654 Advanced Computing Concepts
Assignment3 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. Consider the graph stored in largeDG.txt (download it from Resources). Run DFS on that graph and show the vertices of the graph in pre-order and post-order. Compute the CPU time and report the worst-case complexity of DFS.
Answer:

```
Preorder: 1481
Postorder: 2284
Reverse postorder: 1257
```

As a result, the vertices of graph in pre-order takes more time than post-order.
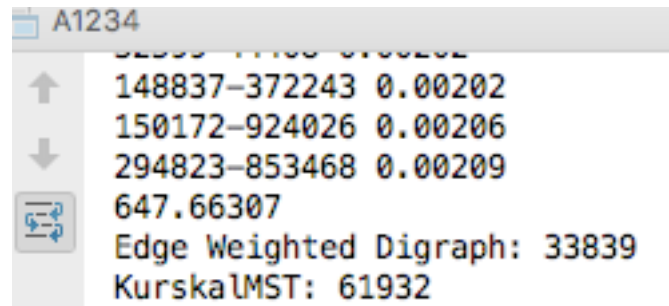For Depth First Search, the worst-case complexity is $O(|V|+|E|)$ since it visits every node and all the edges once.

Task 2. Consider the graph stored in largeEWG.txt (download it from Resources).
      a. Write a program that finds the shortest path for all pairs of nodes (you choose the algorithm). Calculate the CPU time and report the complexity of the algorithm you chose.
      b. Write a program that finds the MST (you choose the algorithm). Calculate the CPU time and compare it with the complexity of the algorithm you chose.
Answer:

```
A1234
148837-372243 0.00202
150172-924026 0.00206
294823-853468 0.00209
647.66307
Edge Weighted Digraph: 33839
KurskalMST: 61932
```

For part a, I chose EdgeWeightedDigraph algorithm, the CPU time for finding the shortest path was 33839ms. The complexity of this algorithm is $O(|E|+|V| \log |V|)$.
For part b, I chose KruskalMST algorithm, the CPU time for finding the Minimum spanning trees was 61932ms. The complexity of this algorithm is $O(|E| \log |E|)$.

Task 3. Consider the movie database stored in movie.txt, and SymbolGraph.java. Write a program that uses DFS to find all connected components. Use CC.java as a template. Show the CPU time and report the worst-case complexity of DFS.

```
Done reading document/Assignment3/data/movies.txt
SymbolGraph: 1659
33 components
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
   60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75
   112 113 114 115 116 117 118 119 120 121 122 123
   154 155 156 157 158 159 160 161 162 163 164 165
```

As a result, there are 33 components have been found.
It cost 1659ms to find all connected components.
The worst-case complexity for Depth First Search is O(|V|+|E|).

Task 4. Write a program that finds the movies starred by a particular actor. Show the movies starred by Leonardo DiCaprio. Show the movies starred by Julia Roberts, by Hugh Grant, and by both of them.
Answer:

```
Done reading document/Assignment3/data/movies.txt
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)
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)
   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)
   Grant, Hugh (I) and Roberts, Julia (I) both:
      Notting Hill (1999)
```

Task 5. Consider the one million Chip-seq reads given in the files called "Chip-seq-reads-1M.dat". Write a program that partitions the list of reads into 4 sublists. Save each sublist in a separate file (called A.dat, B.dat, C.dat, and D.dat). Sort each sublist and store it in a file (AS.dat, BS.dat, CS.dat, DS.dat). Take the 4 sorted sublists from the files and merge them in to a sorted list. Store the sorted list in a file (called "Chip-seq-reads-1M-sorted.dat").
Answer:

A.dat	Today, 22:07	9.3 MB	DocumentType
AS.dat	Today, 22:07	9.3 MB	DocumentType
B-tree.dat	Today, 22:07	35.5 MB	DocumentType
B.dat	Today, 22:07	9.3 MB	DocumentType
BS.dat	Today, 22:07	9.3 MB	DocumentType
C.dat	Today, 22:07	9.3 MB	DocumentType
Chip-seq-reads-1M.dat	Aug 5, 2014, 12:47	38 MB	DocumentType
Chip-seq-reads-1M.sorted.dat	Today, 22:07	37 MB	DocumentType
Chip-seq-reads-5M.dat	Aug 5, 2014, 12:04	205.2 MB	DocumentType
CS.dat	Today, 22:07	9.3 MB	DocumentType
D.dat	Today, 22:07	9.3 MB	DocumentType
DS.dat	Today, 22:07	9.3 MB	DocumentType

Task 6. Create a B-tree and insert all the reads from the original list (Chip-seq-reads-1M.dat) as they appear in the file. List the B-tree in in-order traversal and save the output all keys in a file (called B-tree.dat).

Task 7. Record total CPU times for #5 and #6. Comment on the obtained CPU times and compare them with the corresponding complexities as discussed in class.
Answer:

```
A567
    start...
    separte files: 4581ms
    sort files: 9548ms
    merge files: 17640ms
    end...
    build B-tree cost: 790ms
    size: 909261
    height: 18
    Task5: 17641
    Task6: 4546
    Total For T5&T6: 22187
```

It cost 4581ms to separte files into 4 pices, 9548ms for sorting these sub-files and 17640ms to merge these files into a sorted file.

The building time of the B-tree is 790ms, size is 909661(different keys) and height is 18.

The complexity is given in terms of n and B. if B is constant, $f(b)=g(b)=O(1)$. The complexity of B-tree is $O(\log_B n)$ in general.