

Information Retrieval (CSD358)

Assignment 1

Instructions

- There are **2** questions in this assignment.
- Upload a zipfile containing your code, output screenshots and corpus.
- Properly **document/comment** your code.
- You can do the assignment in group of 2 students.

Due Date and Time: 10:00 pm, September 11, 2024.

Submitting this Assignment

You will submit (upload) this assignment in Blackboard. **Email submissions will not be accepted.**

Grading Criteria

This assignment has 8 marks (with weightage of 8% in your overall 100 points).

1. Write the program to construct an Inverted Index for a given document collection comprising of at least 40 documents with a total vocabulary size of at least 500 words. The program should take the input as boolean search queries using AND, OR and NOT operators and return the list of the documents satisfying the query need. While generating the tokens do case folding, normalization, stop word removal and lemmatization/stemming. (4)
2. Extend the Boolean retrieval model implemented in part (1) to Extended Boolean Retrieval model with following features:
 - a) Phrase Queries search using Biword index
 - b) Proximity queries search by adding positional index
 - c) Use Soundex algorithm for spelling matching of names (4)