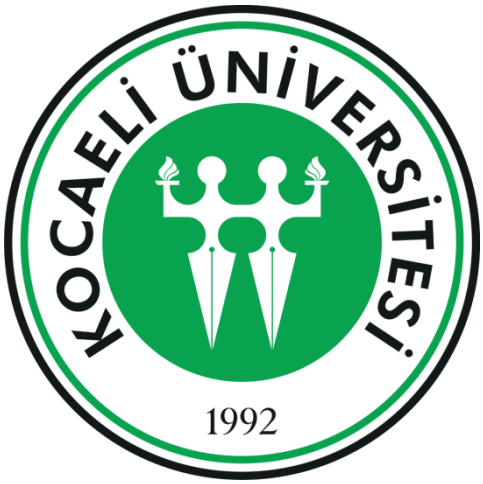


KOCAELİ ÜNİVERSİTESİ
MÜHENDİSLİK FAKÜLTESİ
BİLGİSAYAR MÜHENDİSLİĞİ

YAZILIM LABORATUVARI - II PROJE - I
WEB INDEKSLEME PROJESİ

ENGİN YENİCE
190201133

CEMRE CAN KAYA
190201137

KOCAELİ 2020

Web İndeksleme Uygulaması

Cemre Can Kaya
Bilgisayar Mühendisliği
Kocaeli Üniversitesi
190201137

Engin Yenice
Bilgisayar Mühendisliği
Kocaeli Üniversitesi
190201133

***Özet*— 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.**

Anahtar Kelimeler—anahtar, kelime, frekans, havuz, liste, website, link, url, semantic, benzerlik

I. GİRİŞ

Programın arka planı (backend) **C#**, ön planı(front-end) **Angular Framework**'ü ile geliştirilmiştir. Verilen bir URL'deki web sayfa içeriğine göre diğer birden fazla web sayfasını benzerlik bakımından indeksleyip sıralayan web tabanlı bir uygulama geliştirilmiştir. Bu proje sayesinde web indeksleme yöntemleri hakkında bilgi edinilmiştir.

II. PROJE MİMARİSİ

Projenin arka planı (backend) kurumsal mimariye uygun bir şekilde geliştirilmiştir. Proje yapısı gereği 5 temel katmana parçalanmıştır.

Projenin ön planı (front-end) parçala yönet mantığı ile geliştirilmiş olup her bir işlem parçacığı ayrı bir bileşende (component) üzerinde yapılmaya çalışılmıştır.

A. Arka Plan (Backend) Yapısı

Proje yapısı gereği 5 temel katmana parçalanmıştır. Bu katmanların detayları bu başlık altında açıklanmıştır.

1) Core Katmanı

Bu katman proje içerisinde bulunması gereken temel bileşenleri bulundurmaktadır.

a) ***Entities:***

Oluşturduğumuz nesnelerin daha somut ve yönetilebilmesi için temel arayüzleri (interface) bu klasör içerisinde tutulmaktadır.

- **IEntity:** Temel nesnelerimizin temel arayüz (interface) sınıfıdır.
- **IDto:** Uzun ismi **Veri iletim nesnesi (Data transfer object)** olarak geçmektedir. Temel nesnelerimizden Kullanıcı arayüzüne (UI (User Interface (Front-end))) göndermek istediğimiz nesne tanımlamalarının temel arayüz (interface) sınıfıdır.

b) ***Utilities:***

Projenin genelinde kullanabileceğimiz araçlar bu klasör altında tutulmaktadır.

- **Results:** İş (Business) katmanında kullanacağımız metotların geriye dönüş değerlerinin daha yönetilebilir ve daha düzenli olması için oluşturduğumuz bir araç sınıfıdır. Bu sınıfı temel olarak özetlenecek olursak. İçerisinde temel olarak 2 adet değişken bulunmaktadır. Bu değişkenler mesaj ve başarı (message, success) durumu olarak isimlendirilmektedir. Veri gönderilmesi durumunda miras verdiği alt sınıfta ise data (veri) değişkeninin bulunduğu ayrı bir dönüş tipi bulunmaktadır. Sınıf içerisinde ki değişkenlerin daha kolay yönetilmesi için başarı (success) durumuna göre alt sınıflar oluşturulmuştur. Bu sınıfların çağrılması durumunda başarı durumu otomatik olarak belirlenmektedir.

