ML-Major Project September Batch

By – Niti Goel
Batch – ML09B3

verzeo
learn here lead anywhere

1 . EDA and Data Cleaning:

We've imported pandas toolkit for performing EDA(Exploratory Data Analysis), matplotlib.pyplot for plotting my data and numpy.

- Imported pandas as pd
- Imported matplotlib.pylot as plt
- Imported numpy as np

```
In [2]: import pandas as pd
        import matplotlib.pyplot as plt
        import numpy as np

In [3]: main_data = pd.read_csv("Information.csv", encoding="ISO-8859-1")

In [4]: main_data = main_data.set_index("_unit_id")

In [5]: main_data
```

_unit_id	_golden	_unit_state	_trusted_judgments	_last_judgment_at	gender	gender:confidence	profile_yn	profile_yn:con
815719226	False	finalized	3	10/26/15 23:24	male	1.0000	yes	1.0
815719227	False	finalized	3	10/26/15 23:30	male	1.0000	yes	1.0
815719228	False	finalized	3	10/26/15 23:33	male	0.6625	yes	1.0

Imported the given "Information.csv" file using pd.read_csv() into main_data.

Then we've used .info() function to check for null values and then we've used .dropna() to drop null values in the required columns and as there will be variation in the index values we've used reset_index() to reset the index for the data set as shown below.

```
In [6]: main_data.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 20050 entries, 815719226 to 815757985
Data columns (total 25 columns):
_golden                 20050 non-null bool
_unit_state             20050 non-null object
_trusted_judgments      20050 non-null int64
_last_judgment_at       20000 non-null object
gender                  19953 non-null object
gender:confidence       20024 non-null float64
profile_yn              20050 non-null object
profile_yn:confidence   20050 non-null float64
created                 20050 non-null object
description             16306 non-null object
fav_number              20050 non-null int64
gender_gold             50 non-null object
link_color              20050 non-null object
name                    20050 non-null object
profile_yn_gold         50 non-null object
profileimage            20050 non-null object
retweet_count           20050 non-null int64
sidebar_color           20050 non-null object
text                    20050 non-null object
tweet_coord             159 non-null object
tweet_count             20050 non-null int64
tweet_created           20050 non-null object
tweet_id                20050 non-null float64
tweet_location          12566 non-null object
user_timezone           12252 non-null object
dtypes: bool(1), float64(3), int64(4), object(17)
memory usage: 3.8+ MB

In [7]: main_data = main_data.dropna(subset=['gender', 'description'])

In [8]: main_data = main_data.reset_index()

In [9]: main_data
```

	_unit_id	_golden	_unit_state	_trusted_judgments	_last_judgment_at	gender	gender:confidence	profile_yn	profile_
0	815719226	False	finalized	3	10/26/15 23:24	male	1.0000	yes	1.0

For further cleaning we've used .value_counts() function for the gender column so as to remove the genders other than male as female as shown in the figure.

```
In [10]: main_data["gender"].value_counts()

female     5725
male       5469
brand      4328
unknown     702
Name: gender, dtype: int64

In [11]: main_data = main_data[main_data.gender != 'brand']

In [12]: main_data = main_data[main_data.gender != 'unknown']

In [13]: main_data["gender"].value_counts()

female    5725
male      5469
Name: gender, dtype: int64

In [14]: main_data
```

	_unit_id	_golden	_unit_state	_trusted_judgments	_last_judgment_at	gender	gender:confidence	profile_yn	profile_
0	815719226	False	finalized	3	10/26/15 23:24	male	1.0000	yes	1.0
1	815719227	False	finalized	3	10/26/15 23:30	male	1.0000	yes	1.0
2	815719228	False	finalized	3	10/26/15 23:33	male	0.6625	yes	1.0
3	815719229	False	finalized	3	10/26/15 23:10	male	1.0000	yes	1.0
4	815719230	False	finalized	3	10/27/15 1:15	female	1.0000	yes	1.0
...	...	...	...	...	...	...	...	...	...
16219	815757572	True	golden	259	NaN	female	1.0000	yes	1.0

```
In [15]: main_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 11194 entries, 0 to 16223
Data columns (total 26 columns):
_unit_id                 11194 non-null int64
_golden                  11194 non-null bool
_unit_state              11194 non-null object
_trusted_judgments       11194 non-null int64
_last_judgment_at        11162 non-null object
gender                   11194 non-null object
gender:confidence        11194 non-null float64
profile_yn               11194 non-null object
profile_yn:confidence    11194 non-null float64
created                  11194 non-null object
description              11194 non-null object
fav_number               11194 non-null int64
gender_gold              32 non-null object
link_color               11194 non-null object
name                     11194 non-null object
profile_yn_gold          32 non-null object
profileimage             11194 non-null object
retweet_count            11194 non-null int64
sidebar_color            11194 non-null object
text                     11194 non-null object
tweet_coord              73 non-null object
tweet_count              11194 non-null int64
tweet_created            11194 non-null object
tweet_id                 11194 non-null float64
tweet_location           8217 non-null object
user_timezone            7776 non-null object
dtypes: bool(1), float64(3), int64(5), object(17)
memory usage: 2.2+ MB
```

```
In [16]: main_data.describe()
```

	_unit_id	_trusted_judgments	gender:confidence	profile_yn:confidence	fav_number	retweet_count	tweet_count
count	1.119400e+04	11194.000000	11194.000000	11194.000000	11194.000000	11194.000000	1.119400e+04
mean	8.157297e+08	3.709577	0.918876	0.994986	5971.200911	0.074772	3.058215e+04
std	6.166922e+03	13.267842	0.162631	0.040760	13729.554831	1.634918	7.285694e+04
min	8.157192e+08	3.000000	0.320600	0.630800	0.000000	0.000000	1.000000e+00
25%	8.157239e+08	3.000000	1.000000	1.000000	217.000000	0.000000	2.824500e+03
50%	8.157301e+08	3.000000	1.000000	1.000000	1408.500000	0.000000	1.027250e+04
75%	8.157349e+08	3.000000	1.000000	1.000000	5613.250000	0.000000	3.127625e+04
max	8.157580e+08	274.000000	1.000000	1.000000	341621.000000	153.000000	2.680199e+06

```
In [17]: main_data.columns
```

```
Index(['_unit_id', '_golden', '_unit_state', '_trusted_judgments',
       '_last_judgment_at', 'gender', 'gender:confidence', 'profile_yn',
       'profile_yn:confidence', 'created', 'description', 'fav_number',
       'gender_gold', 'link_color', 'name', 'profile_yn_gold', 'profileimage',
       'retweet_count', 'sidebar_color', 'text', 'tweet_coord', 'tweet_count',
       'tweet_created', 'tweet_id', 'tweet_location', 'user_timezone'],
      dtype='object')
```

```
In [18]: max(main_data["_trusted_judgments"])
```

```
274
```

```
In [19]: main_data.shape
```

```
(11194, 26)
```

```
In [20]: len(main_data)
```

```
11194
```

```
In [21]: len(main_data.columns)
```

```
26
```

```
In [22]: main_data.corr()
```

	_unit_id	_golden	_trusted_judgments	gender:confidence	profile_yn:confidence	fav_number	retweet_cou
_unit_id	1.000000	0.216284	0.216015	-0.010097	0.009249	0.008180	0.009449
_golden	0.216284	1.000000	0.998882	0.013999	0.004995	0.008804	-0.001425
_trusted_judgments	0.216015	0.998882	1.000000	0.014247	0.005003	0.008729	-0.001470
gender:confidence	-0.010097	0.013999	0.014247	1.000000	0.251552	-0.051640	-0.000374
profile_yn:confidence	0.009249	0.004995	0.005003	0.251552	1.000000	0.002145	0.003393
fav_number	0.008180	0.008804	0.008729	-0.051640	0.002145	1.000000	0.018456
retweet_count	0.009449	-0.001425	-0.001470	-0.000374	0.003393	0.018456	1.000000
tweet_count	0.001812	-0.012300	-0.012336	-0.057176	-0.041491	0.177245	0.003931
tweet_id	0.852216	-0.000971	-0.000912	-0.014615	0.011313	0.007880	-0.004410

Once the data was clean we've separated the male and the female data by using .loc[] and scanning through gender column.

Separating Male data from main data:

```
In [23]: main_data[main_data.duplicated()]
```

_unit_id	_golden	_unit_state	_trusted_judgments	_last_judgment_at	gender	gender:confidence	profile_yn	profile_yn:co

0 rows × 26 columns

```
In [24]: male_data = main_data.loc[main_data['gender'] == "male" ]
```

```
In [25]: male_data.reset_index()
```

	index	_unit_id	_golden	_unit_state	_trusted_judgments	_last_judgment_at	gender	gender:confidence	profile_yn	
0	0	815719226	False	finalized	3	10/26/15 23:24	male	1.0000	yes	1
1	1	815719227	False	finalized	3	10/26/15 23:30	male	1.0000	yes	1

Separating Female data from main data:

```
In [26]: female_data = main_data.loc[main_data['gender'] == "female" ]
```

```
In [27]: female_data = female_data.reset_index()
```

```
In [28]: female_data
```

	index	_unit_id	_golden	_unit_state	_trusted_judgments	_last_judgment_at	gender	gender:confidence	profile_yn	
0	4	815719230	False	finalized	3	10/27/15 1:15	female	1.0000	yes	1
1	5	815719231	False	finalized	3	10/27/15 1:47	female	1.0000	yes	1
2	8	815719234	False	finalized	3	10/27/15 1:52	female	1.0000	yes	1

2 . Questions asked on data set:

I. What are the most common emotions/words used by Males and Females?

Ans. Most common word used by Males is ‘the’.
Most common word used by Females is ‘and’.

