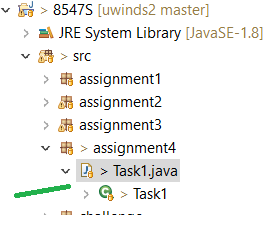
Assignment 3

As a student at the University of Windsor, I pledge to pursue all endeavours with honour and integrity and will not tolerate or engage in academic or personal dishonesty. 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.

Wen Dong #110057395

**Task #1**

1. Source code



1. Output of the java file looks as below:

Indices found by BoyerMoore:

hard: [159, 212, 442, 465, 476, 3585, 3787, 8850, 8948, 9158, 11982, ...

disk: [5, 140, 217, 447, 470, 496, 639, 3590, 3792, 4571, 4940, 4980 ...

hard disk: [212, 442, 465, 3585, 3787, 9158, 11982, 17219, 17559, 19252, ...

hard drive: [476, 8850, 8948, 18019, 25221, 30034, 34130, 51400, 56579]

hard dist: []

xltpru: []

Searching patterns for 100 times with each algorithm, average CPU time:

BoyerMoore: 1790893 ns

KMP: 2205835 ns

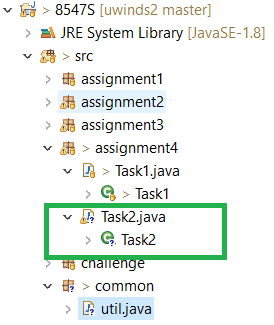
BruteForceMatch: 2301495 ns

Comments:

As we learned from class, BruteForceMatch runs at time complexity O(nm) where n is the length of search string and m is the length of pattern. BoyerMoore is of worst-case time complexity O(nm + s) where s is the size of alphabet, the worst-case may occur in images and bio-sequences, but unlikely in English text, for English text, it runs significantly faster than BruteForceMatch. KMP algorithm runs in optimal worst-case time O(m+n). as we can see from the output of the program, it is consistent with what we learned in regard to the algorithms performance, i.e. BoyerMoore runs much faster than BruteForceMatch, and KMP runs a bit faster then BruteForceMatch.

**Task #2 - a**

1. Source code



1. Output of the java file looks as below:

Searching keys - [protein, complex, PPI, prediction, interaction]

Get keys by complete match------------

protein - [97, 207, 237, 245, 418, 426, 721, 729, 885, 2362, 2370]

complex - [159, 1211, 1331, 1564]

PPI - [1619]

prediction - null

interaction - [105, 253, 1283, 1963, 2031]

Get keys by prefixMatch------------

protein - [97, 207, 237, 245, 418, 426, 721, 729, 885, 2362, 2370]

proteins - [186, 804, 957, 1172, 1845]

complex - [159, 1211, 1331, 1564]

PPI - [1619]

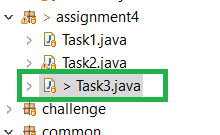
predictions - [1304]

interaction - [105, 253, 1283, 1963, 2031]

interactions - [434, 737, 893, 1825, 2378]

**Task #3**

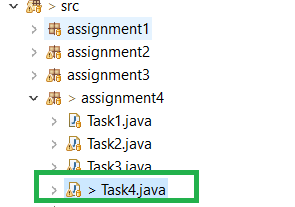
1. Source code



1. Run the java file and generate text file in folder output/

**Task #4**

1. Source code



1. Output of the java file looks as below:

Found 173 emails, 104 unique emails:

[public-pfwg-comments@w3.org, www-multimodal@w3.org, www-multimodal-request@w3.org, public-gld-comments@w3.org, ohto@w3.org, lalitha@tri.sbc.com, johan.hjelm@nrj.ericsson.se, www-mobile@w3.org, www-component-extension@w3.org, liam@w3.org, …]

Found 107 phone numbers, 68 unique phone numbers:

[185788651, 1065075071, 2147483648, +2147483647, 2147483647, 4294967295, 9223372036, 854775808, 854775807, 1844674407, 3709551615, 4294967296, 3709551616, (650) 812-4763, (650) 812-4777, 0123456789, 3398648111, 121.8850708, 201-555-0111, 2343543645, …]

33 connected components found in movies.txt

Task #3 - It took **49 ms** to use DFS identifying the components and 33 connected components identified in movies.txt.

