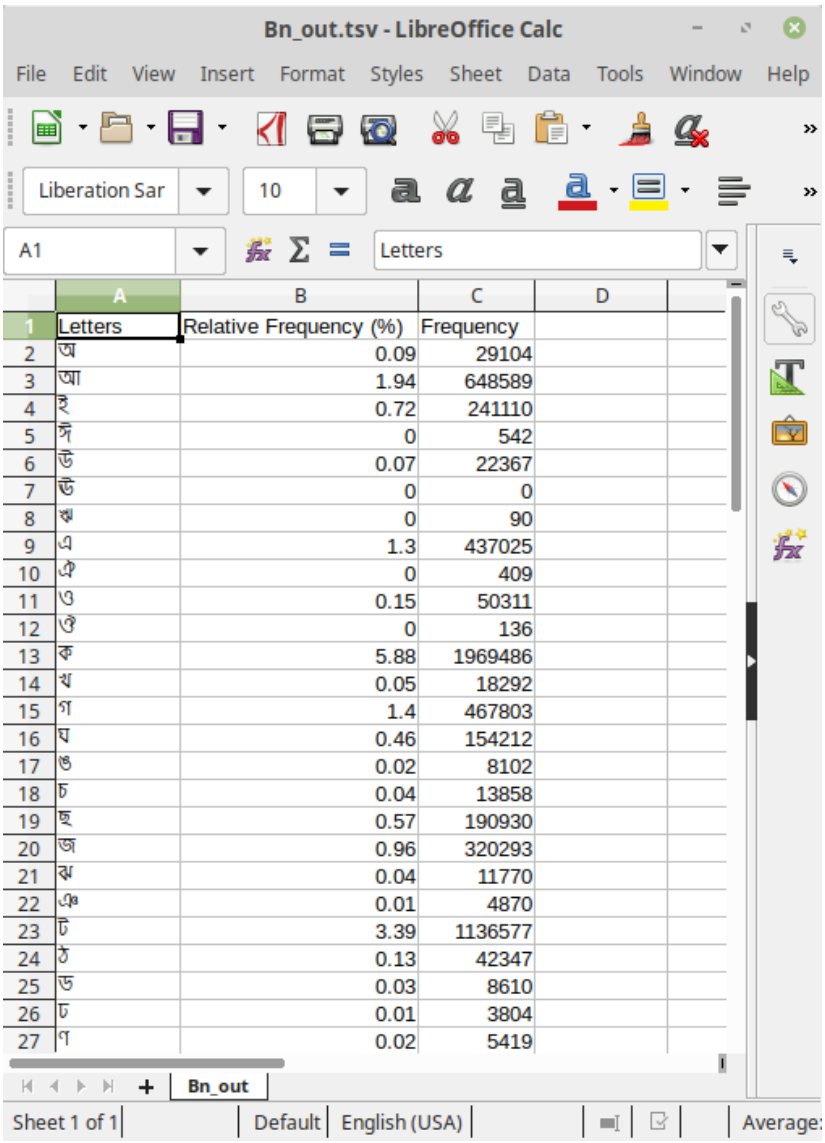
should be written in the command line and enter-key is to be pressed.



