실 세계 데이터셋 추출 및 모델링(1)

학습내용

1. 실 세계 데이터셋 수집, 가공, 모델링 과정
2. 실 세계 데이터셋 실습 준비

학습목표

- 실 세계 데이터셋 수집 과정을 설명할 수 있다.
- 실 세계 데이터셋 가공 과정을 설명할 수 있다.
- 실 세계 데이터셋 모델링 과정을 설명할 수 있다.
- 실 세계 데이터셋 실습에 대한 준비를 마칠 수 있다.


세종대학교
Data Up

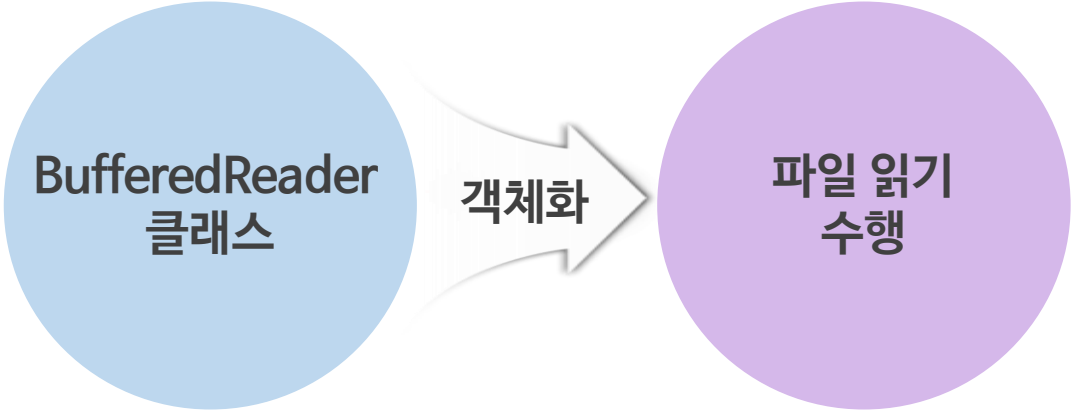
실 세계 데이터셋 프로그래밍 언어

Java

객체지향 프로그래밍 언어

개념을 추상화 해놓은 클래스의
객체들을 조합하여 목적 달성

파일 읽기



파일 읽기

한 줄 읽기를 반복하여 전체 파일 읽기 수행

Files

File	Description
email-EuAll.txt.gz	Email network of a large European Research Institution

```
BufferedReader br =
    new BufferedReader(new FileReader(
        "C:₩₩Users₩₩Sejong₩₩DA₩₩email.txt"));

String line = br.readLine();
```

출처 #01 snap.stanford.edu

생성자

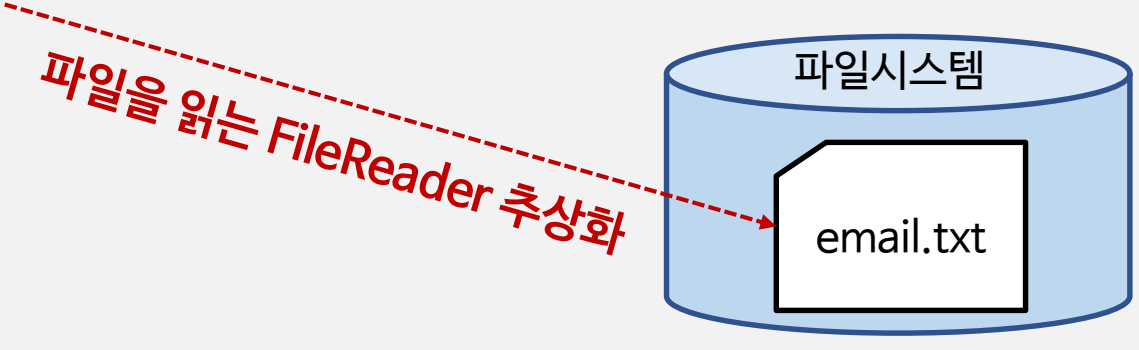
Class FileReader()

문자로 이루어진 파일을
읽을 수 있는 클래스

FileNotFoundException

파일을 찾지 못하였을 때 발생

```
FileReader f =
        new FileReader("C:₩₩Users₩₩Sejong₩₩DA₩₩email.txt");
```

