Programming Assignment #2:

Collecting web page titles

Due: 28th Oct. (Tue), 11:59 PM

1. Introduction

이번에선 인터넷 웹 사이트의 title을 수집하는 프로그램을 제작한다. 이 과제의 주요 목표는 프로세스 생성과 시그널을 익히는 것이다.

2. Problem specification

파일 `collect.c` 을 생성하여, `collect()` 함수를 구현한다. `collect()` 함수의 원형은 다음과 같다:

```
void collect (void);
```

표준 입력(`stdin`)으로부터 웹 페이지 주소나 명령을 입력 받아 처리하는 프로그램을 작성한다. 프로그램의 주요 기능은 `wget` 프로그램을 이용하여 입력된 웹 페이지를 다운로드 받은 다음, 해당 웹 페이지 파일에서 **웹 사이트의 title**을 수집하는 것이다.

처리할 문자열은 웹 페이지 주소이나 명령, 두 종류가 있다. 웹 페이지 주소는 2.1절에 지정된 대로, 명령은 2.2절에 지정된 대로, 이 두 가지가 아닌 경우는 잘못된 입력이고, 2.3절에 지정된 대로 처리한다.

2.1. Web pages

웹 페이지 주소는, "http://" 혹은 "https://" 로 시작되는 문자열을 뜻한다. 이 경우 수행할 작업은 다음과 같다:

- 자식 프로세스를 생성하여, 해당 페이지를 `wget` 프로그램을 이용해 다운로드한다.
  - 다운로드 경로는 `wget`의 option으로 조절할 수 있는데, <현재 처리 순서> 에 해당하는 이름으로 저장한다.
  - 예를 들어, 현재 2번째 명령을 처리한다면, 저장될 파일의 이름은 `"2"` 이다.
  - `wget` 이 표준 출력이나 에러로 출력을 생성하지 않도록 한다.
- 해당 파일을 열어서 title을 찾는다. 만약 title이 두 개 이상 존재한다면, 제일 먼저 나오는 것을 메모리에 수집한다. Title이 존재하지 않는 경우는 없다.
- 수집한 title 문자열을 다음 형태에 맞춰 표준 출력으로 출력한다.
- <현재 처리 순서>는 1부터 시작하며, 명령을 처리할 때마다 1씩 증가한다.

```
Sequence #>"Domain Name":"title"

/* printf format */
printf("%d>%s:%s\n", seq, domain, title);

/* Possible inputs & outputs */
http://www.skku.edu/eng
1>skku.edu:Sungkyunkwan University (SKKU)

http://www.skku.edu
2>skku.edu:성균관대학교

http://icc.skku.ac.kr
3>skku.ac.kr:성균관대학교 정보통신대학
```

여기서 Domain Name은, 어떤 웹 주소의 <대표 도메인>을 의미한다. 직접적으로 이야기하면, 도메인 주소의 최우측(TLD) 주소가 ".kr" 이 아닐 경우, 바로 하위 주소까지가 <대표 도메인>이고, 도메인 주소의 최우측이 ".kr" 일 경우, 바로 하위 주소의 하위 주소까지가 <대표 도메인>이라 간주한다. (참고: 3.3. 절)

2.2. Commands

처리할 명령은 다음과 같이 4 종류가 있다: `print`, `stat`, `load`, `quit`

`print`: 인자로 도메인을 하나 받아서, 지금까지 수집한 해당 도메인의 title중 가장 긴 것을 출력
`stat`: 지금까지 수집한 title 개수를 출력
`load`: 인자로 파일 이름을 하나 받아서, 해당 파일에서 열어 주소나 명령을 읽어 처리한다.
즉, 표준 입력에서 읽어서 수행할 일을 해당 파일을 읽어서 수행한다고 생각해도 좋다.
`quit`: 프로그램을 종료한다

2.2.1. `print`

`print` 명령은 어떤 특정 도메인의 title 중 가장 긴 title만 출력하는데, 위의 예시를 바탕으로 보면 skku.edu란 도메인에 대해서 가장 긴 값은 Sungkyunkwan University (SKKU)가 될 것이다. 여기서 가장 길다의 의미는 [문자열을 구성하기 위해 더 많은 byte를 필요로 한다]란 뜻으로, 한글의 경우 한 문자를 표현하는데 1바이트 이상이 필요하니, 영어와 한글의 경우 눈에 보이는 것과 실제 길이는 다를 수 있다. 만약 길이가 같은 title이 수집된다면, 먼저 수집한 것을 출력한다.