2) Entities Katmanı

Entity ve Dto arayüzlerinden (interface) örnek alan veri sınıfları bulunmaktadır.

a) ***Concrete***

Entity arayüzünden (interface) örnek alınan veri sınıfları bulunmaktadır.

b) ***Dto:***

Dto arayüzünden (interface) örnek alınan veri sınıfları bulunmaktadır.

3) DataAccess Katmanı

Veri havuzunun yönetilmesinden görevli katmandır. Bu katmanımızı ileride geliştirmeye açık olması için arayüzler (interface) kullanarak geliştirdik. Bu sayede ileride bir gerektiğinde başka yapılara geçmeyi planladık. Şuanda veri havuzumuzu bellekte (In Memory) olarak tutuyoruz.

a) Abstract

Veri sınıflarımızın arayüzlerinin tutulduğu klasör.

b) Concrete

Veri sınıflarımızın tutulduğu klasör.

- **WordToExcludeDal:** Anahtar kelime olmasını istemediğimiz kelimelerin listesinin tutulduğu bir listedir.
- **TagAndPointDal:** Site içerisinde belirli etiketlere ait özel puanlama yapılmaktadır. Bu puanlama listesinin puanlarının belirlendiği liste bu sınıfta tutulmaktadır.

HTML Etiketler	Puan (Her bir kelime için)
Title	10
H1	**9**
H2	**8**
H3	**7**
H4	**6**
H5	**5**
H6	**4**
B	**3**
Strong	**3**
U	**2**
P	**2**

- **MemoryGlobalSemanticWord:** Semantik kelimelerin bulunduğu bir listeyi bellekte tutmaktadır. Program ilk açıldığında bu liste hafızaya alınmaktadır.

4) Business Katmanı

Projenin iş kodlarının yazıldığı katmandır. Bu gerekli işlemlerin yönetildiği katmandır. Temel olarak 4 klasöre bölünmüştür.

a) Abstract

Business sınıflarımızın arayüzlerinin tutulduğu klasör.

b) Concrete

Business sınıflarımızın tutulduğu klasör.

- **IndexerManager:** WebAPI tarafından gönderilen isteklerin gerekli helper sınıflarına yönlendirilmesini sağlayan ve dönen sonuçların geri WebAPI tarafına yönlendiren sınıftır.

c) DependencyResolvers

Bağımlılıkların çözülmesi ve isim havuzuna aktarılması için kullanılmaktadır. (Paket olarak Autofac kullanılmıştır.)

d) Helpers

Projenin akışını sürdürecek helper sınıfları yazılmıştır. Helper sınıflarımızın arayüzlerinin tutulduğu abstract klasörü ve örnek alındığı concrete klasörü bulunmaktadır.

- **WebSiteOperation:** Gönderilen web site adresi ile iletişim kurmakla görevlidir. Gönderilen web sitelerinin bilgilerini tespit etmekle görevlidir.
- **HtmlCleaner:** Gönderilen web site nesnesi içerisinde bulunan html kodlarının temizlenmesi görevini yapmaktadır.
- **KeywordOperation:** Gönderilen website nesnesi üzerinde kelime, anahtar kelime, frekans, semantic kelime bulma görevlerini yapmaktadır.