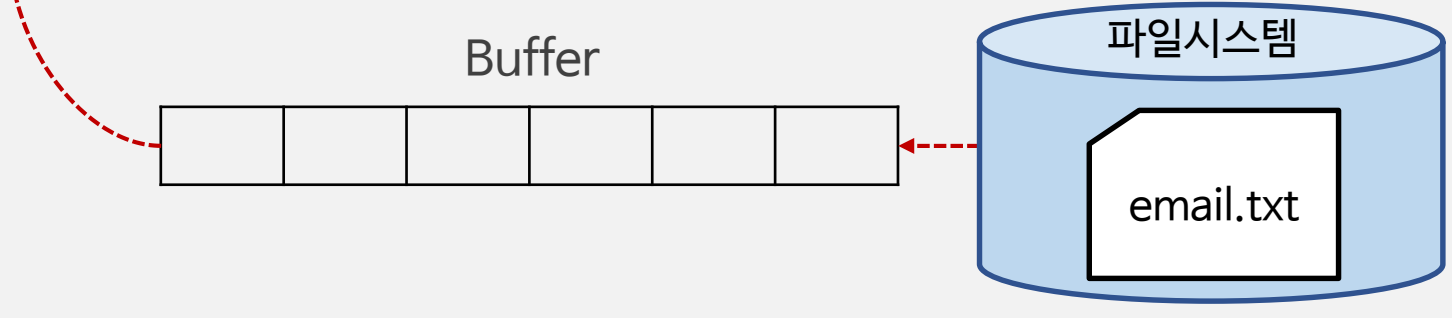


생성자

Class BufferedReader(Reader in)

버퍼를 이용하여 효율적으로 문자열을 읽는 클래스

```
BufferedReader br = new BufferedReader(
new FileReader("C:₩₩Users₩₩Sejong₩₩DA₩₩email.txt"));
```



연산자(BufferedReader)

String ReadLine()

텍스트의 한 줄을 읽어내는 방식

IOException

I/O 에러 발생 시 발생

```
String line;
line = br.readLine();
```

```
# Directed graph (each unordered…
# Email network of a large Europe…
# Nodes: 265214 Edges: 420045
```

연산자(BufferedReader)

String ReadLine()

텍스트의 한 줄을 읽어내는 방식

IOException

I/O 에러 발생 시 발생

```
String line;
line = br.readLine()
line = br.readLine()
```

```
# Directed graph (each unordered…
# Email network of a large Europe…
# Nodes: 265214 Edges: 420045
```

연산자(BufferedReader)

String ReadLine()

텍스트의 한 줄을 읽어내는 방식

IOException

I/O 에러 발생 시 발생

```
String line;
line = br.readLine()
line = br.readLine()
line = br.readLine()
```

```
# Directed graph (each unordered…
# Email network of a large Europe…
# Nodes: 265214 Edges: 420045
```

연산자(BufferedReader)

String ReadLine()

텍스트의 한 줄을 읽어내는 방식
만약 텍스트의 끝에 도달하면
Null을 반환함

IOException

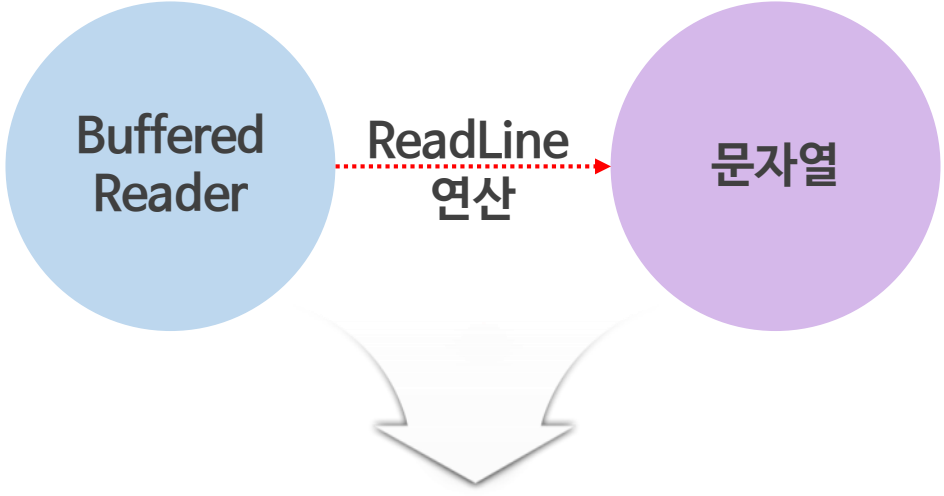
I/O 에러 발생 시 발생

```
String line;
line = br.readLine()
line = br.readLine()
line = br.readLine()
line = br.readLine()
```

```
# Directed graph (each unordered…
# Email network of a large Europe…
# Nodes: 265214 Edges: 420045
```

Null을 반복의 종료 조건으로 이용

데이터 파싱



프로그래밍 언어가
이해할 수 있는 형태

email.txt

```
# Directed graph (each unordered…
# Email network of a large Europe…
# Nodes: 265214 Edges: 420045
# FromNodeId    ToNodeId
0	1
0	4
0	5
0	8
0	11
0	20
0	48
…
```

0이 1에게 메일을 보낸 이벤트

#으로 시작하는 데이터 설명 무시
boolean startsWith("#")

TAB(₩t)으로 문자열을 분리
String [] split("₩t")

배열의 요소를 정수형(Integer)로 변환
static int parseInt("1")

연산자(String)

boolean startsWith(String prefix)

문자열이 특정 prefix로 시작하면 true를 반환함

```
String line = br.readLine();
    # Directed graph (each unordered pair of nodes is …
if(line.startsWith("#"))
   continue;
```