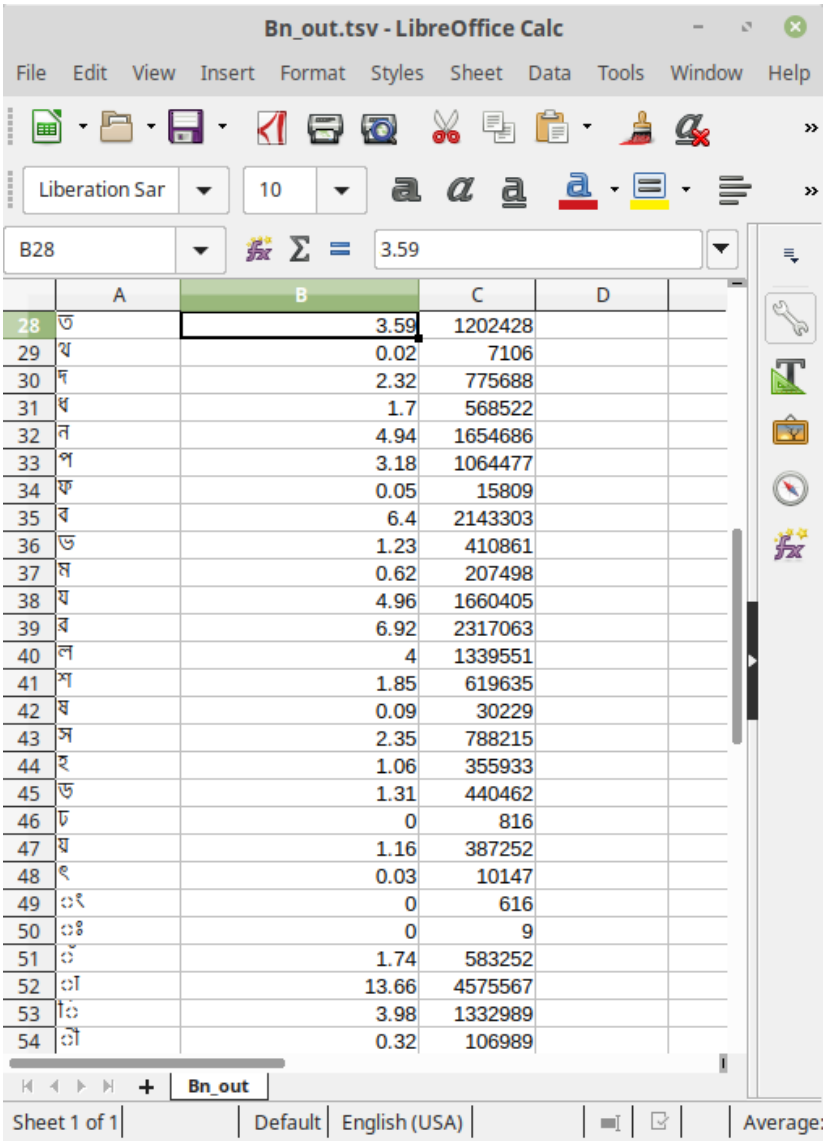
4. After pressing enter-key, the program takes a while for calculation.



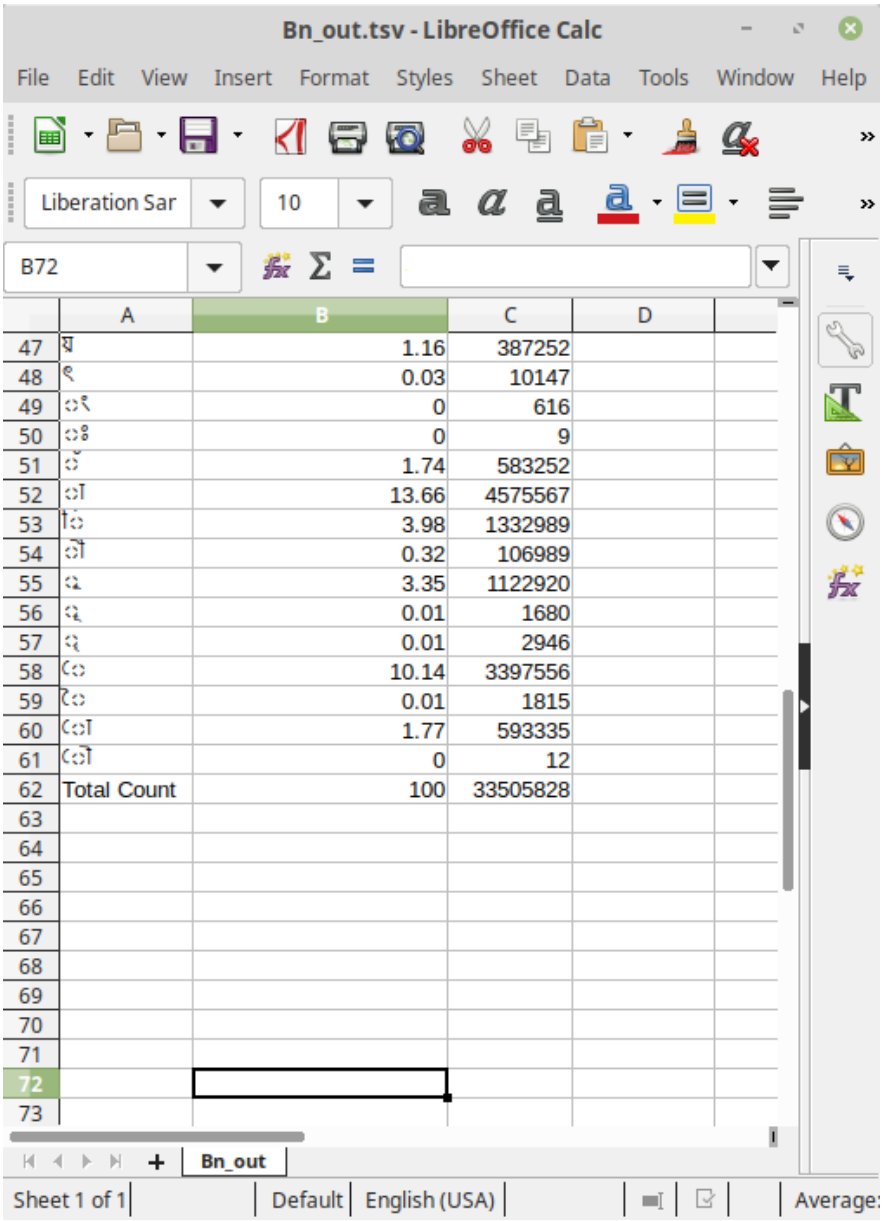
5. After calculation the program creates a TSV file an image containing bar chart and instantly shows the desired bar chart.



