**빅데이터**

**정적 웹데이터 크롤링**



**2016156003**

**소프트웨어 전공**

**김동혁**

**개요**

**목적 : 알바구직 페이지에서 광고나 중복 글들을 피해 원하는 알바자리를 구할 수 있게 하기 위한 프로그램**

**방법 : Crawl.py 실행 후 albaheaven.csv 파일 확인**

**분석대상 페이지 :** 알바천국(http://www.alba.co.kr/)

**크롤링대상 데이터 :**

* 근무지역
* 근무회사
* 근무시간
* 근무급여
* 글 올린 시간

**분석 조건 :**

* 경기도 시흥시를 기준으로 조회할 것’
* 100건이상 데이터를 불러올 것
* **월급 및 연봉으로 급여를 계산하는 곳은 제외할 것** 
  + (알바가 아닌 직원을 구하는 글인 경우가 대부분)
* **중복되는 글들은 필터링 할 것** 
  + (상단에 노출시키기 위해 똑 같은 글을 여러 번 작성하는 경우가 많음)

**사용 라이브러리**

from bs4 import BeautifulSoup

import urllib.request

import pandas as pd

import numpy as np

import re

추가설명 :

자료 수는 기본적으로 100개이상을 보장합니다.

하지만 항상 고정된 수의 데이터를 크롤링하지는 않습니다

기본적으로 상황에 따라 100 ~ 150개 사이의 데이터를 수집하게 됩니다.

**수집대상지역은 경기도 시흥시로 고정되어 있습니다.**

**csv파일은 같은 디렉토리 내 albaheaven.csv 으로 export합니다**