**텍스트 데이터 분석**

**사례연구**

**평가자료**

**-데이터분류 출력-**

수업명 :java,python 기반의 AI 활용 응용

소프트웨어 개발자 양성과정

제출일 : 2020. 11. 24

작성자 : 유승화

□프로젝트 개요

|  |  |  |
| --- | --- | --- |
| **구분** | **내용** | **비고** |
| **제목** | 엑셀 데이터 변환하여 분류하기 |  |
| **프로젝트**  **작성기간** | ~2020.11.24 |  |
| **조원 구성** | 유승화 |  |
| **프로젝트 내용** | 액셀 문서를 csv 형식으로 저장하고 그것을 java를 통해 이름, 전화번호, 이메일, 주민번호로 분류한다 |  |
| **주요기능** | csv파일을 java로 읽어온다 |  |
| csv 파일의 내용을 split을 이용하여 분리한다 |  |
| 분리된 데이터를 정규식을 통해 분석한다 |  |
| **학습내용** | csv 파일 읽기 |  |
| Split() |  |
| 정규식 |  |

□최종 보고서

|  |  |  |
| --- | --- | --- |
| **구분** | **내용** | **비고** |
| **제목** | 데이터 분류하여 출력하기 |  |
| **프로젝트 내용** | 액셀 문서를 csv 형식으로 저장하고 그것을 java를 통해 이름, 전화번호, 이메일, 주민번호로 분류한다 |  |
| **완료내용** | csv파일을 java로 읽어온다 | 프로젝트  개요와 동일 |
| csv 파일의 내용을 split을 이용하여 분리한다 | 프로젝트  개요와 동일 |
| 분리된 데이터를 정규식을 통해 분석한다 | 프로젝트  개요와 동일 |
| **프로젝트 활동 후기** | | |
| **좋았던 점** | 정리되어 있지 않은 데이터를 프로그래밍을 통해 간단히 정리하여 원하는 대로 분류하고 저장할 수 있었다 |  |
| **아쉬웠던 점 및**  **어려웠던 점** | 한글로 된 데이터가 깨져서 따로 UTF-8 설정을 해주어야 했다 |  |
| 인코딩 설정을 했는데도 데이터가 깨져서 csv 데이터의 처음 부분에 영어로 된 데이터를 임시로 삽입하여 해결했다. |  |
| 외부 데이터 csv 파일을 읽어오는 것이 생소하여 다소 헤맸다. |  |
| **학습 효과** | Split, ArrayList 등을 통해 데이터를 자유롭게 정제하고 추가, 삭제할 수 있다.  정규식을 통해 원하는 데이터가 어떤 특징을 갖고 있는지 분석할 수 있다. |  |

□프로젝트 세부 사항 분석 및 구조 설계

|  |  |  |  |
| --- | --- | --- | --- |
| **구분** | **주요 기능/구문** | **내용** | **입출력** |
| **1** | FileInputStream  InputStreamReader  BufferedReader | 파일의 내용을 읽어온다. |  |
| **2** | 데이터.split(“구분할 기호”) | 원하는 글자나 기호를 기준으로 데이터를 잘라낸다. |  |
| **3** | Pattern.matches("정규식", 데이터) | 정규식으로 데이터의 내용을 분석한다. |  |
| **4** | ArrayList | 원하는 데이터를 추가하여 저장할 수 있다. |  |

□작성 과정, 실행결과

|  |  |  |
| --- | --- | --- |
| **구현 내용** | **결과** | **비고** |
| 한 줄로 된 데이터에서 “,”를 기준으로 분리한다 | 완료 |  |
| 분리된 데이터들이 어떤 특징을 가지고 있는지 분석하여 그 특성과 일치하는 ArrayList에 추가한다 | 완료 |  |

□실행결과-1

테이블이(가) 표시된 사진

자동 생성된 설명

이름을 모아 출력한 것

텍스트이(가) 표시된 사진

자동 생성된 설명

전화번호를 모아 출력한 것

텍스트이(가) 표시된 사진

자동 생성된 설명

이메일주소를 모아 출력한 것

텍스트이(가) 표시된 사진

자동 생성된 설명

주민번호를 모아 출력한 것

□소스코드-1

package com.human.ex;

import java.io.\*;

import java.util.ArrayList;

import java.util.regex.Pattern;

public class Test1124 {

public static void main(String[] args) {

try {

String fileLocation = "B:\\javaBackupYSH\\1124시험\\testdata.csv";

FileInputStream inputStream = new FileInputStream(fileLocation);

InputStreamReader reader = new InputStreamReader(inputStream,"UTF-8");

BufferedReader br = new BufferedReader(reader);

String line = "";

int row =0;

ArrayList<String> name = new ArrayList<String>();

ArrayList<String> email = new ArrayList<String>();

ArrayList<String> phone = new ArrayList<String>();

ArrayList<String> jumin = new ArrayList<String>();

while ((line = br.readLine()) != null) {

// -1 은 마지막 "," 이후 빈 공백도 읽기 위함

String[] data = line.split(",", -1);

for(int i=0;i<data.length;i++) {

//이름판별

if(Pattern.matches("^[가-힣]\*$", data[i])) {

name.add(data[i]);

}

//전화번호판별

if(Pattern.matches("^01(?:0|1|[6-9])-(?:\\d{3}|\\d{4})-\\d{4}$", data[i])) {

phone.add(data[i]);

}

//이메일판별

if(Pattern.matches("\\w+@\\w+\\.\\w+(\\.\\w+)?", data[i])) {

email.add(data[i]);

}

//주민번호판별

if(Pattern.matches("\\d{6}\\-[1-4]\\d{6}", data[i])) {

jumin.add(data[i]);

}

}

row++;

}

br.close();

System.out.println("<이름>");

for(int i=0;i<name.size();i++) System.out.println(name.get(i));

System.out.println();

System.out.println("<전화번호>");

for(int i=0;i<phone.size();i++) System.out.println(phone.get(i));

System.out.println();

System.out.println("<이메일>");

for(int i=0;i<email.size();i++) System.out.println(email.get(i));

System.out.println();

System.out.println("<주민번호>");

for(int i=0;i<jumin.size();i++) System.out.println(jumin.get(i));

}

catch (FileNotFoundException e) {

e.printStackTrace();

}

catch (IOException e) {

e.printStackTrace();

}

}

}