"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." + Arun Reddy Nalla + 110088379

Tasks:

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.

Solution :

I had created Task1.java class for this solution
Location : Assignment-4\src\Assignment4\Task1.java

1) First I had printed all the position of the given words from the HardDisk.txt using BruteForceMatch given in source code
2) Then I had written program to find average CPU time taken by performing the task with BruteForceMatch 100 times.

Then I repeated the above 1 and 2 process for BoyerMoore method and KMP ALGORITHM and found the average time taken for respective methods

From the figure we can observe that hard , disk, hard disk, hard drive are found in text file but Hard dist and xltpru are not found in text file

Below are the output for the solution

```
BruteForceMatch
Pattern:hard
151 202 418 441 452 3497 3681 8688 8786 8992 11770 12913 13152 16963 17295 17755 18976 19039 22126 22557
23684 24887 27358 29608 33674 34919 35351 37345 41287 46037 48479 48628 49229 50760 53146 53761 53798 55593 55702 55726
55881 56087 56146 56905
Pattern:disk
5 134 207 423 446 472 615 3502 3686 4441 4810 4846 4895 5155 5731 5838 6153 6366 6554 6613
6727 6784 6905 6947 7051 7088 7687 8843 8997 9643 9756 10036 11775 13236 13570 13664 13675 13724 15759 15813
15916 16267 16304 16968 17300 18279 18407 18501 18627 18894 18981 19044 22131 22562 23689 25366 25499 25737 25857 25904
26123 26501 27363 29768 30690 30859 30909 31552 31712 34924 35356 35540 35656 35732 36144 36607 37350 38727 38754 38834
39253 39816 39922 40479 41292 42973 43290 44134 44301 44553 44680 45032 45952 46042 46139 46225 46615 46691 47088 47191
47290 47415 48139 48221 48484 48633 48758 48910 48964 49234 49583 49806 53766 53803 54024 55598 55653 55707 55731 56092
56151 56910
Pattern:hard disk
202 418 441 3497 3681 8992 11770 16963 17295 18976 19039 22126 22557 23684 27358 34919 35351 37345 41287 46037
48479 48628 49229 53761 53798 55593 55702 55726 56087 56146 56905
Pattern:hard drive
452 8688 8786 17755 24887 29608 33674 50760 55881
Pattern:hard dist
57122
Pattern:xltpru
57122
Average CPU time taken by Brute Force is:897605 Nano seconds
```

```
BoyerMoore
Pattern:hard
151 202 418 441 452 3497 3681 8688 8786 8992 11770 12913 13152 16963 17295 17755 18976 19039 22126 22557
23684 24887 27358 29608 33674 34919 35351 37345 41287 46037 48479 48628 49229 50760 53146 53761 53798 55593 55702 55726
55881 56087 56146 56905
Pattern:disk
5 134 207 423 446 472 615 3502 3686 4441 4810 4846 4895 5155 5731 5838 6153 6366 6554 6613
6727 6784 6905 6947 7051 7088 7687 8843 8997 9643 9756 10036 11775 13236 13570 13664 13675 13724 15759 15813
15916 16267 16304 16968 17300 18279 18407 18501 18627 18894 18981 19044 22131 22562 23689 25366 25499 25737 25857 25904
26123 26501 27363 29768 30690 30859 30909 31552 31712 34924 35356 35540 35656 35732 36144 36607 37350 38727 38754 38834
39253 39816 39922 40479 41292 42973 43290 44134 44301 44553 44680 45032 45952 46042 46139 46225 46615 46691 47088 47191
47290 47415 48139 48221 48484 48633 48758 48910 48964 49234 49583 49806 53766 53803 54024 55598 55653 55707 55731 56092
56151 56910
Pattern:hard disk
202 418 441 3497 3681 8992 11770 16963 17295 18976 19039 22126 22557 23684 27358 34919 35351 37345 41287 46037
48479 48628 49229 53761 53798 55593 55702 55726 56087 56146 56905
Pattern:hard drive
452 8688 8786 17755 24887 29608 33674 50760 55881
Pattern:hard dist
57122
Pattern:xltpru
57122
Average CPU time taken by Boyer Moore is:843571 Nano seconds
```

```
KMP
Pattern:hard
151 202 418 441 452 3497 3681 8688 8786 8992 11770 12913 13152 16963 17295 17755 18976 19039 22126 22557
23684 24887 27358 29608 33674 34919 35351 37345 41287 46037 48479 48628 49229 50760 53146 53761 53798 55593 55702 55726
55881 56087 56146 56905
Pattern:disk
5 134 207 423 446 472 615 3502 3686 4441 4810 4846 4895 5155 5731 5838 6153 6366 6554 6613
6727 6784 6905 6947 7051 7088 7687 8843 8997 9643 9756 10036 11775 13236 13570 13664 13675 13724 15759 15813
15916 16267 16304 16968 17300 18279 18407 18501 18627 18894 18981 19044 22131 22562 23689 25366 25499 25737 25857 25904
26123 26501 27363 29768 30690 30859 30909 31552 31712 34924 35356 35540 35656 35732 36144 36607 37350 38727 38754 38834
39253 39816 39922 40479 41292 42973 43290 44134 44301 44553 44680 45032 45952 46042 46139 46225 46615 46691 47088 47191
47290 47415 48139 48221 48484 48633 48758 48910 48964 49234 49583 49806 53766 53803 54024 55598 55653 55707 55731 56092
56151 56910
Pattern:hard disk
202 418 441 3497 3681 8992 11770 16963 17295 18976 19039 22126 22557 23684 27358 34919 35351 37345 41287 46037
48479 48628 49229 53761 53798 55593 55702 55726 56087 56146 56905
Pattern:hard drive
452 8688 8786 17755 24887 29608 33674 50760 55881
Pattern:hard dist
57122
Pattern:xltpru
57122
Average CPU time taken by KMP is:1871894 Nano seconds
```