Task #4 -

**Movies starred by DiCaprio, Leonardo:**

What's Eating Gilbert Grape (1993)

Total Eclipse (1995)

Titanic (1997)

This Boy's Life (1993)

Romeo + Juliet (1996)

Quick and the Dead, The (1995)

Poison Ivy (1992)

Marvin's Room (1996)

Man in the Iron Mask, The (1998 I)

Gangs of New York (2002)

Departed, The (2006)

Celebrity (1998)

Catch Me If You Can (2002)

Beach, The (2000 I)

Basketball Diaries, The (1995)

Aviator, The (2004)

**Movies starred by Roberts, Julia (I):**

Stepmom (1998)

Steel Magnolias (1989)

Something to Talk About (1995)

Sleeping with the Enemy (1991)

Runaway Bride (1999)

Prêt-à-Porter (1994)

Pretty Woman (1990)

Player, The (1992)

Pelican Brief, The (1993)

Ocean's Twelve (2004)

Ocean's Eleven (2001)

Notting Hill (1999)

Mystic Pizza (1988)

My Best Friend's Wedding (1997)

Mona Lisa Smile (2003)

Michael Collins (1996)

Mexican, The (2001)

Mary Reilly (1996)

I Love Trouble (1994)

Hook (1991)

Full Frontal (2002)

Flatliners (1990)

Everyone Says I Love You (1996)

Erin Brockovich (2000)

Dying Young (1991)

Conspiracy Theory (1997)

Confessions of a Dangerous Mind (2002)

Closer (2004 I)

America's Sweethearts (2001)

**Movies starred by Grant, Hugh (I):**

Two Weeks Notice (2002)

Small Time Crooks (2000)

Sirens (1994)

Sense and Sensibility (1995)

Restoration (1995)

Remains of the Day, The (1993)

Notting Hill (1999)

Nine Months (1995)

Mickey Blue Eyes (1999)

Maurice (1987)

Love Actually (2003)

Lair of the White Worm, The (1988)

Four Weddings and a Funeral (1994)

Extreme Measures (1996)

Englishman Who Went Up a Hill But Came Down a Mountain, The (1995)

Bridget Jones: The Edge of Reason (2004)

Bridget Jones's Diary (2001)

Bitter Moon (1992)

American Dreamz (2006)

About a Boy (2002)

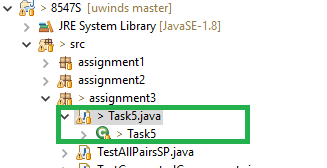
**Movies starred by both Roberts, Julia (I) and Grant, Hugh (I):**

Notting Hill (1999)

1. According to the output, the CPU time is 49 ms for DFS identifying the connected components, while the time complexity of the algorithm DFS which is in my implementation is O(|V|+|E|) as we learned from the class. worst-case is the case O(|V|+|E|) as well.
2. From the output, we can see the movies starred by the given actors respectively, and by both Roberts, Julia (I) and Grant, Hugh (I)

**Task #5**

1. Source code

****

1. Output of the java file looks as below:

Partitioning 'resources/ChIP-seq reads/ChIP-seq-reads-1M.dat' into following files:

output/A.dat

output/B.dat

output/C.dat

output/D.dat

Sorting output/A.dat as output/AS.dat.

Sorting output/B.dat as output/BS.dat.

Sorting output/C.dat as output/CS.dat.

Sorting output/D.dat as output/DS.dat.

Merge sorting following files:

output/AS.dat

output/BS.dat

output/CS.dat

output/DS.dat

Merge Sorting done into 'output/Chip-seq-reads-1M-sorted.dat'

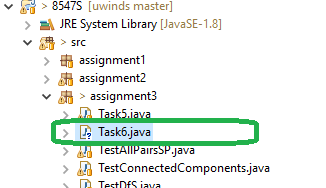
Total CPU time is 4052 ms

1. Time complexity comments:

It uses binary heap for sorting a single file sub file like A.date, the time complexity is O (nlog n) as we learned from class; On the other hand, it uses IndexMinPQ for multiway merge, the time complexity is O (n log d) according to what we learned from class that finding the minimum form the d-way IndexMinPQ is O (log d) and we have in all n elements to sort.

**Task #6**

1. Source code:



Apart from Task6.java, I have modified following file a bit to meet in-order traverse.

memoryManagement/ BTree.java

1. Output of the java file looks as below:

Loading into BTee - 'resources/ChIP-seq reads/ChIP-seq-reads-1M.dat'.

Write the keys (in al 1000000) in in-order to 'output/B-tree.dat'.

Done. Total CPU time is 2405 ms

1. Comment on the obtained CPU times:

As we learn from class, time complexity of search, insertion and removal for B-Tree is O(log*Bn*), while B is the order and N is the total number of elements. So, the insertions during building BTree from is O(Nlog*BN*). In-order traverse is of O(N).

Task #7 – done respectively inline for #5 and #6