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.

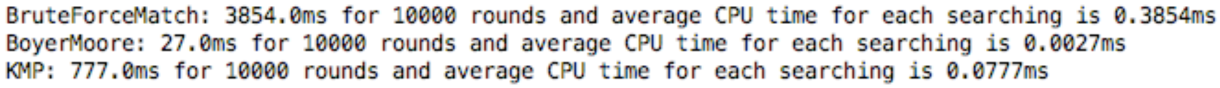
a. Download file Hard disk.txt from the Resources.

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:



```
BruteForceMatch: 3854.0ms for 10000 rounds and average CPU time for each searching is 0.3854ms
BoyerMoore: 27.0ms for 10000 rounds and average CPU time for each searching is 0.0027ms
KMP: 777.0ms for 10000 rounds and average CPU time for each searching is 0.0777ms
```

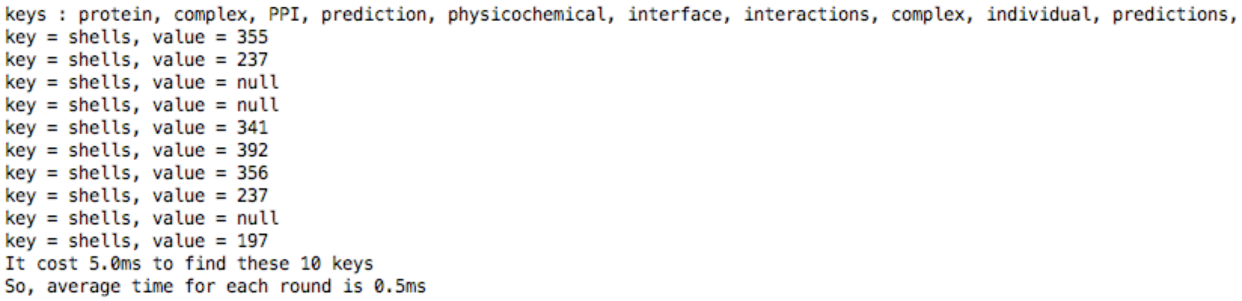
As a result, BruteForceMatch cost 0.3854ms while BoyerMoore use 0.0027ms in average in each search, and KMP cost 0,0777ms in each round.

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

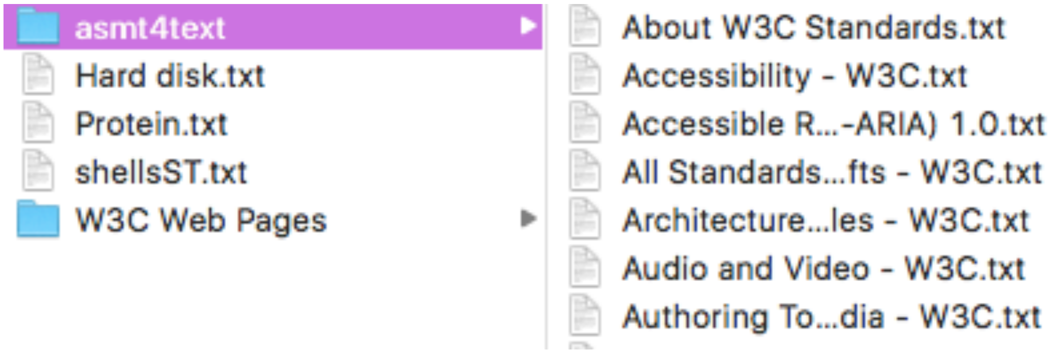
Answer:



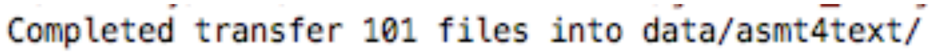
```
keys : protein, complex, PPI, prediction, physicochemical, interface, interactions, complex, individual, predictions,
key = shells, value = 355
key = shells, value = 237
key = shells, value = null
key = shells, value = null
key = shells, value = 341
key = shells, value = 392
key = shells, value = 356
key = shells, value = 237
key = shells, value = null
key = shells, value = 197
It cost 5.0ms to find these 10 keys
So, average time for each round is 0.5ms
```

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.