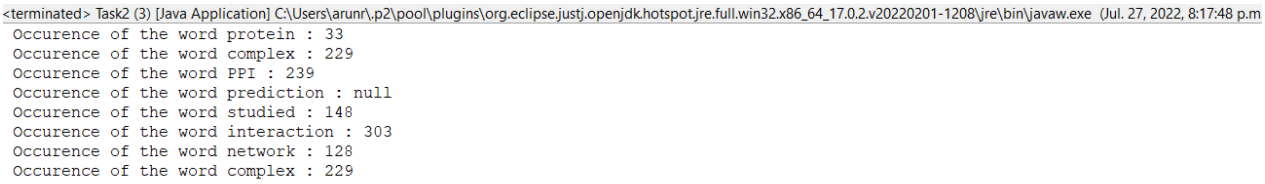
ALGORITHM	Average CPU time	Run time complexity
BruteForceMatch	897605	Brute-force pattern matching runs in time O(nm), (from lecture slides) where n – Size of text and m - Size of pattern
BoyerMoore	843571	The Boyer-Moore algorithm runs in worst-case time O(nm + s), (from lecture slides) where size is s Σ is the alphabet.
KMP Algorithm	1871894	The KMP algorithm runs in optimal worst-case time O(m + n) (from lecture slides)

After testing several times, by looped all the algorithms 100 times to search the same patterns as mentioned above to determine the average CPU time, BruteForceMatch and Boyer-Moore algorithm are providing better results then KMP algorithm.

