Algorithms – HW5

1. Names of team members:

    1. Shashank Shekhar (Unity Id: sshekha4)
    2. Rahul Ravindra(Unity Id: rravind)
    3. Akshay Podila (Unity Id: apodila)

2. NCSU Github URL: https://github.ncsu.edu/sshekha4/HW5

3. Citations for the Code:

    Code is implemented in C++. The source code is obtained from GeeksforGeeks and Tutorialspoint and modified as per requirement. Code Citations:

    LCS Code: https://www.geeksforgeeks.org/printing-longest-common-subsequence/

    File Operations: https://www.tutorialspoint.com/read-file-line-by-line-using-cplusplus

4. **diff** program output for files ex41.txt and ex42.txt

Windows PowerShell

```
PS C:\Users\sshek\Documents\My Documents\Spring 2020 Courses\Algorithms\Homeworks\HW5> ./diff ex41.txt ex42.txt
1 2 3 4 5 6 10 11 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64 66
2 3 4 5 6 7 11 12 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41
```

5. **diff** program output for files ex42.txt and ex41.txt

Windows PowerShell

```
PS C:\Users\sshek\Documents\My Documents\Spring 2020 Courses\Algorithms\Homeworks\HW5> ./diff ex42.txt ex41.txt
2 3 4 5 6 10 11 12 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41
1 2 3 4 5 6 10 11 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64 66
```

6. **diffprint** output

a) for files ex61.txt, ex62.txt

Windows PowerShell

```
PS C:\Users\sshek\Documents\My Documents\Spring 2020 Courses\Algorithms\Homeworks\HW5> ./diffprint ex61.txt ex62.txt
  CONTENTS

- CHAPTER I                                                            Page
+                                                                      Page
  Jonathan Harker's Journal                                               1
- CHAPTER II
  Jonathan Harker's Journal                                              14
- CHAPTER III
  Jonathan Harker's Journal                                              26
- CHAPTER IV
  Jonathan Harker's Journal                                              38
- CHAPTER V
  Letters--Lucy and Mina                                                 51
- CHAPTER VI
  Mina Murray's Journal                                                  59
- CHAPTER VII
  Cutting from "The Dailygraph," 8 August                                71
- CHAPTER VIII
  Mina Murray's Journal                                                  84
- CHAPTER IX
  Mina Murray's Journal                                                  98
- CHAPTER X
  Mina Murray's Journal                                                 111
- CHAPTER XI
  Lucy Westenra's Diary                                                 124
- CHAPTER XII
  Dr. Seward's Diary                                                    136
- CHAPTER XIII
  Dr. Seward's Diary                                                    152
- CHAPTER XIV
  Mina Harker's Journal                                                 167
- CHAPTER XV
  Dr. Seward's Diary                                                    181
- CHAPTER XVI
  Dr. Seward's Diary                                                    194
- CHAPTER XVII
  Dr. Seward's Diary                                                    204
- CHAPTER XVIII
  Dr. Seward's Diary                                                    216
- CHAPTER XIX
  Jonathan Harker's Journal                                             231
- CHAPTER XX
  Jonathan Harker's Journal                                             243
- CHAPTER XXI
  Dr. Seward's Diary                                                    256
```