5) WebAPI katmanı

Ön Plan (Front-end) kısmından gelen istekleri karşılayıp gerekli dönüşleri yapmakla görevli olan katmandır. Ön Plan (Front-end) tarafından 5 farklı isteğe karşılık verebilecek 5 adet Controller bulunmaktadır.

a) StageOneController

Sayfada geçen kelimelerin frekanslarını hesaplayan metotlarını çağırmakla görevlidir. (1. Madde)

b) StageTwoController

Anahtar kelime çıkarma metotlarını çağırmakla görevlidir (2. Madde) [Otomatik olarak 1. Madde ile bağlantılıdır.]

c) StageThreeController

URL ve URL havuzu arasındaki benzerlik skorlaması metotlarını çağırmakla görevlidir (3. Madde) [Otomatik olarak 2. Madde ile bağlantılıdır.]

d) StageFourController

Site indeksleme ve sıralama metotlarını çağırmakla görevlidir. (4. Madde) [Otomatik olarak 3. Madde ile bağlantılıdır.]

e) StageFiveController

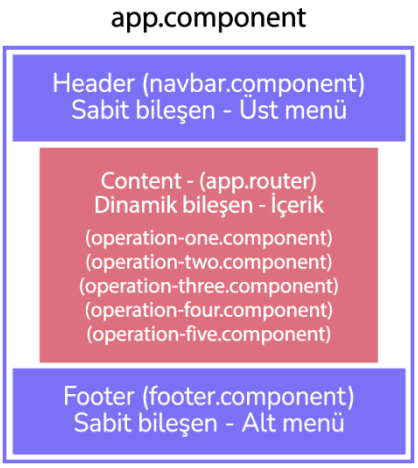
Semantik analiz metotlarını çağırmakla görevlidir (5. Madde) [Otomatik olarak 4. Madde ile bağlantılıdır.]

B. *Ön Plan (Frontend) Yapısı*

Projenin kullanıcı ara yüzü (User Interface) ekranıdır. Arka plan (Backend) tarafında yazılan kodların kullanıcıya görüntüsel olarak aktarıldığı taraftır. Angular Framework ile geliştirilmiştir. NG-ZORRO bileşen (component) paketi kullanılmıştır.

1) *Ana Bileşen (Component) (app.component)*

Projenin çalıştığı ana bileşendir. Bu bileşen içerisinde sabit ve dinamik bileşenler çağrılmaktadır.



Sabit bileşenler direk ana bileşen içerisinden çağrılmaktadır. Dinamik bileşenler ise rotalama (routing) yapısı ile ekrana getirilmektedir.

III. YONTEM

A. *Frekans Hesaplaması*

Frekans hesaplama algoritması web sitesi içerisinde içerik olarak girilen her metni baz alır. Örneğin başlık etiketi içeriği, paragraf etiketi içeriği, buton isimleri gibi. Web sitesi kaynak kodları içerisinde bulunan tüm html etiketleri ve back-end tarafından gelen tüm kodlardan arındırılır. Regex yani Regular Expression (Düzenli ifade) kullanılarak html sayfası içerisindeki tüm etiketler temizlenir. Geriye kalan veri, web sitesinin tarayıcıda kullanıcıya gösterdiği website içeriğidir.

B. *Anahtar Kelime Hesaplaması*

Anahtar kelime hesaplama algoritması web sitesi içerisinde bulunan kelimelerin, sayfa içerisinde bulunduğu etiketlerin önem sırasına göre hesaplanır. Bazı html etiketleri özel olarak puanlandırılır. Bu etiketler dışında kalan kelimeler ise varsayılan olarak 1 puan değerinde hesaplanır.

HTML Etiketler	Puan (Her bir kelime için)
Title	10
H1	**9**
H2	**8**
H3	**7**
H4	**6**
H5	**5**
H6	**4**
B	**3**
Strong	**3**
U	**2**
P	**2**
Varsayılan	**1**

Web site içerisinde bulunan tüm kelimelerin puanı, frekansları ile çarpılarak kelimenin önem skoru elde edilir. Tüm kelimeler büyükten küçüğe sıralanır ve ilk 10 kelime web sitesinin anahtar kelimesi olarak seçilir.

C. *Benzerlik Sıralaması*

Benzerlik sıralamasında girilen iki web sitesi için aşama 2 tekrarlanır ve web sitelerin anahtar kelimeleri oluşturulur. Benzerlik testi iki web sitesi içerisindeki anahtar kelimelerden kaçının ortak olması durumunu inceler. Karşılaştırılan web sitenin içerisinde eşleşen anahtar kelime skoru bölü tüm anahtar kelimelerin skoru, web sitenin benzerlik skorunu belirler.

