Proje 1

WEB INDEXLEME UYGULAMASI

Altan TUĞFAN
170202123
Mühendislik Fakültesi
Bilgisayar Mühendisliği

Arda TALU
170202037
Mühendislik Fakültesi
Bilgisayar Mühendisliği

***Özet*—Bu projede içinde beş adet aşama bulunan bir web indexleme uygulaması istenilmektedir. Birinci aşamada bir URL sayfasında geçen kelimelerin frekansı, yani hangi kelimenin sayfada kaç adet geçtiği hesaplanmaktadır. İkinci aşamada anahtar kelimeleri, yani sayfada en çok yer alan kelimeler bulunup gösterilir. Üçüncü aşamada iki URL'deki en çok geçen kelimeler ve bu kelimelerin benzerliklerine göre sayfalar arası benzerlik skorlaması yapılır. Dördüncü aşamada site indexleme ve sıralama istenilmektedir. Yani bir URL kümesindeki sayfaları sıra, skor ve alt URL'lerin ağaç yapısı ve her sayfadaki anahtar kelimelerin çıktısı istenilmektedir. Son aşamada ise semantic analiz yapılır. Yani eş anlamlı kelimeler bulunup yazdırılır**

Anahtar kelimeler—Word Scan, Kelime indexleme, web geliştirmeı, kelime mimarisi

I. GİRİŞ

Bu Projeyi, web geliştirme yapabileceğimiz, fonksiyonlarıyla kelime indexleme işlemini rahat yapabilme imakanı sağlayan dillerden 'PHP diliyle gerçekleştirdik. Projenin amacı; kullanıcının web arayüzüne girdiği URL'lere göre kelime frekansları, anahtar kelime çıkarma işlemi, İki sayfa arasındaki benzerlik skorlarının çıkarıldığı, verilen URL kümesine göre site indexleme ve sıralama işlemlerinin gerçekleştiği ve aynı zamanda eşanlamlı kelimeleri bulup yazdıran yani semantik analizinin yapıldığı bir sayfa tasarımını yapabilmektir. Proje yukarıda saydığımız gibi toplam beş aşamada yapılan bu projede dört adet kullanıcı arayüzü ve dört adet çalışma mantığının bulunduğu php dosyası, arayüz logosu ve eşanlamlı kelimelerin bulunduğu bir txt dosyası ile yapılmaktadır.

Programı kodlamadan önce projede bizden istenenleri, yani kullanmamız gereken URL indexleme ve kelime mimarisinin kullanılmasını, hangi aşamada URL'den ne verilerin istenildiği, web çıktıları ve bunun gibi diğer yapmamız gereken unsurları netleştirdik. Daha sonra Daha sonra bu arayüzler aracılığıyla sayfalar arası geçişin sağlanma durumlarını düşünüp mantığı ona göre oturttuk. PHP dilinde yazdığımız bu uygulamada SublimeText IDE'si ve server olarak da AppServ kullandık. Localhostta çalışma prensiplerini yerine getirip gerekli kurulumları yaptık.

II. YÖNTEM

A. Aşama 1 – Sayfada Geçen Kelimelerin Frekanslarını Hesaplama

Site açıldığında asamaform.php kısmında istenilen yere URL girilir. Ardından "Ara" butonuna basıldığı anda test1.php sayfasına yönlendirilir. Burda girilen sitenin ismi bir değişkene atanır. Ardından file_get_contents fonksiyonu ile site içeriği contents değişkenine aktarılır.Türkçe karakterlerin okunabilmesi için str_replace ile kelimeler sadeleştirilir. Ardından büyük-küçük harf ayrımı olmaması için strtolower fonksiyonuyla küçük karakterler olarak aynı değişkene kaydedilir. Ardından bir değişkene javascript, head, css ve multi-line kısımlarını temizlemek adına bu özellikteki içerikler atılır. Preg_replace komutu ile bu özelliklere sahip olan içerik silinir. Ardından bir diziye sayfa içeriğini tuttuğumuz değişkenden array_count_values, str_word_count ve strip_tags komutları kullanılarak kelimeleri ve kelimelerin kaçar tane olduğu bilgisini tutarak bu verileri aktarırız. Ardından kelime sayılarını büyükten küçüğe sıralamak adına arsort komutunu kullandık ve yazdırdık.

