[CSEG437/CSE5437] 수치 컴퓨팅 및 GPU 프로그래밍

숙제 4

담당교수: 서강대학교 컴퓨터공학과 임 인 성

2018년 6월 4일

제출 마감: 1. 데이터: 6월 8일 (금요일) 오후 8시 정각, **2. 프로그램 및 설명/분석 자료:** 6월 16일 (토요일) 오후 8시 정각
제출물 및 제출 방법: 1. 자신이 만든 데이터, **2.** 작성한 프로그램/수행 결과 출력물을 포함하는 보고서 형식의 설명 및 분석 자료 등.

[목적] 이번 숙제는 데이터 마이닝 (data mining) 분야에서 문서 추출 기법의 하나인 Latent Semantic Indexing (LSI) 방법의 일부를 구현해봄을 목적으로 한다. 특히 http://www.netlib.org에 공개되어 있는 코드 중, 주어진 행렬에 대한 Singular Value Decomposition을 수행해주는 적절한 FORTRAN 코드를 찾아 자신의 문제 해결에 활용하여 본다.

1. 우선 실험 데이터를 구축하기 위하여, 영문 웹사이트에서 적절한 문서를 찾아 아래와 같은 형식의 문서를 각자 8개씩 만들어 **6월 10일 오후 8시까지** 조교에게 이메일로 제출하라. 제출된 데이터는 조교가 취합하여 정리 후 신속히 여러분들에게 제공해야 하므로, 제출 기한 이후에는 데이터를 받지 않을 예정임 (데이터 생성 및 제출은 전체 점수의 30%의 비중을 차지함).

   - 우선 아래와 같이 각 수강생들에게 IT 기술 관련 주제가 할당되어 있다. (마지막 쪽 표 참조)
     - **스마트폰 기술:** smartphone, processor, apple, iphone, cellphone, samsung, music, galaxy, battery, resolution, cell, app, store, resolution, retina, phone, mobile, operating, 3g, nokia, system, memory, google, blackberry, camera, itunes, wireless, blackbarry gps, video, game, pda, software, internet, android, wifi, $\cdots$
     - **모바일 디바이스 기술:** mobile, device, handheld, pda, tab, smartphone, operating system, internet, drone, apple, ipad, samsung, galaxy, tab, operating, software, system, camera, game, netbook, dvd, notebook, gps, processor, memory, resolution, microsoft, display, music, resolution, wireless, laptop, RAM, PC, wifi, lcd, app, store, wireless, $\cdots$
     - **게임기/게임 SW 기술:** game, nintendo, psp, 3d, battery, vulkan, network, wii, starcraft, OpenGL, mobile, smartphone, nvidia, amd, ati, app, store, intel, processor, graphics, video, card, display, 3ds, stereo, software, rpg, race, geforce, engine, xbox, opengl, portable, directx, entertainment, freeware, microsoft, sony, flash, screen, ps3, console, wifi, wireless, $\cdots$
   - 자신에게 할당된 주제와 관련한 문서 8개를 찾아, 각각을 자신에게 할당된 문서 번호 `***`를 사용하여 `doc***.txt`라는 파일에 텍스트 형태로 저장하라 (`***`은 580에서 1011까지의 번호를 가짐). 가급적 자신의 주제를 중심으로 다른 주제와도 관련이 있는 키워드를 뽑아낼 수 있는 문서를 찾을 것. (예: 스마트폰 상에서의 3D 게임)
   - 다음 문서의 내용을 잘 설명해줄 수 있는 키워드를 각 문서마다 5-10개 정도 선택하여, 하나의 파일에 저장하라. (예들 들어 고윤민의 경우 키워드 파일의 이름이 `key580_587.txt` 이어야 함). 이 파일은 아래와 같은 형식으로 8줄의 키워드 리스트로 구성되어야 함