D. *Alt Url' ler ile Benzerlik Sıralaması*

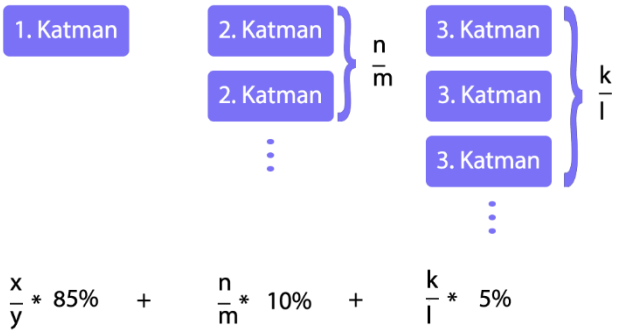
Aşama 4' te Aşama 3' te yapılan işleme ilaveten bir web sitesinin birden çok web site ile benzerlik sıralaması hesaplanır ve benzerlik skoru puanına havuzda bulunan web sitelere ek olarak site içerisinde linklenmiş alt URL'leri de dahil edilir.

Bir web sitesinden linklenmiş maksimum 5 alt url incelenir. Bu alt URL'ler 2. katman URL'lerdir. 2. katman URL'lerin de her birinden maksimum 5 alt URL incelenir. Bu alt URL'ler 3.katman URL'leridir.

Alt URL'leri skor hesaplamasına dahil edildiğinde skor formülü şu şekilde güncellenir;

Havuzda bulunan web sitesinin içeriğinin yüzdelik benzerliğinin yüzde 85 i ile tüm 2.katman alt URL'lerin toplam benzerliğinin yüzde 10 u ile tüm 3.katman alt URL'lerin toplam benzerliğinin yüzde 5' inin toplamı

havuzda bulunan web sitesinin hedef web siteye olan benzerliğini belirler.



$\frac{x}{y}$ = 1. Katmandaki = $\frac{Eşleşen\ anahtar\ kelimeler\ skoru}{Tüm\ anahtar\ kelimeler\ skoru}$

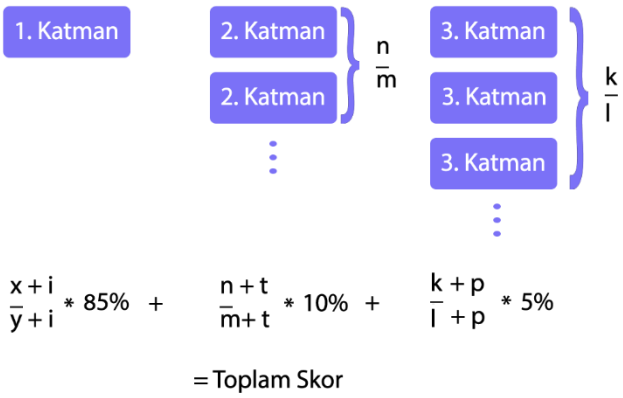
$\frac{n}{m}$ = 2. Katmandaki = $\frac{Eşleşen\ anahtar\ kelimeler\ skoru}{Tüm\ anahtar\ kelimeler\ skoru}$

$\frac{k}{l}$ = 3. Katmandaki = $\frac{Eşleşen\ anahtar\ kelimeler\ skoru}{Tüm\ anahtar\ kelimeler\ skoru}$

E. *Semantik Kelimeler ile Benzerlik Sıralaması*

Aşama 5 te yine aşama 4 te yapılanlar tekrarlanır. Aşama 5 te ekstra olarak skor hesaplamasına, hedef olarak girilen web sitenin anahtar kelimelerinin eş anlamlı kelimeleri ile havuzda bulunan web sitelerin eşleşenleri anlamlı kelimeleri web sitenin skor hesaplamasına eklenir.

Örneğin: Havuzda bulunan ilk web sitesinin eşleşen anahtar kelimeleri $\frac{700}{1000}$ = 0.7 yani %70 olarak benzerlik bulunmuş olsun. Semantik kelimeler bulundukça $\frac{700+50}{1000+50}$ (50 semantik bir kelime skoru, rastgele bir sayı) şeklinde hesaplamaya eklenir yeni değer $\frac{750}{1050}$ = 0.714 yani %71 olarak benzerlik skoru güncellenir.



