

Assignment 3

Huffman coding

|  |  |  |
| --- | --- | --- |
| 과목명 |  | 확률및통계 |
| 담당교수 |  | 심동규 교수님 |
| 학과 |  | 컴퓨터정보공학부 |
| 학년 |  | 3학년 |
| 학번 |  | 2019202009 |
| 이름 |  | 서여지 |
| 제출일 |  | 21.04.25(일) |



1. 과제설명

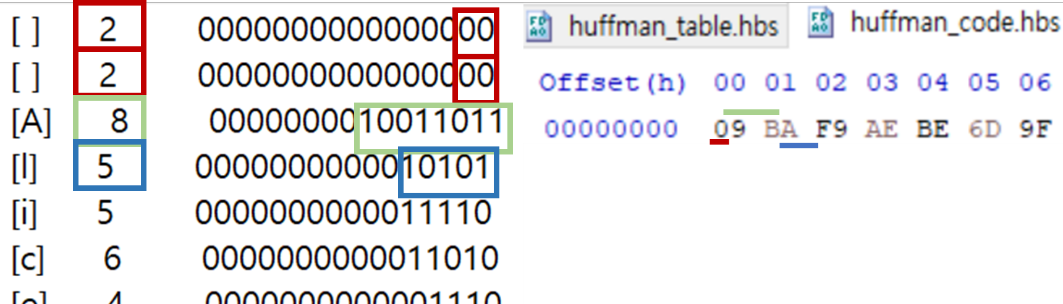
txt 파일을 허프만 코딩을 이용하여 압축하는 프로그램을 작성하는 과제이다. 이번 과제에서는 인코더와 디코더를 분리하여 별도의 프로그램으로 작성한다. 인코더 프로그램은 txt파일을 읽어서 허프만 코드 테이블과 압축 결과를 .hbs로 출력한다. 디코더 프로그램은 인코더 프로그램에서 생성한 두 개의 .hbs 파일의 내용을 이용하여 원본과 동일한 txt 파일을 재생성한다.

1. 접근방법

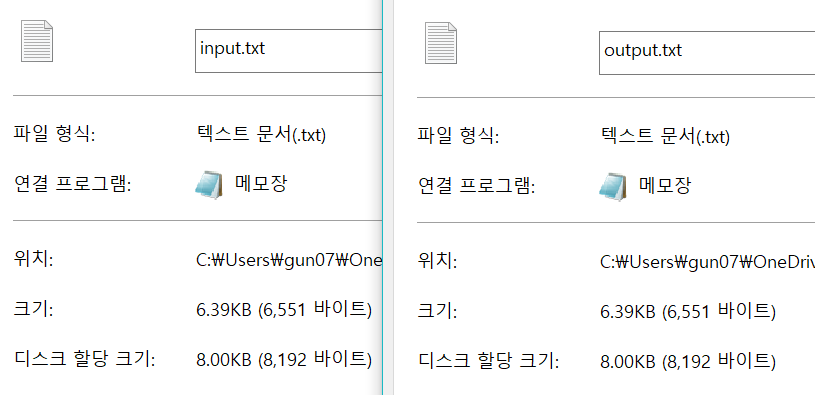
허프만 코딩은 가변길이 압축 알고리즘으로, 문자열에 자주 등장하는 문자를 짧게, 자주 등장하지 않는 문자를 길게 인코딩 한다. 이를 수행하기 위해 가장 먼저 전체 txt파일에 등장하는 각 문자의 수와 전체 문자의 수를 세어 등장 확률을 계산하였다. 문자와 등장 확률을 허프만 트리의 노드에 저장하고, 노드들을 우선순위 큐에 삽입하여 정렬하였다. 정렬된 노드를 순서대로 연결하여 허프만 트리를 생성하고, 완성된 허프만 트리를 root 노드에서부터 순회하며 인코딩 된 codeword를 얻었다. 문자를 key로 이용하여 codeword를 검색하기 위해 문자와 codeword의 길이, codeword내용을 map 컨테이너에 삽입하여 이용하였다. 이것을 바탕으로 허프만 테이블을 작성한다. input txt파일의 내용을 다시 처음부터 한 글자씩 읽고, 해당하는 codeword를 작성하여 압축파일을 생성하였다. 디코딩의 경우 위의 과정을 반대로 시행한다. 허프만 테이블의 정보를 읽어 map 컨테이너에 삽입한다. 이때 사용하는 map 컨테이너는 codeword내용을 key로 한다. 압축파일의 bit를 읽어 key에 해당하는 문자를 txt파일로 출력하는 과정을 반복한다.

과제를 진행하며 이진파일의 입출력 과정에서 어려움을 겪었다. 허프만 코딩의 경우 인코딩 길이가 문자의 등장 빈도수에 따라 다르게 나타난다. 그러나 파일에 값을 쓰거나 읽을 때 byte단위로 동작하기 때문에 한 번의 쓰기 동작에 정보의 일부만 적을 수 있거나, 두 번의 읽기를 통해서 한 문자에 대한 codeword를 읽어올 수 있었다. 이 문제는 파일에 쓰거나 읽을 값을 여러 byte크기의 버퍼에 저장하였다가 1byte씩 쓰거나 읽는 방법으로 해결하였다. bit shift연산과 bit and를 이용한 masking을 통해 버퍼에서 필요한 부분을 추출하여 사용할 수 있었다.

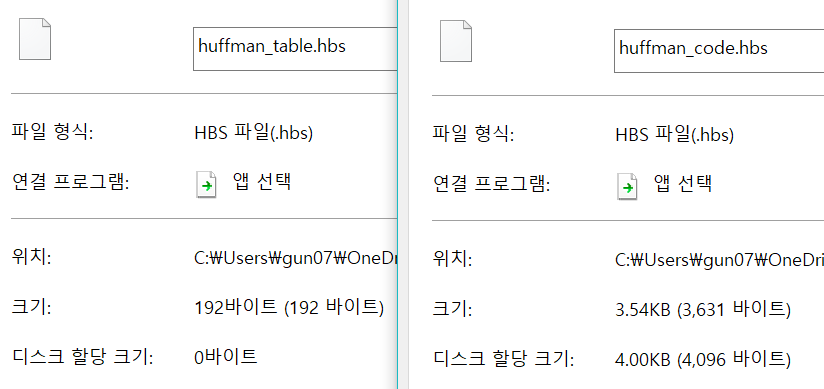
1. 실험 및 검증 결과



txt파일을 Huffman code로 인코딩했을 때 위와 같이 나타났다.



압축 이전 파일과 재생성한 파일의 크기가 같은 것을 확인할 수 있다.



Huffman\_table.hbs와 Huffman\_code.hbs의 파일 크기는 위와 같이 나타났다.

압축률은 41.64%이다.

1. 고찰

과제를 수행하는 과정에서 다양한 stl 자료구조를 이용하였다. 우선순위 큐를 이번 과제에서 처음 이용해보았는데 우선순위를 지정하는 구조체를 오버로딩하여 사용한 것이 기억에 남는다. 이산수학 강의를 수강하며 허프만 알고리즘을 배웠었는데, 직접 프로그램을 작성하며 예상 밖의 문제점을 확인할 수 있었다. 특히 파일 입출력에 txt파일을 주로 이용하고 binary파일을 이용한 적이 없어서 버퍼를 이용한다는 발상을 해낸 것에 오랜 시간이 걸린 것 같다.

1. 참고

우선순위 큐, cppreference.com

<https://en.cppreference.com/w/cpp/container/priority_queue>

구조체 priority queue를 사용하기

https://unluckyjung.github.io/cpp/2020/06/27/Priority\_Queue\_Compare/