KOCAELİ ÜNİVERSİTESİ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ
BLM306 YAZILIM LAB. II
PROJE 1

Web Scraping Akademi Uygulaması

Proje İlan Tarihi: 21/02/2024
Proje Teslim Tarihi: 15/03/2024

Google Akademik gibi akademik arama motorları üzerinden web scraping (web kazıma) yöntemiyle aratılan akademik yayınlara ait bilgilerin kaydedildiği bir veritabanıyla birlikte bu bilgilerin webden aratılması, görüntülenmesi ve istenilen özelliklere göre sorguların yapılmasına imkan sağlayacak bir web arayüzü geliştirmeniz beklenmektedir.

Amaç: Proje sayesinde web scraping ile bir web sayfasından bilgiye erişim sağlama; MongoDB veritabanı ile Elasticsearch sorgu yapılarını kullanma ve web kodlama becerilerinin geliştirilmesi amaçlanmaktadır.

Programlama Dili: Proje veritabanı ve sorguları için MongoDB ile Elasticsearch yapısı kullanılmalıdır. Web arayüzü, istenilen bir web programlama dili kullanılarak oluşturulacaktır..

Proje aşağıda detayları verilen 3 ana kısımlardan oluşmaktadır:

1. Web Scraping:

- En az bir akademik arama motorundan web scraping kullanılarak kullanıcının gireceği anahtar kelimelere göre listelenmiş en az ilk 10 akademik yayının bilgileri, kendi oluşturacağınız web sayfasında görüntülenmelidir. Kullanıcının arama yapmak için kullanacağı anahtar kelimeler oluşturacağınız kendi web sayfanızdaki bir text alanı üzerinden girilecektir.
- Web scraping işlemi için siteye ait html bilgiler kullanılarak veya siteye request isteği yapılarak istenen veriye erişilmelidir. (Erişilecek siteye yönelik hazır web API ler **kullanılmamalıdır**.)
- İstenen yayına ilişkin bilgiler doğrudan akademik arama motorunun sayfasından çekilebileceği gibi arama motoru sayfasındaki link üzerinden yönlendirilecek diğer bir web sayfasından da elde edilebilir. (İkincil sitelere geçiş yapılarak web scraping yapılması durumunda artı puan verilecektir.)
- İstenen her yayın için pdf dosyası mutlaka indirilmelidir. Daha sonra tercihe göre yayın bilgileri ya web sayfası üzerinden ya da indirilmiş pdf dosyasının içeriğinden elde edilebilir.

2. Veritabanı:

- Web scraping ile elde edilen veriler MongoDb kullanılarak veritabanına kaydedilecektir.
- Veritabanında tutulması gereken bilgiler şunlardır:
  - Yayın id,
  - Yayın adı,
  - Yazarların isimleri,
  - Yayın türü (araştırma makalesi, derleme, konferans, kitap vb.),
  - Yayımlanma tarihi,
  - Yayıncı adı (yayının yayımlandığı konferans ismi; dergi veya kitap yayınevi),
  - Anahtar kelimeler (Arama motorunda aratılan),
  - Anahtar kelimeler (Makaleye ait),
  - Özet,
  - Referanslar,
  - Alıntı sayısı,
  - Doi numarası (Eğer varsa),
  - URL adresi

3. Web Sayfası:

- Erişilen yayınların bilgilerinin kullanıcıya gösterilmesi için bir web sayfası oluşturmanız beklenmektedir.
- Web sayfasında aratılacak yayınlar için bir text alanı oluşturulmalı ve bu text alanı girilecek anahtar kelimeler üzerinden ilgili arama motorunun yayınları aratıp bilgilerini web sayfasına getirmesi sağlanmalıdır.
- Web sayfası ilk açıldığında veritabanında bulunan tüm kayıtlar sayfaya getirilmelidir.
- Listeleme işleminde yayınların isimleri sırasına uygun biçimde listelenmelidir. Listelenen bir makale ismine tıklandığında o makaleye özgü bilgiler ayrı bir sayfada gösterilmelidir.
- Web sayfasından herhangi bir çalışmaya yönelik dinamik arama işlemi yapılabilmelidir. Ayrıca arama sırasında yazım yanlışı olması durumunda sistem düzeltilmiş öneride bulunmalıdır. Örnek: deep learrning -- yazımı düzeltilmiş sonuçları görüyorsunuz: deep learning şeklinde olmalıdır.
- Web sayfasında dinamik filtreleme işlemi de ayrıca yer almalıdır. Filtreleme işlemi veritabanında yer alan yayının tüm özellikleri için hem ayrı ayrı hem de birlikte uygulanabilir olmalıdır.
- Web sayfasında en son veya en önce yayımlanma tarihine göre sıralama yapılabilmeli ayrıca yine atıf sayısına göre de en az veya en çok olacak şekilde sıralama yapılabilmelidir.

ÖDEV TESLİMİ

- Proje raporu IEEE formatında (önceki yıllarda verilen formatta) 4 sayfa uzunluğunda olmalıdır. Rapor; akış diyagramı veya yalancı kod içermeli, özet, giriş, yöntem, deneysel sonuçlar, sonuç ve kaynakça bölümünden oluşmalıdır.
- Dersin takibi projenin teslimi dâhil edestek.kocaeli.edu.tr sistemi üzerinden yapılacaktır. edestek.kocaeli.edu.tr sitesinde belirtilen tarihten sonra teslim edilen projeler kabul edilmeyecektir.
- Proje ile ilgili sorular edestek.kocaeli.edu.tr sitesindeki forum üzerinden Arş. Gör. Gamze Korkmaz Erdem veya Arş. Gör. Abdurrahman Gün'e sorulabilir. **Proje teslimine 2 gün kala sorulan hiçbir soruya cevap verilmeyecektir.**
- Sunum tarihleri daha sonra duyurulacaktır.
- Sunum sırasında;
    - Algoritma, geliştirdiğiniz kodun çeşitli kısımlarının ne amaçla yazıldığı ve geliştirme ortamı hakkında sorular sorulabilir.
    - Kullandığınız herhangi bir satır kodu açıklamanız istenebilir.

Projenin tanıtım toplantısı 23 Şubat 2024 Cuma günü saat 15:00'da bölüm duyurularında ve e-destekte duyurulacak derslikte yapılacaktır.

Proje grupları en fazla 2 kişiden oluşmalıdır. Proje grup bilgileri e-destekte paylaşılacak link üzerinden en geç 1 Mart Cuma gününe kadar girilmelidir. Bu tarihten sonra gruplarda herhangi bir değişiklik yapılmayacaktır.