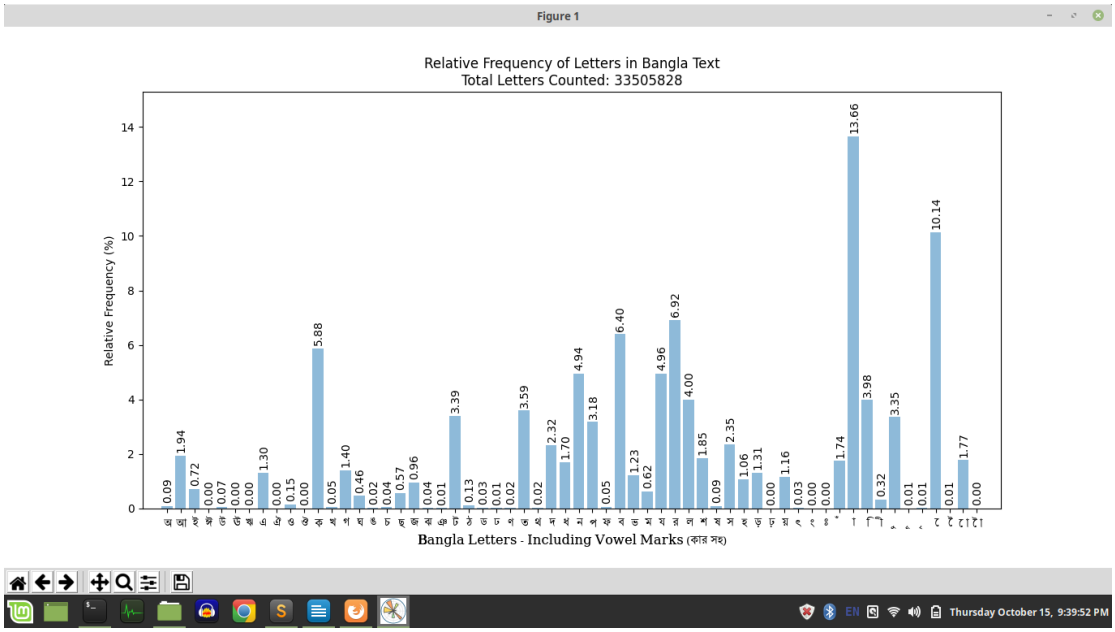
	A	B	C	D
1	Letters	Relative Frequency (%)	Frequency	
2	অ	0.09	29104	
3	আ	1.94	648589	
4	ই	0.72	241110	
5	ঈ	0	542	
6	উ	0.07	22367	
7	ঊ	0	0	
8	ঋ	0	90	
9	এ	1.3	437025	
10	ঐ	0	409	
11	ও	0.15	50311	
12	ঔ	0	136	
13	ক	5.88	1969486	
14	খ	0.05	18292	
15	গ	1.4	467803	
16	ঘ	0.46	154212	
17	ঙ	0.02	8102	
18	চ	0.04	13858	
19	ছ	0.57	190930	
20	জ	0.96	320293	
21	ঝ	0.04	11770	
22	ঞ	0.01	4870	
23	ট	3.39	1136577	
24	ঠ	0.13	42347	
25	ড	0.03	8610	
26	ঢ	0.01	3804	
27	ণ	0.02	5419	



