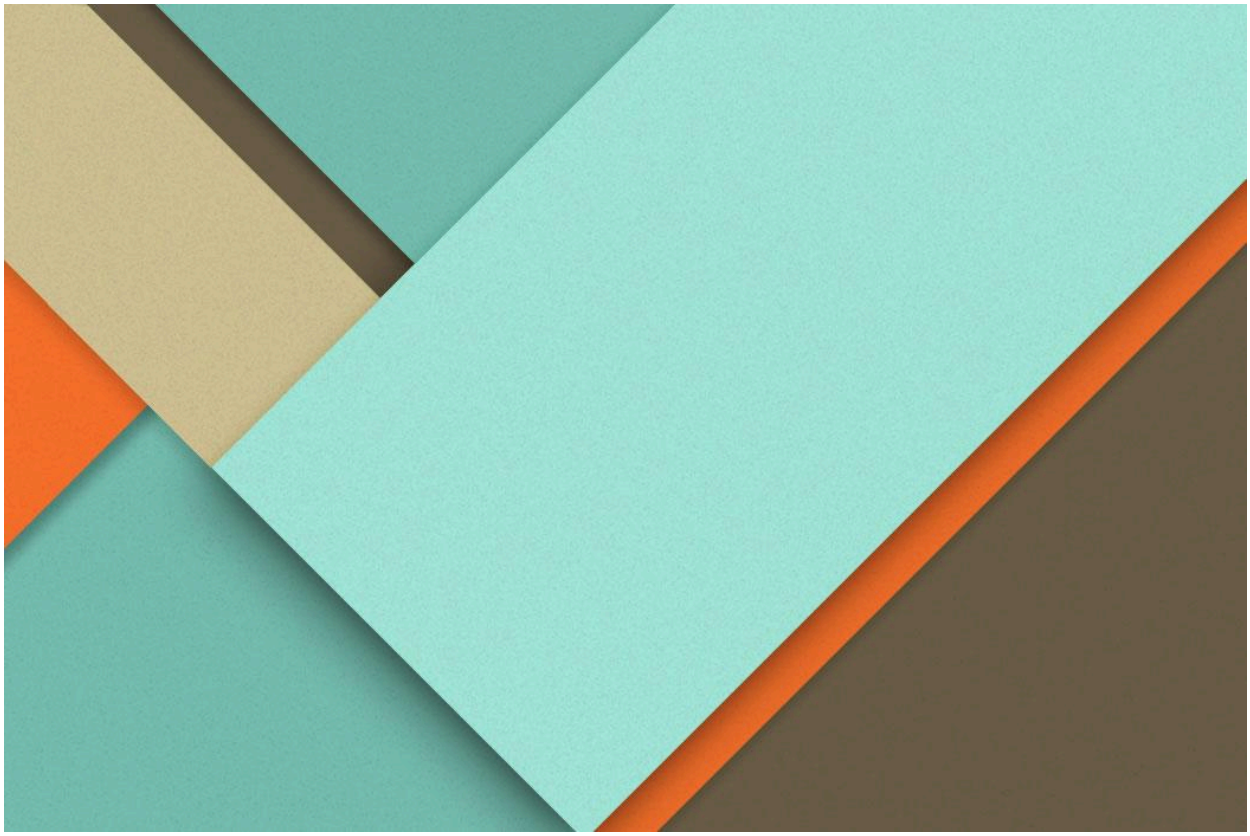

CarrotDB and Analysis

[Project GitHub]

2023年/08月 - 2024年/12月

이건희 (イ·ゴニ / LEE KUNHEE)

sacredcrawler@gmail.com (business)
knutvikings@gmail.com (private-not business)
Seoul, Korea

Overview

평소 데이터의 수집, 수집된 데이터의 흐름과 관리에 관심이 많다. 더불어 데이터를 어떻게 활용하면 좋을 지에 대해서도 고민하는 것을 좋아한다. 그래서 이 모두를 실행해 볼 수 있는 재미있을 프로젝트를 진행하게 되었다.

Milestones

1. 실제 온라인 환경에서 떠다니는 정보를 수집한다.
2. 수집된 정보를 유효화된 데이터로 정제한다.
3. 정제된 데이터를 프로젝트 목적에 맞게 활용한다.

Description

Python, ***C++***, ***SQL*** 을 메인 언어로 채택했으며, ***Python*** & ***C++*** 는 데이터 추출, 정제, 분석에 사용했고 ***SQL*** 은 데이터베이스 관리에 사용했다. 개발 환경으로는 **Google CoLab**, **Microsoft Visual Studio**, **SQLite** 를 이용했다.

Details and Goal

한국 내에서 가장 규모가 큰 중고 거래 시장 플랫폼은 '당근 마켓' 이다. 이 플랫폼에 매일 흐르고 있는 전국의 물건 현황 데이터를 자동화된 웹 스크래퍼를 통해 매일 수집한다. 수집된 데이터는 정제 과정을 거쳐 데이터베이스 파일에 저장되고, 해당 데이터베이스 파일은 **SQLite** 로 관리됨과 동시에 ***Python*** & ***C++*** 로 분석이 이루어진다.

이 프로젝트의 목표는, 실제 세계의 데이터를 직접 추출, 적재, 변형, 활용하는 경험을 쌓는 것과, 전체 과정 중 문제가 발생하면 어떤 방향으로 해결해 가면 좋을 지 숙고하는 능력을 기르는 것이 주(主)이다. 그 안에서의 목적, 즉 분석의 목적은 중고 거래 시장 플랫폼의 동향과 사회적 관계에 따른 구조적 특징을 파악하는 것이다.

Project Schema

I. Web Scraper

BeautifulSoup & **requests** 를 활용해 지역별로 분류된 각 웹 상의 특정 데이터를 반복적으로 끌어 모아 텍스트 파일로 내보낸다. 이후 17개의 텍스트 파일을 읽어들여 **KoNLPy**(한국어 자연어 처리 패키지)를 사용해 한국어 문장에 포함된 단어들만 추출한다. 이들은 명사로 구분되며, **Counter** 패키지를 거쳐 가장 많이 사용된 단어 10개를 리스트에 저장한다.

II. Database

리스트에 저장된 데이터는 **sqlite3**(***Python***) 를 통해 하나의 데이터베이스 파일(**.db**)로 저장되며, **SQLite**(응용 프로그램), **Microsoft Visual Studio**(**sqlite3** ***C++*** Interface) 로 접근하여 관리를 해 준다. 오직 ***Python*** 만으로 데이터베이스 관리가 가능하지만 다양한 인터페이스를 활용해 보고 싶어서 이들을 선택했다.

III. Analysis

여기에 텍스트를 입력하세요.

IV. Management

개발 환경의 전반적인 관리는 모두 수동으로 이루어졌다. **CoLab** 세션 관리, **Microsoft Visual Studio** 솔루션 관리, 그리고 **SQLite** 를 통한 데이터베이스 관리. 그리고 모든 것의 통합은 **GitHub** 상에서 개발 환경 간 연동으로 진행되었다.