654 Advanced Computing Concepts

Assignment4 Report

I confirm that I will keep the content of this assignment confidential. I confirm that I have not received any unauthorized assistance in preparing for or writing this assignment. I acknowledge that a mark of 0 may be assigned for copied work. **Tengxiaoyao (Tab) Tu, #104518447**

**Task 1.** Use classes BruteForceMatch, BoyerMoore and KMP provided in the source code.

a. Download file Hard disk.txt from the Resources.

b. Find all occurrences of patterns “hard”, “disk”, “hard disk”, “hard drive”, “hard dist” and “xltpru”, and show the offsets.

c. Repeat (b) 100 times and record the average CPU time for each case.

d. Compare the CPU times with the running times of the three algorithms (discussed in class) and comment on asymptotic running time of the corresponding algorithms.

**Answer:**



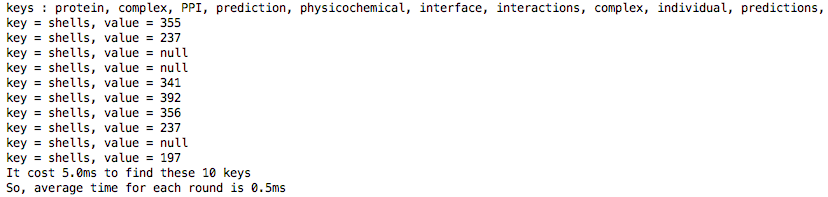
As a result, BruteForceMatch cost 0.3854ms while BoyerMoore use 0.0027ms in average in each search, and KMP cost 0,0777ms in each round.

**Task 2.** Download file Protein.txt from the Resources. Using class TST provided in the source code:

a. Write a program that reads file “Protein.txt” and creates a trie using TST. Use StringTokenizer, Jsoup or a similar API to extract the words from the file.

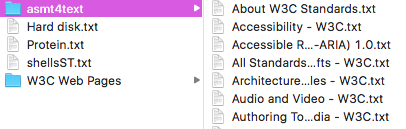
b. Do several searches of keys “protein”, “complex”, “PPI”, “prediction”, and others, and show the occurrences of these words in file Protein.txt

**Answer:**

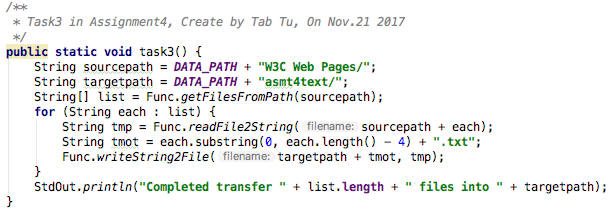


**Task 3.** HTMLtoText converter: Write a program that takes the 100 given Web pages, and using Jsoup, converts all files into text. The text files should be saved as individual files for use in the next tasks of this assignment. Keep good OO design practice by creating a method processes one file. That method will then be called 100 times.

**Answer:**

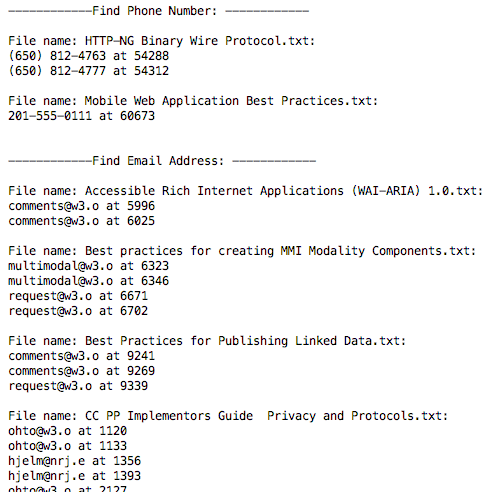






**Task 4.** Pattern finder: Using Java Regex, find phone numbers and email addresses in the 100 text files.

**Answer:**



**Task 5.** URL finder: Using Java Regex, write a program that finds Web links (URLs) in a Web file. Test your program with the 100 HTML files to find the following:

a. Links with domain w3.org

b. Links that contain folders: e.g., www.w3.org/TR/owl-features/

c. Links that contain references in a Web page and may contain folders, for example: www.w3.org/TR/owl-features/#DefiningSimpleClasses

d. Links with domain .net, .com, .org

**Answer:**