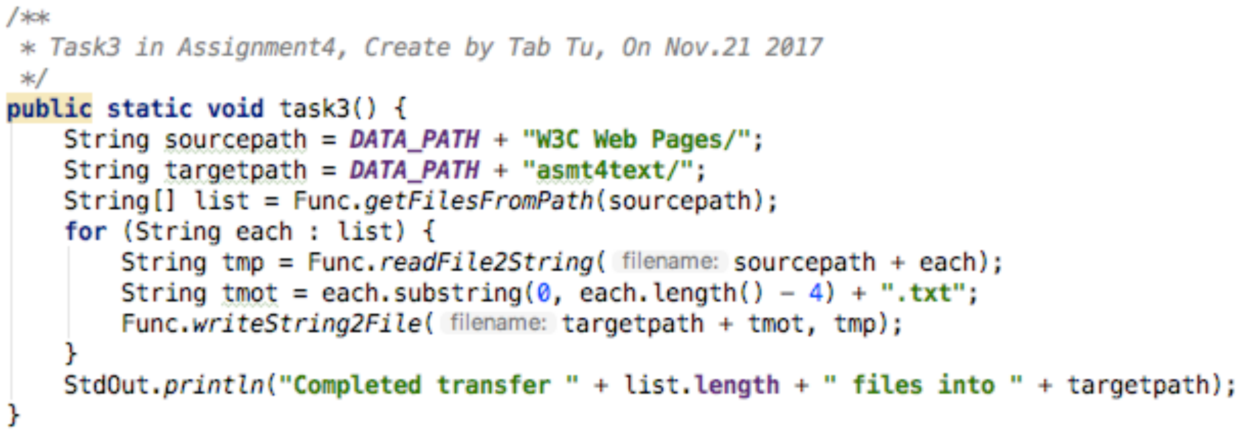
Answer:



```
asmt4text              About W3C Standards.txt
Hard disk.txt          Accessibility - W3C.txt
Protein.txt            Accessible R...-ARIA) 1.0.txt
shellsST.txt           All Standards...fts - W3C.txt
W3C Web Pages          Architecture...les - W3C.txt
                       Audio and Video - W3C.txt
                       Authoring To...dia - W3C.txt
```



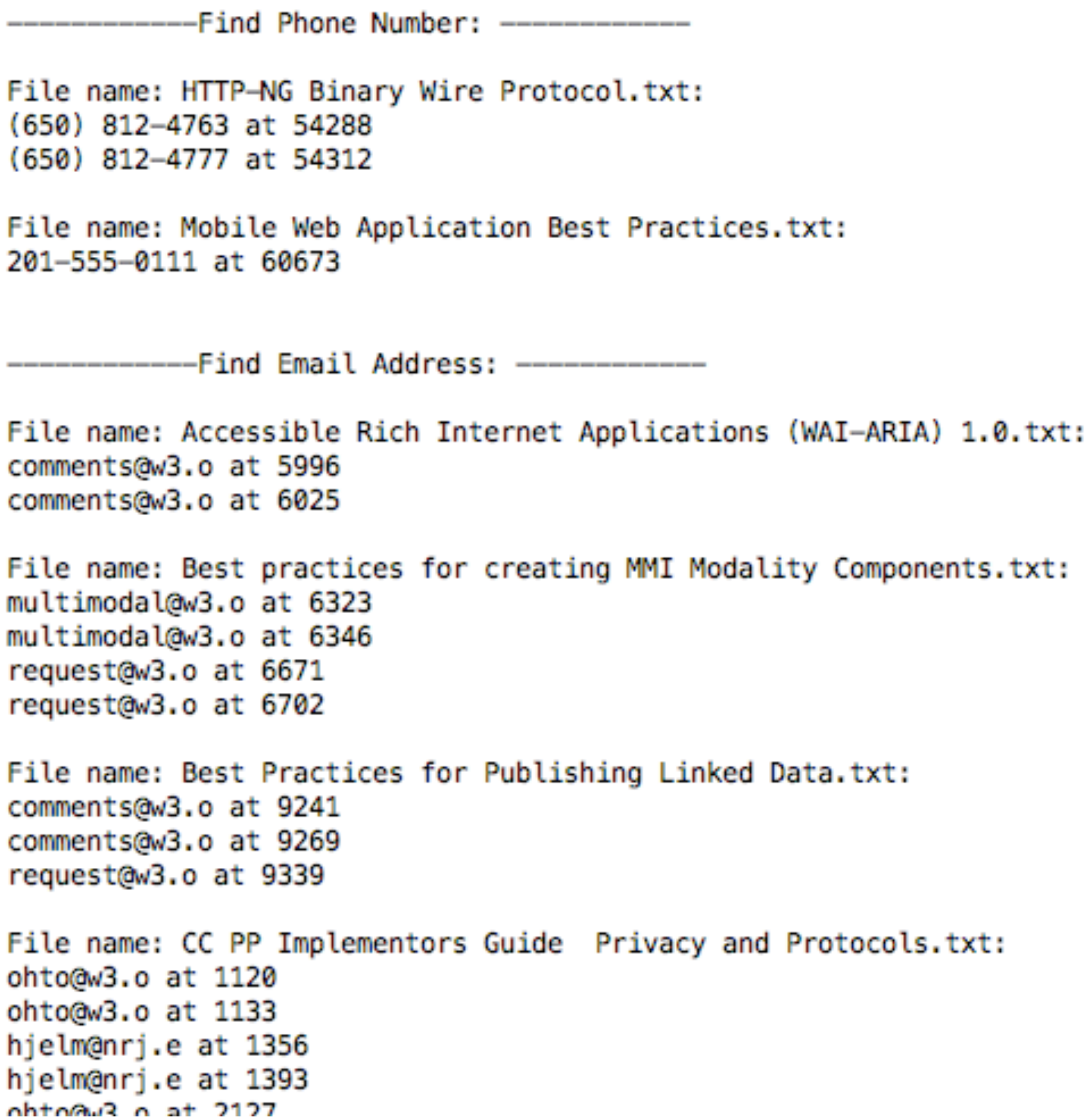
```
Completed transfer 101 files into data/asmt4text/
```



```java
/**
 * Task3 in Assignment4, Create by Tab Tu, On Nov.21 2017
 */
public static void task3() {
    String sourcepath = DATA_PATH + "W3C Web Pages/";
    String targetpath = DATA_PATH + "asmt4text/";
    String[] list = Func.getFilesFromPath(sourcepath);
    for (String each : list) {
        String tmp = Func.readFile2String( filename: sourcepath + each);
        String tmot = each.substring(0, each.length() - 4) + ".txt";
        Func.writeString2File( filename: targetpath + tmot, tmp);
    }
    StdOut.println("Completed transfer " + list.length + " files into " + targetpath);
}
```

