Computational Thinking

2017-2학기

8장-데이터 다루기

박종혁 교수 (컴퓨터공학과)

jhpark1@seoultech.ac.kr

http://www.parkjonghyuk.net

학습개요

- 8장 데이터 다루기
  1. 폰 노이만 구조
  2. 스프레드시트 프로그램 (스킵)
  3. 문장처리
  4. 패턴
- 조별과제
  - 다음 주 강의 시간 발표

학습목표

- 폰 노이만 원리를 이해한다.
- 이차원 데이터 배치와 표 모양 정보 검색의 개념을 이해한다.
- 이산 함수와 표와의 관계를 이해한다.
- 문장 처리에 계산식이 어떻게 사용되는지를 이해한다.
- 계산식이 패턴을 정의하고 처리하는 데 어떻게 사용되는지를 이해한다.

❖ 프로그램 고정형 구조

- 컴퓨터의 명령어들이 하드웨어로 구현됨
- 예) 탁상용 계산기

❖ 프로그램 내장형 구조

- 폰 노이만 구조라고 불림
- 컴퓨터를 처리유닛, 제어유닛, 외부/내부 메모리로 구성함
- 컴퓨터는 컴퓨터 명령어와 데이터를 같은 방식으로 다룸
- 데이터는 프로그램에 의해 변경될 수 있으며 프로그램 역시 명령어를 결과값으로 만들 수 있음.
- 프로그램이 프로그램을 만듦
  - 자기 수정 코드

❖ **문자열 이란?**

- 키보드에서 입력되는 문자들 및 기타 기호들로 구성된 데이터

❖ **대부분의 프로그래밍 언어에서 문자열 자료를 지원함**

- 최신 프로그래밍 언어에서는 문자열 처리 연산을 지원하고 있음.

❖ **문자열 상수**

- "" 로 둘러싸인 데이터
- 예) "Hello World► ► ►" "여러분 안녕하세요!!!"

❖ **문자열 합치기**
- + 연산을 사용

❖ **문자열 길이 구하기**
- length 함수 사용

❖ **문자열 인덱스 구하기**
- 문자열[index]

```
Name = "Dongguk University"
Location = "Seoul, Korea"
School_name = Name + ", " + Level
```

"Dongguk University, Seoul, Korea"

```
School_name.length
```

31

```
School_name[3]
```

"Dongguk University, Seoul, Korea"
01234 567 89.....

❖ 부분 문자열 구하기

- 문자열.subString[index1, index2]

```
School_name.subString( 4, 8 )
```

“guk U”

“Dongguk University, Seoul, Korea”

그림 8.28 이메일 주소는 로컬(사용자 이름)과 도메인(호스트 사이트)의 두 파트로 나누어져 있다.

bob.dylan 이메일 주소의 이메일 분석

```
address ← readAddressFromUser()
username ← address.substring(0, 9)
hostsite ← address.substring(10,28)
extension ← address.substring(25,28)
```

그림 8.29 이메일 주소에서 정보를 추출하기

유럽식: day, month, year

미국식: month, day, year

유럽식 날짜 표기를 미국식으로 전환하기

```
edate ← readDateFromUser()
day ← edate.substring(0, 2)
month ← edate.substring(3, 5)
year ← edate.substring(6, 10)
adate ← month + "/" + day + "/" + year
```

그림 8.31 유럽식 날짜를 미국식으로 전환하기

- ❖ 어떤 문자열은 가지고 있지만 다른 문자열을 가지고 있지 않는 특성
- ❖ 정규식으로 정의됨

자주 쓰이는 문장 패턴	
문자열 상수	**패턴 종류**
“123-45-6789”	미국 주민번호
“999-999-9999”	전화번호
“beatles@sub.edu”	이메일 주소
“04/13/1963”	날짜
KMOX	라디오 방송국 약자
“1234-1234-1234-1234”	신용카드 번호
“XOXO”	허그와 키스
PG-13	미국 영화협회 영화 등급