b) for files ex62.txt, ex61.txt

Windows PowerShell

```
PS C:\Users\sshek\Documents\My Documents\Spring 2020 Courses\Algorithms\Homeworks\HW5> ./diffprint ex62.txt ex61.txt

-                                                                      Page
+ CHAPTER I                                                            Page
  Jonathan Harker's Journal                                               1
+ CHAPTER II
  Jonathan Harker's Journal                                              14
+ CHAPTER III
  Jonathan Harker's Journal                                              26
+ CHAPTER IV
  Jonathan Harker's Journal                                              38
+ CHAPTER V
  Letters--Lucy and Mina                                                 51
+ CHAPTER VI
  Mina Murray's Journal                                                  59
+ CHAPTER VII
  Cutting from "The Dailygraph," 8 August                                71
+ CHAPTER VIII
  Mina Murray's Journal                                                  84
+ CHAPTER IX
  Mina Murray's Journal                                                  98
+ CHAPTER X
  Mina Murray's Journal                                                 111
+ CHAPTER XI
  Lucy Westenra's Diary                                                 124
+ CHAPTER XII
  Dr. Seward's Diary                                                    136
+ CHAPTER XIII
  Dr. Seward's Diary                                                    152
+ CHAPTER XIV
  Mina Harker's Journal                                                 167
+ CHAPTER XV
  Dr. Seward's Diary                                                    181
+ CHAPTER XVI
  Dr. Seward's Diary                                                    194
+ CHAPTER XVII
  Dr. Seward's Diary                                                    204
+ CHAPTER XVIII
  Dr. Seward's Diary                                                    216
+ CHAPTER XIX
  Jonathan Harker's Journal                                             231
+ CHAPTER XX
  Jonathan Harker's Journal                                             243
+ CHAPTER XXI
  Dr. Seward's Diary                                                    256
+ CHAPTER XXII
  Jonathan Harker's Journal                                             269
+ CHAPTER XXIII
  Dr. Seward's Diary                                                    281
```

c) for files ex66a.txt, ex66b.txt

Windows PowerShell

```
PS C:\Users\sshek\Documents\My Documents\Spring 2020 Courses\Algorithms\Homeworks\HW5> ./diffprint ex66a.txt ex66b.txt

- A
- B
+ A
+ B
  C
  D
  E
  F
  G
- H
+ H
+ I
+ J
```

7. My implementation of LCS requires Θ(m*n) space. Here, m and n are the number of lines in the two text files respectively.

8. The 10-bit hash code required by this assignment is sufficient to allow the **diff** program to work correctly. This is because the number of lines provided in the test files are less than 1024. In fact, they are less than 70% of 1024 (maximum integer held by 10-bits). Hence, the probability of collision is very less. Also, in the case of our hash function, the probability of collision is also reduced because our hash function takes into account the position of the characters while generating the hash value for a given line of text. This ensures that even if the characters in the two lines are the same, the hashes generated by them will be different. This has a very small collision probability (even for two different values, there is a slight chance that their modulo reduces to the same hash value. Although, the probability of this happening is very small).

   To check for insufficiency, we can use a Set data structure and keep pushing the hashes generated for each string in it. At the same time, we also create a HashMap of <HashValue, String> for each string. This HashMap will be used for lookup when we need to retrieve the String corresponding to a given hash value. Before the push operation of the hash value into the Set, we look for the hash value (to be pushed) in the Set. If the hash value to be pushed in the Set is already present in the Set, we compare the original string (by looking it up in the HashMap using the HashValue) with the current string character by character. If there is a difference at any character position during the match, then it implies that there is a collision since the hash function is generating the same hash values for two different strings. To resolve this, we resort to methods for collision resolution like Open Addressing and Separate Chaining.

9. Implications for the choice of the hash function when input files are long include taking modulo with a bigger number to generate more slots than the number of lines present in the file to avoid collision (ie. h(key) = key mod M where M >= 10000). Also, 14 bits will be needed to store the generated hash values (2^{14} > 10,000).

   Implications for the choice of the hash function when the input files have long lines include using a simple hash function. This ensures efficient computation of the hash values. For example: Instead of using exponents (r^i where I refers to the character position) for hash value calculation (in polynomial hashing) which is an expensive operation, one can use the character position in the string to generate near unique hash values and resort to separate chaining and open addressing methods for collision resolution.

a) Implications for the Hash Function - A good hash function should have the following properties:
   I. Efficiently computable.
   II. Should uniformly distribute the keys
b) Implications for actual memory usage -In the worst case, my program implementation will need O(m*n) space where m and n are the number of lines in the two text files respectively. In today's scenario, it is difficult to allocate 10000*10000 space. This space requirement can be reduced by using an optimized algorithm which takes O(n) space where n refers to the number of lines in the file.

Reference:

- https://www.geeksforgeeks.org/what-are-hash-functions-and-how-to-choose-a-good-hash-function/
- https://www.geeksforgeeks.org/hashing-set-3-open-addressing/
- https://www.geeksforgeeks.org/hashing-set-2-separate-chaining/
- https://www.geeksforgeeks.org/space-optimized-solution-lcs/