- 인자로 요청되는 주소는 항상 <대표 도메인>이다.
- 출력 형태는 웹 페이지를 처리하는 것과 동일하다.
- 인자로 수집되지 않은 도메인을 요청 받으면, Not Available이란 문자열을 출력한다.

```
print skku.edu
4>skku.edu:Sungkyunkwan University (SKKU)

print google.com
5>Not Available
```

2.2.2. stat

지금까지 에러 없이 성공적으로 수집한 title 개수를 출력한다. 이 개수는 같은 페이지를 여러 번 요청한 것/같은 title을 중복적으로 수집한 것을 모두 포함한다. 출력 형태는 다음과 같다.

```
Sequence #>"Count" titles

/* printf format */
printf("%d>%d titles\n", seq, count);

/* Possible input & output */
stat
6>3 titles
```

2.2.3. load

인자로 받은 파일을 열어서, 해당 파일을 처리한다. 해당 파일이 EOF를 반환할 때까지 표준 입력 대신 해당 파일을 읽어서 title을 수집한다. 명령을 처리할 때마다 <현재 처리 순서> 또한 증가시킨다. load 명령의 인자로 넘어오는 파일은 항상 존재하며, load 중간에 load 또한 일어날 수 있다.

```
$ cat command.txt
http://skku.edu
print skku.edu
stat
================================

/* Possible inputs & outputs */
load command.txt
7>skku.edu:성균관대학교
8>skku.edu:Sungkyunkwan University (SKKU)
9>4 titles
```

2.2.4. quit

본 함수의 실행을 종료하고 반환한다. 만약 자식 프로세스가 존재한다면, SIGKILL 시그널을 통해 모든 자식 프로세스를 강제로 종료시킨다.

2.3. Errors

본 과제의 에러는 크게 두 종류가 있는데, 첫째로 웹 주소나 명령이 아닌 다른 입력값이 들어온 경우와, 둘째로 `wget` 프로세스에서 에러가 발생한 경우이다.

- 웹 주소나 명령이 아닌 경우, 해당 입력을 무시하고, <현재 처리 순서>도 유지한다.
- `wget` 프로세스에서 에러가 발생하면, Error occurred! 란 문자열을 다음과 같이 출력한다.

```
/* Possible input & output */
http://ThisSiteDoesNotReallyExist.com/haha
10>Error occurred!
```

2.4. Child processes

자식 프로세스를 생성하여 파일을 다운로드 받기 때문에, 크게 두 가지 방식 중 하나를 선택할 수 있다.

- 항상 한 개의 자식 프로세스만 만든다.
- 여러 개의 자식 프로세스를 생성해 병렬적으로 동시에 더 많은 명령을 처리한다.

어떤 특정 시간에 한 개의 명령(하나의 자식 프로세스)만 처리하면 상대적인 구현 난이도는 쉬워지겠지만 더 오랜 시간이 걸릴 것이고, 여러 자식 프로세스를 만든다면 더 복잡한 대신 더 빠른 처리가 가능할 것이다.

만약 다음과 같이 입력을 받아 위에서부터 순서대로 3개의 자식 프로세스를 만들었다고 가정하자.

```
http://www.skku.edu/eng
http://www.skku.edu
http://icc.skku.ac.kr
```

만약 둘째, 셋째, 첫째 사이트 순으로 title이 수집되었다면, 다음과 같이 출력해도 좋다.

```
http://www.skku.edu/eng
http://www.skku.edu
http://icc.skku.ac.kr

2>skku.edu:성균관대학교
3>skku.ac.kr:성균관대학교 정보통신대학
1>skku.edu:Sungkyunkwan University (SKKU)
```

여기서 <현재 처리 순서>는 입력 순서대로 유지되어야 한다

`print`나 `stat` 명령의 경우, 항상 동기화가 되어야 한다. 예를 들어, `"print skku.edu"` 명령이 들어온다면, 여태까지 요청된 모든 skku.edu에 대한 title을 수집한 다음에 `print`가 수행되어야 한다. `stat` 명령의 경우, 여태까지 요청된 모든 사이트에 대한 결과가 나온 다음 수행되어야 한다.