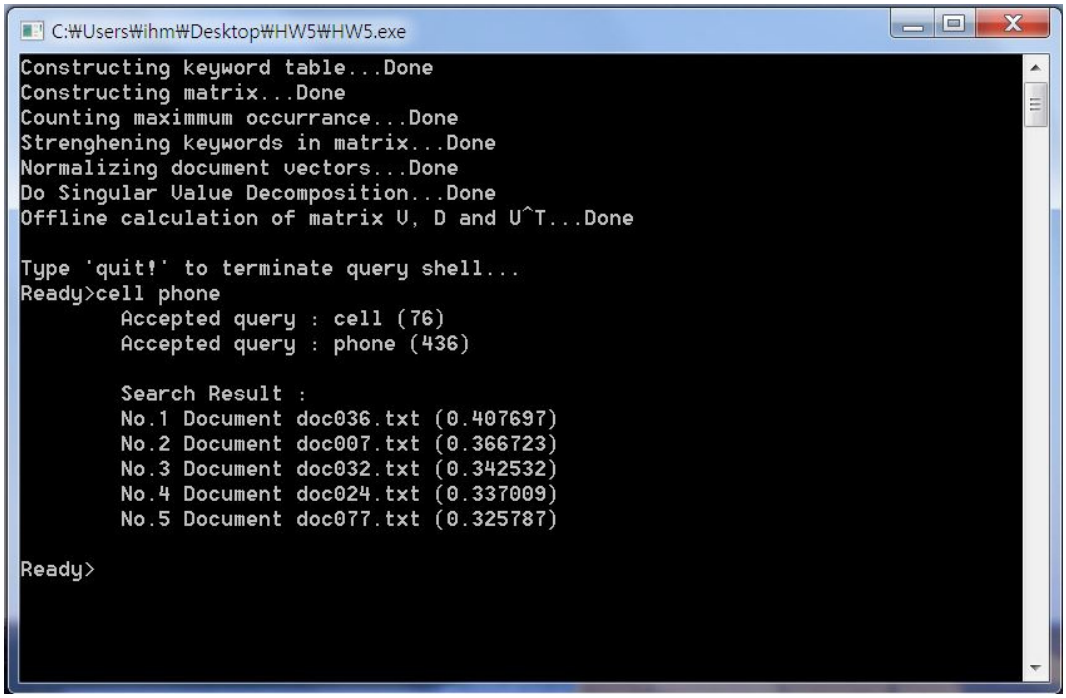

Figure 1: 문서 추출 S/W 사용자 인터페이스 예

```
580 : iphone apple itunes camera
581 : samsung galaxy resolution display processor
 ⋮
```

이때 키워드로 IT 기술과 관련하여 널리 쓰이는 단어를 선택하고(각 주제마다 나열한 추천 키워드를 가급적 많이 사용), 소문자로 가급적 사전에 있는 원형을 사용하라.

- 각자 찾은 8개의 문서와 1개의 키워드 파일을 한 개의 zip 파일로 묶어 조교에게 이메일로 제출할 것.

2. 일단 모든 데이터가 수집되면, 조교가 취합하여 알파벳 순서로 word 리스트와 document-keyword 리스트 데이터를 각자의 이메일 주소로 발송할 예정임.

3. 채점은 올해에 취합한 데이터 (약 1000개의 문서를 포함하는)를 사용하여 진행을 할 예정이며, 새 데이터 배포 이전에 자신의 프로그램을 작성할 수 있도록 기존의 소규모 데이터를 이메일로 제공할 예정임.

4. 다음 반드시 수업 시간에 설명한 LSI 방법 이론에 기반을 두어 문서 추출 S/W를 개발하라. 이 S/W는 대략적으로 다음과 같은 방식으로 작동해야 한다.

   (a) 문서 추출 프로그램을 수행시킨 후,

   (b) 대기 상태에서 word 리스트에 있는 단어를 적절히 사용하여 질의를 하면 (단어는 반드시 키워드 리스트에서 선택을 하나 정렬된 순서일 필요는 없음),

   (c) LSI 방법을 사용하여 최대 5개까지의 문서번호와 점수에 대한 랭킹 결과를 출력하고,

   (d) 다시 대기 상태로 들어감.

5. 자신이 개발한 S/W의 사용자 인터페이스를 사용자 입장에서 편리하게 사용할 수 있도록 설계한 후, 그에 대한 사용법 메뉴얼을 보고서 파일에 추가하라. 그림 1은 사용자 인터페이스 대한 한 예를 보여주고 있음.

6. S/W 개발이 끝난 후, 다양한 방식으로 질의 실험을 해본 후, 그에 대한 실험/분석 내용을 보고서 파일에 기술하라. 최소한 다음과 같은 내용을 포함해야 한다.

   - 자신의 문서 추출 S/W가 제대로 작동하고 있는가? 그렇다는 사실을 어떻게 보일 수 있는가?
   - LSI 방법의 구현 과정에서 SVD 문제를 풀기 위하여 어떤 공개 코드를 어떻게 사용했는가?
   - 수업 시간에 설명한 LSI 방법은 w차원의 문제를 k 차원의 문제로 축소하여 적은 비용으로 문제를 풀려고 하고 있는데, 서로 다른 크기의 k 값에 대해 속도 및 정확성 등의 관점에서 본 S/W의 특성이 어떻게 변하는가?
   - 자신이 문서 추출과 관련하여 본 S/W의 질을 높이기 위하여 어떤 특별한 노력을 기울였는가?
   - 기타

[주의] (a) 본인이 본 숙제에서 요구하는 것 중 무엇을 어디까지 완성했는지 보고서 파일에 정확히 기술하라.

(b) 위의 요구 사항 외에 본인이 추가적으로 구현한 내용이 있을 경우 보고서 파일에 정확히 기술하라 (추가 점수를 부여할 수 있음).

(c) 기본적으로 프로그래밍은 C/C++ 언어를 사용하고, SVD 문제 해결에는 반드시 공개 FORTRAN 함수를 그대로 사용하라 (f2c와 같은 변환 툴을 사용하지 말것).

(d) 조교는 자신이 만든 질의 리스트들을 사용하여 문서가 잘 추출되는지 확일할 예정임.

(e) 제출 방식은 조교가 과목 게시판에 공고할 예정임.

(f) 제출 파일에서 바이러스 발견 시 **만점 X (-1)**이고, 다른 사람의 숙제를 복사할 경우 관련된 사람 모두에 대하여 **만점 X (-10)**임.