Web Scraper Projekt Dokumentáció

Projekt Struktúra

```
root/
├── docs/
│   └── api_doc.md
├── src/
│   ├── main.py
│   ├── database.py
│   ├── requirements.txt
│   └── utils/
│       └── web_scrape.py
│   └── static/
│       ├── index.html
│       ├── styles.css
│       ├── script.js
│       └── view_result.html
│   └── extension/
│       └── firefox-ext/
│           ├── manifest.json
│           └── background.js
├── venv/
├── .gitignore
└── README.md
```

API végpontok

`/scrape/{url}`

- **Leírás:**
  Megadott URL-t letölti és visszaadja a kinyert szöveget.
  Ha az URL már le volt töltve, a gyorsítótárazott eredményt adja vissza.

- **Válasz:**

```
{
  "already_scraped": true | false,
  "result": {
    "time": "ÉÉÉÉ-HH-NN ÓÓ:PP:MM",
    "url": "https://example.com",
    "text": "Kinyert szöveg..."
  }
}
```

- Ha még nem volt letöltve, a `result` lehet egy egész szám (beszúrt id).

`/search_url/{url}`

- **Leírás:**
  Megkeresi, hogy a megadott URL már le lett-e töltve (kis- és nagybetű érzékeny).

- **Válasz:**

```
{
  "already_scraped": true | false,
  "result": {
    "time": "ÉÉÉÉ-HH-NN ÓÓ:PP:MM",
    "url": "https://example.com",
    "text": "Kinyert szöveg..."
  } | null
}
```

`/search_keyword/{keyword}`

- **Leírás:**
  Kulcsszóra keres az összes letöltött szövegben (kis- és nagybetű érzékeny).

- **Válasz:**

```
{
  "already_scraped": true | false,
  "results": [
    {
      "id": 1,
      "time": "ÉÉÉÉ-HH-NN ÓÓ:PP:MM",
      "url": "https://example.com",
      "view_text_link": "/view_result/1"
    },
    ...
  ] | "Nem található eredmény a backend keresése során"
}
```

`/view_result/{id}`

- **Leírás:**
  Visszaadja a teljes letöltött szöveget az adott azonosítóhoz.

- **Válasz:**

```
{
  "already_scraped": true | false,
  "id": 1,
  "time": "ÉÉÉÉ-HH-NN ÓÓ:PP:MM",
  "url": "https://example.com",
  "text": "Kinyert szöveg..."
}
```

Megjegyzések

- Minden végpont HTTP GET-et használ, és JSON-t vár/válaszol, hacsak nincs másképp jelezve.
- A `/scrape` végpont támogatja a query paramétereket is: `/scrape?url=https://example.com`
- Az API kis- és nagybetű érzékeny mind URL-ekre, mind kulcsszavakra.
- Az `already_scraped` jelzi, hogy az adat gyorsítótárból vagy frissen lett letöltve.
- Az időbélyeg formátuma: `ÉÉÉÉ-HH-NN ÓÓ:PP:MM`.
- A frontend a `/` címen érhető el, a statikus fájlok a `/static/` alatt vannak.

Példa használat

URL letöltése

```
GET /scrape/https://example.com
```

Keresés URL alapján

```
GET /search_url/https://example.com
```

Keresés kulcsszó alapján

```
GET /search_keyword/valamiKulcsszo
```

Teljes eredmény megtekintése

```
GET /view_result/1
```

Frontend

- Főoldal: `/static/index.html`

- Eredmény megtekintése: `/static/view_result.html?id={id}`

Függőségek

Lásd `src/requirements.txt`:

- fastapi
- uvicorn
- databases
- aiomysql
- beautifulsoup4
- requests

Környezeti változók

`.env` fájlban (git-ben nem követett):

- `DB_USER`
- `DB_PASSWORD`
- `DB_HOST`
- `DB_NAME`
- `DB_PORT`

Naplózás

- A letöltési naplók a `src/logs/scrape.log` fájlba íródnak.

Böngésző kiegészítő

- A Firefox kiegészítő a `src/extension/firefox-ext/` mappában található, és lehetővé teszi az aktuális böngészőfül URL-jének elküldését a backendnek letöltésre.
