Kocaeli Üniversitesi
Bilgisayar Mühendisliği Bölümü
BLM 307:Yazılım Lab.II, 2020-2021 Bahar
Proje I

Web İndeksleme Uygulaması

Proje Başlangıç Tarihi .....05 Mart 2021
Proje Teslim Tarihi:.....28 Mart 2021

Amaç: Verilen bir URL'deki web sayfa içeriğine göre diğer birden fazla web sayfasını benzerlik bakımından indeksleyip sıralayan web tabanlı bir uygulama geliştirmek. Böylece bu proje sayesinde web indeksleme yöntemleri hakkında bilgi edinilmesini ve web tabanlı uygulama yazma becerisinin geliştirilmesi amaçlanmaktadır.

Dil: Serbest

1 Sayfada Geçen Kelimelerin Frekanslarını Hesaplama (%10)

Projenin içeriği ve adımları aşağıda gösterilmiştir:

Verilen bir URL için:

- URL içeriğinde (URL'in gösterdiği sayfa içeriğinde) her kelimenin kaçar defa yer aldığını (frekansını) bulunuz.

Çıktı: Bu aşamada projenin web sitesindeki bir sayfada (örn: asama1.php, asama1.asp, ...) URL girilecek bir alan oluşturulacaktır. Sonrasında girilen bu URL'nin metnindeki her kelimenin sayfada kaç defa yer aldığı bilgisi hesaplanarak yazdırılacaktır.

2 Anahtar Kelime Çıkarma (%20)

Verilen 1 URL için,

- URL metninde geçen kelimelerden en önemli kelimelerin (sayfanın içerik özelliklerini tanımlayan ve kategorik özelliklerini yansıtan kelime kümesi) belirlenerek anahtar kelimelerin çıkartınız **(Aşağıda verilen anahtar kelime oluşturma kriteri tamamen örnek gösterim amaçlıdır. Uygulamada başarılı bir sonuç vermesi beklenmemektedir. Bu kısımda sizin kendi kriterlerinizi doğru bir yaklaşımla oluşturmanız beklenmektedir.)**

  - ❖ Örneğin, en yüksek frekansa sahip 5 kelimenin seçilmesi.

3 İki Sayfa (URL) Arasındaki Benzerlik Skorlaması (%20)

Verilen 2 URL için,

- İkinci bölümde 1. URL için elde edilen anahtar kelimelerin 2. URL'nin içeriğinde yer alma sayısına dayalı bir benzerlik skor formulü tanımlayınız. **(Aşağıda gösterilen skor formülü tamamen örnek gösterim amaçlıdır. Uygulamada anlamlı bir sonuç vermesi beklenmez. Bu kısımda sizin kendi formülünüzü yapacağınız araştırmalar sonucunda doğru bir yaklaşımla oluşturmanız beklenmektedir.)**

  - ❖ Örneğin, 1. Url deki anahtar kelimeler a, b, c, d ve e ile gösterelim. Bu anahtar kelimelerden 2. Url metninde geçenler a, c ve e olsun. Bu 3 anahtar kelimenin 2. Url'de tekrarlanma sayıları (frekansları) da f_{2a},f_{2c} ve f_{2e} ile gösterelim. Her iki URL' de geçen ortak anahtar kelimeler (a, c ve e) üzerinden aşağıdaki gibi örnek bir benzerlik skor formülü tanımlayalım:

    ($f_2 =$ 2. URL sayfa içeriğinde geçen tüm kelimelerin toplam sayısı olmak üzere)

    $$\text{Skor}_{12} = \frac{f_{2a} \text{ x } f_{2c} \text{ x } f_{2e}}{f_2}$$

- 1. URL'nin tüm anahtar kelimeleri ve 1. URL içeriğinde geçme sayısı (frekansını), bu iki URL için benzerlik skorunu, 2. URL'de geçen tüm anahtar kelimeleri ve bunların 2. URL içeriğindeki frekanslarını, yazdırınız (2. ve 3. Bölümler için tek bir web sayfasında (örn: asama23.php, asama23.asp,…) sonuçlar yazdırılmalıdır).

