MINI PROJECT (SEMESTER 5th)

Unintentional Information Leakage on Twitter
“Prediction of Gender, Interest and Occupation using Twitter”

Members -

Shruti Reddy (IIT2016019)

Garima Chadha (IIT2016020)

Vaibhav Srivastava (IIT2016034)

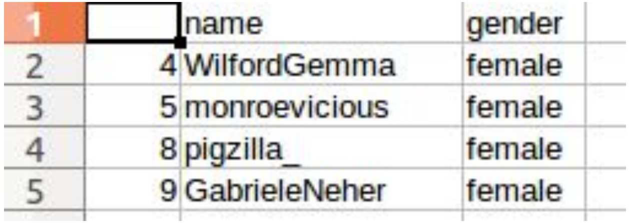
Niharika Shrivastava (IIT2016501)

Guidance By:-

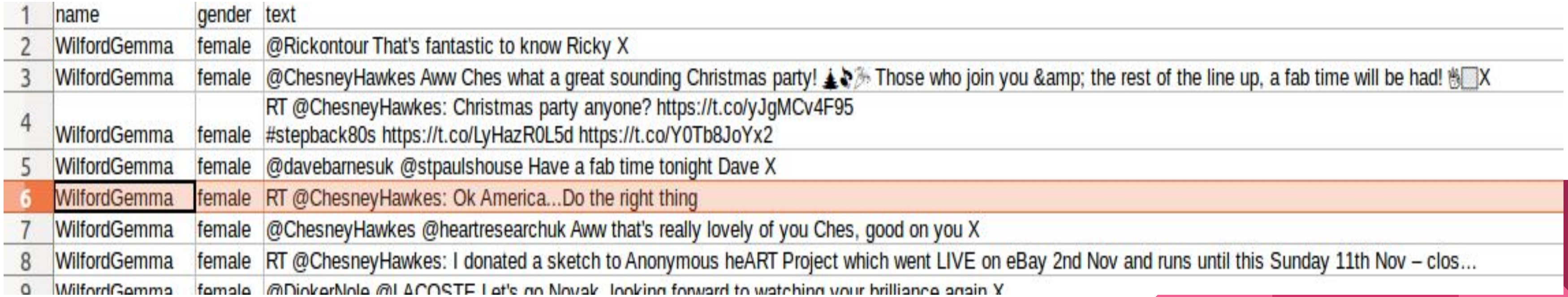
Dr. Bibhas Ghoshal

GENDER

DATA-SET :



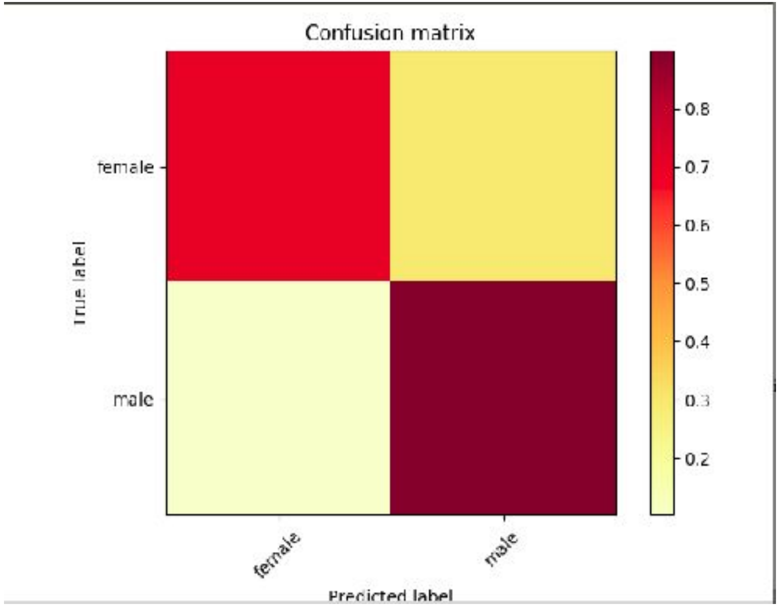
1		name	gender
2	4	WilfordGemma	female
3	5	monroevicious	female
4	8	pigzilla_	female
5	9	GabrieleNeher	female



