Pre-requisite programming experience for Data Sciences 774 and 874

Programming experience at 1st year university level is a pre-requisite for the module Data Sciences 774 and 874

Practical knowledge of a programming language forms part of the prior knowledge required for the module. There is however no specific prescribed programming language for the course. Students are free to use a suitable high-level programming language of their choice for the practicals and assignments. The most common data science languages are Python and R. The lecturer will assume all students are comfortable in at least one high-level programming language (e.g. Java, C++, Python etc) – students should be able to read/write from/to a file, write functions, display output, use libraries, and use an IDE for example. It is very difficult to get meaning productivity in data science without practical implementation of tasks. Competence in one language generally makes it easy to comprehend code written in another language and learn it.

Example exercise:

Programming Practice Question

A data mining expert wants to create a sentence retrieval system based on the Bag of Words (BoW) framework. This framework allows each sentence to be represented as a vector indicating the frequency of words from a fixed vocabulary. For example, consider the following sentences:

1. "Data mining finds patterns in data."
2. "Data mining uses data for patterns."
3. "Patterns in data are important."

The BoW representation of these sentences can be created by following the steps below.

1. Vocabulary Construction

The vocabulary consists of all unique words, sorted alphabetically:
["are", "data", "finds", "for", "important", "in", "mining", "patterns", "uses"]

2. Bag of Words Matrix Representation

Each sentence is transformed into a word-frequency matrix, where rows represent words from the vocabulary, and columns represent sentences. The (i, j) entry represents how many times word i appears in sentence j.

Word	Sentence 1	Sentence 2	Sentence 3
are	0	0	1
data	2	2	1
finds	1	0	0
for	0	1	0
important	0	0	1
in	1	0	1
mining	1	1	0
patterns	1	1	1
uses	0	1	0

Write a program in your preferred programming language that:

1. Reads a text file containing multiple sentences.
2. Tokenizes the sentences into words, removing punctuation and converting letters to lowercase.
3. Constructs a fixed vocabulary from unique words across all sentences.
4. Computes a Bag of Words matrix, where rows represent words and columns represent sentences.
5. Outputs the BoW matrix in tabular format.

Your program should observe the following constraints:

- Treat words in a case-insensitive manner.
- Remove punctuation such that words with hyphens are split into separate words (e.g., "machine-learning" is split into "machine" and "learning").
- Ensure the vocabulary is sorted alphabetically for consistency in row ordering.