Mini Project - (Semester 5)

UNINTENTIONAL PERSONAL INFORMATION LEAKAGE VIA ONLINE SOCIAL NETWORK

SHRUTI REDDY (IIT2016019)

GARIMA CHADHA (IIT2016020)

VAIBHAV SRIVASTAVA (IIT2016034)

NIHARIKA SHRIVASTAVA (IIT2016501)

MENTORED BY -

DR. BIBHAS GHOSHAL

INTRODUCTION

- Due to cyber crimes happening these days, people generally tend to hide their personal information like age, gender etc.
- In recent years, there has been an exponential increase in user generated text content, mainly in the form of blogs, tweets, reviews, and messages on social networks.
- This increase in textual information has sparked interest in automatically predicting user attributes such as gender, profession of users etc.
- In this project, we decided to extract hidden attributes like gender, interests and profession of a user belonging to an OSN (Online Social Network).

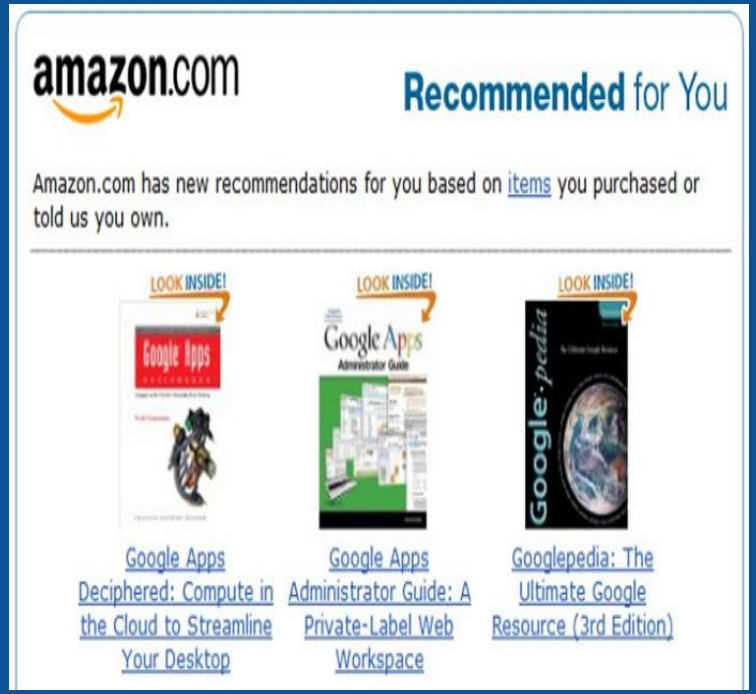
HOW TO USE THE PREDICTED DATA ?

- Abundance of data on the internet, i.e. : Big Data
- Variety of OSNs, e.g. : Facebook, LinkedIn, Twitter
- Huge user base (in billions)
- User interaction possible in the form of messages, tweets, blogs, likes, comments, connections, etc.


```
        "lang": "en",
        "metadata": {
            "iso_language_code": "en",
            "result_type": "recent"
        },
        "place": null,
        "possibly_sensitive": false,
        "retweet_count": 8,
        "retweeted": false,
        "source": "<a href=\"https://mobile.twitter.com\" rel=\"nofollow\">Twitter Lite</a>",
        "text": "Crack the code on Manafort's secret ties to Moscow -- that's coming into focus now -- and you've gone a long way to\u2026
https://t.co/uDUMJoVYZj",
        "truncated": true,
        "user": {
            "contributors_enabled": false,
            "created_at": "Sat Jul 14 19:35:48 +0000 2012",
            "default_profile": true,
            "default_profile_image": false,
            "description": "#Natsec @observer, historian, security consultant, author, provocateur, bon vivant, polyglot,
counterintelligencer, cat guy. Former NSA, NAVSECGRU, NWC.",
```

