ชนสิทธ์ เกิดเปี่ยม 65050316

Lab 12: Huffman Coding

Mathematics for Computer Science

Huffman Coding เป็นการเข้ารหัสข้อมูลแบบเต็มบิท (จำนวนบิทเป็นจำนวนเต็ม) โดยการเข้ารหัสจะให้ตัวอักษรที่มีความถี่สูง มีความยาวของบิทต่ำ เพื่อให้โดยรวมแล้วการเข้ารหัสจะให้จำนวนบิททั้งหมดน้อยลง

การเขียนโปรแกรมเพื่อเข้ารหัส/ถอดรหัส Huffman เป็นการฝึกเขียนโปรแกรมที่ดี แต่เนื่องจากวิชานี้เป็นวิชาปี 1 นักศึกษาบางคนจึงอาจมีพื้นฐานไม่พอที่จะสร้างโปรแกรมจากศูนย์ ในแล็บนี้ จึงจะเป็นการศึกษาโค้ดโปรแกรมการเข้ารหัส Huffman โดยจะมีโปรแกรม Huffman.java มาให้ ให้นักศึกษาคอมไพล์และรันโปรแกรม Huffman.java ที่ให้ ซึ่งข้อมูลใน charArray จะเป็นข้อมูล char ที่มีค่า ASCII ตรงกับค่า index และข้อมูลใน charFreq จะเก็บค่าความถี่ของตัวอักษร ณ index นั้นๆ โดยส่วนของโปรแกรมด้านล่าง จะกำหนดค่าความถี่ของทุกตัวอักษรเป็น 1 ทั้งหมด

```
        // remvoe code below ********************
        for (int i = 0; i < charArray.length; i++) {
            charArray[i] = (char) i;
            charFreq[i] = 1;
        }
        // remove code above ********************
```

ถ้าไม่มีอะไรผิดพลาด ผลจากการรันโปรแกรม จะได้เป็นรหัสไบนารีของอักษรทั้งหมด และรหัสไบนารีทั้งหมดจะมีความยาวเท่ากันที่ 8 ทำให้ค่าเฉลี่ยออกมาเป็น 8 ซึ่งเท่ากับการเข้ารหัส ASCII เลย เพียงแค่สลับไบนารีกันเท่านั้นเอง

Task 1: หากเปลี่ยน charFreq[i] = 1 เป็น charFreq[i] = i นักศึกษาคาดว่า ค่าเฉลี่ยจำนวนบิทต่อตัวอักษรจะมากขึ้นหรือน้อยลง เพราะเหตุใด

ตำตอบ น้อยลง เพราะ จำนวนตัวอักษรที่คำนวณมีมากขึ้น

ในไฟล์ที่ให้ จะมีไฟล์หนังสือเรื่อง Harry Potter เล่น 1-7 ในรูปแบบของ text file โดยไฟล์เหล่านี้เป็นไฟล์ที่มีลิขสิทธิ์ ห้ามนำไปเผยแพร่ แต่เราสามารถนำมาทดสอบการเข้ารหัสของเราได้

Task 2: ให้นักศึกษาเปิดไฟล์แต่ละไฟล์ และทำการเขียนโปรแกรมนับความถี่ของแต่ละตัวอักษรโดยแก้ไขเมธอด buildCharFreqFromFile จากนั้นให้รันโปรแกรมและบันทึกว่า ค่าเฉลี่ยของจำนวนบิตต่อตัวอักษรของแต่ละไฟล์เป็นเท่าใด (บันทึกเป็นทศนิยม 4 ตำแหน่ง) และนักศึกษาคิดว่า ตัวเลขดังกล่าว บ่งบอกอะไรเกี่ยวกับปริมาณข้อมูลที่อยู่ในหนังสือแต่ละเล่ม

หนังสือ	จำนวนตัวอักษรทั้งหมด	ค่าเฉลี่ยบิทต่อตัวอักษร
Book 1	492,161	4.7352
Book 2	551,186	4.7491
Book 3	702,656	4.7651
Book 4	1,226,024	4.7245
Book 5	1,666,165	4.7383
Book 6	1,096,150	4.7313
Book 7	1,267,060	4.7155

คำตอบ ปริมาณจำนวนตัวอักษรและที่วางทั้งหมดมีเท่าไหร่

Task 3: นักศึกษาคิดว่า หากนับความถี่ของตัวอักษรทั้ง 7 ไฟล์รวมกัน ค่าเฉลี่ยบิทต่อตัวอักษรจะมากขึ้นหรือน้อยลง เพราะเหตุใด

คำตอบ น้อยลง เพราะ มีจำนวนตัวอักษรที่คำนวณมหาศาล

Task 4: ให้นักศึกษาแก้ไขเมธอด buildCharFreqFromFile เพื่อนับความถี่ตัวอักษรของทั้ง 7 ไฟล์รวมกัน และทำการบันทึกลงตารางด้านล่าง

หนังสือ	จำนวนตัวอักษรทั้งหมด	ค่าเฉลี่ยบิทต่อตัวอักษร
ทั้ง 7 ไฟล์	7,212,437	4.7853

สิ่งที่ต้องส่ง: Huffman.java ที่นักศึกษาทำการแก้ไข และคำตอบที่กรอกในไฟล์นี้