CSE648: Privacy and Security in Online Social Media

Mid-Sem Exam (165 Marks) Deadline:23:59hrs March 6, 2021

Instructions:

- Use Jupyter Notebooks / Google Colab for doing this assignment and showing analysis, graphs and code. Name the file as **name_rollnumber_section.ipynb.**
- Write all inferences and include the plots in a separate file named **name_rollnumber.pdf**.
- Please cite any sources that you might use in the process.
- Please write your own code. All code will be tested for plagiarism and if found, institute policy will be followed.
- *Attempt the mid-sem exam individually.*
- Submit your code and report as individual files. **Do not** zip your submissions.

Section - I (Reddit + NLP/ML) [65 Marks]

For this section, make use of the pickled reddit comments dataset from two categories of subreddits (humour and news) given in the '.zip' file. It can be read using the pandas library as follows:

```
reddit_df = pandas.read_pickle('redditDataset.pkl')
```

Q1. Across the two categories of subreddits, preprocess the text corpus and analyse it using the following methods:

a. Use a sentiment analysis library such as textblob/VADER to extract the polarity and subjectivity of all comments. Plot distribution of polarity and subjectivity for both categories (x-axis: polarity/subjectivity, y-axis: # of comments). Also report their mean and standard deviation. What do you observe?
b. Plot distributions of the top 10 most occurring (i) unigrams, (ii) bi-grams across the two categories. What are your inferences?

Deliverables: {a. 4 Plots, 4 Statistics, Observations, b. 4 Plots, Inferences}

[2.5*4 + 2 + 3 + 2.5*4 + 5 = 30 Marks]

Q2. In this question, you are required to build a machine learning classifier to classify the comments into the two categories **"news"** and **"humour".**

a. Split the data into train and test sets in a 70-30 ratio. Preprocess the data and represent the comments as vectors using a bag of words approach. Describe your steps and elaborate on your rationale for each one of them.
b. Using sklearn, train a machine learning model of your choice to classify the comments as **"news"** and **"humour".** Pickle it to ensure the reproducibility of results.

c. Evaluate your model on the test-set and report its accuracy, precision, recall and F1-score. Plot a confusion matrix to demonstrate your model's performance.
d. Enumerate 5 examples of comments where your model performed well and 5 examples of comments where it failed. Give reasons for the same.

Deliverables: {a. Preprocessing + Justifications, b. Model, c. Evaluation, Confusion Matrix, d. Examples, Reasons}

[10 + 5 + 5 + 5 + 10 = 35 Marks]

Section - II (Location-Based Social Networks) [50 Marks]

Q1. Brightkite was once a location-based social networking service provider where users shared their locations by checking-in. On this link, you can find a dataset of the social network which contains both edge data and check-in data. Use the **check-in data** for the following questions:

a. Plot the distribution of (i) locations, (ii) users in terms of number of checkins (x axis: locations/users, y axis: # of check-ins). What can you infer about the network?
b. Using the geospatial information in the data, plot the check-in locations of users on a geographic heatmap. What do you observe?
c. Make interactive time series plots for
    i. The user with the most number of check-ins
    ii. The location with the most number of check-ins

    Comment on how the temporal information can be used to target users. Do you see periodicity?

Deliverables: {a. 2 Plots, Inferences, b. Heatmap, Observations, c. Time Series Plots, Inferences}

[2.5*2 + 5 + 10 + 5 + 5*2 + 5 = 40 Marks]

Q2. Given that this dataset has been collected using a public API in the past, how can you leverage the information in this dataset to target users on the platform? List down some of the challenges in terms of privacy and security in the context of location-based social networks like Brightkite?

[5 + 5 = 10 Marks]

Section - III (Identity Resolution) [35 Marks]

Q1. Identity Resolution in Online Social Media is a major problem because it allows the aggregation of data on users across multiple platforms. Given in the .zip file is another ground truth dataset of 324 users with their usernames across three different social networks, in the following format:

Name	Twitter Username	Facebook Username	Instagram Username

Use the dataset to answer the following questions:

a. Use two commonly used distance metrics to study whether their usernames across different pairs of platforms depict some similarity. What is the best metric to use for
    i. Twitter-Facebook Identity Resolution
    ii. Facebook-Instagram Identity Resolution
    iii. Twitter-Instagram Identity Resolution
b. Plot the distribution for both the metrics across all three pairs of platforms from (a). What can you infer?

Deliverables: {a. Distance Metrics + Best Metrics, b. 6 Plots + Inferences}

[3*5 + 6*2.5 + 5 = 35 Marks]

Section - IV (Privacy Principles) [15 Marks]

Q1. We saw in class on ways to predict the SSN number in the US, if you were to take up a project / research / work for a company that wants to predict aadhaar number in India, what are the steps that you can try? Explain if it is possible or not possible? If you can, how and if you can't, why not?

[5 Marks]

Q2. From the industry point of view, implementing the “Use limitation” principle is very hard; why is it so? Give at least 5 reasons for the same. Please keep the answers short and to the point.

[5 Marks]

Q3. There is a general notion that Large social network organisations like Facebook, Twitter, use our information against us, like shown in the Social Dilemma. If you were working for one of these organizations and you had to defend that point, and argue that companies are doing this for users’ benefit. List down 5 different ways by which you will argue for this? Please keep the answers short, to the point and from a privacy / security standpoint.

[5 Marks]