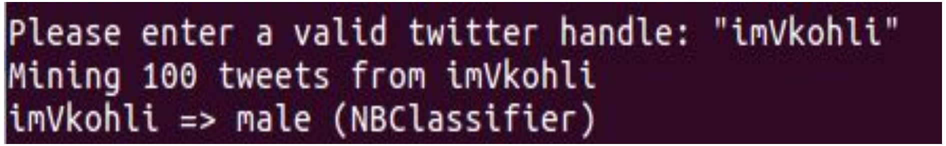
1	name	gender	text
2	WilfordGemma	female	@Rickontour That's fantastic to know Ricky X
3	WilfordGemma	female	@ChesneyHawkes Aww Ches what a great sounding Christmas party! 🎄🎶✨ Those who join you &amp; the rest of the line up, a fab time will be had! ✌□X
4	WilfordGemma	female	RT @ChesneyHawkes: Christmas party anyone? https://t.co/yJgMCv4F95 #stepback80s https://t.co/LyHazR0L5d https://t.co/Y0Tb8JoYx2
5	WilfordGemma	female	@davebarnesuk @stpaulshouse Have a fab time tonight Dave X
6	WilfordGemma	female	RT @ChesneyHawkes: Ok America...Do the right thing
7	WilfordGemma	female	@ChesneyHawkes @heartresearchuk Aww that's really lovely of you Ches, good on you X
8	WilfordGemma	female	RT @ChesneyHawkes: I donated a sketch to Anonymous heART Project which went LIVE on eBay 2nd Nov and runs until this Sunday 11th Nov – clos…
9	WilfordGemma	female	@DjokerNole @LACOSTE Let's go Novak, looking forward to watching your brilliance again X

GENDER

CONFUSION MATRIX :

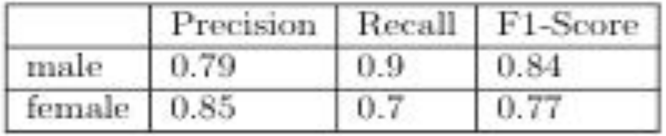


OUTPUT :



```
Please enter a valid twitter handle: "imVkohli"
Mining 100 tweets from imVkohli
imVkohli => male (NBClassifier)
```

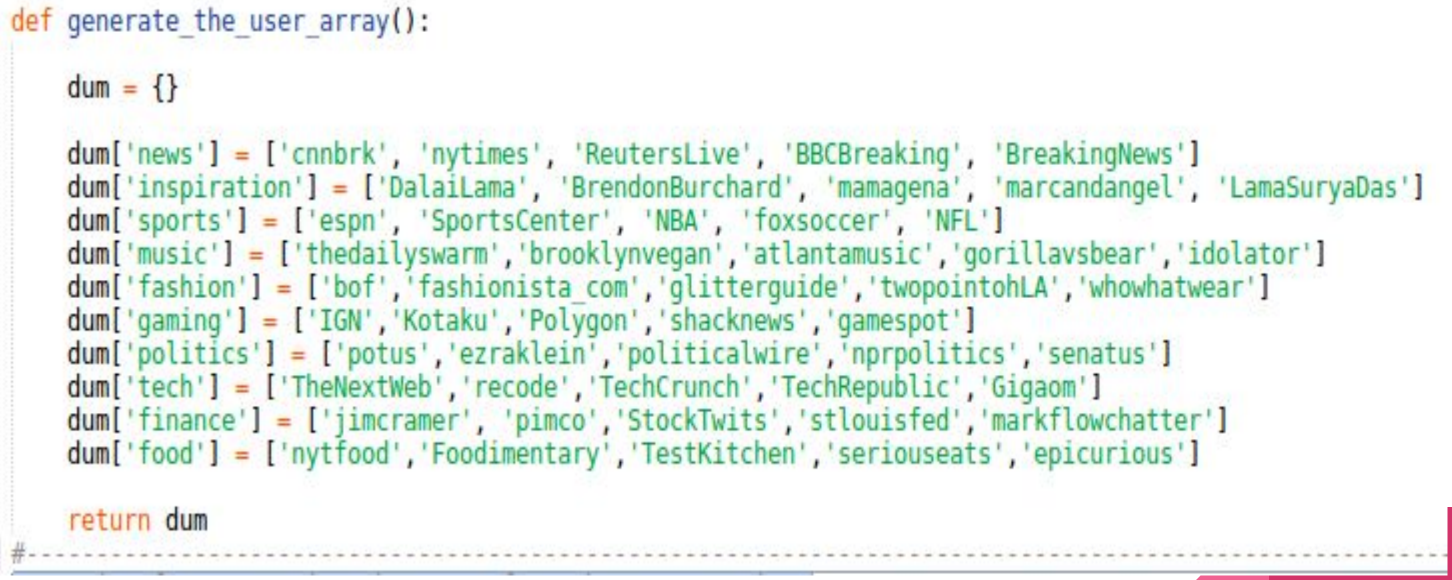
PRECISION, RECALL, F1-SCORE :



