CSE 2011 자료구조와 실습

설계문제 #2: 단일 검색어 기반 단어 검색 엔진 설계

단원	그래프	난이도	초급 /중급 /고급
참여인원	1명 /4주 (man/month)	마감	6월 14일 오후 5시

1. 과제 개요

(1) 정의

두 번째 과제는 단일 검색어 기반 단어 검색 엔진을 설계하는 것입니다. '.txt' 확장자를 가진 ASCII 코드의 문서 파일에서 문자를 읽어 자료구조를 이용해 저장하여 사용자가 원하는 단어를 입력하면 미리 만들어 놓은 자료구조에서 해당 단어를 검색하고 그 단어의 파일의 출현 빈도와 위치를 출력해주는 단어 검색 엔진을 설계합니다.

(2) 필요성

탐색은 일상 생활 뿐만 아니라 컴퓨터에서 가장 많이 이루어지는 작업 중 하나입니다. 탐색은 컴퓨터 프로그램에서 가장 많이 사용되는 작업이고 또 가장 시간을 효율적으로 수행해야 되는 작업입니다. 이번 설계 문제에서는 지금까지 수업을 통하여 숙지한 트리, 그래프, 해싱과 같은 자료구조를 이용하여 검색 횟수를 최소화 하며 메모리를 절약할 수 있는 단어 검색 엔진을 설계 합니다.

2. 과제 목표 및 주요 내용

(1) 과제목표

- 트리와 해싱을 이용한 단어 검색 엔진 설계 및 구현
- 텍스트 문서(txt) 내부의 단어 검색이 가능한 검색 엔진을 구현
- 단어가 파일 내에 출현한 빈도와 라인 위치, 근사화된 시간복잡도 출력 구현

(2) 주요 내용

- ASCII 코드로 내용이 이루어진, 확장자가 ".txt"인 하나의 문서파일의 문자들을 읽어온 후, 그 정보를 적당한 자료구조를 이용해 저장합니다. 이 후 사용자가 찾고자 하는 단어를 입력하면, 해당 단어를 미리 저장해놓은 자료구조를 활용해 검색한 후 그 단어의 파일 내 출현 빈도와 위치들(Line number, Word order)을 출력해주는 단어 검색 엔진을 설계합니

다. 어떤 자료구조 및 알고리즘을 사용하는가는 설계자에 의해 자유롭게 선택할 수 있습니다. 단, 설계과제의 채점기준으로 결과물로서 제출될 소프트웨어의 '효율성'을 체크하므로 가급적 연산횟수가 작아지도록 자료구조 및 알고리즘을 개선하도록 합니다.

예) 단어 및 단어에 대한 라인 정보를 자료구조 내부에 저장

<table>
<tr><th>검색할 텍스트</th><th>내부 저장 모양</th></tr>
<tr><td>Line 1: i i i you me
Line 2: me me you me
Line 3: and me</td><td>
<table>
<tr><th rowspan="2">단어</th><th colspan="2">노드 내용</th></tr>
<tr><th>라인 넘버</th><th>빈번도</th></tr>
<tr><td rowspan="3">me</td><td>1</td><td>1</td></tr>
<tr><td>2</td><td>3</td></tr>
<tr><td>3</td><td>1</td></tr>
<tr><td>i</td><td>1</td><td>3</td></tr>
<tr><td>and</td><td>3</td><td>1</td></tr>
<tr><td rowspan="2">you</td><td>1</td><td>1</td></tr>
<tr><td>2</td><td>1</td></tr>
</table>
</td></tr>
</table>

(3) 제한 사항

- 검색 단어에서 대문자와 소문자는 같은 것으로 취급하며, 검색단어를 포함하는 문서 내의 단어는 검색결과에서 제외시킨다. (즉, "computer"으로 검색할 경우 검색 대상 문서 내에 있는 "computerengineering"은 "computer"을 포함하고는 있지만 별개의 단어로 인식해 검색하지 않는 것으로 한다.)
- 검색대상이 될 문서파일은 ".txt" 확장자를 가진 하나의 파일로서 그 파일의 내용은 ASCII코드의 영문자 (a~z, A~Z), 개행문자 ('\n', '\t', ' ', …) 및 문장부호 (", ', ?, …)들로 이루어져 있으며, 파일 내의 총 line 수 및 단어 수는 미리 정해져 있지 않다. 각 line의 구분은 '\n (new line)' 또는 '. (마침표)' 로 이루어지며 파일 내에 각 line마다 line의 위치(Line number)가 따로 표기되어 있지는 않다.
- 출력 시 사용자가 원하는 단어가 파일 내에 출현한 빈도(Frequency)와 출현한 line의 위치들(Line number) 과 그 단어가 그 line에서 몇 번째 단어 순서에 위치하는지 (Word order)를 출력한다.

3. 리포트 제출 시 유의 사항

(1) 제출 목록

- 리포트(*.doc or *.hwp), 소스 코드(*.c or *.cpp)

(2) 제출 방식

- 기한 내 e-class 등록

(3) 리포트 포함 내용

- 프로그램에서 사용된 자료구조와 알고리즘 설명
- 프로그램에서 사용된 함수 설명
- 프로그램 테스트 결과 캡처 파일 및 기타 내용 설명에 필요하다고 판단되는 내용 포함