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

Answer:



```
------------Find Phone Number: ------------

File name: HTTP-NG Binary Wire Protocol.txt:
(650) 812-4763 at 54288
(650) 812-4777 at 54312

File name: Mobile Web Application Best Practices.txt:
201-555-0111 at 60673


------------Find Email Address: ------------

File name: Accessible Rich Internet Applications (WAI-ARIA) 1.0.txt:
comments@w3.o at 5996
comments@w3.o at 6025

File name: Best practices for creating MMI Modality Components.txt:
multimodal@w3.o at 6323
multimodal@w3.o at 6346
request@w3.o at 6671
request@w3.o at 6702

File name: Best Practices for Publishing Linked Data.txt:
comments@w3.o at 9241
comments@w3.o at 9269
request@w3.o at 9339

File name: CC PP Implementors Guide  Privacy and Protocols.txt:
ohto@w3.o at 1120
ohto@w3.o at 1133
hjelm@nrj.e at 1356
hjelm@nrj.e at 1393
ohto@w3.o at 2127
```

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:

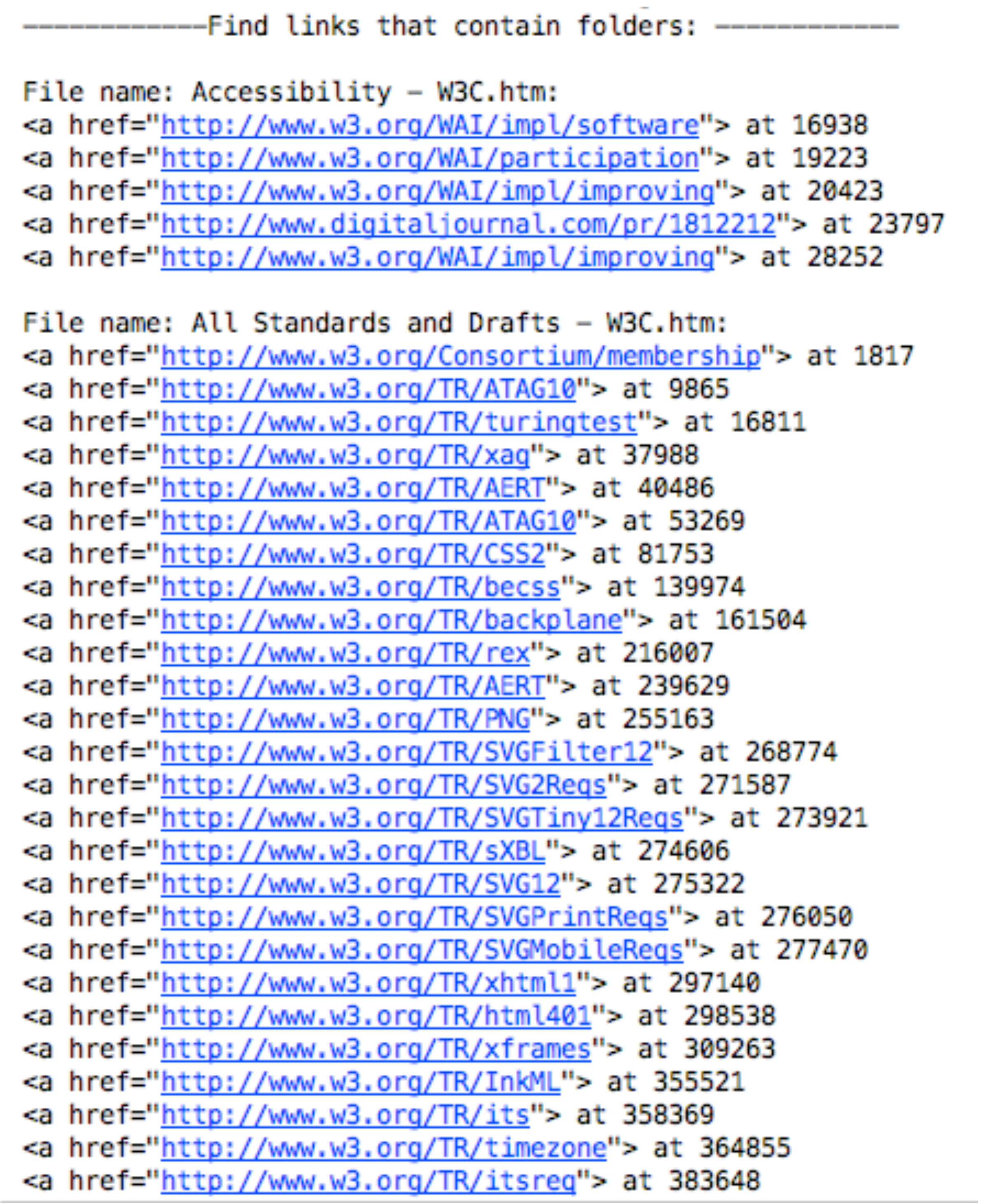
a. Links with domain w3.org

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:



```
------------Find links that contain folders: ------------

File name: Accessibility - W3C.htm:
<a href="http://www.w3.org/WAI/impl/software"> at 16938
<a href="http://www.w3.org/WAI/participation"> at 19223
<a href="http://www.w3.org/WAI/impl/improving"> at 20423
<a href="http://www.digitaljournal.com/pr/1812212"> at 23797
