CMSC 401 – Fall 2016

Assignment 3
(due Tuesday 11/15 – 11:59pm)

Dr. Eyuphan Bulut

CMSC 401- Algorithm Analysis with Advanced Data Structures

VCU School of Engineering | Computer Science

Word Mining

- Assume that you are given a huge file consisting of billions of emails
- You are asked to point some of the words in this file by providing the whole sentence that contains these words
- Assume that you already found out the start and end positions of the sentences
  - A sorted array S[] of numbers indicating positions from the start of the text
  - First letter is at index 1
- Assume that you also parsed the collection and found the positions of the first letter of the words
  - A sorted array of W[] of numbers that contain the positions

Assignment 3

- Your task is, given S[] and W[],
  - Find starting and ending position of each sentence that contains the word of interest
  - Design a quick algorithm using one of the data structures you learned. Assume size N of S[] is much greater than size M of W[].
    - You should not just go through S[] in a loop to find the sentences, as it will be O(N). It will be slow for huge data even it is linear.
    - Hint: A balanced Binary Search Tree can do searches in O(log(N)). But
      - how to insert data to make the tree balanced?
      - how to modify search to get the answer after tree is constructed?

Assignment 3

Write a program cmsc401.java that

- takes as input
  - Line 1: single integer N>=1, size of array S[]
  - Line 2: single integer M>=1, size of array W[]
  - Lines 3..(N+2): N positive integers in array S[] indicating the ends of sentences
  - Lines (N+3)…(N+M+2): M positive integers in array W[], positions of the words searched
- returns as output
  - Line 1: number of sentences that contain the words searched
  - Following lines: two integers: i) start and end of the sentence that contains the word
    - end: (previous sentence end)+1

-Input:

```
3
2
25
33
74
19
40
```

Output:

```
2
1  25
34 74
```

The example corresponds to the word "output" and the text:

No human-oriented output. Please. Such output creates problems in grading.

Submission

- **Date due:** Tuesday, Nov 15th, 11:59 pm
- Upload through Blackboard
  - Your submission should be a zip archive 3_FamilyName_FirstName.zip containing
    - Java source code in a file cmsc401.java (**all low case letters!**)
      - The file should have your name in a comment in the first line
      - If you use multiple files, the main file must be cmsc401.java
      - Remember: in Java, class name should match the file name, and is case sensitive
- Please do NOT create your own packages
- Do NOT place the file into a folder – just zip the file
- Use standard I/O to read input (System.in, System.out) and output
- Make sure the program compiles
- Do not use any library for BST, create your own.