LAB ASSIGNMENT V

ADVANCED PREDICTIVE ANALYTICS (MDI3003)

CLASS ID - VL2023240104377

MAXIMUM MARKS: 10 **DUE DATE**: 13 OCTOBER 2023

Consider the following **sentiment140** dataset. It contains 1,600,000 tweets extracted using the twitter api. The tweets have been annotated (0 = negative, 4 = positive) and they can be used to detect sentiment. It contains the following 6 fields:

1. **target**: the polarity of the tweet (0 = negative, 4 = positive)
2. **ids**: The id of the tweet (2087)
3. **date**: the date of the tweet (Sat May 16 23:58:44 UTC 2009)
4. **flag**: The query (lyx). If there is no query, then this value is NO_QUERY.
5. **user**: the user that tweeted (robotickilldozr)
6. **text**: the text of the tweet (Lyx is cool)

https://drive.google.com/file/d/1pTIFG44Jjcq7MUBU51HJCEaAzlPgXo_j/view?usp=drive_link

Use Jupyter Notebook and perform the following tasks –

1. Convert the **text** field into lower case.
2. Remove punctuations from **text**.
3. Remove Stopwords from **text**.
4. Remove the words with frequency 1 from **text**.
5. Perform Stemming on **text** using PorterStemmer.
6. Perform Lemmatization on **text** using WordNetLemmatizer.
7. Plot WordCloud for negative tweets.
8. Plot WordCloud for positive tweets.
9. Generate a histogram for 50 most common words in negative tweets.
10. Generate a histogram for 50 most common words in positive tweets.