B. Aşama 2 – Anahtar Kelime Çıkarma

İlk aşamada olduğu gibi burada da Site açıldığında asamaform2.php kısmında istenilen yere URL girilir. Ardından "Ara" butonuna basıldığı anda test2.php sayfasına yönlendirilir. Burda girilen sitenin ismi bir değişkene atanır. Ardından file_get_contents fonksiyonu ile site içeriği contents değişkenine aktarılır.Türkçe karakterlerin okunabilmesi için str_replace ile kelimeler sadeleştirilir. Ardından büyük-küçük harf ayrımı olmaması için strtolower fonksiyonuyla küçük karakterler olarak aynı değişkene kaydedilir. Ardından bir değişkene javascript, head, css ve multi-line kısımlarını temizlemek adına bu özellikteki içerikler atılır. Preg_replace komutu ile bu özelliklere sahip olan içerik silinir. Aynı zamanda yaygın kelimeler olan stopwords composer'ı kurulur. Bu composer yardımıyla bu yaygın kelimeler bir dizide tutularak dizimizde bulunan bu kelimeler silinir. Ardından bir diziye sayfa içeriğini tuttuğumuz değişkenden array_count_values, str_word_count ve strip_tags komutları kullanılarak kelimeleri ve kelimelerin kaçar tane olduğu bilgisini tutarak bu verileri aktarırız. Ardından kelime sayılarını büyükten küçüğe sıralamak adına arsort komutunu kullandık. En yaygın beş kelimeyi yazdırmak için dizimizi, beş elemanlı dizi yaptık ve ilk beş elemanı array_splice komutu ile tekrar atadık. Daha sonra da print_r komutuyla da beş yaygın kelimeyi yazdırdık.

C. Aşama 3 – İki Sayfa (URL) Arasındaki Benzerlik Skorlaması

İlk iki aşamada olduğu gibi burada da Site açıldığında asamaform3.php kısmında istenilen yere bu kez iki URL girilir. Çünkü amaç bu iki URL arasındaki benzerlikeri

skorlamakdır. “Ara” butonuna basıldığı anda test3.php sayfasına yönlendirilir. Burda girilen iki sitenin de isimleri iki farklı değişkene atanır. Ardından file_get_contents fonksiyonu ile site içerikleri contents ve contents2 değişkenlerine aktarılır. Türkçe karakterlerin okunabilmesi için str_replace ile kelimeler sadeleştirilir. Ardından büyük-küçük harf ayrımı olmaması için strtolower fonksiyonuyla küçük karakterler olarak aynı değişkene kaydedilir. Ardından bir değişkene javascript, head, css ve multi-line kısımlarını temizlemek adına bu özellikteki içerikler atılır. Preg_replace komutu ile bu özelliklere sahip olan içerik silinir. Aynı zamanda yaygın kelimeler olan stopwords composer’ı kurulur. Bu composer yardımıyla bu yaygın kelimeler bir dizide tutularak dizilerimizde bulunan bu kelimeler silinir. Ardından iki farklı diziye sayfa içeriğini tuttuğumuz değişkenden array_count_values, str_word_count ve strip_tags komutları kullanılarak kelimeleri ve kelimelerin kaçar tane olduğu bilgisini tutarak bu verileri aktarırız. Ardından kelime sayılarını büyükten küçüğe sıralamak adına arsort komutunu kullandık. En yaygın beş kelimeyi yazdırmak için dizimizi, beş elemanlı dizi yaptık ve ilk beş elemanı array_splice komutu ile tekrar atadık. Daha sonra da print_r komutuyla da beş yaygın kelimeyi yazdırdık. Son olarak da skorlama işlemi yapmak adına array_sum komutu içinde array_intersect_key kumutu kullanarak bir değişkene birinci sayfadaki anahtar kelimelerden ikinci sayfada kaçar adet olduğu bilgisi gönderilir. Ardından başka bir değişkene ikinci sayfadaki anahtar kelimelerin toplam kaç adet olduğu bilgisi yine array_sum komutu ile gönderilir. Yazdıracağımız son değişkende skorlama mantığı olarak (benzer kelimelerin toplamı/toplam anahtar kelime)*100 işlemi yapılarak skorlamayı tamamlarız.

D. Aşama 5 - Semantik Analiz

Bu bölümde iki URL içeriğindeki eş anlamlı kelimeleri bulup yazdırma işlemi gerçekleştirmek amaçlanmıştır. Mantık olarak yine diğer adımlardaki işlemler uygulanarak test5.php dosyasına ulaşılır ve sitedeki kelimeler değişkenlere aktarılır. Ardından bir dizi oluşturulur ve esanlamlilar.txt dosyasındaki kelimeleri fopen ile okutup while döngüsüyle bu dizimize txt dosyamızın içindeki kelimeler satır satır alınarak aktarılır. Amaçlanan dizilerdeki eşleştirmeler bulunarak eş anlamlısı olan kelimeleri ve eşanlamlarını yazdırmaktı. Fakat dizideki kelimeler index olarak alındığı için tam anlamıyla çalıştırılamadı.

III. SONUÇ

Sonuç olarak kazanımlar;

- “PHP” dilinde yazılmıştır.
- Web dilinde kelime mimarisi kullanımları öğrenildi
- Kelimeleri hızlı taratıp performans arttırımı sağlandı.
- Metin indexleme ve metinde bulunan kelimelerin frekanslarını hesaplama kısımları için hangi fonksiyonlara ihtiyacımız olduğunu ve bu fonksiyonları nasıl kullanmamız gerektiği öğrenildi.
- Formlar arası geçiş ve form dizaynları yapım incelikleri öğrenildi.