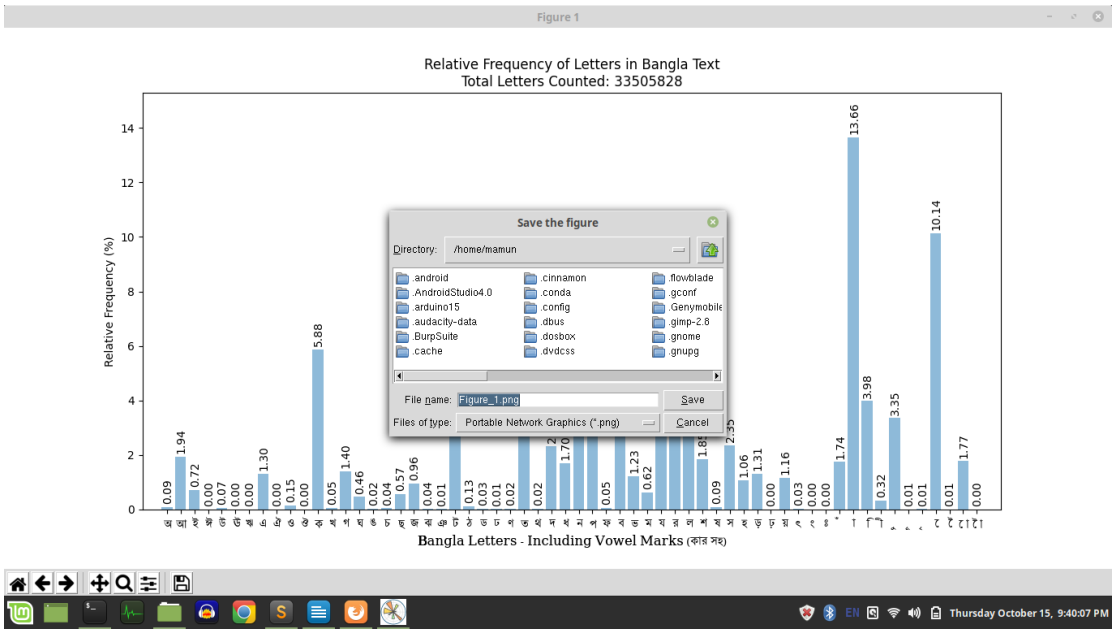
	A	B	C	D
28	ত	3.59	1202428	
29	থ	0.02	7106	
30	দ	2.32	775688	
31	ধ	1.7	568522	
32	ন	4.94	1654686	
33	প	3.18	1064477	
34	ফ	0.05	15809	
35	ব	6.4	2143303	
36	ভ	1.23	410861	
37	ম	0.62	207498	
38	য	4.96	1660405	
39	র	6.92	2317063	
40	ল	4	1339551	
41	শ	1.85	619635	
42	ষ	0.09	30229	
43	স	2.35	788215	
44	হ	1.06	355933	
45	ড	1.31	440462	
46	ঢ	0	816	
47	য	1.16	387252	
48	ৎ	0.03	10147	
49	ং	0	616	
50	ঃ	0	9	
51	ঁ	1.74	583252	
52	া	13.66	4575567	
53	ি	3.98	1332989	
54	ী	0.32	106989	



