Birla Institute of Technology & Science, Pilani

Work Integrated Learning Programmes Division

Second Semester 2024

Course: AIML-CZG-567 (AI & ML Techniques for Cyber Security)
Assignment-01: Text Summarization using Python & NLTK: TF-IDF Algorithm

A. Problem Statement

Build. A summary of long pieces of text keeping key information content and overall meaning. The summary must represent the most important or relevant information within the Text.

B. Process Steps

Steps involved to create the text summary

- Tokenize sentences
- Create frequency matrix of words in each sentence
- Calculate Term Frequency and Generate matrix
- Create a table for documents per words
- Calculate IDF and generate matrix
- Calculate TF-IDF and generate matrix
- Score the sentences
- Find the threshold
- Generate the summary

C. Perquisites

- Python 3
- NLTK Toolkit
- IDE or Text Editor

D. Submission Instructions

A PDF document has to be uploaded on Canvas under 'Assignment' covering following:

- Overall process description & solution approach
- Tool used and reasons to use this specific tool
- Source code snippets
- Final output results and analysis of results

Note: Each document page should have student's BITS Id.

E. References

Refer following for detailed steps and examples of text summarization case studies.

https://towardsdatascience.com/text-summarization-using-tf-idf-e64a0644ace3?gi=ebc5e81b9984

https://medium.com/@ashins1997/text-summarization-f2542bc6a167

F. Evaluation Criteria

The assignment is for 10 marks. Following evaluation scheme will be used to grade the assignments:

S.No.	Evaluation Task	Marks
1	Overall solution design and process architecture	3
2	Tool used and reasons to use this specific tool	2
3	Final output results and analysis of results	3
4	Document quality (structure, detailing, presentation etc)	2