	Precision	Recall	F1-Score
male	0.79	0.9	0.84
female	0.85	0.7	0.77

INTEREST

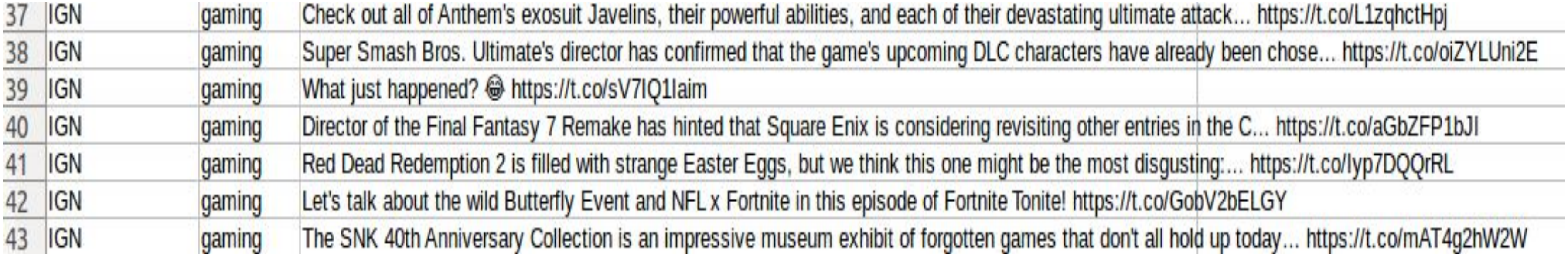
MULTI CLASS DICTIONARY :



```
def generate_the_user_array():

    dum = {}

    dum['news'] = ['cnnbrk', 'nytimes', 'ReutersLive', 'BBCBreaking', 'BreakingNews']
    dum['inspiration'] = ['DalaiLama', 'BrendonBurchard', 'mamagena', 'marcandangel', 'LamaSuryaDas']
    dum['sports'] = ['espn', 'SportsCenter', 'NBA', 'foxsoccer', 'NFL']
    dum['music'] = ['thedailyswarm','brooklynvegan','atlantamusic','gorillavsbear','idolator']
    dum['fashion'] = ['bof','fashionista_com','glitterguide','twopointohLA','whowhatwear']
    dum['gaming'] = ['IGN','Kotaku','Polygon','shacknews','gamespot']
    dum['politics'] = ['potus','ezraklein','politicalwire','nprpolitics','senatus']
    dum['tech'] = ['TheNextWeb','recode','TechCrunch','TechRepublic','Gigaom']
    dum['finance'] = ['jimcramer', 'pimco','StockTwits','stlouisfed','markflowchatter']
    dum['food'] = ['nytfood','Foodimentary','TestKitchen','seriouseats','epicurious']

    return dum
#-------------------------------------------------------------------------------------------------
```

