Assignment File Finder and Merge Sort

Object Orientation

Spring 2021

1 Goals

This exercise consists of two parts. Both ask you to write multi-threaded programs. Use the approach for designing concurrent programs outlined in the lectures. After this exercise you should be able to:

- Create threads to execute parts of a computation.
- Synchronize threads after termination and combine their result.
- Speed-up a Java program by using multiple cores.

2 Find File

The class `java.io.File` represents files and directories on a computer. We will use this class to locate a file with a given name somewhere in a directory tree. The program below prints the full path of all files with a given name. `FileFinder` is a class that executes this search recursively. The constructor sets the initial directory if it exists and throws an exception otherwise. The recursive search by the private method `find` is initiated by a call to the public method `findFile` of a `FileFinder` object.

```
public class FileFinder {
    private final File rootDir;

    public FileFinder(String root) throws IOException {
        rootDir = new File(root);
        if (! (rootDir.exists() && rootDir.isDirectory())) {
            throw new IOException(root+" is not a directory.");
        }
    }

    public void findFile(String file) {
        find(rootDir, file);
    }

    private void find (File currentDir, String fileName) {
        File [] files = currentDir.listFiles();
        if (files != null) {
            for (File file: files) {
                if (file.getName().equals(fileName)) {
                    System.out.println("Found at: " + file.
                        getAbsolutePath());
                } else if (file.isDirectory()) {
                    find(file, fileName);
                }
            }
        }
    }
}
```

The class `Main` searches for the file named “needle.txt” in the directory “haystack”.

```
public class Main {
    public static void main(String[] args) {
        try {
            String goal = "needle.txt";
            String root = "haystack";
            FileFinder ff = new FileFinder(root);
            ff.findFile(goal);
        } catch (IOException e) {
            e.printStackTrace();
        }
    }
}
```

Since there can be many directories and files, such a search can take a lot of time. Make a parallel version of the file finder to speed up the search.

3 Merge Sort

Merge sort is sorting algorithm with time complexity $O(N \log N)$. In this algorithm the array to be sorted is split into two parts of equal size. These parts are sorted recursively. Next, these sorted arrays are merged into one sorted array. To make this final step easy, the sub-arrays in the first step of the algorithm are copies of the original array. Since the arrays to be merged are sorted, merging them into a single sorted array is easy. In contrast to the sorting algorithms discussed in the courses *Programming for AI* and *imperatief programmeren*, this implementation of merge sort is not *in place*; it does not run in constant space. The arrays `firstHalf` and `secondHalf` do require additional temporary memory. The class `MergeSort` provides an implementation of this algorithm.

```
public class MergeSort {
  /**
   * sort the given array in O(N log N) time
   * The array is split into two parts of equal size.
   * These parts are sorted recursively and merged.
   * @param array
   */
  public static void sort(int[] array) {
    if (array.length > 1) {
      int[] firstHalf  = Arrays.copyOf(array, array.length / 2)
          ;
   int[] secondHalf = Arrays.copyOfRange(array, array.length
        / 2, array.length);
   sort(firstHalf);
   sort(secondHalf);
   merge(firstHalf, secondHalf, array);
  }
 }

 /**
  * merge two sorted arrays: time O(N)
  * @param part1 a sorted array
  * @param part2 a sorted array
  * @param dest  destination, length must be >= part1.length +
     part2.length
  */
 public static void merge(int[] part1, int[] part2, int[] dest
    ) {
  int part1Index = 0, part2Index = 0, destIndex = 0;
  while (part1Index < part1.length && part2Index < part2.
     length) {
   if (part1[part1Index] < part2[part2Index])
    dest[destIndex ++] = part1[part1Index ++];
   else
    dest[destIndex ++] = part2[part2Index ++];
  }
  // copy elements when at most one of the parts contains
     elements
  while (part1Index < part1.length)
   dest[destIndex ++] = part1[part1Index ++];
  while (part2Index < part2.length)
   dest[destIndex ++] = part2[part2Index ++];
 }
}
```

This algorithm is very suited for parallelization. The arrays `firstHalf` and `secondHalf` have the same size, do not overlap, and can be sorted independently. Since creating a thread to sort such an array is also work, you should only create such a thread when the size of the array is bigger than some threshold value, for example 1000. When the size of the array is below the threshold, the sequential version of the sorting algorithm should be used.

This idea of parallelization of large problems and solving small problems sequentially is an instance of a general *divide and conquer pattern* for recursive problems. The general pattern looks as follows.

```
if (size < threshold) {
    sequential_solution;
} else {
    divide_in_nonoverlapping_subproblems;
    solve_subproblems_in_parallel;
    combine_results;
}
```

4 Your Tasks

1. The file haystack.zip contains a file needle.txt somewhere in a large tree of directories. Unpack it somewhere outside your project directory and use it to try the parallel version. Depending on how fast your hard drive is you might or might not notice a big difference. **Please do not submit this directory as part of your submission.**
2. Speed-up the file search by creating a new searching thread for each directory encountered. In the code above you have to replace the recursive call `find(file, fileName)` by the creation of a thread that tries to find the file `fileName` in the directory `file`.
3. This way, your program can create a huge numbers of threads. Since each thread is a Java object, it occupies some memory. With a huge number of threads the memory of your program might become exhausted. Java will generate an exception when there is not enough memory left to create a thread. You can limit the size of the search by choosing a better starting directory. Moreover, you can give your Java program more execution space to accommodate more threads. By extending the heap space of the Java virtual machine it can create more threads before it runs out of memory. You can set the maximum heap space of your Java virtual machine to 2048 MB by the option `-Xmx2048m` in the properties of your project. In NetBeans, this option can be found in the `Run` section of the project options.
4. Make a parallel version of the merge sort algorithm where sub-arrays are sorted in parallel by different threads. The class should be called `ParallelMergeSort`. You can only start merging the sub-arrays when their sorting threads are finished. Your program needs an explicit `join`.
5. In order to measure the effect of parallelization you apply the sequential and parallel version of merge sort to a long array of pseudo random integers, say 10,000,000 elements.
6. Include the timing results and the number of cores of your computer as a comment in your program. You can obtain the current time in milliseconds by `System.currentTimeMillis()`. The Java runtime method `Runtime.getRuntime().availableProcessors()` returns the number of processors or cores available on your machine. By adjusting the threshold of your program, it should be possible to obtain a speed-up of 2 on a 4-core processor.

5 Submit Your Project

To submit your project, follow these steps.

1. Use the **project export** feature of NetBeans to create a zip file of your entire project: File → Export Project → To ZIP.
2. **Submit this zip file on Brightspace**. Do not submit individual Java files. Do not submit any other archive format. Only one person in your group has to submit it. Submit your project before the deadline, which can be found on Brightspace.