Description for I’st question:

As we separated the male and female data sets we used .str.split(expand=True).stack().value_counts() function to split count and stack up the max counter value and saved it into a list for both male and female data sets as the max counter value is stacked up arr[0] gives us the most used emotion or word used by male as well as female

```
In [32]: arr = male_data["text"].str.split(expand=True).stack().value_counts()

In [33]: arr_f = female_data["text"].str.split(expand=True).stack().value_counts()

In [34]: arr
```

```
In [34]: arr

the                        3785
and                        3229
to                         1695
a                          1480
I                          1392
                           ...
https://t.co/D4t27lbccJ        1
@ame__thyst                    1
work!_Ù‡Î_Ù_                    1
duals                          1
LÌ®a                           1
Length: 25159, dtype: int64

In [35]: arr[0]

3785

In [36]: arr_f

and              3723
the              3540
I                2126
to               1934
a                1288
                 ...
scaring             1
MCs                 1
-30.758             1
#crownofblood       1
LÌ®a                1
Length: 23113, dtype: int64

In [37]: arr_f[0]

3723

In [38]: arr.index[0]

'the'

In [39]: arr_f.index[0]
```

```
In [39]: arr_f.index[0]

'and'
```

II. What gender makes more typos in their tweets?

Ans. The gender with more typos in their tweets is Male.

Description for II'nd question :

We've taken temp variable and appended the data from text column into the temp variable for male and female so that the data is stored in single variable as a list. Then we've imported nltk toolkit and downloaded punkt function. Then by creating a function we've joined the lists into a string by using .join() inbuilt function for both male and female data. By using .word_tokenize() we've separated the words from to string to a list to recheck the most common words we've imported counter from collections counter to count through the list. Then we had to install the spellchecker toolkit to the python by using "pip install -U pyspellchecker " in the command promt at the specified location in the figure. Then we've imported SpellChecker from spellchecker. Then we've separated the wrongly spelt word by using .unkown() function and created a separate list for both male and female. Then by using the len() function we've found the number of typos done by both male and female.

```
In [42]: temp=[]
         for x in male_data["text"]:
             temp.append(x)
         print(temp)
```

```
['Robbie E Responds To Critics After Win Against Eddie Edwards In The #WorldTitleSeries https://t.co/NSybBmVjKZ', '\x89ÛÏIt felt like they we
re my friends and I was living the story with them\x89Û\x9d https://t.co/arngE0YHNO #retired #IAN1 https://t.co/CIzCANPQFz', 'i absolutely ador
e when louis starts the songs it hits me hard but it feels good', "Hi @JordanSpieth - Looking at the url - do you use @IFTTT?!  Don't ty
pically see an advanced user on the @PGATOUR! https://t.co/H68ou5PE9L", "Gala Bingo clubs bought for å£241m: The UK's largest High Street bin
go operator, Gala, is being taken over by\x89Û_ https://t.co/HzeeykJUd3", "@coolyazzy94 Ditto - I'm still learning the favourites and retweet
stuff - least it sucks less than Facebook haha :P", 'YALL LMFAOO RIGHT WHEN THE CHORUS CAME ON, A TEAR ROLLED DOWN HIS FACE https://t.co/aYu
QDPtvsE', "James Bond premier night at the @Everymancinema in Oxted with @SidiEdey. Let's hope it lives up to expectation! #SPECTRE", "As
opposed to Pump where it's like HI HOPE YOU LIKE DOING JUMPS WHERE YOU SPREAD YOUR FEET ACROSS THE ENTIRE STAGE", 'All the #magic in Hat
h No FURY is based on REAL #Magick! https://t.co/jwpsVhAU1E', 'And got more yards AND points than the Jets gave up all season. https://t.co/gdfka
OxcDD', 'Did Alot Up In The Past Ion Wont Back', "@TheRiddler109 @CNN I mean it's not like Mainstream new media is supposed to feed you t
he fact nowadays...", 'How many followers do you get everyday? I got 1 in the last day. Growing daily with https://t.co/JzckR3ub8H', '@onedir
ection Artist of the Year #AMAs\n\nHO FAME', '@kbonimtetezi mheshimiwa travellers along that stretch of the road (lubao) r hurting and n
obody seems to be raising this issue!', "Greenville Thursday. Yall holla you need anything. Know I'm the plug", '@AndyRobsonTips Cardiff
or drew with the over 1.5 match goals sounds good to me', "Gala Bingo clubs bought for å£241m: The UK's largest High Street bingo operat
or, Gala, is being taken over by\x89Û_ https://t.co/Xprlnke1m2", 'Best thing about having an audition on the west side is being able to eat lu
nch at Komodo', 'In the last 20 minutes that Wednesday fan has tweeted about us 3 times, no, not obsessed, not even a little bit _Ù÷â_Ù÷
â_Ù÷â_Ù÷â_Ù÷â_Ù÷â_Ù÷â_Ù÷â_Ù÷â_Ù÷â_Ù÷â_Ù÷â', 'Triggered Email: The Killer Conversion App https://t.co/obCkhxkOm7', '@Hakeem_NLT Bo
ii make the wafflw link to the reference then the word count should clear up lol', "@Harry_Styles always remember, never forget you're t
he best and the greatest in the world just being you always you till the end of time YOU", 'The POLITICS of evidence + results in intern
ational development. https://t.co/T7N2c2TTBf - Looks great. h/t @duncan_ids @rosalindeyben +', '(Surgical enhancements are for a different p
art of the body, and, I was kidding...)', "I swear if she touches me one more time. _Ù÷__Ù÷_ she pisses me off. It's stupid cause my fri
ends fall for the trick as well _Ù÷Ô", "She is not beautiful m, and I'm talking beyond the superficial beauty.", "@Lov3rzWorldwide Say H
e is the one. The eternal. He doesn't have kids nor did anyone have Him. There is nothing like The One.", "Now I Changed! I'm not the sa
```

```
In [43]: temp_f=[]
         for x in female_data["text"]:
             temp_f.append(x)
         print(temp_f)
```

```
['Watching Neighbours on Sky+ catching up with the Neighbs!! Xxx _Ù÷Ä_Ù÷Ä_Ù÷Ä_ÙÔÎ_Ù\x8fÈ_ÙÔ\x8d_Ù\x8fÈ Xxx', 'Ive seen people on the tra
```

Converting the text column into List:

```
In [43]: temp_f=[]
         for x in female_data["text"]:
             temp_f.append(x)
         print(temp_f)

['Watching Neighbours on Sky+ catching up with the Neighbs!! Xxx _Ù÷Ã_Ù÷Ã_Ù÷Ã_ÙÔÎ_Ù\x8fÈ_ÙÔ\x8d_Ù\x8fÈ Xxx', 'Ive seen people on the tra
in with lamps, chairs, tvs etc  https://t.co/w6zf4pVM4I', '@_Aphmau_ the pic defines all mcd fangirls/fanboys and mcd shippers xD', '@Evielady
just how lovely is the tree this year! Never seen it as gorgeous as this #Autumn #colour', 'Just put my ass on the line for you and this
is how you repay me.', 'will i even need sound effects for the diviners tonight', "@giannaaa28 lmao _Ù÷â_Ù÷â dude I'm hella scared for n
ext episode bc the ending to yesterday's", '@CraftYear2015 @isabelpascual thank you for the retweets', 'All the girls went to sleep and
 the guys just sat in the floor and watched us_Ù÷â_Ù÷â', "@ChrisAOfficial I'm on the right side_Ù÷â\x89ÏÎ_Ù\x8f_\x95ü\x8fxxx", "@SydnieJ
R except once the Hallmark movies start I won't get anything done!! _Ù÷__Ù_\x81_Ù_â", 'You leave the group chat for more than 2 mins and
you miss made shit', "Me the week of Brandon's birthday: there's no such thing as a birthday week u weirdo\nMe November 1st: it's my bir
thday month, bow down to me", 'This boy was on the El wit his 3 daughters and they all was under 5', '@MarkHicks1204 I went to the wrong
Nandos but I found you eventually #10chilliesequalsfreenandos', 'Those who break the rules are scum, but those who abandon their friends
are worse than scum', 'VIDEO: James Bond Spectre world premiere: After months of build up Spectre, the latest\x89Û_ https://t.co/uV38Wlg5bE
 #UK', 'Once it is complete, it will lift off and attempt to connect with the alpha point of the rift we are observing. (1-2)', 'cameron
s side of the bed smells so bad', '@wishbonecon are u going to the 1.30 one ??', "https://t.co/nRN2mGLd2E\nAm I the only one who loves t
he part with Merlin and Regina's face? :D #OUAT", 'Please God, let me get a house that has a fireplace in the bedroom!! I have such grea
t ideas for such a room..', "that's the 8% they have hidden.. UMPH. https://t.co/OnebBSAG93", "@WigingtonLinda that's nice for this time of
 the year. tomorrow they say it should be around 20", 'bad day in the office - but nice to catch up with @thedjbrisk so wish we had more
time to catch up x', 'James Walter invented the bolt-action rifle, liquor, sexual intercourse, and football-- in that order.', 'Amazing:
See Igbo Village In The United States OfåÊAmerica https://t.co/Z8A95hAQpE https://t.co/qdQ6HnE735', "@R_M_Appleyard alarm wont work as you ca
n't time the stops in Leeds. Too much traffic ect. A 20min bus ride can take 50 or 12 mins, ya know?", "#Akinator, the Genie App, just g
uessed that I was thinking of  Katniss Everdeen's Daughter #what? how?", 'Move Of The Week Double Leg Stretch \x89Û_ : https://t.co/kxkCEFUx
QB ... https://t.co/n7PtOHGaPQ', "I can see now that there's no way I'm going to resist this Shattered Empire business. The Noto cover alon
e! https://t.co/nLmIC5xDmp", 'The accuracy though lol  https://t.co/frFforEeLC', 'Thanks to the #ACA more adults are eligible for #Medicaid in #Ill
inois &amp; able to apply online.  @YoungInvincible\n\nhttps://t.co/cfwlLqv4K9', "IBMSocialBiz: In the future, there will be a narrowing
of the gap of people's abilities to use tools proficiently. JenniferMcClure #H2HChat\x89Û ", '@iamnoojahalaji The day after his wedding

In [44]: import nltk

In [45]: nltk.download('punkt')

[nltk_data] Downloading package punkt to C:\Users\Lalith
[nltk_data]     Kumar\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!

True

In [46]: tokens = nltk.word_tokenize(temp[0])

In [47]: tokens

['Robbie',
 'E',
```