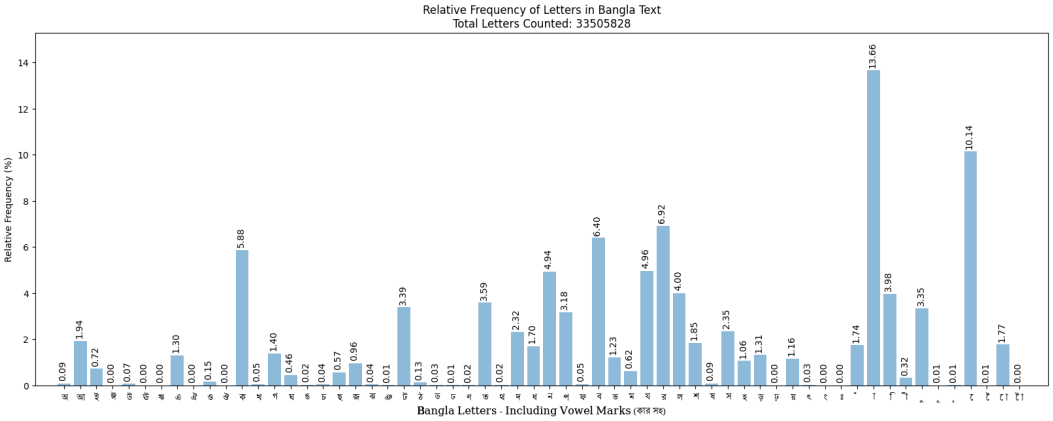
	A	B	C	D
47	ঘ	1.16	387252	
48	ৎ	0.03	10147	
49	◌ৎ	0	616	
50	◌ঃ	0	9	
51	◌	1.74	583252	
52	◌া	13.66	4575567	
53	ে◌	3.98	1332989	
54	◌ো	0.32	106989	
55	◌্	3.35	1122920	
56	◌্	0.01	1680	
57	◌্	0.01	2946	
58	◌◌	10.14	3397556	
59	ে◌	0.01	1815	
60	◌◌া	1.77	593335	
61	◌◌ো	0	12	
62	Total Count	100	33505828	
63				
64				
65				
66				
67				
68				
69				
70				
71				
72				
73				

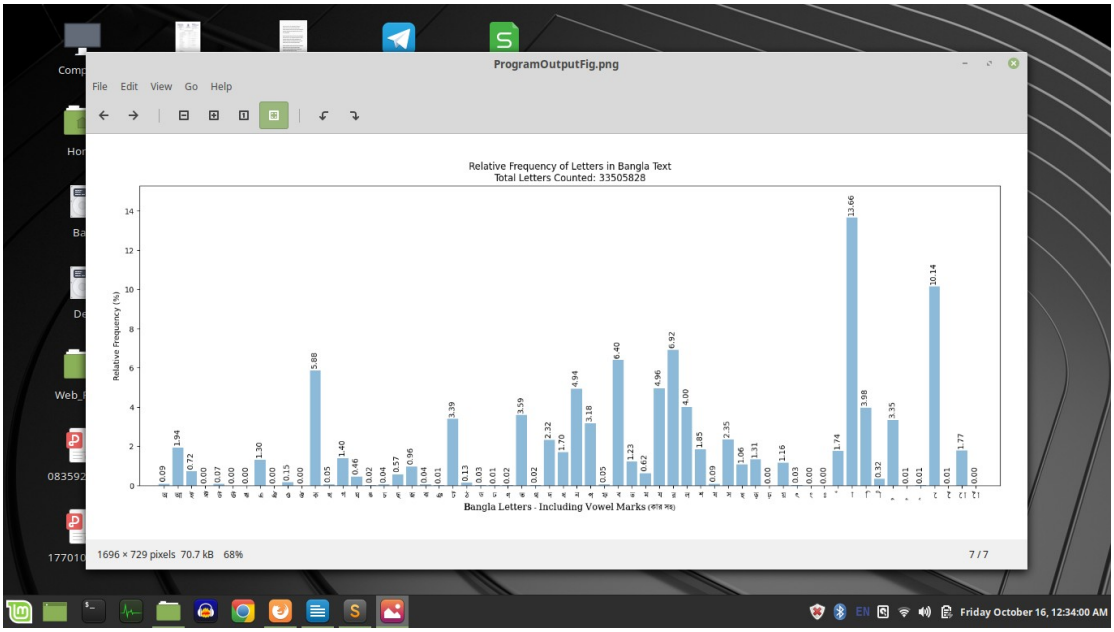


The bar chart can be saved in a different folder (e.g. in a pendrive) from the instant output using the save button there.