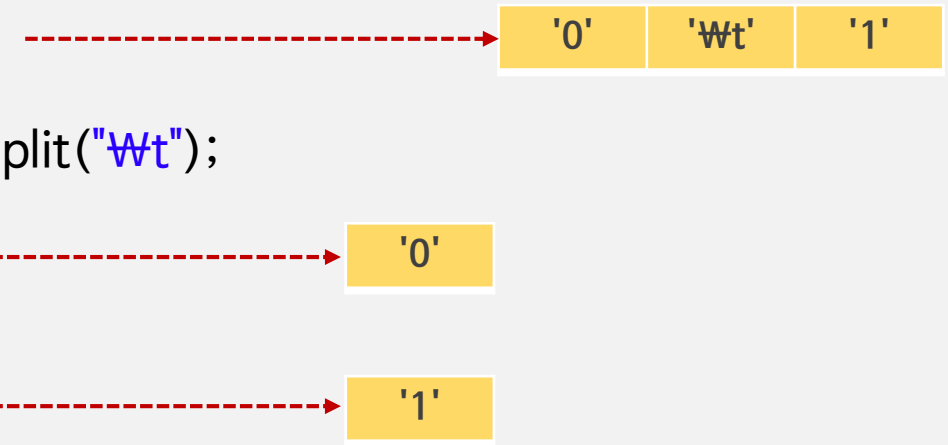
→ true

연산자(String)

String[] split(String regex)

문자열을 주어진 정규표현식에 맞춰 분리함

```
String line = "0    1";
String [] splited =line.split("₩t");
splited [0]
splited [1]
```



연산자(String)

static int parseInt(String s)

숫자형의 문자열 ‘s’ 를 10진수의 정수형으로 반환함

```
int num1 = Integer.parseInt("3");
int num2 = Integer.parseInt("5");
num1 + num2 ----------------------> 8

int num = Integer.parseInt("a");
Exception in thread "main" java.lang.NumberFormatException : For input string: "a"
```

예외사항

NumberFormatException

숫자형의 문자열이 아닐 때 발생함

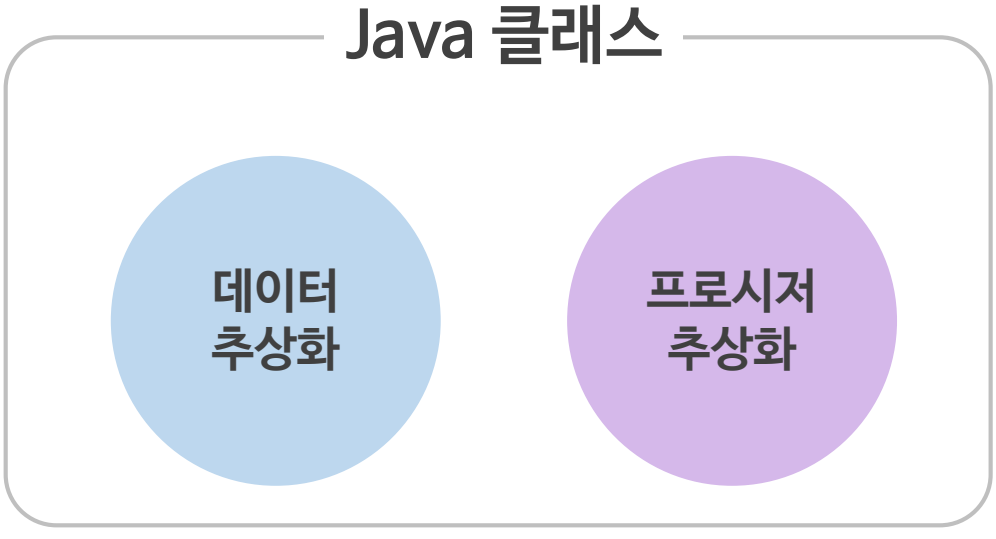
```
int num1 = Integer.parseInt("3");
int num2 = Integer.parseInt("5");
num1 + num2                                    8

int num = Integer.parseInt("a");
Exception in thread "main" java.lang.NumberFormatException: For input string: "a"
```

이메일을 보낸 사실

=

이벤트 추상화


Java 클래스
데이터
추상화
프로시저
추상화

email.txt

```
# Directed graph (each unordered…
# Email network of a large Europe…
# Nodes: 265214 Edges: 420045
# FromNodeId        ToNodeId
0           1
0           4
0           5
0           8
0           11
0           20
0           48
…
```

0이 1에게 메일을 보낸 이벤트

```
public class Email {
 //데이터 추상화
 int from;    // 보낸 사람
 int to;      // 받는 사람
 //프로시저 추상화
 public Email(int from, int to) {
   this.from = from; this.to = to;
 }
 public String toString() {
   return from + "->" + to;
 }
}
```

인스턴스화

```
Email e1 = new Email(0,1);
System.out.println(e1);
```

콘솔 추상화 매개변수를 콘솔에 출력

콘솔 결과 : 0 → 1

Remind

실 세계 데이터셋
수집, 가공, 모델링 과정의 개념

자료 출처

#01 snap, 2021, URL : http://snap.stanford.edu/data/email-EuAll.html