

Projektmenedzsment Terv

1. Követelményanalízis

A projekt célja egy webes scraper rendszer fejlesztése, amely képes weboldalakról szöveges tartalmat kinyerni, azokat adatbázisban tárolni, keresni bennük, valamint egy böngésző kiegészítőn keresztül is elérhetővé tenni a scraping funkciót. A rendszerhez frontend felület is tartozik, amelyen keresztül a felhasználó egyszerűen használhatja a szolgáltatásokat.

Főbb követelmények:

- Weboldalak szövegének letöltése és feldolgozása.
- Eredmények tárolása adatbázisban.
- Keresés URL vagy kulcsszó alapján.
- Eredmények megtekintése azonosító alapján.
- Frontend felület a funkciók eléréséhez.
- Firefox böngésző kiegészítő integráció.
- Naplózás a scraping műveletekről.
- Környezeti változók támogatása (adatbázis eléréshez).

2. Követelményspecifikáció

Funkcionális követelmények

- A rendszer képes legyen tetszőleges URL-t letölteni és a szöveget kinyerni.
- Az eredményeket MySQL adatbázisban tárolja.
- A felhasználó kereshet URL vagy kulcsszó alapján.
- A keresési eredményekből az adott rekord teljes szövege megtekinthető.
- A frontend felületen keresztül minden funkció elérhető.
- A böngésző kiegészítő egy kattintással elküldi az aktuális URL-t a backendnek scraping céljából.

Nem-funkcionális követelmények

- A rendszer REST API-t biztosít JSON válaszokkal.
- A rendszer naplózza a scraping eseményeket.
- A rendszer támogatja a környezeti változókból történő konfigurációt.
- A frontend reszponzív és felhasználóbarát legyen.
- A kód legyen jól dokumentált és karbantartható.

3. Részletes tervezés

Architektúra

- **Backend:** Python (FastAPI), adatbázis-kezelés (databases, aiomysql), scraping (BeautifulSoup, requests).
- **Frontend:** Statikus HTML, CSS, JavaScript.
- **Adatbázis:** MySQL, `scraped_text` tábla (id, url, text, time).

- **Böngésző kiegészítő:** Firefox extension (manifest.json, background.js).

Fő modulok

- **main.py:** API végpontok, statikus fájlok kiszolgálása.
- **database.py:** Adatbázis kapcsolat kezelése.
- **web_scrape.py:** Weboldalak letöltése és szöveg kinyerése.
- **static/:** Frontend fájlok (index.html, view_result.html, styles.css, script.js).
- **extension/firefox-ext/:** Böngésző kiegészítő forrásai.

Folyamatok

1. **Scraping:**
 - Felhasználó megad egy URL-t → backend letölti, feldolgozza, eltárolja (ha új).
2. **Keresés:**
 - Felhasználó keres URL vagy kulcsszó alapján → backend visszaadja a találatokat.
3. **Eredmény megtekintése:**
 - Felhasználó kiválaszt egy találatot → backend visszaadja a teljes szöveget.
4. **Böngésző kiegészítő:**
 - Felhasználó elküldi az aktuális URL-t → backend feldolgozza.

Adatbázis séma (példa)

```
CREATE TABLE scraped_text (  
  id INT AUTO_INCREMENT PRIMARY KEY,  
  url VARCHAR(2048) NOT NULL,  
  text LONGTEXT NOT NULL,  
  time DATETIME DEFAULT CURRENT_TIMESTAMP  
);
```
