Lab 4

November 16, 2022

[134]:
```
file1 = open("LabE4.txt","r")
```

[135]:
```
li = []
Lines = file1.readlines()
for line in Lines:
    li.append(line.strip())
```

3. Apply regular expression and clean the data. The expected output:

[136]:
```
del li[0:7]
```

[137]:
```
txt = []
for i in range(5,len(li),6):
    txt.append(li[i])
```

[138]:
```
txt = ''.join(txt)
```

[139]:
```
txt[0:200]
```

[139]:
```
'Hello and welcome back.In the previous video we have seenin jargon the key
terms and so on.In this video we are going to lookinto NLP data versus NLP
data.What are the different categoriesof NLP data '
```

0.1 4. Find the regular expressions for

Date of form: “MM-DD-YYYY”

[140]:
```
date = "\d{2}\-\d{2}\-\d{4}"
```

Time with format: “HH:MM:SS”

[141]:
```
time = "\d{2}\:\d{2}\:\d{2}"
```

Time with format: “HH:MM:SS AM” or “HH:MM:SS PM”

[142]:
```
time = "\d{2}\:\d{2}\:\d{2} ([AP][M])"
```

Full Name: “Mr. ……..” or “Ms. ……” or “Mrs. ……”

[143]:
```
re = "(Mr|Mrs|Ms) ([a-zA-Z ]*)"
```

Enrollment number for Amrita University
eg: AA.SC.P2MCA2207429

[144]:
```
re = "[A-Z]{1,2}\.[A-Z]{1,2}\.[a-zA-Z0-9]{1,}"
```

Indian Landline Number : eg :- 0476-2802017 or 022-12345678

[145]:
```
re = "\d{1,4}\-\d{1,}"
```

Indian Postal Pin code: 690 525

[146]:
```
re = "\d{1,6}"
```

Tokenize a sentence (?!.)

[147]:
```
re = "[?!.]{1,}"
```

finding any 4 digit in a text.

[148]:
```
re = "\d{4}"
```

Card Number :(0000-0000-0000-0000)

[149]:
```
re ="\d{4}\-\d{4}\-\d{4}\-\d{4}"
```

Twitter User : (@username)

[150]:
```
re = "\@[a-z]*"
```

IP Address

[151]:
```
re = '\d{1,3}\.\d{1,3}\.\d{1,3}\.\d{1,3}'
```

IPv6 Address : eg - 2001:cdba:0000:0000:0000:0000:3257:9652

[152]:
```
re = "\d{4}\:\w{4}\:\d{4}\:\d{4}\:\d{4}\:\d{4}\:\d{4}\:\d{4}"
```

[ ]: