Genome Regions – Coding Project

Background

Consider we have a number of regions, each region corresponding to a portion of the human genome. Each region is specified by a start and end position, both of which are positive integers.

We have provided two text files, Regions_Small.txt and Regions_Large.txt, which give example data.

Each line of the files specifies a region and contains a pair of integers separated by a tab. These are the start and end coordinates of each region. Please use these region files in the following questions.

Part 1

We would like to draw the regions in the following style where overlapping regions stack up. An example is shown in the diagram below where the y-axis represents the row number, and x-axis the position on the genome.

To do this, each region must be assigned to a drawing row. Because on-screen real estate is limited, we want to draw the stack of regions using the smallest number of rows.

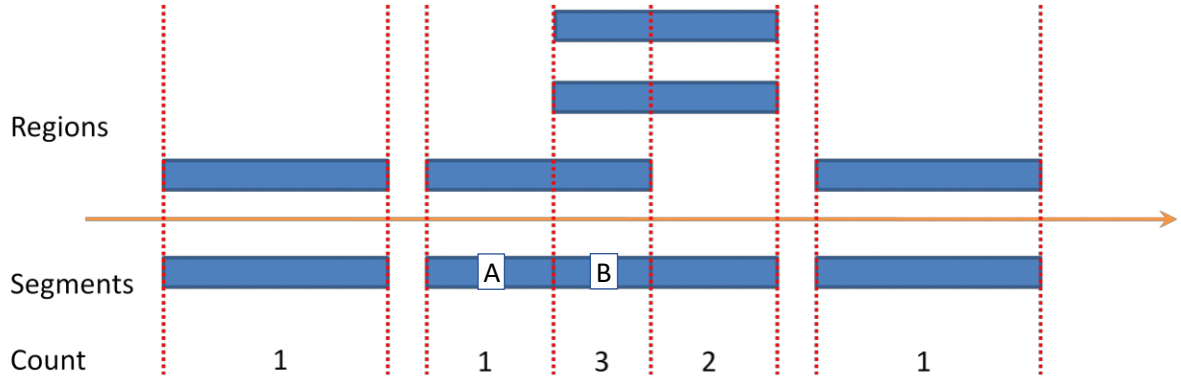
Please write code that reads a regions file and produces as output a tab-delimited text file. The output file should contain the same number of rows as the input file. On each line of the output file there should be an integer, which indicates the drawing row to which the corresponding region in the input file was assigned, followed by the region start and end coordinates e.g.

```
1	28777761	28778321
```

Part 2

When there are many overlapping regions, the stack of regions can become too deep to display in limited space. Instead, we can summarise the data and display it in the style of a histogram.

To summarise the regions data, the regions must be split into non-overlapping segments. Each segment has associated with it a count of the number of regions that overlap that segment. This idea is illustrated below.



In the above example, if segment B has start position X, then segment A has end position X-1. The segments do not overlap.

Please write code that reads a regions file and produces a tab-delimited text file. The output file should contain one line per segment, with each line containing three numbers: start coordinate, end coordinate and count.

Instructions

The choice of programming language is yours, but please provide any instructions for building and running the code, and an output file for each input file. Please also provide a short description of your algorithm and how you developed it.

Please report the total time you spent on the project. It should not be more than 1 day.

If anything is unclear, simply state your assumptions in the submission.

Note: You must not re-distribute this document.