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

   Solution :

I had created Task2.java class for this solution
Location : Assignment-4\src\Assignment4\Task2.java

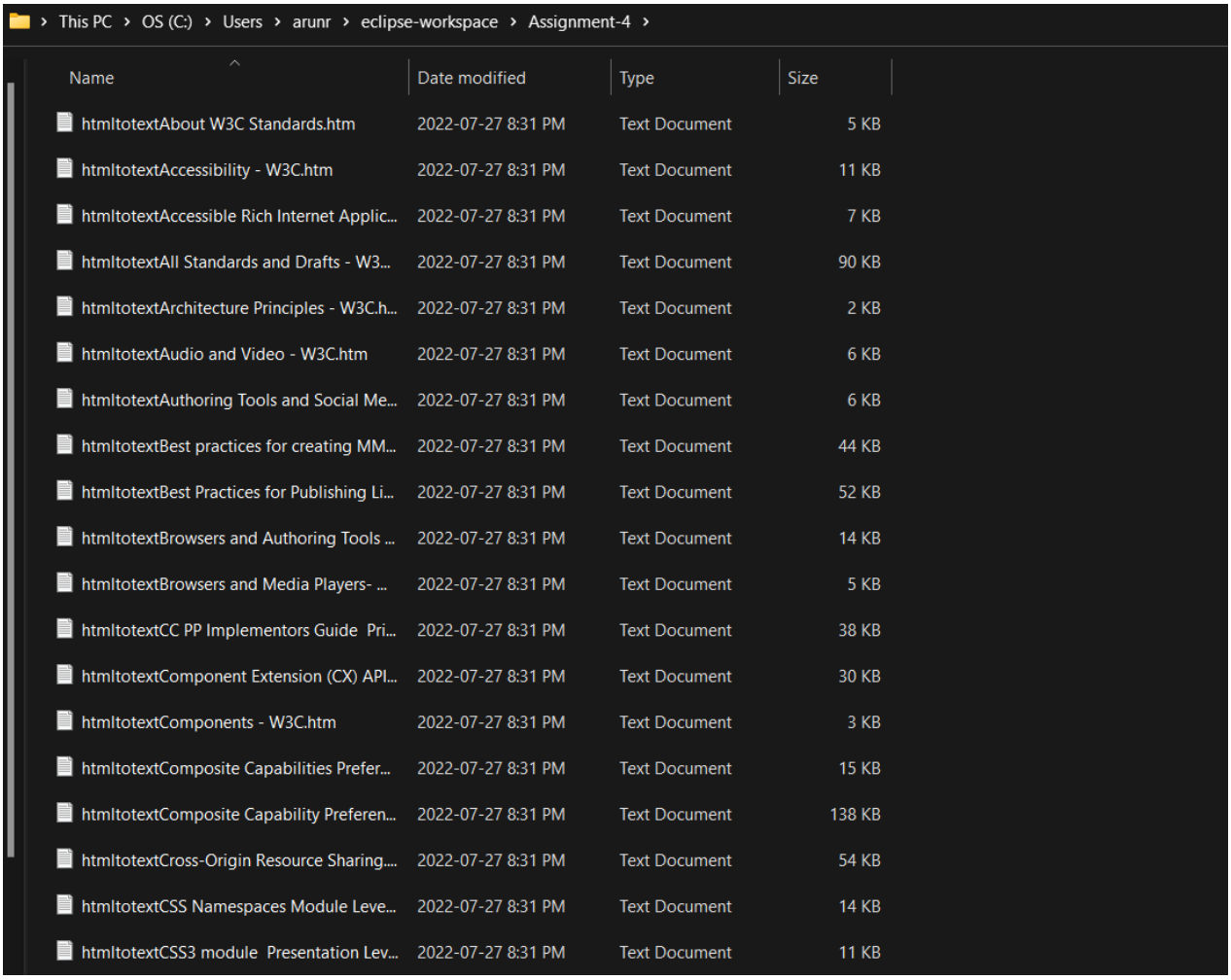


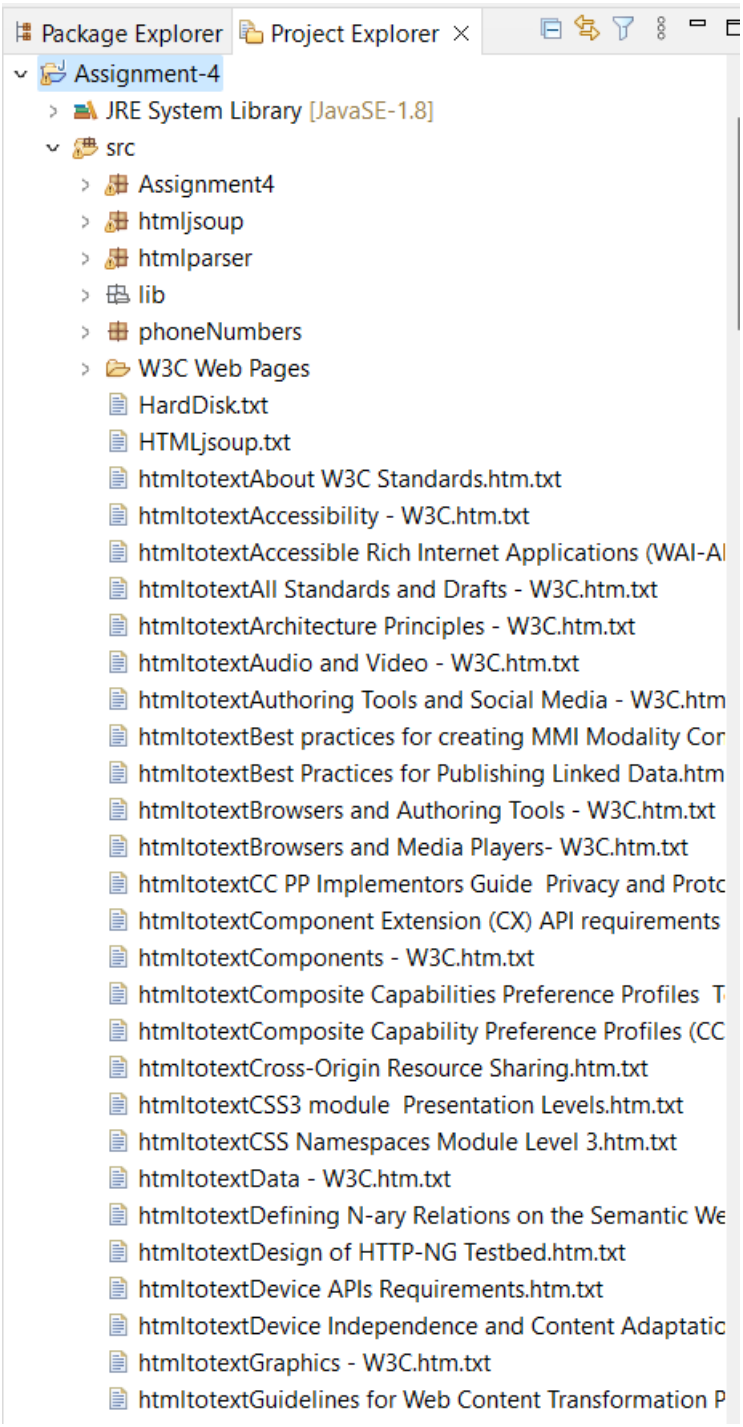
I had used TST algorithm given in the resources to extract the words from the file Protein.txt and also printed the occurrence of key given in the question

3. HTMLtoText converter: Write a program that takes the 100 given Web pages (W3C Web Pages.zip), 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.

Solution :

I had created Task3.java class for this solution
Location : Assignment-4\src\Assignment4\Task3.java





Using the HtmltoText.java source code from the resources, I read all the input HTML files and takes 100 web pages from W3 web pages folder with the help of BufferedReader. Then converted all the webpages into text files and saved as individual files in in Assignment-4 folder.

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

   Solution:

I had created Task4.java class for this solution
Location : Assignment-4\src\Assignment4\Task4.java

I had used the 100 text files the I got from task 3 as input files and found phone and email addresses using the java regex.