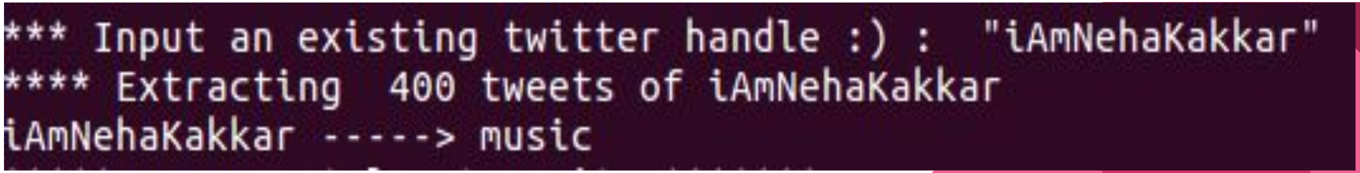
INTEREST

DATASET :



37	IGN	gaming	Check out all of Anthem's exosuit Javelins, their powerful abilities, and each of their devastating ultimate attack... https://t.co/L1zqhctHpj
38	IGN	gaming	Super Smash Bros. Ultimate's director has confirmed that the game's upcoming DLC characters have already been chose... https://t.co/oiZYLUni2E
39	IGN	gaming	What just happened? 😂 https://t.co/sV7IQ1laim
40	IGN	gaming	Director of the Final Fantasy 7 Remake has hinted that Square Enix is considering revisiting other entries in the C... https://t.co/aGbZFP1bJI
41	IGN	gaming	Red Dead Redemption 2 is filled with strange Easter Eggs, but we think this one might be the most disgusting:... https://t.co/Iyp7DQQrRL
42	IGN	gaming	Let's talk about the wild Butterfly Event and NFL x Fortnite in this episode of Fortnite Tonite! https://t.co/GobV2bELGY
43	IGN	gaming	The SNK 40th Anniversary Collection is an impressive museum exhibit of forgotten games that don't all hold up today... https://t.co/mAT4g2hW2W

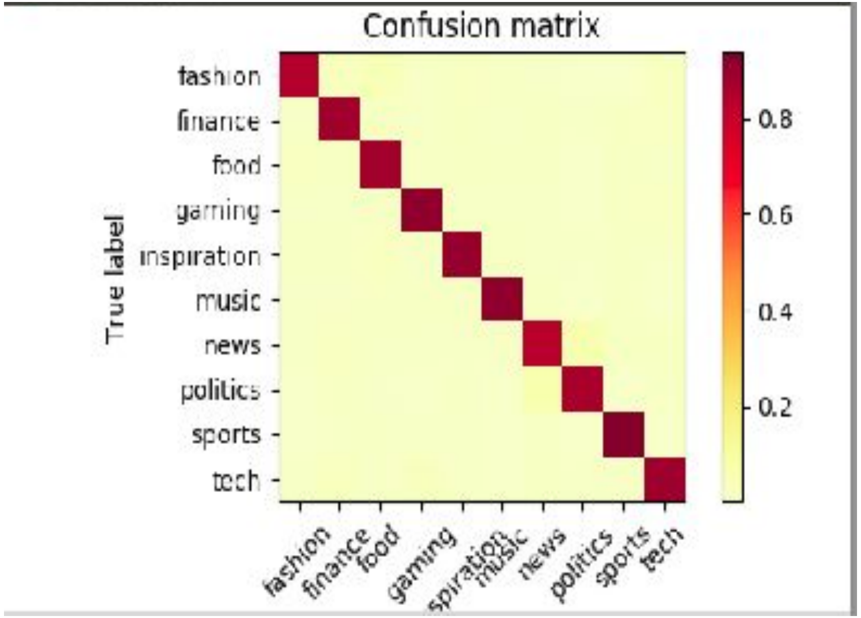
OUTPUT :



