SWE2007: System Software Experiment2 (Fall 2018)

Programming Assignment #1

Due: October 17, 11:59 PM

1. Introduction

과제를 통해 File I/O 및 Data structure사용에 익숙해지도록 한다.

2. Specification

2.1 Index Builder

주어진 성서 창세기를 이용하여 단어별로 index를 만든다.

- 입력 파일은 http://www.stewartonbibleschool.org.uk/bible/text/에서 Genesis를 다운로드 받을 수 있다.
- 입력 파일에서 장:절: 이 포함된 라인에 대해서만 index를 만들고, 장:절: 이 포함되지 않은 라인은 index 및 검색에 대해서 제외한다.
- 성서 이름은 입력 파일에서 확장자를 제외한 이름으로 한다.
  예를 들어, genesis.txt 파일의 성서 이름은 genesis로 한다.
- Index 생성 시간, 탐색 시간, 크기를 고려하여 data structure를 디자인 한다.

Index 마다 단어, 나타난 장/절, 횟수, 절 내에서의 위치 등을 관리한다.

Index는 파일로 저장한다.

Bonus

1) Implementation(10%)

   Buffering을 이용한 I/O를 구현한다.

   코드 내에서 lseek을 사용하지 않고, read / write의 length가 항상 1보다 커야 인정.

2) Runtime(10%)

   남들보다 runtime 시간을 줄여 보너스 점수를 얻는다.

   Hint : Hash 알고리즘을 이용하거나 Tree를 구현해 index를 관리하도록 한다.

2.2 Index Printer

Index builder에서 생성된 index 파일을 읽어 들여 인덱스 된 내용을 정해진 형식에 따라 출력한다

- 출력 파일의 첫 부분에는 성서 이름: 장 수, 총 절 수, 총 인덱스 수, 총 워드 수를 출력해야 한다.
- 이 후에는, 매 index 마다 다음의 포맷으로 출력한다.
- 단어: 총 출현 횟수, 장:절:위치, 장:절:위치, ...
- 단어를 sorting하여 출력할 필요는 없다.

3. Restriction

과제는 본인이 직접 설치한 리눅스 환경에서 수행하고, 실행한 화면을 캡처하여 추가한다.

예제 코드에 include된 것 이외의 library는 사용하지 않는다.

파일의 입출력은 open(), close(), read(), write(), lseek() 등의 system call 중 필요한 것을 선택하여 사용해야 한다.

단어의 대, 소문자는 구별할 필요가 없으며 영문 알파벳, – (하이픈), ' (어퍼스트로피)를 제외한 문자는 단어에서 제외하여 인덱스에 추가한다.

예를 들어, index로 관리되는 단어들은 다음과 같다.
god, and, adam, brother's, priests', kirjath-arba, sons'
genesis.txt 뿐만 아니라 http://www.stewartonbibleschool.org.uk/bible/text/에 있는 임의의 파일에 대해서도 동작해야 한다.

4. Hand in instructions

과제 및 스크린샷은 제출시 "학번.tar.gz" 파일로 압축하여 제출한다. (예: 2012345678.tar.gz)
과제는 기본적으로 아이캠프서에 제출하고,
기한이 지나 제출이 안 된다면 tkroh0198@gmail.com이나 hsewan@gmail.com으로 아래와 같은 형식을 맞춰 보낸다.
[SWE2007] PA1, 학번, 이름

5. Logistics

과제 제출 결과는 아이캠퍼스에서 확인할 수 있다.
과제 제출 시간은 메일 도착시간을 기준으로 하며, 기한 이후엔 10%씩, 최대 60%까지 감점 될 수 있다.
과제를 compile할 때, "-Wall –W" 옵션 기준으로 warning이나 error가 전혀 없어야 한다. Warning 하나 당 1%씩, 최대 10%까지 감점될 수 있다.

```
$ ./indexBP genesis.txt
** Index Builder : Start       **         // index builder 수행시간이 usec 단위로 나타남
Elapsed Time  : 96043(usec)
** Index Builder : End         **         // indexBP를 실행하면 index file인 genesis_index와 output file인
                                          // genesis_ouput이 생성됨
** Index Printer : Start       **
** Index Printer : End         **

$ ./cat genesis_output
genesis: 50 1533 2510 38260               // 성서: genesis, 장수: 50, 절수: 1533, Index수: 2510, 단어수: 38260
leaves: 1, 3:7:94                         // 단어: 총 개수, 장:절:위치
ephraim's: 3, 48:14:58, 48:17:153, 50:23:15
hadar: 2, 25:15:0, 36:39:43               // 위치는 절 내에서 단어가 나타나는 위치로, Byte 단위로 한다.
depart: 2, 13:9:149, 49:10:22
parted: 1, 2:10:73
…

$
```