Below is the output for the TASK4

```
<terminated> Task4 (3) [Java Application] C:\Users\arunr\.p2\pool\plugins\org.eclipse.justj.openjdk.hotspot.jre.full.win32.x86_64_17.0.2.v20220201-1208\jre\bin\javaw.exe  (Jul. 27, 2
list of Phone Numbers :
(650) 812-4763 Phone number is found at 43434
(650) 812-4777 Phone number is found at 43454
201-555-0111 Phone number is found at 39605

list of Email id :
public-pfwg-comments@w3.org Email id is found at 3114
www-multimodal@w3.org Email id is found at 2715
www-multimodal-request@w3.org Email id is found at 2943
public-gld-comments@w3.org Email id is found at 2098
ohto@w3.org Email id is found at 289
johan.hjelm@nrj.ericsson.se Email id is found at 357
www-mobile@w3.org Email id is found at 1743
www-component-extension@w3.org Email id is found at 2387
mikael.nilsson@ks.ericsson.se Email id is found at 311
www-mobile@w3.org Email id is found at 1551
franklin.reynolds@nokia.com Email id is found at 467
woodroc@metaphoria.net Email id is found at 533
ohto@w3.org Email id is found at 595
ohan.hjelm@ericsson.com Email id is found at 659
mark-h_butler@hp.com Email id is found at 709
luu.tran@sun.com Email id is found at 757
www-mobile@w3.org Email id is found at 3720
public-webappsec@w3.org Email id is found at 3098
www-style@w3.org Email id is found at 395
www-style@w3.org Email id is found at 2543
www-style@w3.org Email id is found at 13327
www-style@w3.org Email id is found at 1677
public-swbp-wg@w3.org Email id is found at 2556
semantic-web@w3.org Email id is found at 2719
public-device-apis@w3.org Email id is found at 1537
public-bpwg-comments@w3.org Email id is found at 3191
public-content-transformation-conformance@w3.org Email id is found at 3253
public-bpwg-comments@w3.org Email id is found at 3548
public-content-transformation-conformance@w3.org Email id is found at 18170
public-webapps@w3.org Email id is found at 350
public-webapps@w3.org Email id is found at 1383
```

public-webapps@w3.org Email id is found at 468
public-webapps@w3.org Email id is found at 1801
janssen@parc.xerox.com Email id is found at 43476
rguha@us.ibm.com Email id is found at 333
phayes@ihmc.us Email id is found at 371
www-rdf-comments@w3.org Email id is found at 3713
www-rdf-logic@w3.org Email id is found at 3878
nt@inf.ed.ac.uk Email id is found at 371
richard@inf.ed.ac.uk Email id is found at 428
norman.walsh@marklogic.com Email id is found at 488
xml-editor@w3.org Email id is found at 2156
public-gld-comments@w3.org Email id is found at 1594
public-mbui@w3.org Email id is found at 1990
public-media-annotation@w3.org Email id is found at 2649
public-media-annotation@w3.org Email id is found at 91728
public-bpwg-comments@w3.org Email id is found at 2443
public-html-comments@w3.org Email id is found at 1455
public-rif-comments@w3.org Email id is found at 2406
public-rif-dev@w3.org Email id is found at 2657
public-owl-comments@w3.org Email id is found at 2606
public-owl-dev@w3.org Email id is found at 2831
public-owl-comments@w3.org Email id is found at 2569
public-owl-dev@w3.org Email id is found at 2794
public-owl-comments@w3.org Email id is found at 2689
public-owl-dev@w3.org Email id is found at 2914
public-owl-comments@w3.org Email id is found at 2435
public-owl-dev@w3.org Email id is found at 2660
public-owl-comments@w3.org Email id is found at 2651
public-owl-dev@w3.org Email id is found at 2876
ivan@w3.org Email id is found at 15531
sandro@w3.org Email id is found at 15559
public-owl-comments@w3.org Email id is found at 15632
armelen@cs.vu.nl Email id is found at 765
public-webont-comments@w3.org Email id is found at 3180
www-rdf-logic@w3.org Email id is found at 3279
connolly@w3.org Email id is found at 39168
khare@w3.org Email id is found at 39332

