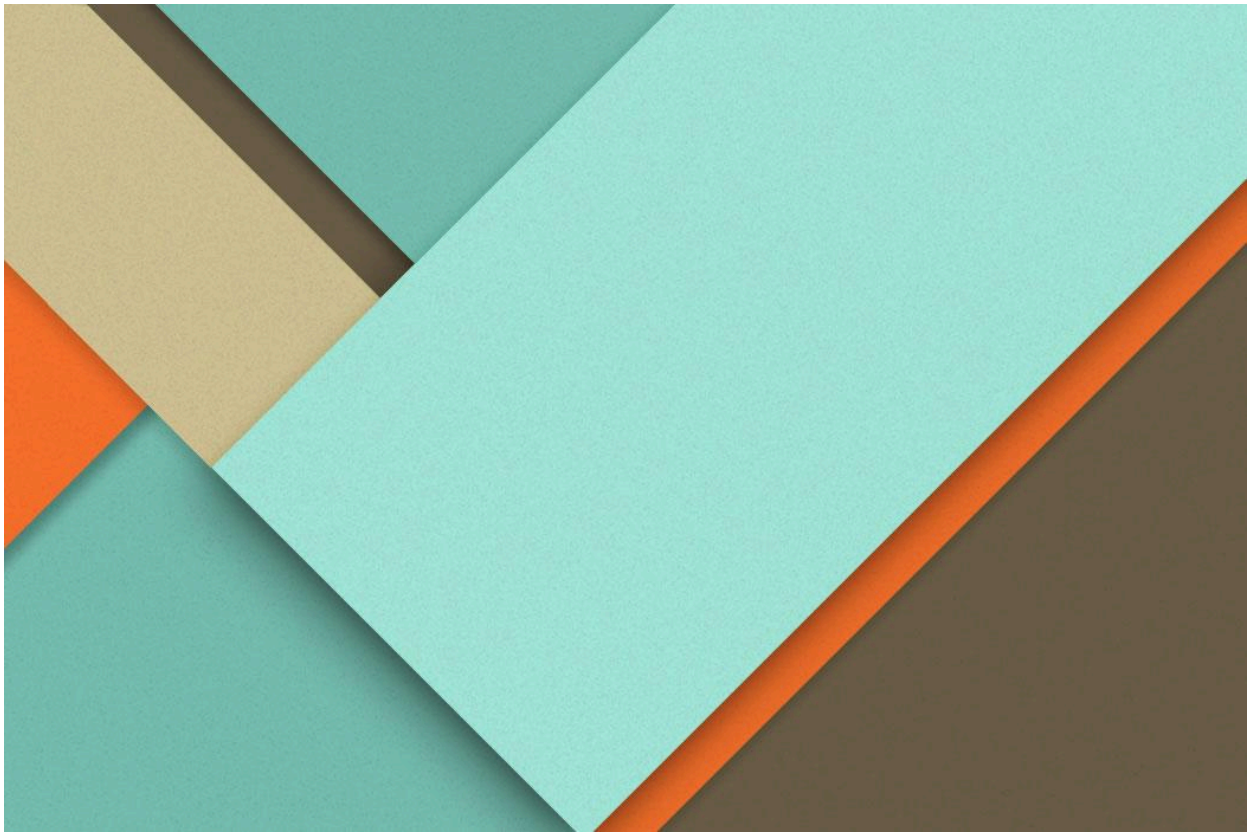

CarrotDB and Analysis

[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

日頃からデータの収集や、収集したデータの流れと管理に関心がある。さらに、データをどのように活用すればよいかについても考えることが好きだ。このすべてを実行できる面白いプロジェクトを進めることにした。

Milestones

1. 実際のオンライン環境から情報を収集する。
2. 収集した情報を、有効なデータに精製する。
3. 精製されたデータをプロジェクトの目的に合わせて活用する。

Description

Python、***C++***、***SQL***をメインの言語として採用。***Python***と***C++***はデータの抽出、精製、分析に使用し、***SQL***はデータベース管理に使用した。開発環境としては、**Google Colab**、**Microsoft Visual Studio**、**SQLite**を利用した。

Details and Goal

韓国国内で最も規模の大きい中古取引市場プラットフォームは「ダングンマーケット」である。このプラットフォームで日々流れている全国の商品の現状データを自動化されたウェブスクレイパーを使って毎日収集する。収集したデータは精製過程を経てデータベースファイルに保存され、そのファイルは**SQLite**で管理され、同時に***Python***と***C++***で分析が行われる。

このプロジェクトの主な目的は、実際の世界のデータを直接抽出、蓄積、変換、活用する経験を積むこと、そして全体の過程で問題が発生した場合にどのように解決していくかを考える能力を養うことである。その中での分析の目的は、中古取引市場プラットフォームの動向と、社会的関係による構造的特徴を把握することである。

Project Schema

I. Web Scraper

BeautifulSoupと**requests**を使用して、地域ごとに分類されたウェブ上の特定データを繰り返し収集し、テキストファイルとして出力する。その後、17個のテキストファイルを読み込み、**KoNLPy**（韓国語自然言語処理パッケージ）を使って韓国語の文章に含まれる単語を抽出する。これらは名詞として分類され、**Counter**パッケージを通じて最も多く使われた単語10個がリストに保存される。

II. Database

リストに保存されたデータは**sqlite3**（***Python***）を通じて1つのデータベースファイル（**.db**）として保存され、**SQLite**（アプリケーション）や**Microsoft Visual Studio**（**sqlite3** ***C++***インターフェース）でアクセスし、管理される。***Python***だけでもデータベース管理は可能だが、さまざまなインターフェースを試してみたかったため、これらを選択した。

III. Analysis

ここにテキストを入力してください。

IV. Management

開発環境の全般的な管理はすべて手動で行われた。**CoLab**セッション管理、**Microsoft Visual Studio**ソリューション管理、そして**SQLite**を通じたデータベース管理。すべての統合は**GitHub**上で開発環境間の連携によって行われた。