$\frac{x}{y}$ = 1. Katmandaki = $\frac{Eşleşen\ anahtar\ kelimeler\ skoru}{Tüm\ anahtar\ kelimeler\ skoru}$

i = 1. Katmandaki semantik kelimeler skoru

$\frac{n}{m}$ = 2. Katmandaki = $\frac{Eşleşen\ anahtar\ kelimeler\ skoru}{Tüm\ anahtar\ kelimeler\ skoru}$

t = 1. Katmandaki semantik kelimeler skoru

$\frac{k}{l}$ = 3. Katmandaki = $\frac{Eşleşen\ anahtar\ kelimeler\ skoru}{Tüm\ anahtar\ kelimeler\ skoru}$

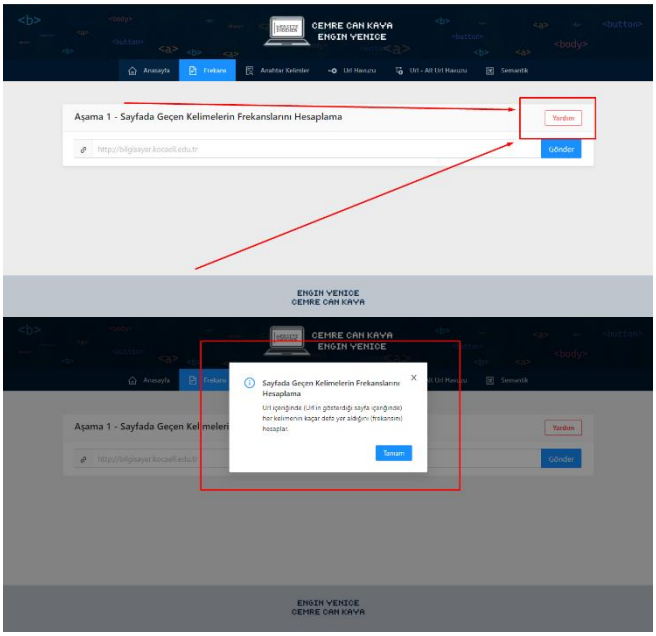
p = 1. Katmandaki semantik kelimeler skoru

IV. NASIL KULLANILIR

A. *Temel Birleşenler*

1) *Yardım*

Kullanıcının bulunduğu sayfanın temel işleyişiyle ilgili bilgiler vermektedir.



B. *Anasayfa*

Tüm sayfalar hakkında temel bilgiler vermektedir.

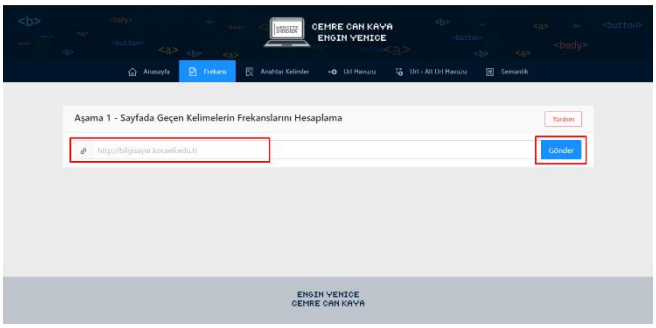
C. *Frekans (Ödev Madde 1)*

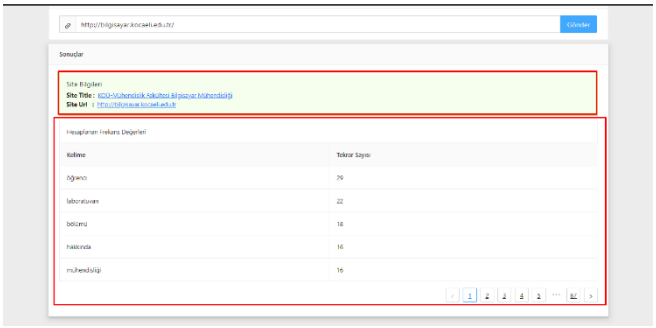
Sayfanın ortasında bulunan veri girişi nesnesi (input) içerisine kontrol edilmesi istenen link girilmektedir. Ardından gönder butonuna tıklayarak site hakkında sonuçlara ulaşabilirsiniz.