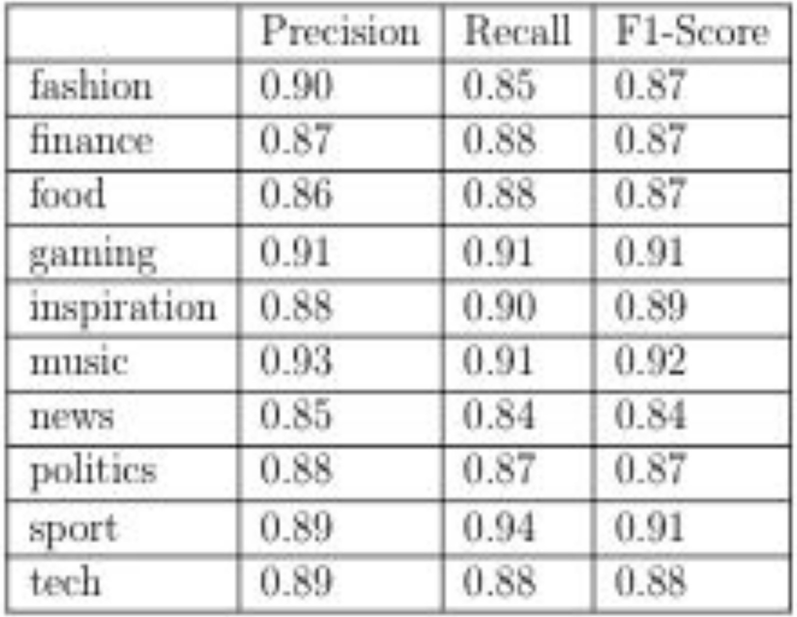
```
*** Input an existing twitter handle :) :  "iAmNehaKakkar"
**** Extracting  400 tweets of iAmNehaKakkar
iAmNehaKakkar -----> music
```

INTEREST

CONFUSION MATRIX :



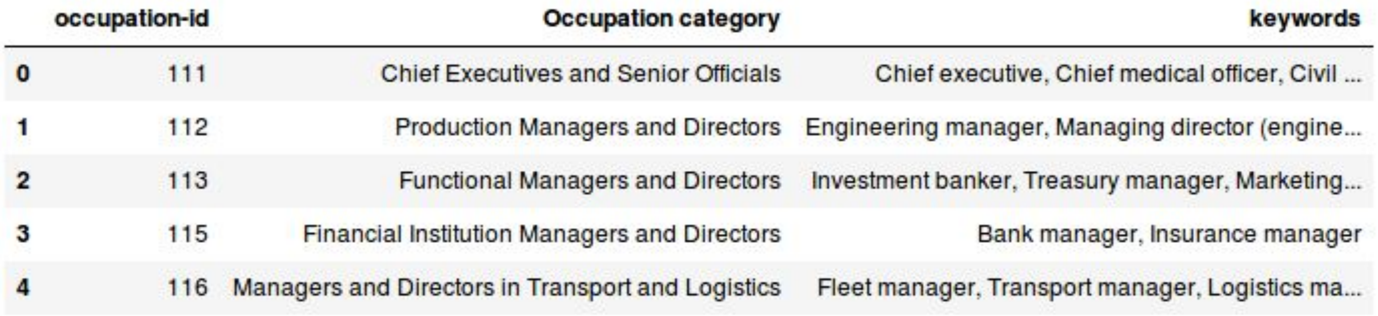
PRECISION, RECALL, F1-SCORE



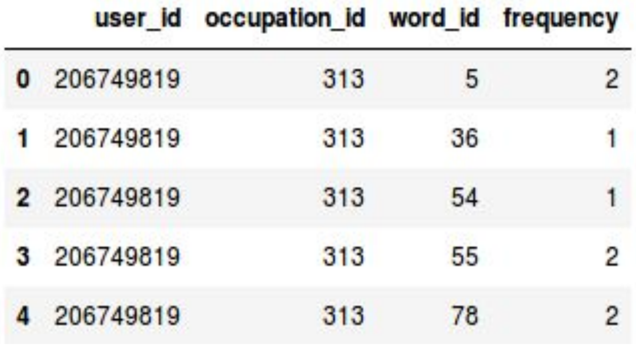
	Precision	Recall	F1-Score
fashion	0.90	0.85	0.87
finance	0.87	0.88	0.87
food	0.86	0.88	0.87
gaming	0.91	0.91	0.91
inspiration	0.88	0.90	0.89
music	0.93	0.91	0.92
news	0.85	0.84	0.84
politics	0.88	0.87	0.87
sport	0.89	0.94	0.91
tech	0.89	0.88	0.88