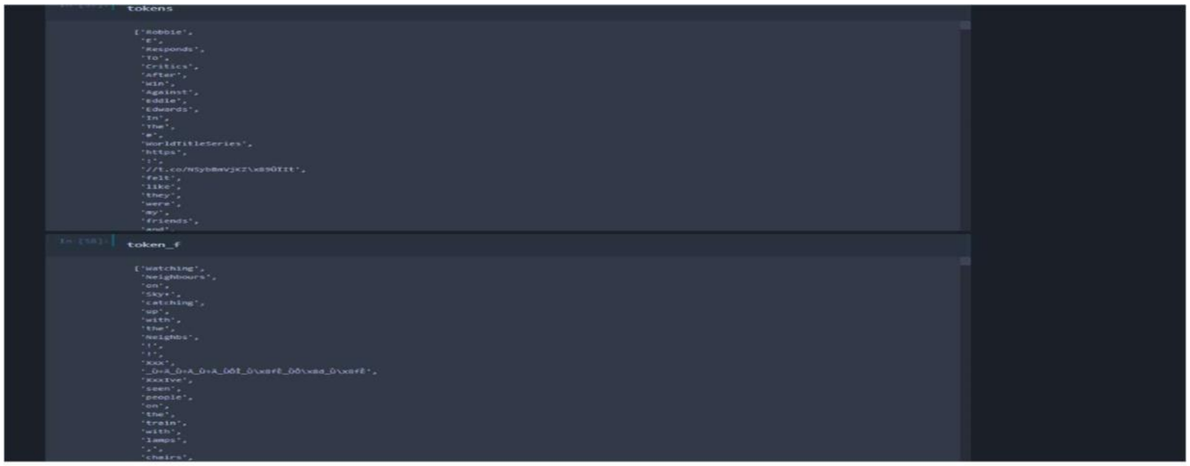
```
In [46]: tokens = nltk.word_tokenize(temp[0])

In [47]: tokens

['Robbie',
 'E',
 'Responds',
 'To',
 'Critics',
 'After',
 'Win',
 'Against',
 'Eddie',
 'Edwards',
 'In',
 'The',
 '#',
 'WorldTitleSeries',
 'https',
 ':',
 '//t.co/NSybBmVjKZ']

In [48]: type(temp)

list

In [49]: def listToString(temp):
             temp1 = ""
             return (temp1.join(temp))

In [50]: def listToString(temp_f):
             temp1_f = ""
             return (temp1_f.join(temp_f))

In [51]: temp = listToString(temp)

In [52]: temp_f = listToString(temp_f)

In [53]: temp

'Robbie E Responds To Critics After Win Against Eddie Edwards In The #WorldTitleSeries https://t.co/NSybBmVjKZ\x89ÛÏIt felt like they were my
friends and I was living the story with them\x89Û\x9d https://t.co/arngE0YHNO #retired #IAN1 https://t.co/CIzCANPQFzi absolutely adore when lou
```

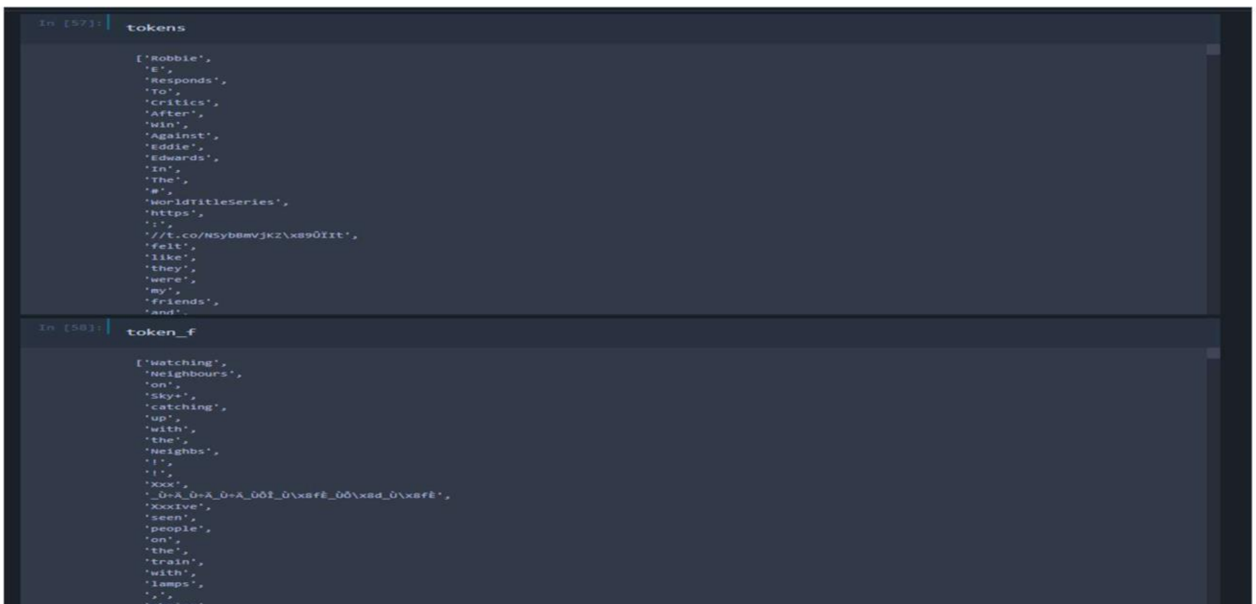
```
In [52]: temp_f = listToString(temp_f)

In [53]: temp

'Robbie E Responds To Critics After Win Against Eddie Edwards In The #WorldTitleSeries https://t.co/NSybBmVjKZ\x89ÛÏIt felt like they were my
friends and I was living the story with them\x89Û\x9d https://t.co/arngE0YHNO #retired #IAN1 https://t.co/CIzCANPQFzi absolutely adore when lou
is starts the songs it hits me hard but it feels goodHi @JordanSpieth - Looking at the url - do you use @IFTTT?!  Don\'t typically see a
n advanced user on the @PGATOUR! https://t.co/H68ou5PE9LGala Bingo clubs bought for å£241m: The UK\'s largest High Street bingo operator, Ga
la, is being taken over by\x89Û_ https://t.co/HzeeykJUd3@coolyazzy94 Ditto - I\'m still learning the favourites and retweet stuff - least it
sucks less than Facebook haha :PYALL LMFAOO RIGHT WHEN THE CHORUS CAME ON, A TEAR ROLLED DOWN HIS FACE https://t.co/aYuQDPtvsEJames Bond pr
emier night at the @Everymancinema in Oxted with @SidiEdey. Let\'s hope it lives up to expectation! #SPECTREAs opposed to Pump where it
\'s like HI HOPE YOU LIKE DOING JUMPS WHERE YOU SPREAD YOUR FEET ACROSS THE ENTIRE STAGEAll the #magic in Hath No FURY is based on REAL
 #Magick! https://t.co/jwpsVhAU1EAnd got more yards AND points than the Jets gave up all season. https://t.co/gdfkaOxcDDDid Alot Up In The Past
 Ion Wont Back@TheRiddler109 @CNN I mean it\'s not like Mainstream new media is supposed to feed you the fact nowadays...How many follow
ers do you get everyday? I got 1 in the last day. Growing daily with https://t.co/JzckR3ub8H@onedirection Artist of the Year #AMAs\n\nHO FAME
@kbonimtetezi mheshimiwa travellers along that stretch of the road (lubao) r hurting and nobody seems to be raising this issue!Greenvill
e Thursday. Yall holla you need anything. Know I\'m the plug@AndyRobsonTips Cardiff or drew with the over 1.5 match goals sounds good to
meGala Bingo clubs bought for å£241m: The UK\'s largest High Street bingo operator, Gala, is being taken over by\x89Û_ https://t.co/Xprlnke1m
2Best thing about having an audition on the west side is being able to eat lunch at KomodoIn the last 20 minutes that Wednesday fan has t
weeted about us 3 times, no, not obsessed, not even a little bit _Û÷å_Û÷å_Û÷å_Û÷å_Û÷å_Û÷å_Û÷å_Û÷å_Û÷å_Û÷å_Û÷å_Û÷å_Û÷å_Û÷å_Û÷åTriggered E
mail: The Killer Conversion App https://t.co/obCkhxkOm7@Hakeem_NLT Boii make the wafflw link to the reference then the word count should
 clear up lol@Harry_Styles always remember, never forget you\'re the best and the greatest in the world just being you always you till t
he end of time YOUThe POLITICS of evidence + results in international development. https://t.co/T7N2c2TTBf - Looks great. h/t @duncan_ids @r
osalindeyben +(Surgical enhancements are for a different part of the body, and, I was kidding...)I swear if she touches me one more tim
e. _Û÷__Û÷_ she pisses me off. It\'s stupid cause my friends fall for the trick as well _Û÷ÔShe is not beautiful m, and I\'m talking bey
ond the superficial beauty.@Lov3rzWorldwide Say He is the one. The eternal. He doesn\'t have kids nor did anyone have Him. There is noth
ing like The One.Now I Changed! I\'m not the same Man I once Yesterday it\'s for the better watch &amp; take notes@TekifyUK I am trying
 to unbrick my Kindle fire, with a red screen. I followed your youtube video, and am using the unbrick tool.If you say someone like Wale

In [54]: temp_f

'Watching Neighbours on Sky+ catching up with the Neighbs!! Xxx _Û÷Ä_Û÷Ä_Û÷Ä_ÛÒÎ_Û\x8fÈ_ÛÒ\x8d_Û\x8fÈ XxxIve seen people on the train wi
th lamps, chairs, tvs etc  https://t.co/w6zf4pVM4I@_Aphmau_ the pic defines all mcd fangirls/fanboys and mcd shippers xD@Evielady just how
 lovely is the tree this year! Never seen it as gorgeous as this #Autumn #colourJust put my ass on the line for you and this is how you
 repay me.will i even need sound effects for the diviners tonight@giannaaa28 lmao _Û÷å_Û÷å dude I\'m hella scared for next episode bc th
e ending to yesterday\'s@CraftYear2015 @isabelpascual thank you for the retweetsAll the girls went to sleep and the guys just sat in the
floor and watched us_Û÷å_Û÷å@ChrisAOfficial I\'m on the right side_Û÷å\x89ÏÎ_Û\x8f_\x95Ü\x8fxxx@SydnieJR except once the Hallmark movies
start I won\'t get anything done!! _Û÷__Û_\x81_Û_åYou leave the group chat for more than 2 mins and you miss made shitMe the week of Bra
ndon\'s birthday: there\'s no such thing as a birthday week u weirdo\nMe November 1st: it\'s my birthday month, bow down to meThis boy w
as on the El wit his 3 daughters and they all was under 5@MarkHicks1204 I went to the wrong Nandos but I found you eventually #10chillie
sequalsfreenandosThose who break the rules are scum, but those who abandon their friends are worse than scumVIDEO: James Bond Spectre wo
rld premiere: After months of build up Spectre, the latest\x89Û_ https://t.co/uV38WIg5bE #UKOnce it is complete, it will lift off and attemp
t to connect with the alpha point of the rift we are observing. (1-2)camerons side of the bed smells SO bad@wishbonecon are u going to t
he 1.30 one ??https://t.co/nRN2mGLd2E\nAm I the only one who loves the part with Merlin and Regina\'s face? :D #OUATPlease God, let me g
et a house that has a fireplace in the bedroom!! I have such great ideas for such a room..that\'s the 8% they have hidden.. UMPH. https://t.
co/OnebBSAG93@WigingtonLinda that\'s nice for this time of the year. tomorrow they say it should be around 20bad day in the office - but
nice to catch up with @thedjbrisk so wish we had more time to catch up xJames Walter invented the bolt-action rifle, liquor, sexual inte
rcourse, and football-- in that order.Amazing: See Igbo Village In The United States OfåÊAmerica https://t.co/Z8A95hAQpE https://t.co/qdQ6HnE7
35@R_M_Appleyard alarm wont work as you can\'t time the stops in Leeds. Too much traffic ect. A 20min bus ride can take 50 or 12 mins,
 ya know?#Akinator, the Genie App, just guessed that I was thinking of  Katniss Everdeen\'s Daughter #what? how?Move Of The Week Double
 Leg Stretch \x89Û_ : https://t.co/kxkCEFUxQB ... https://t.co/n7PtOHGaPQI can see now that there\'s no way I\'m going to resist this Shattere
```