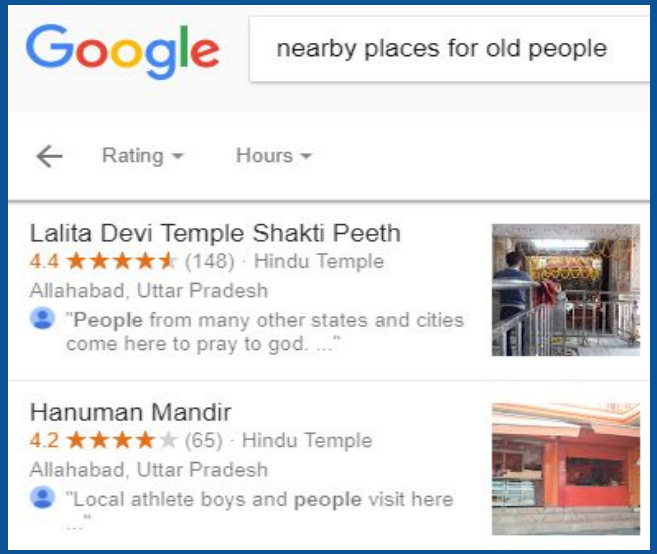
POSSIBLE USES OF OUR PROJECT

- Advertisement

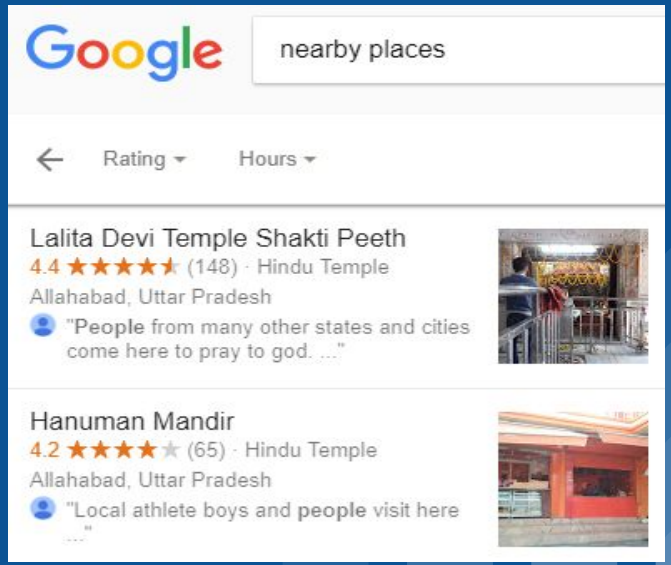


- Prevent Cyber Crime

- **Enhancement in search engine results**



PRESENT SEARCH

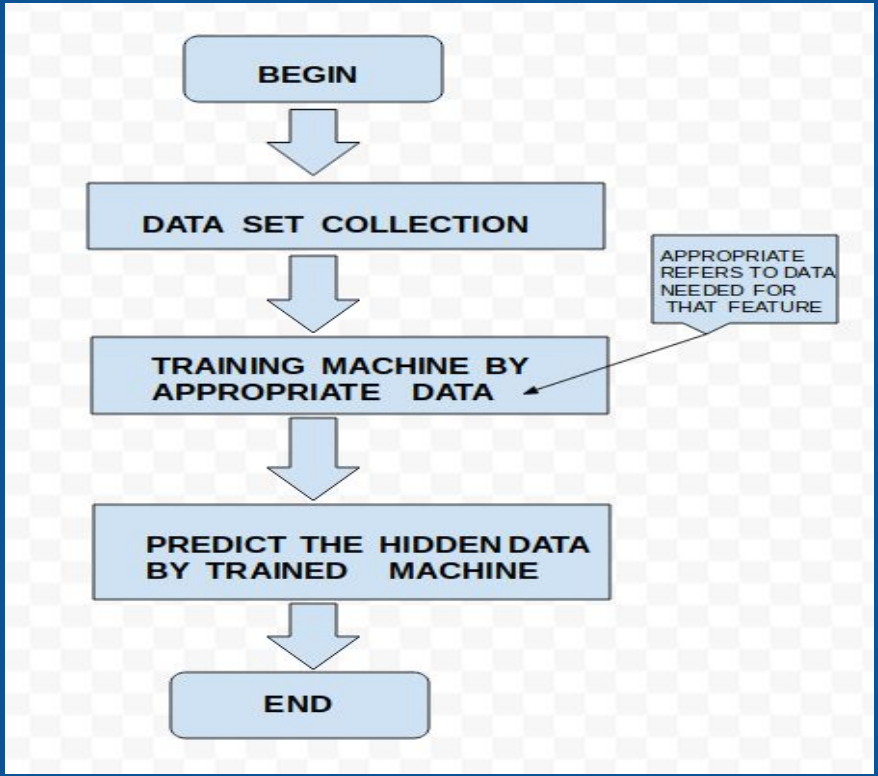


ENHANCED SEARCH

OVERVIEW

- Different people tweet differently based on their interests, gender etc.
- Example : Males generally tweet more about technology, sports and politics, whereas Females tweet more about fashion and food.
- Example : Different people have varying interests. Teenagers tweet about trending memes, music etc whereas tweets of Elderly People generally promote calm and motivating thoughts.
- Example : A Musician is expected to tweet more about new or old songs, whereas a Technocrat will generally tweet about latest gadgets and technological developments.