<a href="http://www.w3.org/WAI/impl/improving"> at 28252

File name: All Standards and Drafts - W3C.htm:
<a href="http://www.w3.org/Consortium/membership"> at 1817
<a href="http://www.w3.org/TR/ATAG10"> at 9865
<a href="http://www.w3.org/TR/turingtest"> at 16811
<a href="http://www.w3.org/TR/xag"> at 37988
<a href="http://www.w3.org/TR/AERT"> at 40486
<a href="http://www.w3.org/TR/ATAG10"> at 53269
<a href="http://www.w3.org/TR/CSS2"> at 81753
<a href="http://www.w3.org/TR/becss"> at 139974
<a href="http://www.w3.org/TR/backplane"> at 161504
<a href="http://www.w3.org/TR/rex"> at 216007
<a href="http://www.w3.org/TR/AERT"> at 239629
<a href="http://www.w3.org/TR/PNG"> at 255163
<a href="http://www.w3.org/TR/SVGFilter12"> at 268774
<a href="http://www.w3.org/TR/SVG2Reqs"> at 271587
<a href="http://www.w3.org/TR/SVGTiny12Reqs"> at 273921
<a href="http://www.w3.org/TR/sXBL"> at 274606
<a href="http://www.w3.org/TR/SVG12"> at 275322
<a href="http://www.w3.org/TR/SVGPrintReqs"> at 276050
<a href="http://www.w3.org/TR/SVGMobileReqs"> at 277470
<a href="http://www.w3.org/TR/xhtml1"> at 297140
<a href="http://www.w3.org/TR/html401"> at 298538
<a href="http://www.w3.org/TR/xframes"> at 309263
<a href="http://www.w3.org/TR/InkML"> at 355521
<a href="http://www.w3.org/TR/its"> at 358369
<a href="http://www.w3.org/TR/timezone"> at 364855
<a href="http://www.w3.org/TR/itsreq"> at 383648
```