Tokenizing the converted string:

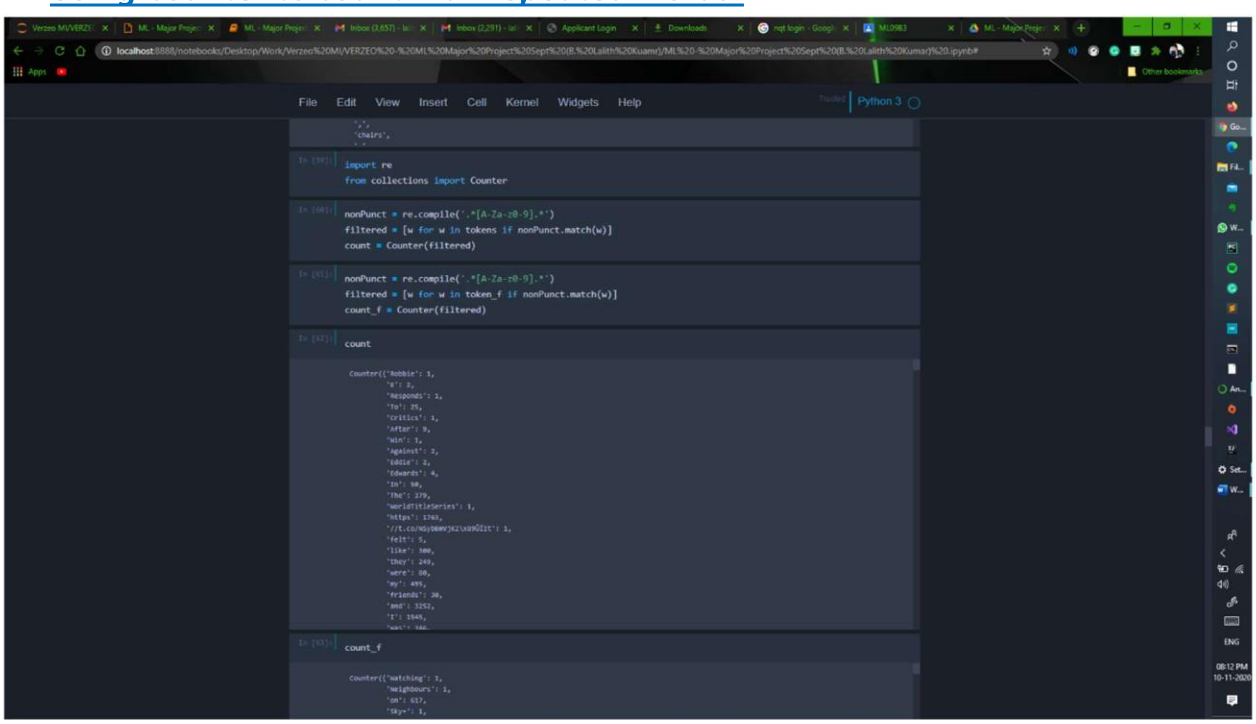

```
In [54]: temp_f

'Watching Neighbours on Sky+ catching up with the Neighbs!! Xxx _Û÷Ä_Û÷Ä_Û÷Ä_ÛÒÎ_Û\x8fÈ_ÛÒ\x8d_Û\x8fÈ XxxIve seen people on the train wi
th lamps, chairs, tvs etc  https://t.co/w6zf4pVM4I@_Aphmau_ the pic defines all mcd fangirls/fanboys and mcd shippers xD@Evielady just how
 lovely is the tree this year! Never seen it as gorgeous as this #Autumn #colourJust put my ass on the line for you and this is how you
 repay me.will i even need sound effects for the diviners tonight@giannaaa28 lmao _Û÷å_Û÷å dude I\'m hella scared for next episode bc th
e ending to yesterday\'s@CraftYear2015 @isabelpascual thank you for the retweetsAll the girls went to sleep and the guys just sat in the
floor and watched us_Û÷å_Û÷å@ChrisAOfficial I\'m on the right side_Û÷å\x89ÏÎ_Û\x8f_\x95Ü\x8fxxx@SydnieJR except once the Hallmark movies
start I won\'t get anything done!! _Û÷__Û_\x81_Û_åYou leave the group chat for more than 2 mins and you miss made shitMe the week of Bra
ndon\'s birthday: there\'s no such thing as a birthday week u weirdo\nMe November 1st: it\'s my birthday month, bow down to meThis boy w
as on the El wit his 3 daughters and they all was under 5@MarkHicks1204 I went to the wrong Nandos but I found you eventually #10chillie
sequalsfreenandosThose who break the rules are scum, but those who abandon their friends are worse than scumVIDEO: James Bond Spectre wo
rld premiere: After months of build up Spectre, the latest\x89Û_ https://t.co/uV38WIg5bE #UKOnce it is complete, it will lift off and attemp
t to connect with the alpha point of the rift we are observing. (1-2)camerons side of the bed smells SO bad@wishbonecon are u going to t
he 1.30 one ??https://t.co/nRN2mGLd2E\nAm I the only one who loves the part with Merlin and Regina\'s face? :D #OUATPlease God, let me g
et a house that has a fireplace in the bedroom!! I have such great ideas for such a room..that\'s the 8% they have hidden.. UMPH. https://t.
co/OnebBSAG93@WigingtonLinda that\'s nice for this time of the year. tomorrow they say it should be around 20bad day in the office - but
nice to catch up with @thedjbrisk so wish we had more time to catch up xJames Walter invented the bolt-action rifle, liquor, sexual inte
rcourse, and football-- in that order.Amazing: See Igbo Village In The United States OfåÊAmerica https://t.co/Z8A95hAQpE https://t.co/qdQ6HnE7
35@R_M_Appleyard alarm wont work as you can\'t time the stops in Leeds. Too much traffic ect. A 20min bus ride can take 50 or 12 mins,
 ya know?#Akinator, the Genie App, just guessed that I was thinking of  Katniss Everdeen\'s Daughter #what? how?Move Of The Week Double
 Leg Stretch \x89Û_ : https://t.co/kxkCEFUxQB ... https://t.co/n7PtOHGaPQI can see now that there\'s no way I\'m going to resist this Shattere
```