WORK FLOW


BEGIN
DATA SET COLLECTION
APPROPRIATE REFERS TO DATA NEEDED FOR THAT FEATURE
TRAINING MACHINE BY APPROPRIATE DATA
PREDICT THE HIDDEN DATA BY TRAINED MACHINE
END

RETRIEVAL OF DATA

- OSN Selected - TWITTER
- Create a Twitter application.
- Using the API Key, API Secret, Access Token and Access Token Secret, access Twitter application to collect data
- API used: Tweepy.
- Data extracted format: JSON

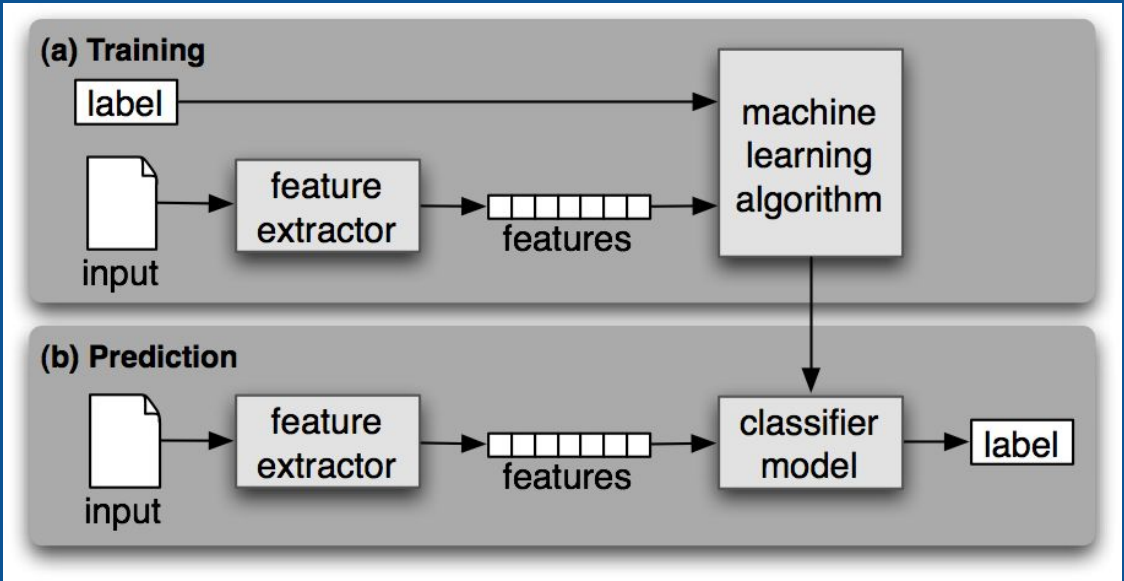
- NEXT GOAL - To create huge dataset(10 GB) on Hadoop Server using Flume.

TRAINING AND PREDICTION

- The next step is to extract hidden attributes.
- The 3 major attributes extraction on which we are focussing on :-

- GENDER
- INTEREST
- OCCUPATION

DIAGRAMMATIC REPRESENTATION



GENDER

TRAINING : Done using Naive Bayes Classifier.

Input: Dataset comprising of Name, Gender and Frequency of Name.
Eg: The names : Mary Kom, Mary Trump, Mary Obama;
Will be stored as : (Mary,F,3)

