CSE648: Privacy and Security in Online Social Media

Homework - 1 (95 Marks) Deadline: Feb 3, 2021

Instructions:

- Jupyter Notebooks / Google Colab are strongly recommended for doing this assignment and showing analysis, graphs and code.
- 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.
- *Do this assignment individually.*
- Submit your code as individual files. **Do not** zip your submissions.
- The Parler dataset given for this homework **must not be shared** publicly or used beyond the scope of this course.

Since the Parler platform and its API are no longer accessible, please refer to the following tables of column details to interpret the datasets:

Important Information about Parler Posts Dataset:

Impressions	# of Impressions on the post.
Id	Id of the post.
Upvotes	# of Upvotes on the post.
At	Set of users mentioned in the post.
Comments	# of Comments on the post.
Reposts	# of Reposts of the post.
CreatedAt	Time of creation of post in 'YYYYMMDDHHMMSS' format
Body	Text within the post.
Creator	Id of the User who created the post.

Important Information about Parler Users Dataset:

Name	Name as used by the user.
Username	Username of the user on Parler.

Score	Sum of upvotes of the user.
Id	User Id.
Bio	Bio of the user.
Joined	Date on which the user account was created (in 'YYYYMMDDHHMMSS' format).
Interactions	# of interactions of the user.
Human	Whether the user has been verified by Parler to be a human.
Verified	Whether the user is a prominent personality and has been verified by Parler.

Q1. Collect the top 10 usernames responsible for generating the most content.

a) Create a bar plot denoting their individual percentages. What percentage of the total content is generated by the top 10 users cumulatively? What inference can you draw from the plot?
b) How many of these users are verified/human? Highlight them in the plot generated and draw inferences.

Deliverables: {Plot, Inference} **[10 + 5 = 15 Marks]**

Q2. Identify the top 10 usernames with i) most number of upvotes, ii) most number of interactions and (iii) most number of mentions.

a) Represent (i), (ii) and (iii) on a bar plot. How many of them are verified/human on Parler? Highlight them on the plots generated.
b) For (ii), make a word cloud from the text in the bios of users. Can you comment on their demographics?

Deliverables: {3 Bar Plots, World Cloud, Inference} **[5x3 + 5 + 5 = 25 Marks]**

Q3. Identify the major topics being discussed in the content (posts) by extracting words and hashtags from the content.

a) Make a word cloud to qualitatively describe the data. Also plot the top 10 words based on number of occurrences. What do you observe?
b) Plot the top 10 hashtags based on the following and draw inferences:
    i) Number of occurrences.
    ii) Average post length (number of words).

Deliverables: {Word Cloud, Plot of Top 10 Words, 2 Plots of Top 10 Hashtags, Inferences}

[5 + 5 + 5x2 + 5 = 25 Marks]

Q4. Draw time-series graphs using an interactive visualization library (for example - plotly) to study granular patterns of the following:

a) Content generation on Parler.
b) User account creation on Parler.

What do you observe?

Deliverables: {Interactive Time Series Plots, Inferences} **[10 + 10 + 5x2 = 30 Marks]**