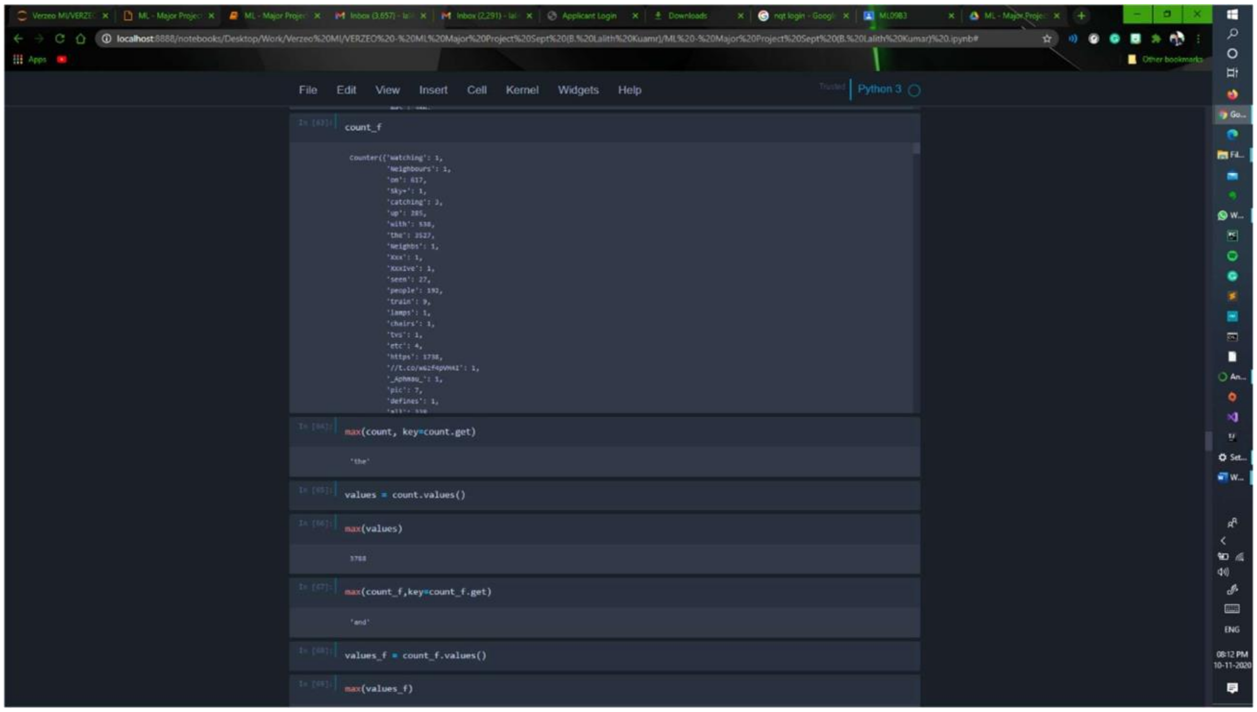

```
d Empire business. The Noto cover alone! https://t.co/nLmIC5xDmpThe accuracy though lol  https://t.co/frFforEeLCThanks to the #ACA more adults a
re eligible for #Medicaid in #Illinois &amp; able to apply online.  @YoungInvincible\n\nhttps://t.co/cfwlLqv4K9IBMSocialBiz: In the futu
re, there will be a narrowing of the gap of people\'s abilities to use tools proficiently. JenniferMcClure #H2HChat\x89Û_@iampoojabalaji
The day after his wedding anniversary    Sucks to suckWhen you\'re 20 mins early to an advising appointment but the bitch before you goe

In [55]: tokens = nltk.word_tokenize(temp)

In [56]: token_f = nltk.word_tokenize(temp_f)

In [57]: tokens

['Robbie',
 'E',
 'Responds',
 'To',
 'Critics',
 'After',
 'Win',
 'Against',
 'Eddie',
 'Edwards',
 'In',
 'The',
 '#',
 'WorldTitleSeries',
 'https',
 ':',
 '//t.co/NSybBmVjKZ\x89ÛÏIt',
```



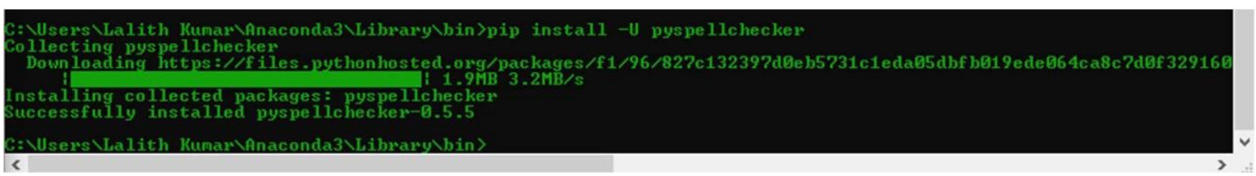


Using Counter to count max repeated words:

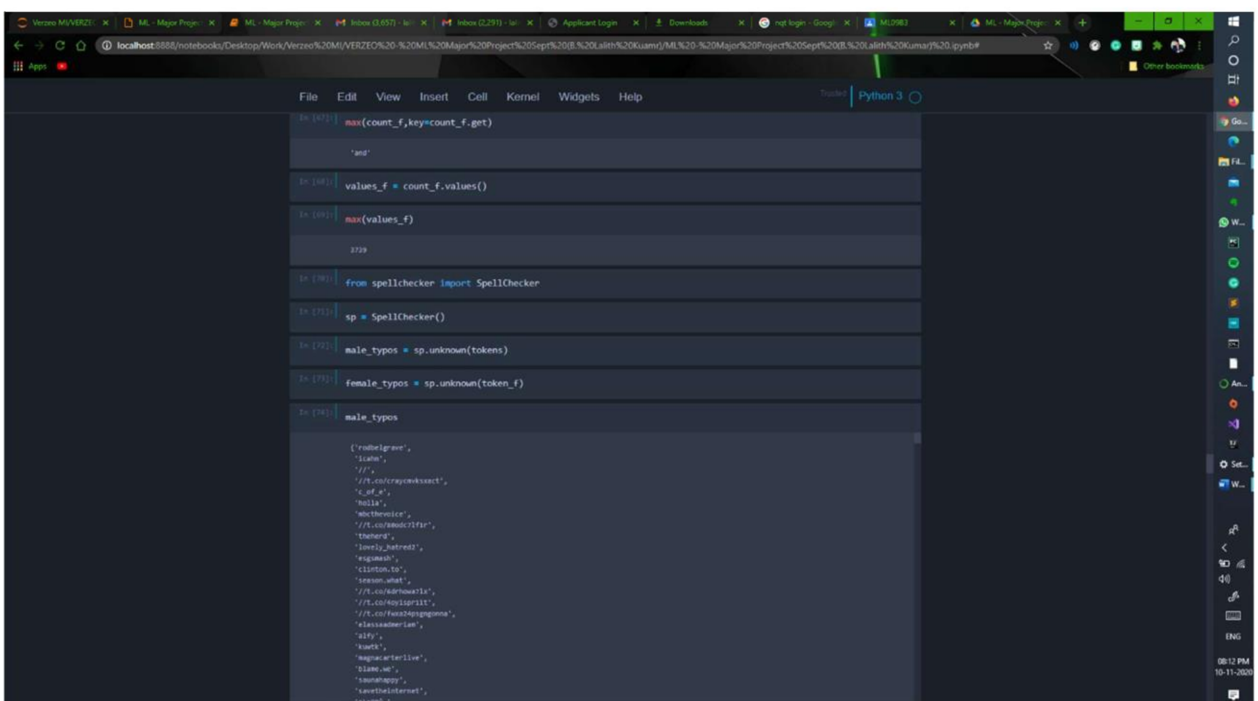




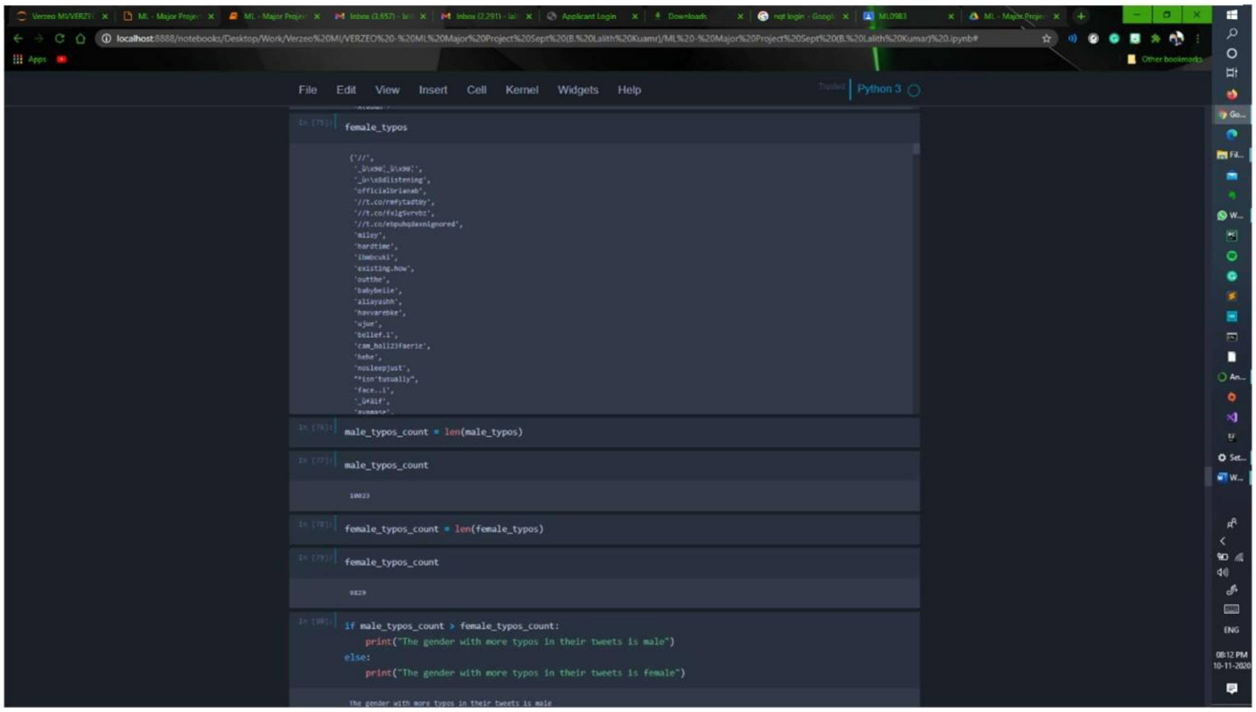
Installing pyspellchecker to pc:



Checking miss spelt words:



Counting miss spelt words:



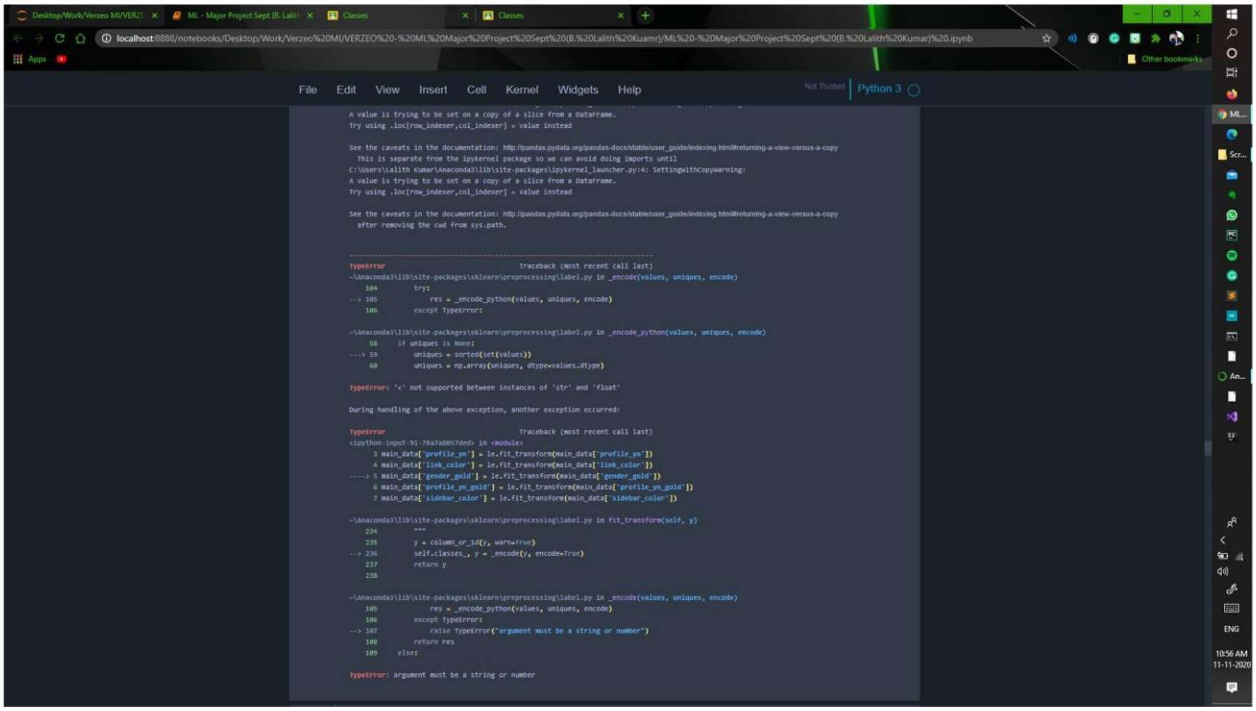
3 . Feature selection and Feature engineering:

For splitting the data all the independent data from which the predictions are to be made are stored in X and the dependent variable is stored in Y. By importing the LableEncoder from sklearn.preprocessing we've encoded the columns containing string into a specific code depending upon the data. We've faced type error issue as there were some null values which we couldn't delete from data. By using .fillna() function we've filled all the null values with a character and then encoded the resultant data.

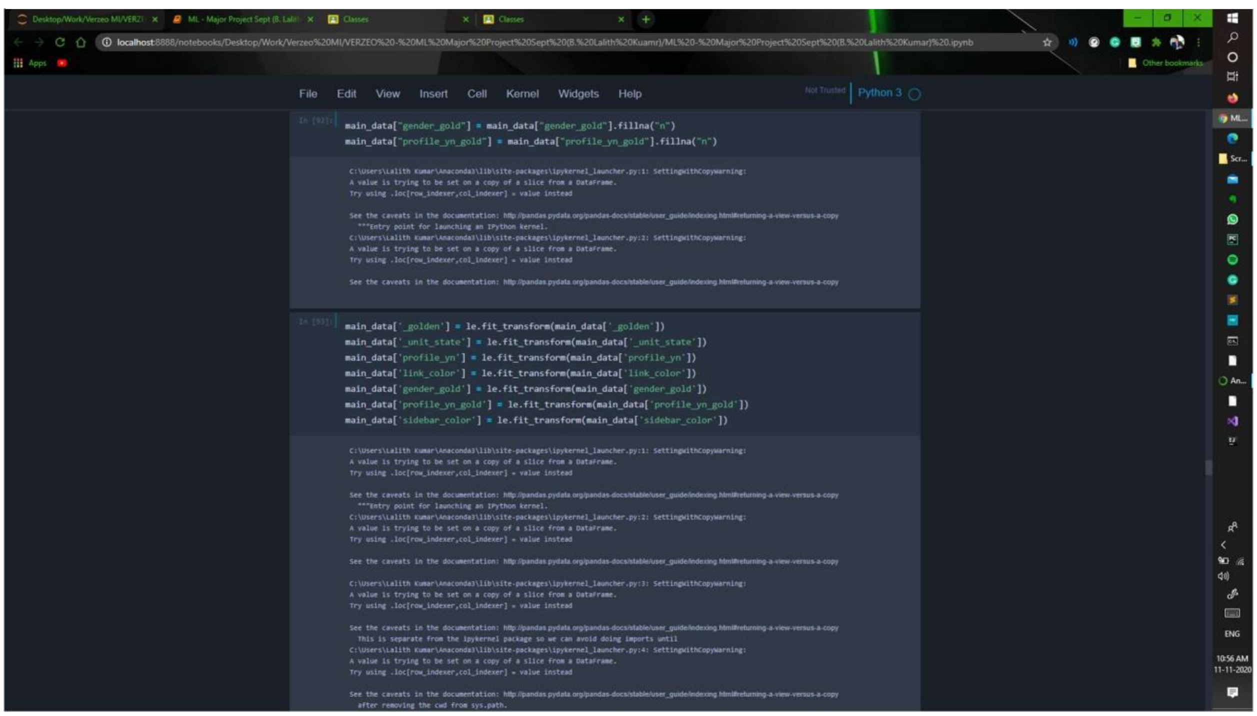
By importing train_test_split from sklearn.model_selection we've set the training and testing data of both independent and dependent variable i.e., X and Y.

```
In [89]: from sklearn.preprocessing import LabelEncoder

In [90]: le = LabelEncoder()

In [91]: main_data['_golden'] = le.fit_transform(main_data['_golden'])
         main_data['_unit_state'] = le.fit_transform(main_data['_unit_state'])
         main_data['profile_yn'] = le.fit_transform(main_data['profile_yn'])
         main_data['link_color'] = le.fit_transform(main_data['link_color'])
         main_data['gender_gold'] = le.fit_transform(main_data['gender_gold'])
         main_data['profile_yn_gold'] = le.fit_transform(main_data['profile_yn_gold'])
         main_data['sidebar_color'] = le.fit_transform(main_data['sidebar_color'])
```

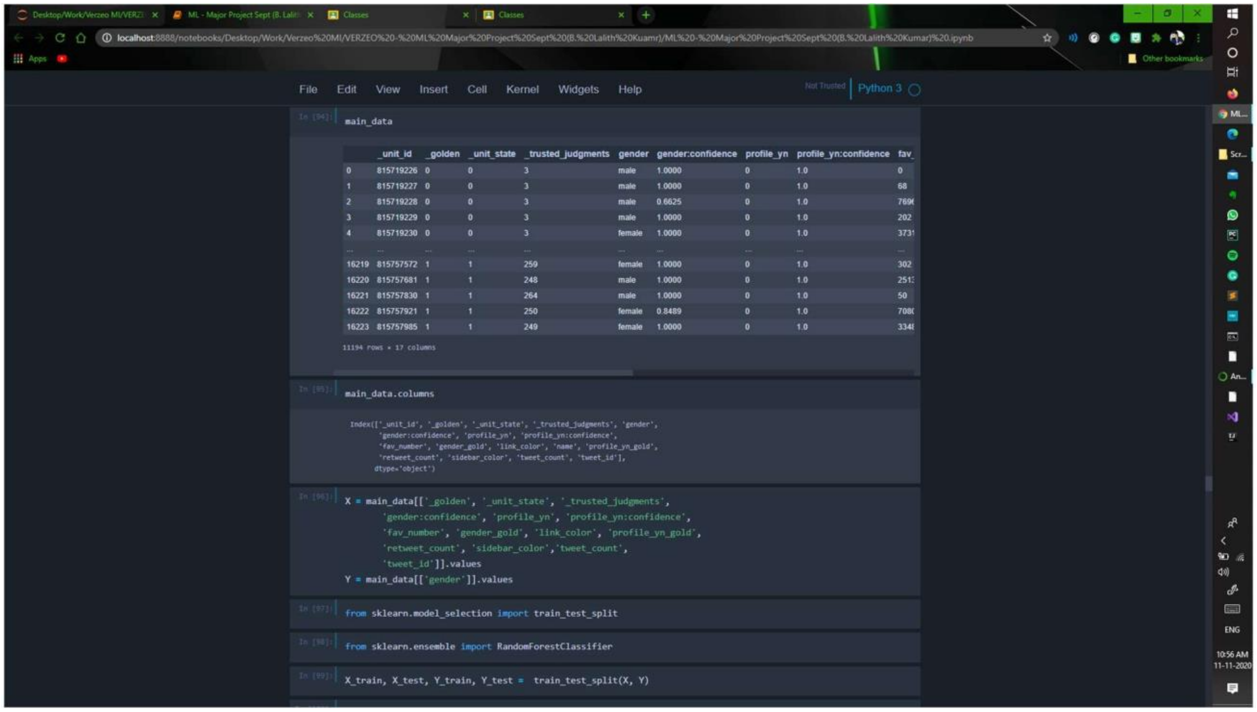
Error while encoding:



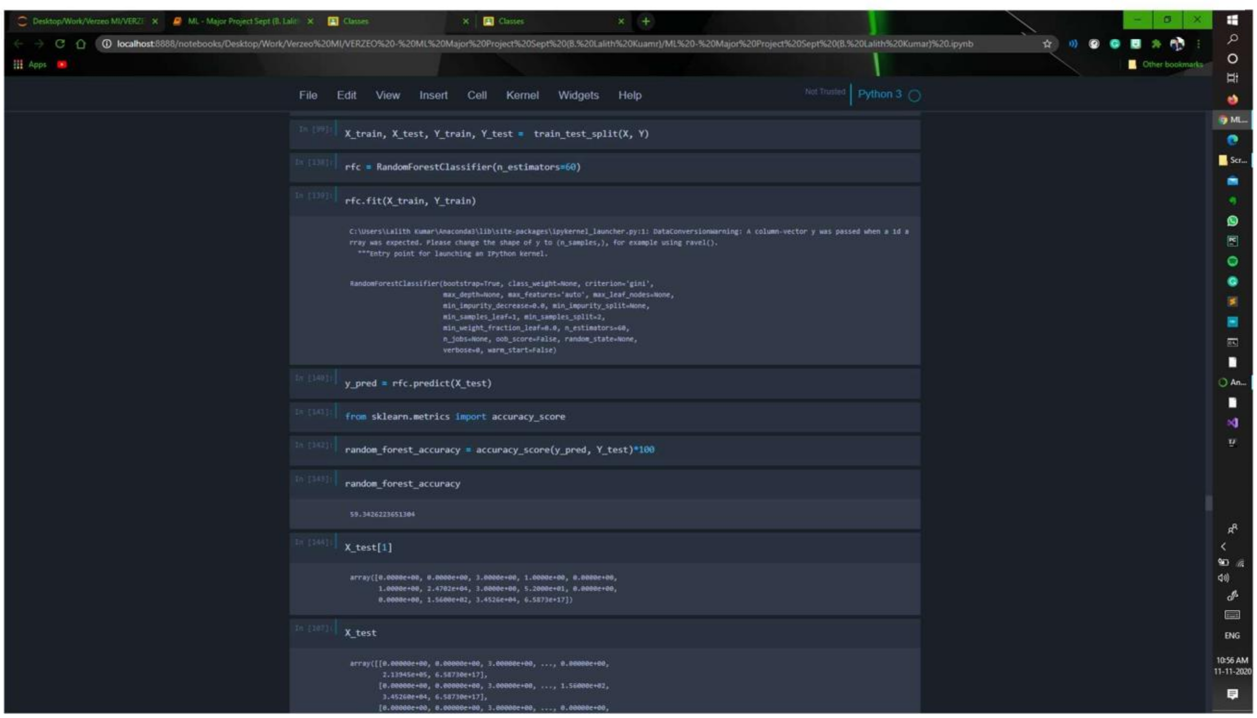
Error rectification while encoding:



Splitting data for training and testing purposes:



Calling Random Forest Algorithm:



4 . Ensemble Machine Learning Modelling:

1. Random Forest Algorithm –

We've imported RandomForestClassifier from sklearn.ensemble . We've fitted the X_train and Y_train to the random forest classifier and by using .predict() we've predicted the values for X_test.

By importing accuracy_scores from sklearn.metrics we've found the accuracy of predicted value i.e. y_pred and test value i.e. Y_test. As shown in the above figure.

2. Logistic Regression –

We've imported LogisticRegression from sklearn.linear_model. We've fitted the X_train and Y_train to the and by using .predict() we've predicted the values for X_test.

By importing accuracy_scores from sklearn.metrics we've found the accuracy of predicted value i.e. y_pred and test value i.e. Y_test. As shown in the figure.

```
In [113]: from sklearn.linear_model import LogisticRegression

In [114]: LogReg = LogisticRegression()

In [115]: LogReg.fit(X_train, Y_train)

C:\Users\Lalith Kumar\Anaconda3\lib\site-packages\sklearn\linear_model\logistic.py:432: FutureWarning: Default solver will be changed to
'lbfgs' in 0.22. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Lalith Kumar\Anaconda3\lib\site-packages\sklearn\utils\validation.py:724: DataConversionWarning: A column-vector y was passed whe
n a 1d array was expected. Please change the shape of y to (n_samples, ), for example using ravel().
  y = column_or_1d(y, warn=True)

LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=100,
                   multi_class='warn', n_jobs=None, penalty='l2',
                   random_state=None, solver='warn', tol=0.0001, verbose=0,
                   warm_start=False)

In [116]: y_pred = LogReg.predict(X_test)

In [117]: from sklearn.metrics import accuracy_score

In [118]: logistic_regession_accuracy = accuracy_score(y_pred, Y_test)*100

In [119]: logistic_regession_accuracy

50.01786352268668
```

3.SVM Algorithm –

We've imported svm from sklearn. We've fitted the X_train and Y_train to the random forest classifier and by using .predict() we've predicted the values for X_test.

By importing accuracy_scores from sklearn.metrics we've found the accuracy of predicted value i.e. y_pred and test value i.e. Y_test. As shown in the figure.

SVM ALGORITHM

```
In [183]: from sklearn import svm
          clf = svm.SVC(kernel='rbf',max_iter=-1)
          clf.fit(X_train, Y_train)

C:\Users\Lalith Kumar\Anaconda3\lib\site-packages\sklearn\utils\validation.py:724: DataConversionWarning: A column-vector y was passed whe
n a 1d array was expected. Please change the shape of y to (n_samples, ), for example using ravel().
  y = column_or_1d(y, warn=True)
C:\Users\Lalith Kumar\Anaconda3\lib\site-packages\sklearn\svm\base.py:193: FutureWarning: The default value of gamma will change from 'aut
o' to 'scale' in version 0.22 to account better for unscaled features. Set gamma explicitly to 'auto' or 'scale' to avoid this warning.
  "avoid this warning.", FutureWarning)

SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,
    decision_function_shape='ovr', degree=3, gamma='auto_deprecated',
    kernel='rbf', max_iter=-1, probability=False, random_state=None,
    shrinking=True, tol=0.001, verbose=False)

In [184]: Y_pred = clf.predict(X_test)

In [185]: from sklearn import metrics

          # Model Accuracy: how often is the classifier correct?
          svm_algorithm_accuracy = metrics.accuracy_score(Y_test, Y_pred)*100

In [186]: svm_algorithm_accuracy

50.80385852090033
```

4.KNN Algorithm –

We've imported KNeighborsClassifier from sklearn.neighbors. We've fitted the X_train and Y_train to the and by using .predict() we've predicted the values for X_test.

By importing accuracy_scores from sklearn.metrics we've found the accuracy of predicted value i.e. y_pred and test value i.e. Y_test. As shown in the figure.

KNN ALGORITHM

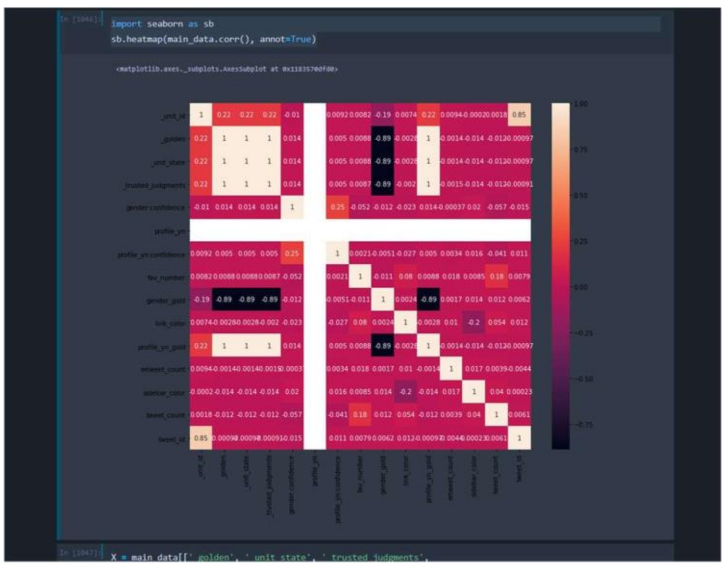
```
In [124]: from sklearn.neighbors import KNeighborsClassifier

In [125]: knn=KNeighborsClassifier(n_neighbors=60)

In [126]: knn.fit(X_train,Y_train)

C:\Users\Lalith Kumar\Anaconda3\lib\site-packages\ipykernel_launcher.py:1: DataConversionWarning: A column-vector y was passed when a 1d a
rray was expected. Please change the shape of y to (n_samples, ), for example using ravel().
  """Entry point for launching an IPython kernel.

KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
                     metric_params=None, n_jobs=None, n_neighbors=60, p=2,
                     weights='uniform')

In [127]: y_pred=knn.predict(X_test)

In [128]: from sklearn.metrics import accuracy_score

In [129]: accuracy_score(y_pred,Y_test)

0.548410146480886
```

5. Accuracy Results:

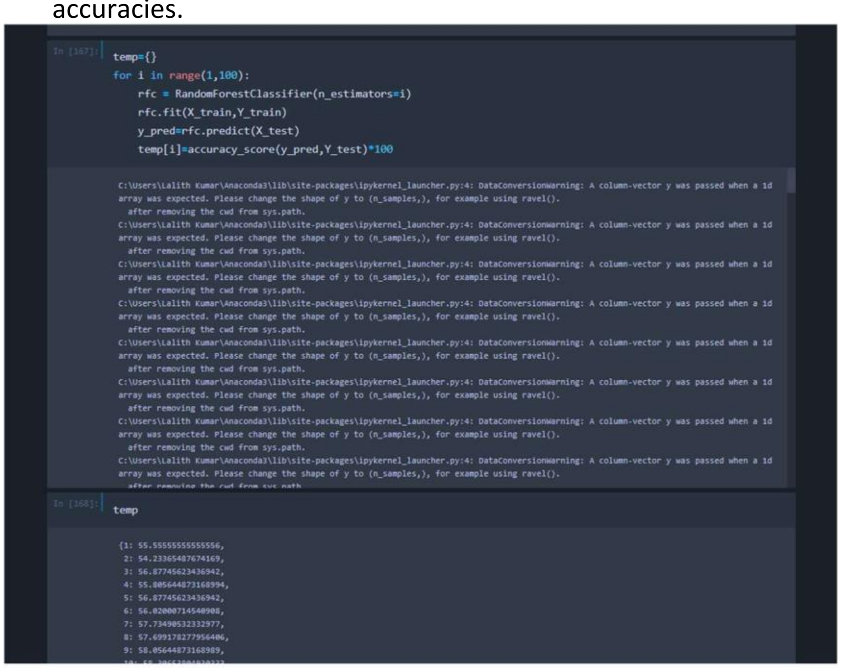
- ☐ ***Random Forests – 59.34***
- ☐ ***Logistic Regression – 50.01***
- ☐ ***SVM Algorithm – 50.80***
- ☐ ***KNN algorithm – 54.84***

Additional Functions Performed:

- ☐ Plotted Heatmap by importing seaborn



- ☐ In Random Forest calculated the accuracy for all the Estimators in the range of 1 to 100, plotted the graph for all the accuracies and found the average of 100 accuracies.