그림 8.32 자주 쓰이는 문장 패턴

❖ 정규식은 문자열의 특성을 표현하는 방법

- 정규식을 작성함에 의해 문자열의 특성을 정의할 수 있다.
- 정규식을 통해 수 많은 문서에서 특정한 특성을 가지는 문자열을 찾을 수 있다.
  - 학생 출석부에서 특정한 성을 가지는 학생들의 숫자를 찾기
  - 문서에서 특정한 단어를 가지는 문서들 찾기
  - 자료 입력 창을 만들 때 요구되는 형식에 어긋나지 않도록 입력을 제한하기

❖ 정규식은 문자열의 패턴을 표현하는 방법

❖ 주민등록번호 문자열을 나타내는 정규식을 작성해 보자

- 980901-1234567

❖ 패턴의 기본 규칙

패턴 작성의 기본 규칙

1) (몇 개의 특수문자들을 제외하고) 모든 단일 문자는 패턴이다. 그 문자는 해당 문자를 표현한다.
2) 마침표(.)는 패턴이다. 마침표는 단일 문자를 표현하며 마침표는 특수문자이다.
3) A와 B가 모두 패턴이라면, 다음의 내용도 특수문자이다.
   a) 순차패턴으로 A다음에 패턴 B가 나온다는 것을 나타낸다.
   b) 교대패턴으로 패턴 A 또는 패턴 B 중 하나가 나타남을 나타낸다.
   c) 패턴 A를 그룹짓는 것을 말한다. 괄호는 특수문자이다.

그림 8.33 패턴 작성에 대한 기본 규칙

❖ 패턴 내에서 반복이 허용되는 횟수를 지정

반복 규칙

4) 만약 A가 패턴이라면
   a) A*: A가 0번 이상 반복된다는 것을 나타낸다. *는 특수문자이다.
   b) A+ : A가 1번 이상 반복된다는 것을 나타낸다. +는 특수문자이다.
   c) A? : A가 0번 또는 1번 나온다는 것을 나타낸다. ?는 특수문자이다.
   d) A가 정확히 m번 반복됨을 나타낸다. {}는 특수문자이다.
   e) A가 최소 m번 최대한 n번 반복된다는 것을 나타낸다.
   f) A가 최소 m번 반복된다는 것을 나타낸다.

그림 8.35 패턴 작성을 위한 반복 규칙

❖ 반복 규칙을 사용한 미국 사회보장번호 패턴

(0|1|2|3|4|5|6|7|8|9){3}- (0|1|2|3|4|5|6|7|8|9){2}- (0|1|2|3|4|5|6|7|8|9){4}

❖ 문자들을 나타내는 규칙

문자 클래스 규칙

5) 대괄호([])는 문자 클래스를 표시한다. 괄호안의 각 문자는 문자 클래스에 속하게 된다. 문자 클래스 안의 대쉬(-)는 문자의 범위를 나타낸다.
   a) \d : 하나의 숫자를 표시한다.
   b) \D : 숫자가 아닌 문자를 표시한다.
   c) \w : 단어를 구성하는 비단어 문자를 표시한다. (a-z, A-Z, 0-9).
   d) \W : 단어를 구성하지 않는 문자를 표시한다.
   e) \s : 스페이스, 탭, 줄바꿈과 같은 공백 문자를 표시한다.
   f) \S : 공백문자가 아닌 문자를 표시한다.