Gelen sonuç içerisinde temel site bilgileri ve frekans değerleri gösterilmektedir.

Frekans tablosu:

Kelime ve site içerisinde geçen tekrar sayısı gösterilmektedir.



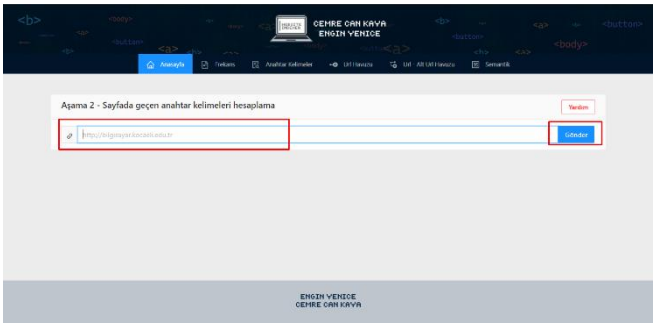


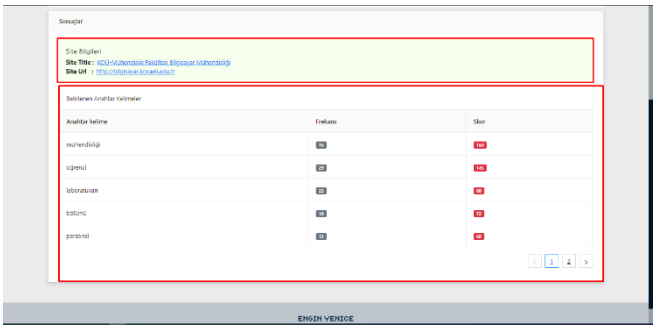
D. *Anahtar Kelimeler (Ödev Madde 2)*

Sayfanın ortasında bulunan veri girişi nesnesi (input) içerisine kontrol edilmesi istenen link girilmektedir. Ardından gönder butonuna tıklayarak site hakkında sonuçlara ulaşabilirsiniz.

Gelen sonuç içerisinde temel site bilgileri ve belirlenen anahtar kelimeler gösterilmektedir.

Anahtar Kelime Tablosu: Anahtar kelime, site içerisinde geçen tekrar sayısı ve hesaplama sonucu aldığı skor değeri gösterilmektedir.





E. *URL Havuzu (Ödev Madde 3)*

Sayfanın ortasında bulunan veri girişi nesnesi (input) içerisine kontrol edilmesi istenen link girilmektedir.

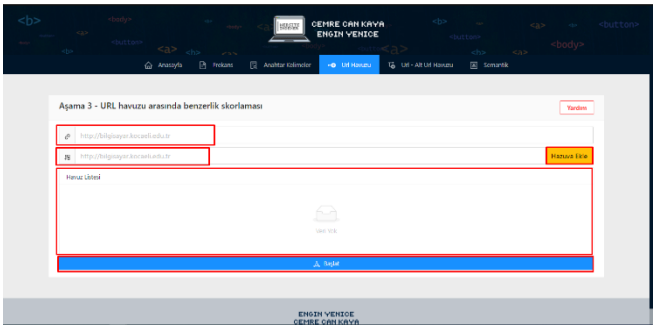
Ardından havuza eklenmesini istediğiniz linkleri havuza ekle butonunu kullanarak ekleyebilirsiniz.