frystyk@w3.org Email id is found at 39504
eric@w3.org Email id is found at 39669
public-device-apis@w3.org Email id is found at 1744
public-rdf-comments@w3.org Email id is found at 2202
public-swbp-wg@w3.org Email id is found at 2215
public-swbp-wg@w3.org Email id is found at 2193
public-pfwg-comments@w3.org Email id is found at 2858
shane@aptest.com Email id is found at 461
public-pfwg-comments@w3.org Email id is found at 3544
public-webappsec@w3.org Email id is found at 1524
wai-uaag-editor@w3.org Email id is found at 3475
w3c-wai-ua@w3.org Email id is found at 3573
public-gld-comments@w3.org Email id is found at 1743
/echo@paul.demos Email id is found at 26581
/dumb@paul.demos Email id is found at 26637
/echo@paul.demos Email id is found at 27188
/dumb@paul.demos Email id is found at 27245
paul@digicool.com Email id is found at 37570
public-gld-comments@w3.org Email id is found at 3148
public-webapps@w3.org Email id is found at 2435
whatwg@whatwg.org Email id is found at 2483
public-tracking-comments@w3.org Email id is found at 2264
www-multimodal@w3.org Email id is found at 4810
www-multimodal-request@w3.org Email id is found at 5038
www-multimodal@w3.org Email id is found at 8511
w3c-ecommerce-ig@w3.org Email id is found at 1694
w3c-ecommerce-ig@w3.org Email id is found at 6667
w3c-ecommerce-ig-request@w3.org Email id is found at 16899
public-pfwg-comments@w3.org Email id is found at 3411
public-audio@w3.org Email id is found at 1521
public-audio@w3.org Email id is found at 44981
public-audio@w3.org Email id is found at 622
public-audio@w3.org Email id is found at 4220
public-usable-authentication@w3.org Email id is found at 2626
www-ws-arch@w3.org Email id is found at 1877
public-ws-policy-comments@w3.org Email id is found at 2782
public-ws-policy@w3.org Email id is found at 75180

```
shane@aptest.com Email id is found at 542
www-html-editor@w3.org Email id is found at 4053
www-html@w3.org Email id is found at 4185
www-xml-linking-comments@w3.org Email id is found at 2016
public-swbp-wg@w3.org Email id is found at 1837
semantic-web@w3.org Email id is found at 1995
d3e3e3@gmail.com Email id is found at 433
reagle@mit.edu Email id is found at 465
dsolo@alum.mit.edu Email id is found at 492
frederick.hirsch@nokia.com Email id is found at 529
mnystrom@microsoft.com Email id is found at 592
tlr@w3.org Email id is found at 638
kelviny@microsoft.com Email id is found at 680
mbartel@adobe.com Email id is found at 730
boyerj@ca.ibm.com Email id is found at 760
bal@microsoft.com Email id is found at 833
edsimon@xmlsec.com Email id is found at 861
public-xmlsec@w3.org Email id is found at 4138
public-webapps@w3.org Email id is found at 2223
whatwg@whatwg.org Email id is found at 7061
public-webapi@w3.org Email id is found at 7081
public-appformats@w3.org Email id is found at 7104
public-webapps@w3.org Email id is found at 7131
public-webapps@w3.org Email id is found at 9787
public-qt-comments@w3.org Email id is found at 4803
```

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**

I had created Task5A.java class for this solution
Location : Assignment-4\src\Assignment4\Task5A.java

I had used this (https?|ftp|file)://[-a-zA-Z0-9+&@#/%?=~_|!:,.;]*[-a-zA-Z0-9+&@#/%=~_|] regex pattern to find links with domain w3.org