OCCUPATION

DATASET :



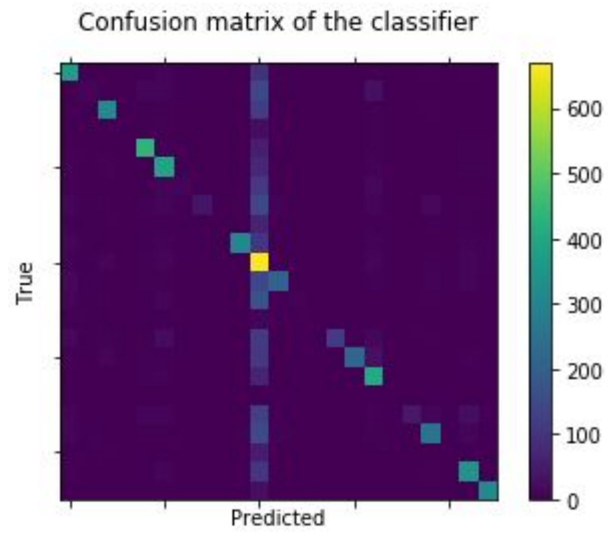
	occupation-id	Occupation category	keywords
0	111	Chief Executives and Senior Officials	Chief executive, Chief medical officer, Civil ...
1	112	Production Managers and Directors	Engineering manager, Managing director (engine...
2	113	Functional Managers and Directors	Investment banker, Treasury manager, Marketing...
3	115	Financial Institution Managers and Directors	Bank manager, Insurance manager
4	116	Managers and Directors in Transport and Logistics	Fleet manager, Transport manager, Logistics ma...



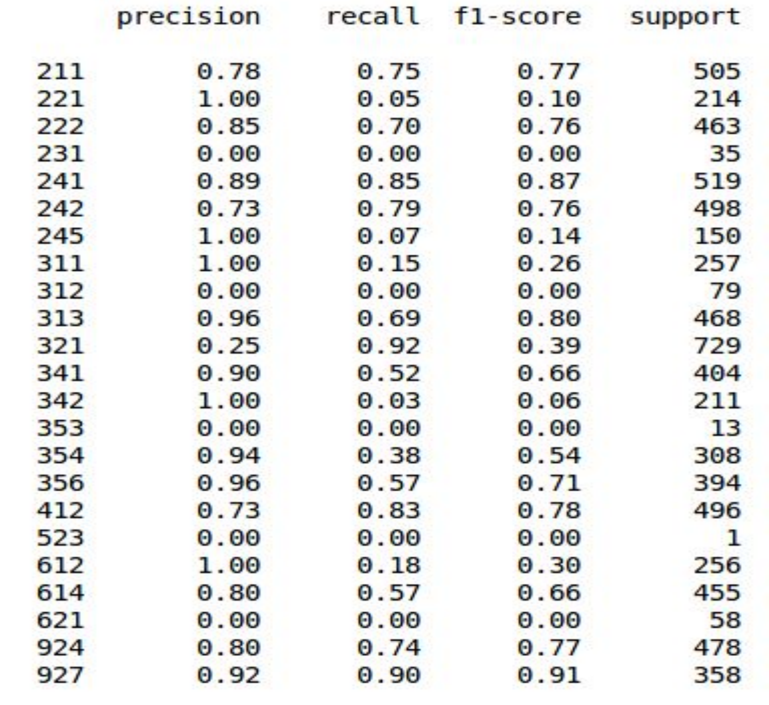
	user_id	occupation_id	word_id	frequency
0	206749819	313	5	2
1	206749819	313	36	1
2	206749819	313	54	1
3	206749819	313	55	2
4	206749819	313	78	2

OCCUPATION

CONFUSION MATRIX :