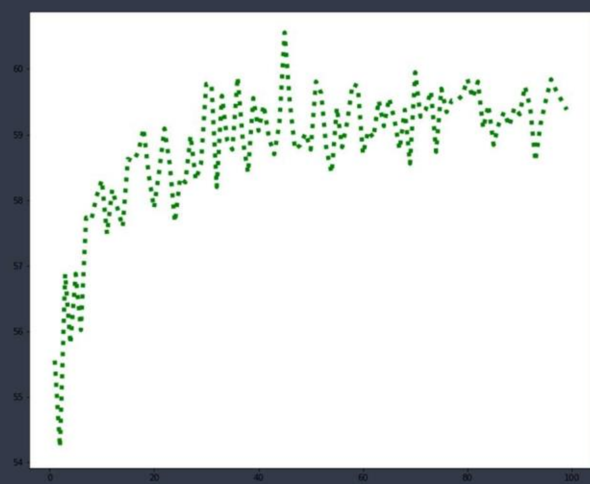
```
99: 59.37834941050375}
```

```
In [169]: x=list(range(1,100))
```

```
In [170]: y=[]
          for i in temp:
              y.append(temp[i])
```

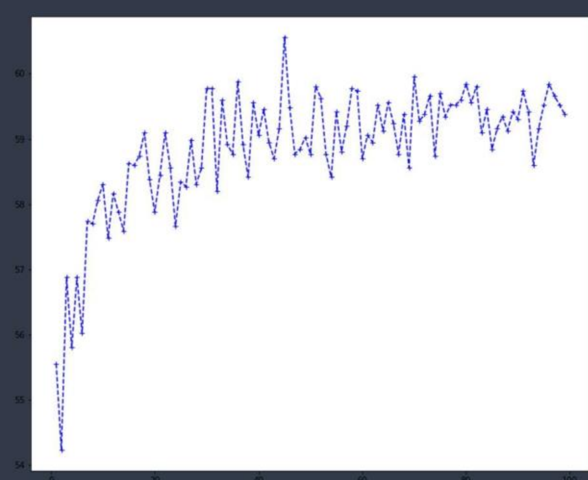
```
In [188]: plt.plot(x,y,color='green',linewidth=5,linestyle='dotted')
```

```
[<matplotlib.lines.Line2D at 0x2d19ed3f548>]
```




```
In [172]: plt.plot(x,y,'b+--')
```

```
[<matplotlib.lines.Line2D at 0x2d19ec55b08>]
```




```
In [173]: random_forest_accuracy= sum(y)/len(y)
```

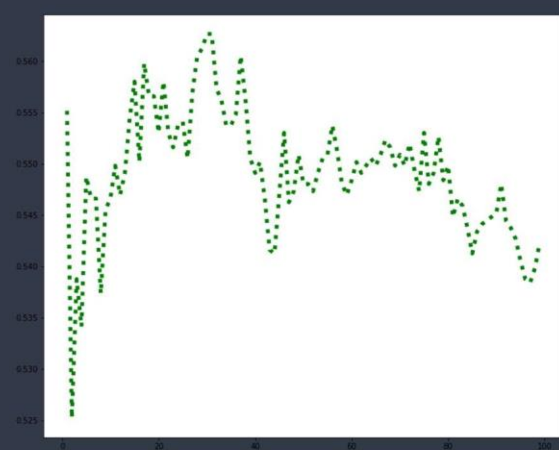
```
In [174]: random_forest_accuracy
```

```
58.84063933367257
```

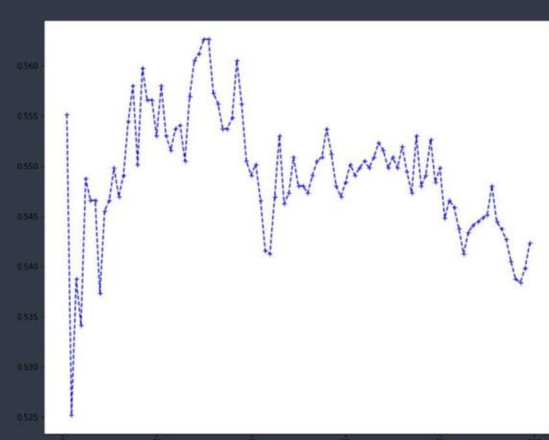
☐ In KNN Algorithm calculated the accuracy for all the Neighbours in the range of 1 to 100, plotted the graph for all the accuracies and found the average of 100 accuracies.

```
In [130]: temp={}
          for i in range(1,100):
              knn=KNeighborsClassifier(n_neighbors=i)
              knn.fit(X_train,Y_train)
              y_pred=knn.predict(X_test)
              temp[i]=accuracy_score(y_pred,Y_test)

C:\Users\Lalith Kumar\Anaconda3\lib\site-packages\ipykernel_launcher.py:4: DataConversionWarning: A column-vector y was passed when a 1d
array was expected. Please change the shape of y to (n_samples, ), for example using ravel().
  after removing the cwd from sys.path.
C:\Users\Lalith Kumar\Anaconda3\lib\site-packages\ipykernel_launcher.py:4: DataConversionWarning: A column-vector y was passed when a 1d
array was expected. Please change the shape of y to (n_samples, ), for example using ravel().
  after removing the cwd from sys.path.
C:\Users\Lalith Kumar\Anaconda3\lib\site-packages\ipykernel_launcher.py:4: DataConversionWarning: A column-vector y was passed when a 1d
array was expected. Please change the shape of y to (n_samples, ), for example using ravel().
```

```
In [132]: x=list(range(1,100))

In [133]: y=[]
          for i in temp:
              y.append(temp[i])

In [134]: plt.plot(x,y,color='green',linewidth=5,linestyle='dotted')

[<matplotlib.lines.Line2D at 0x2d19e1a0c08>]
```



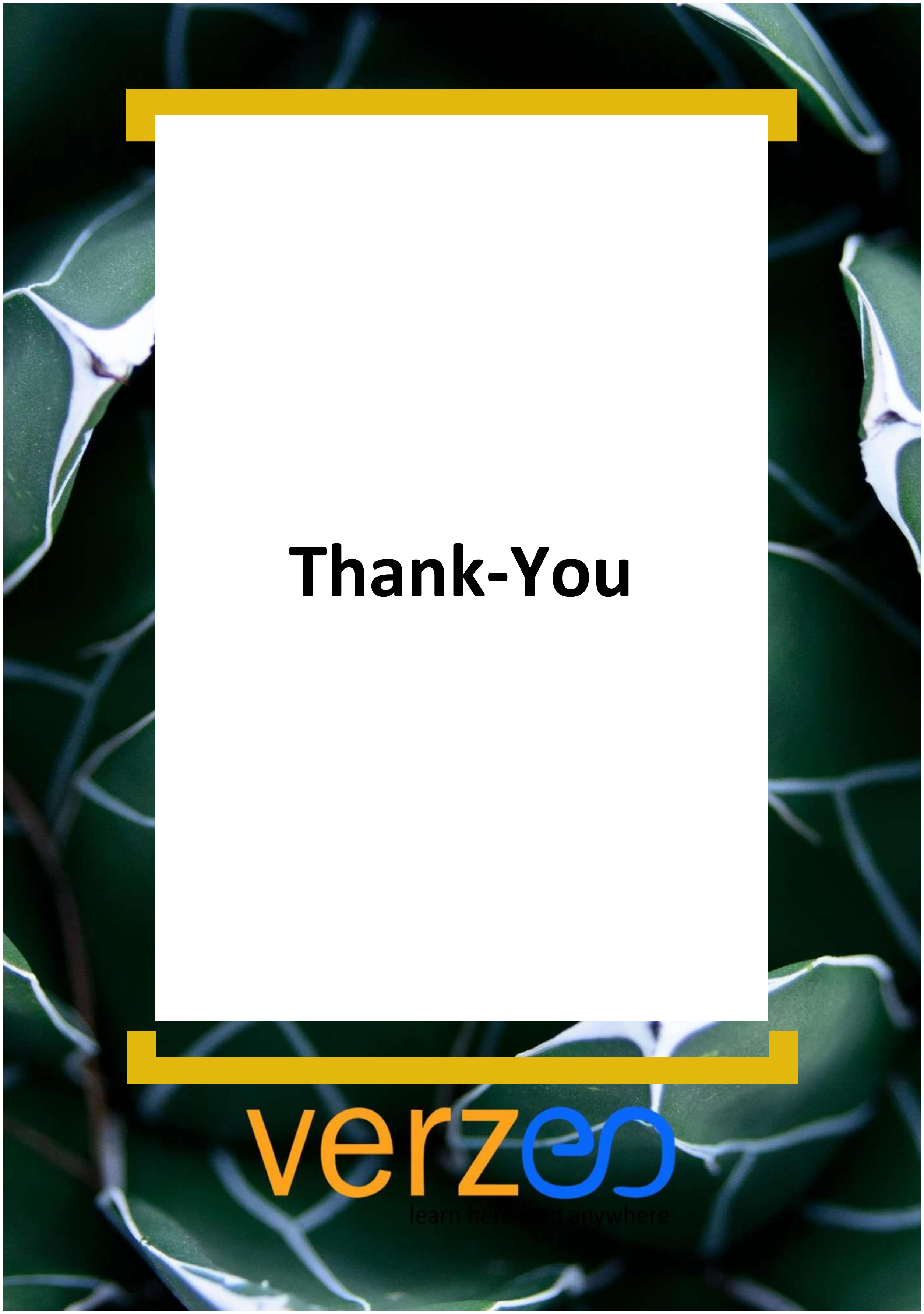

```
In [135]: plt.plot(x,y,'b+--')

[<matplotlib.lines.Line2D at 0x2d19e1be188>]
```




```
In [136]: knn_algorithm_accuracy = sum(y)/len(y)

In [137]: knn_algorithm_accuracy

0.5492185159923635

In [ ]:
```

Team – ML09B3

- B. Lalith Kumar
- Niti Goel
- Aswin CA
- Himani Thakkar
- Aishwary Krishna Singh
- Sreeja Samanthapuri
- Utkarsh Soni
- Ruchita Singh
- Santosh Reddy
- Mohammad Abdul Saleem
- Jay Chandra



Thank-You