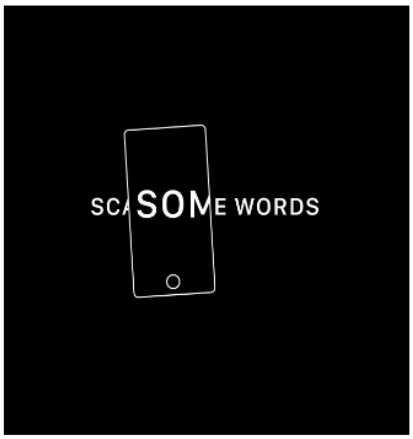
Kullanılan diller, formlar ve composerlar

- PHP
- Asamaform.php
- Asamaform2.php
- Asamaform3.php
- Asamaform5.php
- `Test1.php`
- `Test2.php`
- `Test3.php`
- `Test5.php`
- `stopWords composer`
- `logo.gif`
- `stopwords.txt`
- `esanlamlilar.txt`

Karşılaştığımız problemler

Programı yazarken bazı sıkıntılarla karşılaştık. Bunların başında kelime grubunu bir diziye attığımızda index ve içerik kısımları ters atıldı. Bunu daha sonra çözsem de bu kez indexleri kelimelerden kaçar tane olduğu olarak kaldı. Yani index’i aynı olan, unique olmayan bir çok dizi elemanı olduğundan bu elemanlar arasında sağlıklı işlemler yapmamız engellendi.

IV. ARAYÜZ GÖRÜNTÜLERİ



İlk Url'yi girin...

İkinci Url'yi girin...

Ara

Aşama 1 - Aşama 2 - Aşama 3 - Aşama 5

```
Array
(
    [ve] => 248
    [isvec] => 181
    [bir] => 144
    [bu] => 124
    [-] => 108
    [olarak] => 83
    [degistir] => 75
    [tarihinde] => 73
    [arsivlendi] => 70
    [en] => 66
    [ile] => 60
    [the] => 57
    [olan] => 55
    [isvec'in] => 54
    [of] => 52
    [sweden] => 46
    [da] => 46
    [in] => 41
    [kadar] => 40
    [ulkenin] => 40
    [de] => 40
    [buyuk] => 40
    [kaynagi] => 38
    [and] => 36
    [wayback] => 36
    [sitesinde] => 35
    [machine] => 35
    [kaynagindan] => 35
    [arasinda] => 34
    [swedish] => 33
    [yer] => 33
    [isvec'te] => 31
    [tarihi] => 28
    [ulke] => 26
    [tarafindan] => 26
    [yilinda] => 25
    [daha] => 25
    [a] => 24
    [mayis] => 24
    [statistics] => 23
    [icin] => 23
    [ancak] => 23
    [diger] => 23
    [v] => 23
    [erisim] => 23
    [yine] => 23
    [cok] => 22
    [ayni] => 21
    [kuzey] => 21
    [onemli] => 21
    [gibi] => 21
    [vikipedi] => 20
    [ayrica] => 20
    [ise] => 19
    [her] => 18
    [isvecce] => 18
    [ilk] => 18
    [sahip] => 18
```

Url girin...

Ara

Aşama 1 - Aşama 2 - Aşama 3 - Aşama 5

```
Array
(
    [ve] => 91
    [norvec] => 62
    [degistir] => 45
    [bir] => 34
    [-] => 26
)
```

Aşama 1 - Aşama 2 - Aşama 3 - Aşama 5

```
Array
(
    [ve] => 91
    [norvec] => 62
    [degistir] => 45
    [bir] => 34
    [-] => 26
)

Array
(
    [ve] => 248
    [isvec] => 181
    [bir] => 144
    [bu] => 124
    [-] => 108
)
```

http://tr.wikipedia.org/wiki/%C4%B0sve%C3%A7'nin http://tr.wikipedia.org/wiki/Norveç'ye benzerlik skoru : 62.111801242236

Aşama 1 - Aşama 2 - Aşama 3 - Aşama 5

ab-su
aba-üstlük
abartı-mübalağa
abece-alfabe
abecesel-alfabetik
abes-boş
abes-gereksiz
abıhayat-bengisu
abide-anıt
abluka-kuşatma
abuksabuk-anlamsız
abullabut-hantal
abullabut-kaba
abus-somurtkan
acar-cesur
acar-becerikli
acaba-acep
acayip-garip
acele-çabuk
acele-ivedi
aceleci-ivecen
acemce-farsça
acemi-toy
acemi-bilgisiz
acı-üzüntü
acı-ıstırap
acıma-merhamet
acımasız-gaddar
açıkgöz-kurnaz
ad-isim
adale-kas
adalet-hak
adet-tane
âdet-gelenek
âdet-töre
adıl-zamir
adi-bayağı

V. KAYNAKÇA

1. www.php.net/manual/tr/ref.array.php
2. www.php.net/manual/tr/language.types.array.php
3. https://www.webcebir.com/93-php-dizinin-anahtari-ve-degerleri-yazdirma-dersi.html
4. https://www.webcebir.com/91-php-dizide-eleman-sayisini-bulma-count-fonksiyonu-dersi.html
5. https://stackoverflow.com/questions/6159683/read-each-line-of-txt-file-to-new-array-element
6. https://www.php.net/manual/tr/function.array-sum.php
7. https://www.php.net/manual/tr/function.array-intersect-assoc.php