CSC 423: Special Topics in IT

CAT 1: COLLECTING AND PREPROCESSING TEXT DATA FROM SOCIAL MEDIA

Due: 24/2/2023 by 21:00 hrs.

You have been hired (by a company of your choice) to write a Jupyter Notebook for preprocessing text data from user comments on the company's website, blogs, and social media.

1. Identify a real company of your choice that has social media presence. Document the URLs of these sites.
2. Scrap all the user-generated comments and data from their weblogs and social media platforms Store the data in a .csv file.
3. Create a Jupyter notebook and Import necessary libraries
4. Load your social media data
5. Clean the text data
   a. Remove URLs
   b. Remove special characters
   c. Convert text to lowercase
   d. Tokenize the text data
   e. Remove stopwords
6. Stem and Display a sample of the stemmed data
7. Lemmatize and display a sample of the lemmatized data in step 6
8. Use word cloud to visualize the data in step 7
9. Save the cleaned data to a new CSV file
10. Upload your notebook on your enaz portal.