4 Site İndexleme ve Sıralama (%20)

Verilen bir web sitesi kümesi ve farklı bir URL için,

- Bu aşamada projenin web sitesindeki bir sayfada (örn: asama4.php, asama4.asp,…) URL girilecek bir alan oluşturulacaktır. Girilen bu URL'nin içeriği ile web site kümesindeki her bir web sayfasının içeriklerinin benzerlik skorları ayrı ayrı hesaplanacaktır. Ancak bu sefer skor hesaplaması yaparken bu web site kümesinde bulunan web sayfalarının içeriğine ilaveten yine bu sayfalarda bulunan tüm alt URL'leri de dikkate alınmalıdır. Alt URL'lerindeki anahtar kelimelerin yer alma sayılarına dayalı olarak skor formülünü yeniden geliştirin (derinlik max 3 seviye; ana sayfa= 1. seviye derinlik, ana sayfadan linklenmiş bir sayfa: 2. seviye derinlik, ana sayfadan linklenmiş bir sayfadan linklenmiş bir sayfa:3. seviye derinlik olarak kabul edilmelidir).
- Web sitelerini benzerlik skorlarına göre yüksekten düşüğe doğru sıralayınız (tüm alt URL'leri dahil)
- Her URL (bir web sitesi) için, sırasını, skorunu, alt URL'lerin ağaç yapısını ve her düğümdeki her bir anahtar kelimenin yer alma sayısı ile birlikte yazdırın. Bu bölüm için tek bir web sayfasında sonuçlar yazdırılmalıdır (örn: asama4.php, asama4.asp,…).

5 Semantik Analiz (%20)

- Verilen web siteleri içerisinde anahtar kelimelerle semantik olarak alakalı kelimeler olabilir. Örneğin, "ulusal" yerine "milli" kelimesi yer alabilir. Bu kelimelere "alakalı anahtar kelimeler" diyelim.
- Alakalı anahtar kelimeleri bulun ve yazdırın.
- Yinelemeli olarak 4. aşamadaki analizi yapınız.
- Bu bölüm için tek bir web sayfasında sonuçlar yazdırılmalıdır (örn: asama4.php, asama4.asp,…).

6 Çalışmanın Çevrimiçi Yayımlanması ve Rapor (%10)

- Uygulamanız için bir web sitesi geliştirin ve raporunuzda geliştirdiğiniz projenin tüm adımlarını detaylıca anlatınız.

7 Ödev Teslimi

- Dersin takibi projenin teslimi dahil edestek.kocaeli.edu.tr sistemi üzerinden yapılacaktır. edestek.kocaeli.edu.tr sitesinde belirtilen tarihten sonra getirilen projeler kabul edilmeyecektir.
- Proje ile ilgili sorular edestek.kocaeli.edu.tr sitesindeki forum üzerinden Arş. Gör. Abdurrahman GÜN'e veya Arş. Gör. Dilara GÖRMEZ'e sorulabilir.
- Sunum sırasında algoritma, geliştirdiğiniz kodun çeşitli kısımlarının ne amaçla yazıldığı ve geliştirme ortamı hakkında sorular sorulabilir. Kullandığınız herhangi bir satır kodu açıklamanız istenebilir.
- Sunum tarihleri daha sonra ayrıca duyurulacaktır.
- Proje grupları 2 kişiden oluşmalıdır. Proje grup bilgileri aşağıda verilen linke en geç 12 Mart Cuma gününe kadar girilmelidir. Bu tarihten sonra gruplarda herhangi bir değişiklik yapılmayacaktır. Link: https://docs.google.com/spreadsheets/d/1uW93wOxeIyrt2cFN3f0PjtIhvT04euAou32mJY-UP-w/edit#gid=0
- Projenin tanıtım toplantısı pazartesi günü saat 14.00 te bölüm duyurularında ve e-destekte duyurulacak toplantı linki üzerinden uzaktan yapılacaktır.