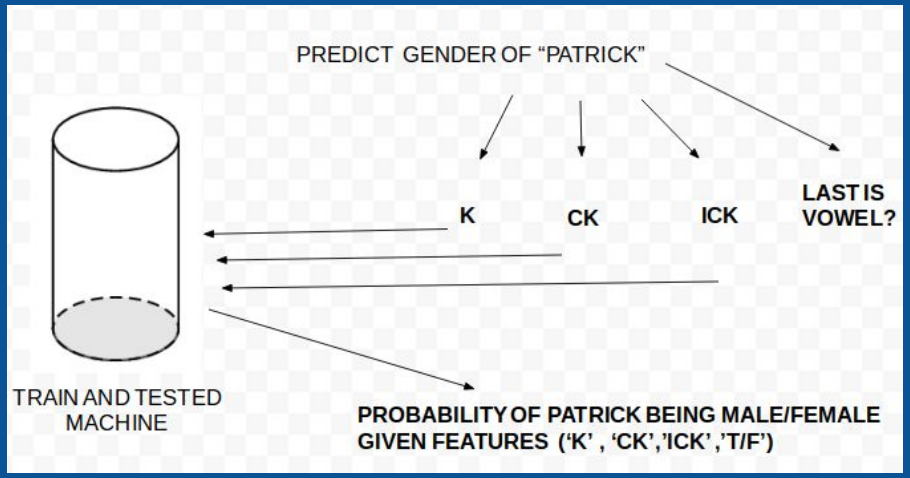
FEATURE EXTRACTOR: Extract features from the dataset. Features are :
(Last Letter, Last Two Letters, Last Three Letters, Last is Vowel (True/False))

Eg: Feature Set for (Mary,F,3) → (Y, RY, ARY, False)
Feature Set for (Robert,M,5) → (T, RT, ERT, False)

Train the machine using this feature vector...

GENDER

PREDICTION :



BAYES THEOREM

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

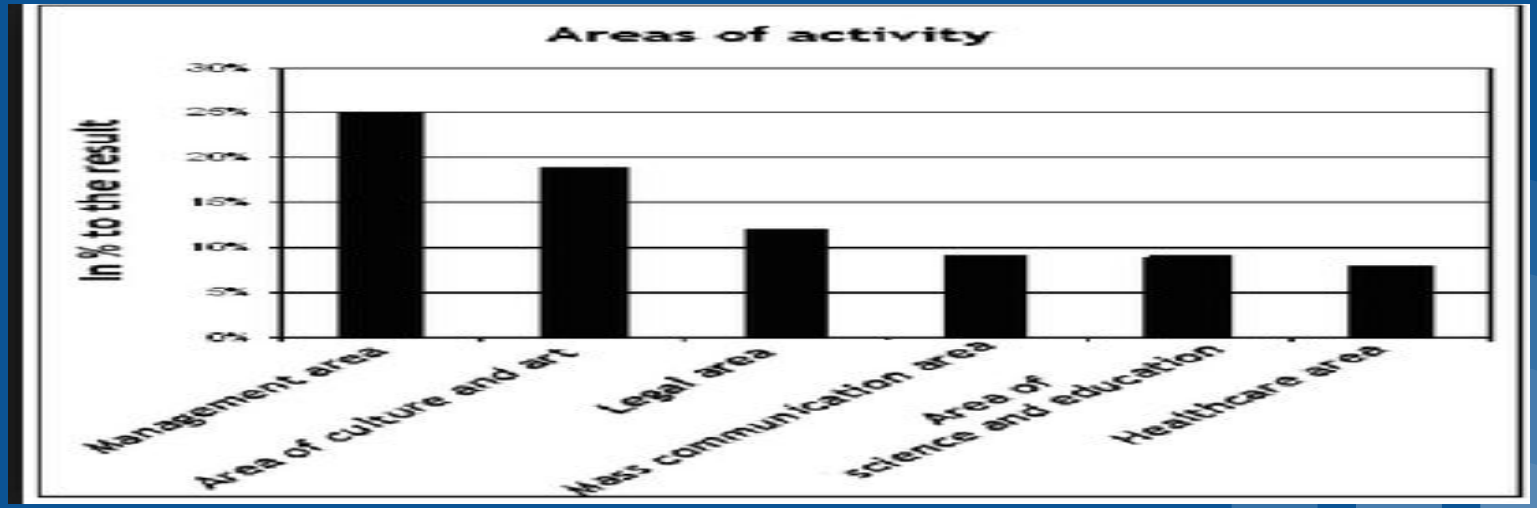
A - MALE/FEMALE
B - ELEMENTS OF FEATURE SET.

```
iiita@iiita-HCL-Desktop:~/project$ python gender.py
Please enter a name: Patrick
M
iiita@iiita-HCL-Desktop:~/project$ python gender.py
Please enter a name: Mary
F
iiita@iiita-HCL-Desktop:~/project$ python gender.py
Please enter a name: Robert
M
```

GOALS FOR FUTURE

- Gender prediction using tweets and description.
- Interest set prediction
- Occupation prediction

Above three attributes will be predicted using the composition of TF/IDF (Term Frequency/Inverse Document Frequency) and N-gram.



THANK YOU !