CSCI 4342 Natural Language Processing
Semester 2 2022/23

Assignment 2: Text Analytics (Group Submission)
Due: Sunday, 04/07/2023 (11.59 pm)

CHOOSE ONE OF THE FOLLOWING:

1) Text Analytics on Twitter Data

i. Register for a **new twitter account** or use any of the group member's **existing twitter account**. Make sure you have **"pip install tweepy"** or other similar packages.
ii. Apply to be a **twitter developer** to request for **twitter developer API** (one member is enough). Complete the application form and wait for your tokens to be delivered (refer slide #26-30). **Application may take 1-3 days to be approved, apply early*.
iii. Choose a suitable keyword related to the growing cases of **child abuse** or **scam** in Malaysia (pick one only) and use it as English hashtags or keywords to filter the tweets. Be creative and look for the most commonly used terms in social media (i.e., #childabuse, #kidsabuse, "child abuse", #scammer, "scam", etc.).
iv. Use your tokens to extract **live tweets** (refer slide #32-34). Choose 'Malaysia' or 'my' as the location to extract the tweets. Limit your data to 100 tweets maximum.
v. Use the uploaded **data files in italeem** containing the list of **positive, negative, neutral and booster words** to check if each tweet contains any of those terms. Assign and accumulate scores according to Slide #35 for each occurrence of the terms. You may also define your own scoring method that make sense/applicable (e.g., scores between the range of 0 – 1.0 as discussed in class).

2) Text Analytics on Telegram Data

i. Register for a **new telegram account** or use an **existing telegram account** of one of your group members. Make sure you have **"pip install telethon"** or other similar packages.
ii. Apply for **telegram API** (one member is enough). Complete the application form and wait for your approval. **Application may take a few days to be approved, apply early.*
iii. Refer to the links given in Slide #44
iv. Try to implement the **"chat-analytics"** or **"Telegrammetry"** project. Find groups that discusses about the growing cases of **child abuse** or **scam** (pick one only), or anything related, etc in Malaysia and get your telegram data. Limit your data to 100 messages at maximum.
v. Use the uploaded **data files in italeem** containing the list of **positive, negative, neutral and booster words** to check if each tweet contains any of those terms. Assign and accumulate scores according to Slide #35 for each occurrence of the terms. You may also define your own scoring method that make sense/applicable (e.g., scores between the range of 0 – 1.0 as discussed in class).