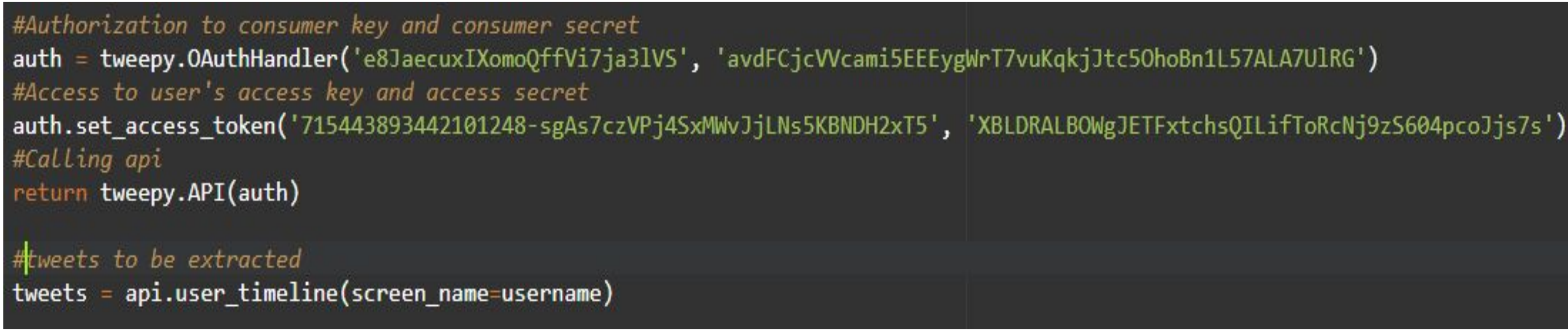


PRECISION, RECALL, F1-SCORE:



```
              precision    recall  f1-score   support

         211       0.78      0.75      0.77       505
         221       1.00      0.05      0.10       214
         222       0.85      0.70      0.76       463
         231       0.00      0.00      0.00        35
         241       0.89      0.85      0.87       519
         242       0.73      0.79      0.76       498
         245       1.00      0.07      0.14       150
         311       1.00      0.15      0.26       257
         312       0.00      0.00      0.00        79
         313       0.96      0.69      0.80       468
         321       0.25      0.92      0.39       729
         341       0.90      0.52      0.66       404
         342       1.00      0.03      0.06       211
         353       0.00      0.00      0.00        13
         354       0.94      0.38      0.54       308
         356       0.96      0.57      0.71       394
         412       0.73      0.83      0.78       496
         523       0.00      0.00      0.00         1
         612       1.00      0.18      0.30       256
         614       0.80      0.57      0.66       455
         621       0.00      0.00      0.00        58
         924       0.80      0.74      0.77       478
         927       0.92      0.90      0.91       358
```

OBTAINING TWEETS - TWEEPY.



```
#Authorization to consumer key and consumer secret
auth = tweepy.OAuthHandler('e8JaecuxIXomoQffVi7ja3lVS', 'avdFCjcVVcami5EEEygWrT7vuKqkjJtc5OhoBn1L57ALA7UlRG')
#Access to user's access key and access secret
auth.set_access_token('715443893442101248-sgAs7czVPj4SxMWvJjLNs5KBNDH2xT5', 'XBLDRALBOWgJETFxtchsQILifToRcNj9zS604pcoJjs7s')
#Calling api
return tweepy.API(auth)

#tweets to be extracted
tweets = api.user_timeline(screen_name=username)
```

EXPERIMENTS USING DATA SET & TWEEPY.

WE INCREASED THE DATASET GRADUALLY FROM 400 TWEETS PER PERSON TO 600 AND MORE…

WE FIND ACCURACY IN CASE OF GENDER INCREASE FROM

77% -> 79% -> 80% ->…..->81%.

THIS LEADS TO A POSSIBLE FUTURESCOPE..

FUTURE PLANS

WE EXTRACT TWEETS THEN PREDICT THE ANY OF THE ABOVE 3 ATTRIBUTE , THEN WE USE THIS ATTRIBUTE AGAIN TO TRAIN THE MACHINE ALONG WITH TWEETS FOR ANY OTHER ATTRIBUTE.

Eg : We find the “interest” using our machine, then we take this as a feature label to train machine for “gender” along with tweets.

The results of this experiments might result in increase/decrease of accuracy that shows “one attribute” affect the prediction of “other attributes”, that can be used to drive some co-relation.. “A WORK NOT DONE YET ….”

Thank You