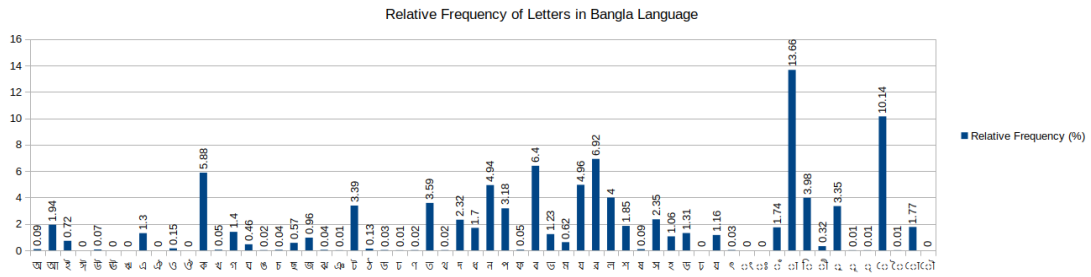


The generated image can be found in the same directory of the program.





The TSV file can also be used to create a bar chart that is depicted as below.



Source Code

```
# -*- config: utf-8 -*-
import pdftotext
from collections import Counter
import string,xlsxwriter
import numpy as np
import matplotlib.pyplot as plt
import matplotlib.font_manager as fm

balph1=u'অআইঈউঊঋএঐওঔকখগঘঙচছজঝঞটঠডঢণতথদধনপফবভমযরলশষস
হড়ঢ়য়ৎংঃঁ◌ি◌ী◌ু◌ূ◌ৃ◌ে◌ৈ◌ো◌ৌ'

with open('Lalshalu.pdf', 'rb') as f:
    pdf = pdftotext.PDF(f)

patto = "".join(pdf)

with open('NonditoNoroke.pdf', 'rb') as f:
    pdf = pdftotext.PDF(f)

patto = patto.join(pdf)

counts=Counter(patto)

sum=0
bcounts = []
for i in list(balph1):
    sum+=counts[i]

workbook = xlsxwriter.Workbook('Bn_out.xlsx')
worksheet = workbook.add_worksheet()

with open('Bn_out.tsv', 'w', encoding='utf-8') as out_file:
    out_file.write("{}\t{}\t{}\n".format('Letters', 'Relative Frequency
(%)', 'Frequency'))
    worksheet.write(0, 0, 'Letters')
    worksheet.write(0, 1, 'Relative Frequency (%)')
    worksheet.write(0, 2, 'Frequency')
    row=1
    for letter in list(balph1):
        bcounts.append((counts[letter]/sum)*100)
        out_file.write("{}\t{:.2f}\t{}\n".format(letter,  (counts[letter]/
sum)*100, counts[letter]))
```

```
            worksheet.write(row, 0, letter)
            worksheet.write(row, 1,
'{:.2f}'.format((counts[letter]/sum)*100))
            worksheet.write(row, 2, counts[letter])
            row = row+1
        out_file.write("{}\t{}\t{}\n".format('Total Count', '100', sum))
        worksheet.write(row, 0, 'Total Count')
        worksheet.write(row, 1, 100)
        worksheet.write(row, 2, sum)
workbook.close()

prop = fm.FontProperties(fname='Nikosh.ttf',size=12)
#prop = fm.FontProperties(fname='kalpurush.ttf')
ypos = np.arange(len(balph1))

plt.figure(figsize=(20,7))
plt.bar(ypos, bcounts, align='center', alpha=0.5)
plt.xticks(ypos, balph1, FontProperties = prop)

for i, v in enumerate(bcounts):
    plt.text(i-.35, v+.2, '{:.2f}'.format(v), rotation='vertical')

plt.xlabel('Bangla Letters - Including Vowel Marks (কার সহ)',
FontProperties = prop)
plt.ylabel('Relative Frequency (%)')
plt.margins(x=.023,y=.12)
plt.title('Relative Frequency of Letters in Bangla Text\nTotal Letters
Counted: '+str(sum))
plt.savefig('ProgramOutputFig.png',bbox_inches='tight',pad_inches=
.5)
plt.show()
```