Gruparea Website-urilor după Similaritatea Logo-urilor

Soluție Tehnică

March 18, 2025

1 Introducere

Acest proiect necesită o combinație de procesare de imagini, învățare automată și clustering pentru a grupa website-urile pe baza similarității logo-urilor.

2 Workflow-ul Soluției

2.1 1. Extracția Logo-urilor

- Utilizarea `BeautifulSoup` + `requests` pentru scraping (dacă ai doar link-uri).
- Folosirea `selenium` pentru site-uri dinamice.
- Detectarea logo-ului cu `OpenCV` sau `EasyOCR` (pentru recunoașterea textului din logo).

2.2 2. Preprocesarea Imaginilor

- Redimensionare și normalizare cu `Pillow` sau `OpenCV`.
- Convertire în grayscale și binarizare, dacă este necesar.
- Eliminare fundal cu `rembg` sau `OpenCV` + `grabCut`.

2.3 3. Feature Extraction și Similaritate

- Extragerea vectorilor de caracteristici cu rețele neuronale pre-antrenate (`ResNet50`, `VGG16` din TensorFlow/Keras).
- Folosirea descriptorilor vizuali clasici: `ORB`, `SIFT` sau `SURF`.

2.4 4. Clustering și Grupare

- Aplicarea de algoritmi de clustering: `KMeans`, `DBSCAN` (din `scikit-learn`).
- Măsurarea similarității cu `cosine distance` sau `euclidean distance`.

2.5 5. Evaluare și Output

- Salvarea grupurilor într-un fișier JSON sau generarea unui raport vizual.
- Vizualizarea datelor cu `matplotlib` sau `seaborn`.

3 Tech Stack Recomandat

- **Python**: TensorFlow, OpenCV, Selenium, scikit-learn.
- **Node.js**: pentru viteză în scraping și `TensorFlow.js` pentru deep learning.
- **Scala/PySpark**: Apache Spark ML pentru scalabilitate.
- **Faiss** (Facebook AI) pentru căutare eficientă a vectorilor.