본 과제를 수행함에 있어 다양한 최적화 요소가 있으니, 이번 과제를 **정상적으로 구현한 사람 중, 프로그램 수행 시간이 빠른 상위 10명은 본 과제의 코드 점수에 대해 20%의 보너스 점수를 받는다.**

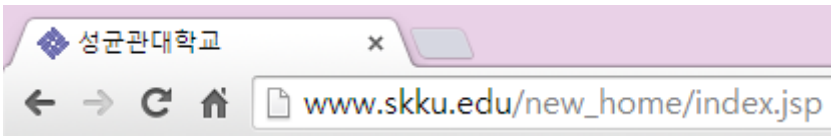
2.5. Signals

- 이 프로그램은 `SIGINT` 시그널을 수신해도 종료되지 않는다.
- `SIGUSR1` 시그널을 수신하면, `stat` 명령이 들어왔다고 간주하고 동일한 행동을 한다.

3. Background

3.1. Web page title

웹 페이지의 title이란 웹 페이지의 제목으로 보통 웹 브라우저의 최상단에 표시된다. 아래 그림은 구글 크롬 브라우저로 www.skku.edu 를 연 것으로, "성균관대학교" 가 해당 페이지의 title임을 알 수 있다.



웹 페이지는 HTML 문서로 구성되는데, HTML 문서는 여러 tag, attribute 등을 이용하여 어떤 웹 사이트를 표현한다. 웹 페이지 제목의 경우 HTML 문서 상 title이란 element에 기술되는데, HTML 파일을 보면 <title>과 </title>이란 태그 사이에 title이 적힌다. 따라서 위 그림의 경우 해당 HTML 파일을 분석하면 <title>성균관대학교</title> 란 부분을 확인할 수 있다.
(참고: http://en.wikipedia.org/wiki/HTML_element)

3.2. wget

`wget`은 주어진 URL로부터 데이터를 받아 저장하는 프로그램이다. http, https, ftp 등의 프로토콜을 사용할 수 있다. 예를 들어, 다음 명령을 내리면, CSL 서버에 있는 signal 교안을 현재 디렉토리에 signal.pdf라는 이름으로 저장한다.

```
$ wget -O signal.pdf http://csl.skku.edu/uploads/SWE2007F14/5-signals.pdf
```
(참고: `$ man wget`)

3.3. Domain

도메인 네임이란 인터넷의 특정 자원을 문자열로 연결하는 주소로, 사람의 편의를 위해 사용한다. 예를 들어, 115.145.179.100 이란 IP 주소를 외우기보단, csl.skku.edu란 주소를 외우는 것이 훨씬 쉽다. 다만 누군가 이 도메인 주소를 IP 주소로 변경해줘야 하고 이런 시스템을 DNS (Domain Name System) 라고 부른다.

특정 인터넷 주소를 봤을 때, 프로토콜(e.g., `http://`)을 제외하고 왼쪽에서부터 처음 '/'가 등장할 때까지가 도메인 주소이다. 3.1.의 그림에 나온 주소인 www.skku.edu/new_home/index.jsp 란 주소를 보면, 처음으로 '/' 문자가 등장할 때까지 추리면 www.skku.edu 이고, 이 문자열이 도메인 주소이다.