그림 8.37 패턴 작성에 사용되는 문자 클래스 규칙

❖ MPAA 영화 등급

- 문법 교대 (|) 사용
- G|PG|PG-13|R|NC-17

❖ 미국 사회보장번호

- 세 개의 숫자 대쉬(-) 두 개의 숫자 대쉬(-) 네 개의 숫자로 구성
- 숫자는 10개로 문법교대를 통해 표현됨
  (0|!|2|3|4|5|6|7|8|9)

미국 사회보장번호 패턴

```
(0|1|2|3|4|5|6|7|8|9) (0|1|2|3|4|5|6|7|8|9) (0|1|2|3|4|5|6|7|8|9)- (0|1|2|3|4|5|6|7|8|9)
(0|1|2|3|4|5|6|7|8|9)-(0|1|2|3|4|5|6|7|8|9) (0|1|2|3|4|5|6|7|8|9) (0|1|2|3|4|5|6|7|8|9)
(0|1|2|3|4|5|6|7|8|9)
```

그림 8.34 기본 규칙을 사용한 미국 사회보장번호 패턴

❖ DNA 염기서열

- 염기서열은 A,C,G,T로 구성되는 긴 길이의 문자열
- 작은 부분의 유전자 코드(예를 들면 CATT)를 찾기 위해 DNA 염기서열 분석
- AC로 시작 AG로 끝나고 GAA를 중간에 담고 있는 염기서열의 패턴
  - AC.*GAA.*AG

DNA 데이터베이스에서 검색하기

CAGACTTTCAGAACTGTCAGTTCCCCGGATTTTACCCATCACATTTTGCTACTACTTTC
TACTACTATATACTTTTCCAATTTCATACGGGTACTATTATCCATACTCTACTATTAC

그림 8.39 CATT 염기서열 찾아보기

❖ 웹에서 원하는 문장이 있는 문서들을 패턴을 사용하여 검색

❖ 대문자만으로 구성된 단어를 찾는 패턴

- 잘못된 패턴: [A-Z]+
- 정확한 패턴: ₩W([A-Z]+₩W){3,}

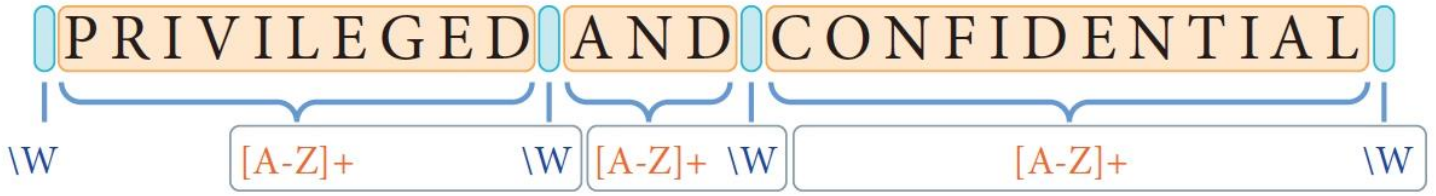

그림 8.45 패턴([A−Z]+)3,과 주어진 문장이 일치함을 보인다.

- 발표방식: ppt 자료 활용 2개 조 각각 10분 발표 (질의응답 포함)

- 문제 1
  - 우리가 많이 접하는 영상에 대한 처리는 많은 계산 작업을 필요로 하고 보조 하드웨어를 통해 성능향상을 이룰 수 있다. 이러한 처리 시스템은 프로그램-고정형 구조를 통해 만들어 진다. 이 구조는 많은 계산 작업에는 유용하게 사용하지만 문장 처리나 메시지 교환같은 작업을 위한 프로그램은 만들 수 없다. 이러한 프로그램-고정형 구조에 대해 책에 나와있는 예시를 제외하고 실생활에 적용되고 있는 예를 조사하고 설명해보자. 또한 고정형 구조의 반대인 내장형 구조에 대한 예도 조사하여라.

- 문제 2
  - 패턴은 문장 자료들을 처리하는 매우 쓸모 있는 기법이다. 패턴에 의해 정의된 집단에 속하는지 아닌지를 결정하는 방법으로 주민등록번호, 전화번호, 웹 검색 등 실생활에서 다방면으로 사용되어지고 있다. 책 8.4에 나와있는 예시를 제외하고 실생활에서 사용되어지는 패턴응용 예와 추후 패턴방식이 다른 응용처에 활용될 수 있는 예에 대해 조사하고 생각하여라.