Ethnicity prediction using popular twitter accounts

Nirmal kumar Ravi

Introduction

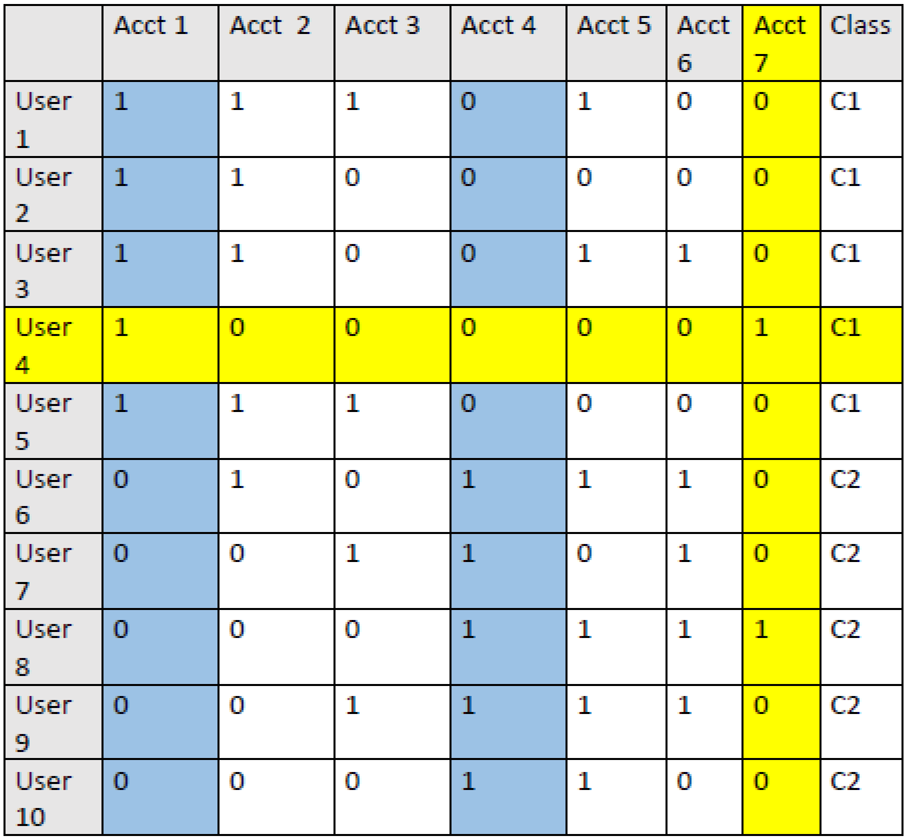
Can we predict ethnicity using most popular twitter accounts?
People follows movie stars, politicians and sports person in twitter. In this project we try to predict ethnicity of users with whom they follow.

Background

We select some of popular accounts from twitter whose followers belongs to some ethnicity, For example people who follows Mr. Narendra modi are mostly Indians. We use this information to label the data and train a model

We call the popular accounts as **prime accounts.** First we came up with set of prime accounts, we use these prime accounts to label the data. Choosing this prime accounts is very important, because these prime accounts help us to learn other features of ethic group. The prime accounts should be popular among particular ethnic group at-least 90% of followers who follows those accounts should belong to same ethnicity and these prime accounts should cover wide area like politics , sports to cover all set of audience. Of course it is difficult to come with the set of prime accounts. With domain knowledge about that ethnic group it is possible

Once we came with prime accounts, we sample 300 followers and 5000 friends of those followers. And then we construct a **Boolean matrix** (refer fig below) with users in y axis and accounts they follow on x axis. After that we label those users with prime accounts they follow.