<terminated> Task5A [Java Application] C:\Users\arunr\.p2\pool\plugins\org.eclipse.justj.openjdk.hotspot.jre.full.win32.x86_64_17.0.2.v20220201

```
Folder: http://www.w3.org/1999/02/22-rdf-syntax-ns#type found at 326295
Folder: http://www.w3.org/2001/XMLSchema#token found at 326402
Folder: http://www.w3.org/2007/rif-builtin-predicate#is-literal-not-token found at 326774
Folder: http://www.w3.org/2007/rif#error found at 327341
Folder: http://www.w3.org/1999/02/22-rdf-syntax-ns#type found at 328055
Folder: http://www.w3.org/2001/XMLSchema#language found at 328162
Folder: http://www.w3.org/2007/rif-builtin-predicate#is-literal-not-language found at 328537
Folder: http://www.w3.org/2007/rif#error found at 329107
Folder: http://www.w3.org/1999/02/22-rdf-syntax-ns#type found at 329821
Folder: http://www.w3.org/2001/XMLSchema#Name found at 329928
Folder: http://www.w3.org/2007/rif-builtin-predicate#is-literal-not-Name found at 330299
Folder: http://www.w3.org/2007/rif#error found at 330865
Folder: http://www.w3.org/1999/02/22-rdf-syntax-ns#type found at 331579
Folder: http://www.w3.org/2001/XMLSchema#NCName found at 331686
Folder: http://www.w3.org/2007/rif-builtin-predicate#is-literal-not-NCName found at 332059
Folder: http://www.w3.org/2007/rif#error found at 332627
Folder: http://www.w3.org/1999/02/22-rdf-syntax-ns#type found at 333341
Folder: http://www.w3.org/2001/XMLSchema#NMTOKEN found at 333448
Folder: http://www.w3.org/2007/rif-builtin-predicate#is-literal-not-NMTOKEN found at 333822
Folder: http://www.w3.org/2007/rif#error found at 334391
Folder: http://www.w3.org/1999/02/22-rdf-syntax-ns#type found at 335105
Folder: http://www.w3.org/2001/XMLSchema#boolean found at 335212
Folder: http://www.w3.org/2007/rif-builtin-predicate#is-literal-not-boolean found at 335586
Folder: http://www.w3.org/2007/rif#error found at 336155
Folder: http://www.w3.org/1999/02/22-rdf-syntax-ns#type found at 336869
```

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

I had created Task5B.java class for this solution
Location : Assignment-4\src\Assignment4\Task5B.java

I had used this regex patten (https?|ftp|file)://[-a-zA-Z0-9+&@#/%?=~_|!:,.;]*[-a-zA-Z0-9+&@#/%=~_|] to find Links that contain folders: e.g., www.w3.org/TR/owl-features/

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

Folder: http://www.w3.org/TR/xhtml-modularization found at 30348
Folder: http://www.w3.org/TR/xhtml-modularization/abstract_modules.html#s_commonatts found at 56776
Folder: http://www.w3.org/TR/xhtml-modularization/abstract_modules.html#s_presentationmodule found at 61251
Folder: http://www.w3.org/TR/xhtml-modularization/abstract_modules.html#s_linkmodule found at 61685
Folder: http://www.w3.org/TR/xhtml-modularization/abstract_modules.html#s_metamodule found at 62925
Folder: http://www.w3.org/TR/xhtml-modularization/abstract_modules.html#s_basemodule found at 63320
Folder: http://www.w3.org/TR/xhtml-modularization/abstract_modules.html#s_scriptmodule found at 64567
Folder: http://www.w3.org/TR/xhtml-modularization/abstract_modules.html#s_stylemodule found at 64966
Folder: http://www.w3.org/TR/xhtml-modularization/abstract_modules.html#s_styleattributemodule found at 65378
Folder: http://www.w3.org/TR/xhtml-modularization/abstract_modules.html#s_imagemodule found at 65784
Folder: http://www.w3.org/TR/xhtml-modularization/abstract_modules.html#s_objectmodule found at 66555
Folder: http://www.w3.org/TR/xhtml-modularization/abstract_modules.html#s_simpletablemodule found at 67716
Folder: http://www.w3.org/TR/xhtml-modularization/abstract_modules.html#s_extformsmodule found at 68219
Folder: http://www.w3.org/TR/xhtml-modularization/abstract_modules.html#s_intrinsiceventsmodule found at 73869
Folder: http://www.w3.org/TR/xhtml-modularization/abstract_modules.html#s_targetmodule found at 74296
```

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

I had created Task5C.java class for this solution
Location : Assignment-4\src\Assignment4\Task5C.java

I had used this (https?|ftp|file)://[-a-zA-Z0-9+&@#/%?=~_|!:,.;]*[-a-zA-Z0-9+&@#/%=~_|] regex pattern to find Links that contain references in a Web page and may contain folders.
for example: www.w3.org/TR/owl-features/#DefiningSimpleClasses

