Task 1

Data Analytics

Description

Apply Text analysis Lifecycle on spam dataset to detect if the email is spam or ham.

The dataset contains 5572 rows × 2 columns (v1, v2), use this dataset to build a prediction model that will accurately classify which texts are spam.

- Apply the most appropriate preprocessing steps (Tokenization, stemming, lemmatization, etc.)
- Apply Feature generation (Bag of words, word embedding)
- Apply Feature Extraction
- Apply the model (select the classifier that is suitable for this data)
- Evaluate the selected model (Accuracy, F1-score, Precision, Recall)
- Use Python and needed libraries like(nltk, pandas, sklearn)

The Dataset file is attached to this file.

Teams:

- Each team consists of 3-6 members.

Deadline& Delivery:

- The task delivery in the lab started on the 2nd of April.
- Only one member can deliver and discuss the task with the assigned TA.