도메인은 '.' 문자를 분리점으로 삼아 오른쪽에서 시작해 왼쪽으로 해석해나가는 구조인데, 첫째 문자열을 top level domain (TLD)라고 부른다. TLD는 사이트의 성격을 기준으로 정해지거나(global TLD; gTLD), 각 국가를 의미하는 TLD를 가질 수도 있다. (country code TLD; ccTLD)
gTLD가 사용될 경우 둘째 문자열부터 해당 서브 도메인을 소유한 기관에서 관리하게 된다. ccTLD가 이용될 경우 보통 각 국가에서 관리하는 둘째 문자열 (second level domain; SLD)이 사이트의 성격을 표현하고, 셋째 문자열부터 해당 서브 도메인을 소유한 기관에서 관리하게 된다.

우리 학교가 보유한 도메인으로 skku.edu 와 skku.ac.kr 가 있는데, 전자의 경우 `".edu"`라는 gTLD이기 때문에 둘째 문자열부터 우리 학교가 소유한다. 후자의 경우, 한국을 의미하는 `".kr"`이란 ccTLD이기 때문에 둘째 문자열로 교육기관을 의미하는 `".ac"` ccSLD가 사용되었고, 셋째 문자열부터 우리 학교가 소유하게 된다.

도메인을 소유하면, 소유자가 자유롭게 하위 도메인을 생성할 수 있기 때문에 www.skku.edu 혹은 csl.skku.edu 와 같은 서브도메인은 우리 학교에서 자체적으로 생성하여 운영할 수 있다.
(참고: http://en.wikipedia.org/wiki/Domain_name and
http://en.wikipedia.org/wiki/Domain_Name_System)

4. Restrictions

- 파일을 다루기 위해 리눅스 시스템 콜을 이용한다.
- 프로세스를 다루기 위해, `wait`, `exec`계열 시스템 콜/라이브러리 함수를 이용한다.
- 시그널을 전송/수신하기 위해 리눅스 시스템 콜을 이용한다.
- `malloc()`, `calloc()`, `free()` 함수를 제외한 다른 라이브러리 함수는 **사용할 수 없다**. 필요하다면, `collect.c` 안에 직접 구현하여 사용한다.
  - 쉬운 디버깅을 하기 위해 라이브러리 함수를 사용할 수 있다. 다만, 제출한 파일엔 라이브러리 함수가 포함되지 않아야 한다. (주석 등으로 처리해도 괜찮음)
- `collect.c` 파일 안에 `main()`를 담지 않는다.
- 어떤 자원을 동적으로 할당 받았다면, 프로그램 종료 전에 반드시 해제해야 한다.
  - 여기서 자원이란 파일이나 메모리를 뜻한다.

5. Hand in instruction

- 작성한 코드 상단의 주석에 이름과 학번을 작성한다.
- `collect.c` 파일의 이름을 "`학번.c`" 로 바꾼다. (e.g., `2008311920.c`)
- 본 과제 수행 시 구현 방법과 디자인을 설명하는 보고서를 PDF 포맷으로 작성하여 "`학번.pdf`" 이란 이름을 붙인다. (가능하면 PDF가 가장 좋지만, 대중적인 문서 포맷은 다른 포맷도 괜찮음)
- 과제를 제출하기 위해 [wooyeong at csl.skku.edu] 주소로 메일을 보낸다. 메일 전송 시 전술한 코드 파일과 문서 파일을 각각 첨부하고(압축하지 말 것!), 메일 제목은 다음과 같이 명명한다:

  [SWE2007] PA #2, 학번, 이름

6. Logistics

- 본 과제는 혼자 수행한다.
- 제출 상태는 과목 홈페이지 http://csl.skku.edu/SWE2007F14/Projects 에 즉각적으로 공지될 것이다.
- 과제 제출 시간은 메일 도착 시간을 기준으로 하며, 과제를 지연 제출하면 기한 직후부터 매 8시간마다 점수를 10%씩 추가로 감점한다.
- 다른 사람의 과제를 copy할 경우, 개입한 사람 전부 해당 과제에 대해 0점 처리되고, 교수님께 보고되며, **성적 산정에 불이익이 있다**. 또한, copy가 두 차례 이상 적발될 경우 **F 학점이 부여될 수 있다**.

Have fun!

정우영, 담당 조교
컴퓨터시스템연구실