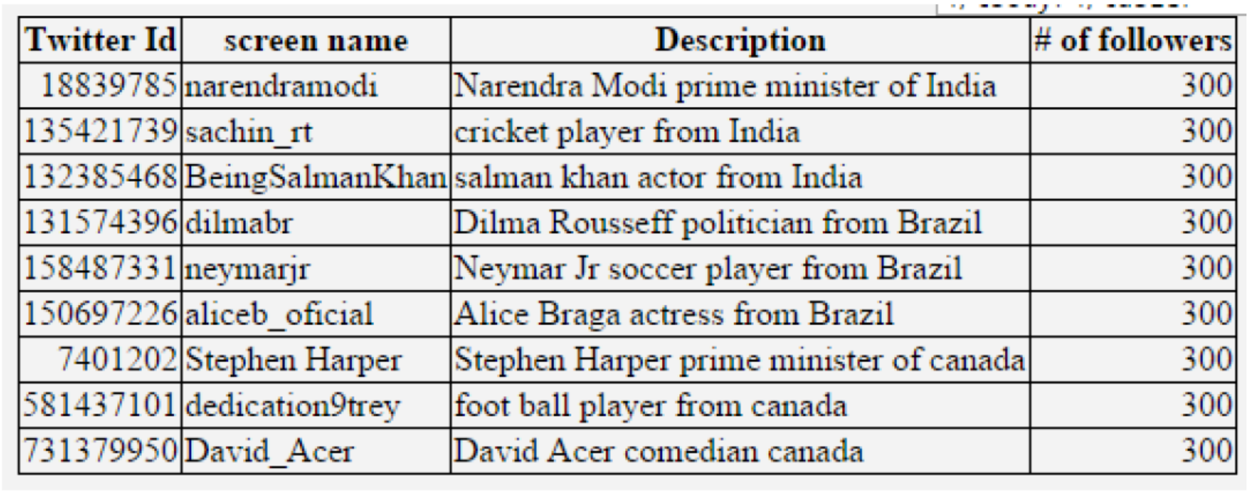


	Acct 1	Acct 2	Acct 3	Acct 4	Acct 5	Acct 6	Acct 7	Class
User 1	1	1	1	0	1	0	0	C1
User 2	1	1	0	0	0	0	0	C1
User 3	1	1	0	0	1	1	0	C1
User 4	1	0	0	0	0	0	1	C1
User 5	1	1	1	0	0	0	0	C1
User 6	0	1	0	1	1	1	0	C2
User 7	0	0	1	1	0	1	0	C2
User 8	0	0	0	1	1	1	1	C2
User 9	0	0	1	1	1	1	0	C2
User 10	0	0	0	1	1	0	0	C2

Now we are ready to fit and predict.

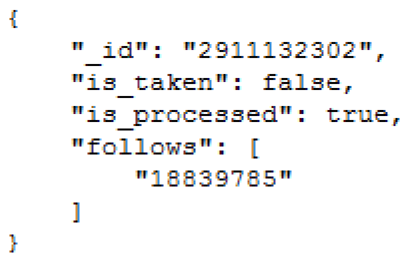
Data Collection

We came up with nine prime accounts with three ethnic group (Canadian, Indian, and Brazilian)



Twitter Id	screen name	Description	# of followers
18839785	narendramodi	Narendra Modi prime minister of India	300
135421739	sachin_rt	cricket player from India	300
132385468	BeingSalmanKhan	salman khan actor from India	300
131574396	dilmabr	Dilma Rousseff politician from Brazil	300
158487331	neymarjr	Neymar Jr soccer player from Brazil	300
150697226	aliceb_oficial	Alice Braga actress from Brazil	300
7401202	Stephen Harper	Stephen Harper prime minister of canada	300
581437101	dedication9trey	foot ball player from canada	300
731379950	David_Acer	David Acer comedian canada	300

We collected data for three days, we used Mongo DB as our data store and run data collection scripts on four nodes to get our Boolean matrix. Each record looks like this

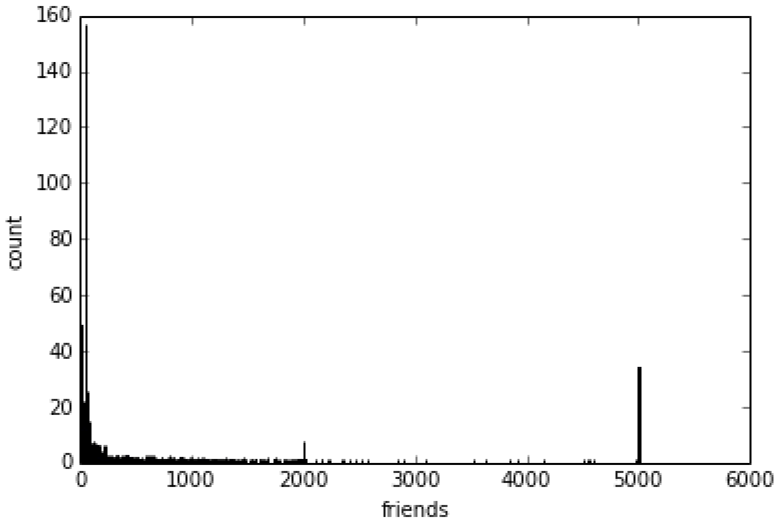


```
{
    "_id": "2911132302",
    "is_taken": false,
    "is_processed": true,
    "follows": [
        "18839785"
    ]
}
```

Keys	Description
_id	Id of twitter user
Is_taken	Flag for data processing
Is_processed	Flag for data processing
follows	Ids of twitter accounts whom the user follows

Methods

After we collect the data, we exported the data as a JSON file. This makes our processing faster. Then we read through each record, we used python Counters and Dictionary to create list of dictionaries. Each element in the list represents row of our Boolean matrix and dictionary represent mapping from column to value. Then we used DictVectorizer to convert it in to sparse matrix.



```
[(40, 157), (41, 73), (1, 49), (2, 34), (5001, 34), (42, 31), (4, 28), (6, 28), (8, 28), (9,
26), (13, 26), (61, 26), (3, 25), (39, 25), (7, 23), (11, 22), (16, 22), (22, 22), (60, 22)
, (43, 21), (5, 20), (12, 19), (14, 19), (18, 19), (28, 19), (38, 19), (46, 19), (17, 18), (
50, 18), (25, 17)]
```

From the graph above, we can see with we have 157 users with 40 friends

Experiments

Once we process the data we then fit **logistic regression.** Once we fit our regression model we did **5 fold cross validation**, which gives 98% of accuracy.

Then we test our model on out **manually labeled test set**. We get 78% accuracy.

Related work

Understanding the Demographics of Twitter Users . In this paper the data is labeled using US census. Our approach is unique as we have used popular accounts to label our data.

Conclusions and Future Work

Our results suggests we can predict ethnicity of user using popular twitter. Our approach is unique. We get training set labeled almost freely. We can improve our model accuracy by collecting more data, by collecting more data we can learn more features about that ethnicity.

We search through the space to learn features about ethnic group. In real world we have dense clusters our approach learns important features of those clusters.

There are many other approach to label the data, for example we can use hash tags to label the data. We can combine models to label the data which also reduces error.