Logo

Description automatically generated **FACULTY OF COMPUTER SCIENCE AND ENGINEERING**

**Time: 1 hours AI 361 Lab Marks: 80**

**Instructor: Muhammad Muneeb Baig Lab #05**

**Task 1:**

You have been given a collection of 100 text documents on various topics. Your task is to find 10 documents that are related to each other based on their content, and to compare the performance of two text similarity measures: minimum edit distance and TF-IDF.

**Part 1: Finding related documents.**

* Choose any topic of your choice and find 10 documents from the collection that are related to that topic.
* Explain your process of selecting the documents and the criteria you used to determine their relevance to the chosen topic.
* Provide a summary of each of the selected documents and describe how they are related to each other.

**Part 2: Implementing minimum edit distance.**

* Implement the minimum edit distance algorithm from scratch in Python.
* Use your implementation to compute the similarity between all pairs of the 10 selected documents.
* Discuss the limitations of the minimum edit distance algorithm in measuring text similarity.
* Compare the results of your implementation with the results obtained using the text similarity function provided by the NLTK library.

**Part 3: Implementing TF-IDF**

* Implement the TF-IDF algorithm from scratch in Python.
* Use your implementation to compute the similarity between all pairs of the 10 selected documents.
* Discuss the advantages and limitations of the TF-IDF algorithm in measuring text similarity.
* Compare the results of your implementation with the results obtained using the TF-IDF function provided by the scikit-learn library.

**Part 4: Conclusion**

* Summarize your findings and discuss which algorithm you think is more appropriate for measuring text similarity, based on your analysis. Also Implement the graphs of your results so that we can have a better idea to understand the difference or similarity between starch and built-in module
* Discuss possible extensions to your work and how they could improve the accuracy of text similarity measures.