# WEB İNDEKSLEME UYGULAMASI

Alparslan Beraat ÖZDEMİR – 170202045
Bilgisayar Mühendisliği Bölümü
Kocaeli Üniversitesi
Kocaeli, Türkçe
beraat78@gmail.com

## ÖZET

Bu proje, 5 farklı PHP programından oluşmaktadır. Her aşama, farklı işlevler yerine getirir.

İlk aşama, verilen bir web sayfasındaki kelimelerin frekanslarını bulur. Bunları ekrana yazdırır.

İkinci aşama, verilen bir web sayfasındaki anahtar kelimeleri ekrana yazdırır.

Üçüncü aşama, iki web sayfası arasındaki benzerlik oranını bulur ve ekrana yazdırır.

Dördüncü aşama, verilen bir URL kümesini benzerliklerine göre sıralar.

Beşinci aşama, verilen bir web sayfasının anahtar kelimelerinin eş anlamlılarını bularak ekrana yazdırır.

## I. GİRİŞ

Web indeksleme uygulaması, 5 adet farklı php dosyası ve 1 adet html2text.php kütüphanesinden oluşmuştur

Bir apache sunucusunda dosyalar tarayıcı yardımı ile ayrı ayrı çalıştırılabilir.

Her sayfa önce html kısmıyla input alanlarını barındırır, daha sonra PHP kodu kısmı yer alır.

Bütün aşamalarda string temizleyici ayikla fonksiyonu standart olarak yer almaktadır.

Temizle fonksiyonu da aynı şekilde, özel karakter içeren kelimeleri temizlemek için standart olarak yer almaktadır.

## II. TEMEL BİLGİLER

Proje gelişiminde;

Tümleşik geliştirme ortamı olarak PHPStorm kullanılmıştır.

Program Windows 10 İşletim sisteminde test edilmiştir. Program PHP dilinde 8.0 versiyonunda geliştirilmiştir.

## III. TASARIM

Web indeksleme projesinin programlanma aşamaları altta belirtilen başlıklar altında açıklanmıştır.

### *A. Kullanıcı arayüzünü oluşturmak.*

Temel HTML ile her sayfada amaca uygun, çok sade şekilde tasarlanmış formlar yer almaktadır. Tasarım kısmına önem verilmemiş, sadece görevini yerine getirecek kadar bir tasarım yapılmıştır.

### *B. Fonksiyonlar*

**Main Fonksiyonu:**

Php dil yapısı gereği özel bir main fonkisyonu tanımlanmamıştır. Programın php taglerinin başladığı yerden itibaren main fonksiyonu olarak sayılabilir, HTML formlarından verilerin alınması genelde bu aşamada yer almaktadır. Ayrıca, bazı aşamalarda gerekli işlemlerin temel kısmı bu kısımda yapılmıştır.

**Ayikla Fonksiyonu:**

Bu fonksiyon, bütün aşamalarda standart olarak yer almaktadır. İşlem yapılacak URL leri parametre olarak alan bu fonksiyon, sayfa kaynak kodunu en ideal hale gelene kadar onlarca filtrasyona tabi tutar. Bu filtrasyonların büyük çoğunluğu, Html2Text kütüphanesi kullanılarak yapılmıştır. Tüm HTML taglerini büyük oranda temizleyen bu kütüphane, sayfa içeriğini büyük oranda temizlenmiş bir string olarak return eder ve diğer filtreler de uygulandıktan sonra, sadece sayfada “anlam ifade eden” kelimeler kümesi kalır.

**Temizle fonksiyonu:**

Bu fonksiyon, içerisinde özel karakter yer alan bütün kelimeleri yok eder. Diğer filtrasyonlar uygulanırken

bozulmuş olarak kalan kelimeleri bu fonksiyon aracılığı ile regexler kullanılarak yok edilir.

**Esanlambul fonksiyonu:**

Bu fonksiyon, sadece asama5.php dosyasında, son aşamada yer alır. Sayfanın anahtar kelimelerini parametre olarak alan bu fonksiyon, online bir sözlük hizmetinin API sini kullanarak parametre olarak aldığı kelimelerin eş anlamlılarını uzak sunucuda sorgular. API, bu kelimenin eş anlamlılarını response olarak döndürür, ve fonnksiyon tarafından anahtar kelimelerin eş anlamlıları eğer bulunabildiyse ekrana yazdırılır.

### *C. Algoritma*

Program, birden fazla aşamadan oluştuğundan dolayı, her aşamada farklı bir algoritma kullanılmıştır. Genel olarak program akışı şu şekildedir:

1- Başla

2- Kullanıcıdan gerekli inputları form üzerinden al.

3- POST isteği ile PHP kodundaki değişkenlere ata.

4- Hedef adreslere istek atarak sayfa kaynaklarını değişkene ata.

5- Sayfa kaynaklarına gerekli filtreleri uygula.

6- Eğer gerekiyorsa, sonuçları birbiriyle kıyasla.

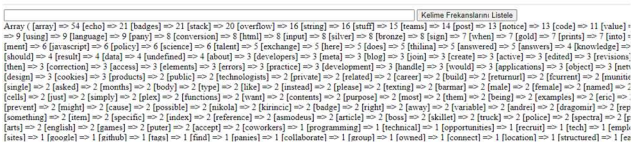
7- İsterleri kaydet.

8- Ekrana çıktı olarak web sayfasında yazdır.

## IV. SONUÇLAR VE EKRAN ÇIKTILARI

Aşama1'in çalışması ve sonucu şekil 1 ve 2'de gösterilmiştir.


Şekil1.

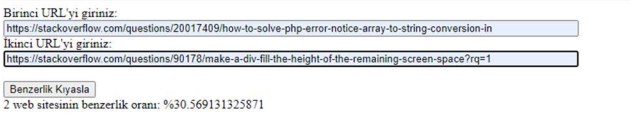



Şekil 2.

Aşama2'nin çalışması ve sonucu şekil 3'de gösterilmiştir.

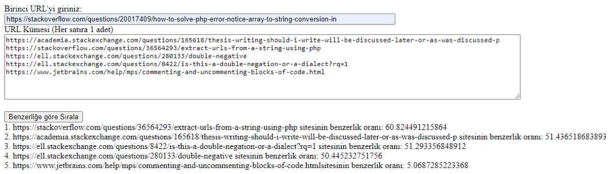

Şekil 3.

Aşama3'ün çalışması ve sonucu şekil 4'de gösterilmiştir.




Şekil 4.

Aşama 4'ün input ve sonucu şekil 5'de gösterilmiştir.




Şekil 5.

Aşama5'in Çalıştırılması ve çıktısı şekil6'da gösterilmiştir.

## KAYNAKÇA


[1] https://github.com/mtibben/html2text

[2] https://www.wordsapi.com/#try

[3] https://stackoverflow.com/questions/9442249/php-regex-remove-words-from-string-which-contain-non-letters-numbers