AI414: Processing of Formal and Natural Languages
Dr. Asmaa Ahmed
TA. Mohab Wahdan

Cairo University, Faculty of Computers
and Artificial intelligence

Assignment 1: Preprocessing Data

Description:

Scrap data(text) from any two different websites then apply preprocessing techniques to it after that get all the unique values.

The minimum preprocessing techniques required:

1. Tokenization
- Splitting text into words, sentences, or subwords.
- Example: "I love NLP" → ["I", "love", "NLP"]

2. Lowercasing
- Converting all text to lowercase to ensure uniformity.
- Example: "Machine Learning" → "machine learning"

3. Stopword Removal
- Removing common words like "the," "is," "and" that do not add much meaning.
- Example: "I love the new AI model" → ["love", "new", "AI", "model"]

4. Removing Special Characters, Numbers and Punctuation
- Example: "Hello!!! NLP is awesome :) " → "Hello NLP is awesome"
- Example: "COVID-19 cases reached 500000" → "COVID cases reached"

Deleviry:

- The assignment will be done in **teams of three**.
- The assignment discussion will be held next Sunday 9/3 at lab time (the grades will be on discussion only there is no submission).