```
Links that contain references in a Web page and may contain folders:
Link found at : http://www.w3.org/2005/xpath-functions/collation/codepoint at 160998
Link found at : http://www.w3.org/2005/xpath-functions/collation/codepoint at 166087
Link found at : http://www.w3.org/2005/xpath-functions/collation/codepoint at 463591
Link found at : http://www.w3.org/2005/xpath-functions/collation/codepoint at 463864
Link found at : http://www.w3.org/2005/xpath-functions/collation/codepoint at 469103
Link found at : http://www.w3.org/2005/xpath-functions/collation/codepoint at 469376
Link found at : http://www.w3.org/2005/xpath-functions/collation/codepoint at 474057
Link found at : http://www.w3.org/2005/xpath-functions/collation/codepoint at 520776
```

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

I had created Task5D.java class for this solution
Location : Assignment-4\src\Assignment4\Task5D.java

I had used this (https?|ftp|file)://[-a-zA-Z0-9+&@#/%?=~_|!:,.;]*[-a-zA-Z0-9+&@#/%=~_|] regex pattern to find the links with domain .net, .com, .org

```
<terminated> Task5D (1) [Java Application] C:\Users\arunr\.p2\pool\plugins\org.eclipse.justj.openjdk.hotspot.jre.full.win32.x86_64_17.0.2.v20220201-1208\jre\bin\javaw.exe  (Jul. 27, 2022, 9:24:08 p.m. – 9
links found at : http://www.w3.org/TR/2014/REC-xslt-xquery-serialization-30-20140408/ at 601590
links found at : http://www.w3.org/TR/xslt-xquery-serialization-30/.XQuery at 601695
links found at : http://www.w3.org/TR/2010/REC-xquery-semantics-20101214/ at 601978
links found at : http://www.w3.org/TR/xquery-semantics/.XQuery at 602071
links found at : http://www.w3.org/TR/2010/REC-xquery-20101214/ at 602319
links found at : http://www.w3.org/TR/xquery/.XQuery at 602402
links found at : http://www.w3.org/TR/2014/REC-xquery-30-20140408/ at 602626
links found at : http://www.w3.org/TR/xquery-30/.XML at 602712
links found at : http://www.w3.org/TR/2006/REC-xinclude-20061115/ at 602971
links found at : http://www.w3.org/TR/xinclude/.XML at 603056
links found at : http://www.w3.org/TR/xmlschema-2/Schema at 603203
links found at : http://www.w3.org/TR/2012/REC-xmlschema11-2-20120405/ at 603434
links found at : http://www.w3.org/TR/xmlschema11-2/.Namespaces at 603524
links found at : http://www.w3.org/TR/2009/REC-xml-names-20091208/ at 603731
links found at : http://www.w3.org/TR/xml-names.A.2 at 603817
links found at : http://www.w3.org/TR/2013/WD-xslt-30-20131212/ at 604036
links found at : http://www.w3.org/TR/xslt-30/.Calendrical at 604119
links found at : http://www.w3.org/TR/2004/WD-charmod-norm-20040225/HTML at 604475
links found at : http://www.w3.org/TR/REC-html40/ISO at 604592
links found at : http://www.iso.org/ at 604750
links found at : http://pubs.opengroup.org/onlinepubs/9699919799/Working at 604863
links found at : http://www.w3.org/TR/2005/NOTE-timezone-20051013/B at 605035
links found at : http://www.twinsun.com/tz/tz-link.htm at 644770
links found at : http://www.unicode.org/standard/versions/ at 645060
links found at : http://www.unicode.org/reports/tr15/ at 645479
```