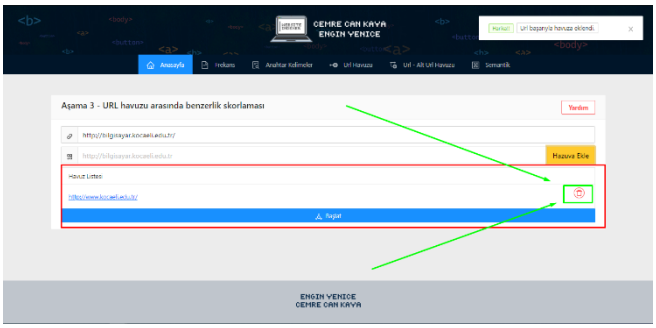
Bir hata ile karşılaşılması durumunda ekranın sağ üst köşesinden çıkan bir açılır pencere (pop-up) ile bilgilendirme sağlanmaktadır.

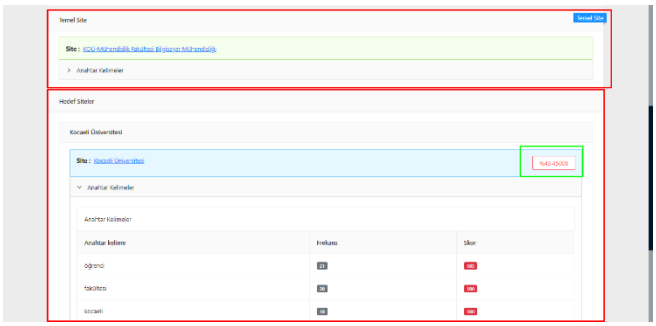
Havuza eklemiş olduğunuz bir linki sağ tarafında bulunan çöp kutusu görünümündeki butona tıklayarak havuzdan kaldırabilirsiniz.

Ardından başlat butonuna tıklayarak gerekli sonuçları elde edebilirsiniz.

Oluşturulan sonuç ekranında üst tarafta hedeflediğiniz ana site ile ilgiler verilmektedir. Alt tarafta ise havuzunuzun içerisinde bulunan tüm sitelerin bilgileri ve benzerlik skorlaması bulunmaktadır.







F. Url – Alt Url Havuzu (Ödev Madde 4)

Sayfanın ortasında bulunan veri girişi nesnesi (input) içerisine kontrol edilmesi istenen link girilmektedir.

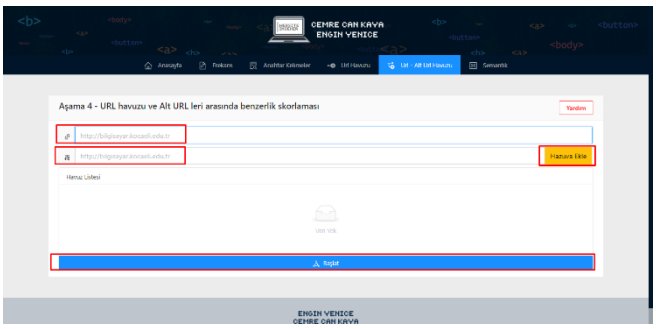
Ardından havuza eklenmesini istediğiniz linkleri havuza ekle butonunu kullanarak ekleyebilirsiniz.

Bir hata ile karşılaşılması durumunda ekranın sağ üst köşesinden çıkan bir açılır pencere (pop-up) ile bilgilendirme sağlanmaktadır.

Havuza eklemiş olduğunuz bir linki sağ tarafında bulunan çöp kutusu görünümündeki butona tıklayarak havuzdan kaldırabilirsiniz.

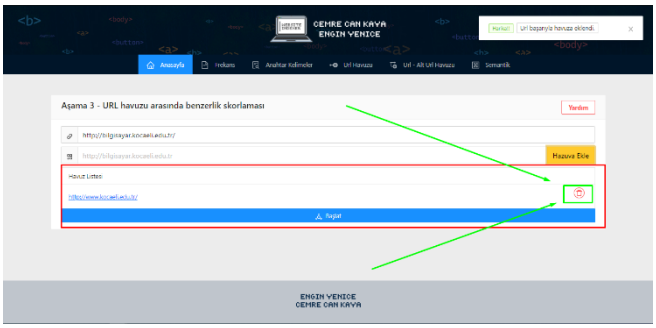
Ardından başlat butonuna tıklayarak gerekli sonuçları elde edebilirsiniz.

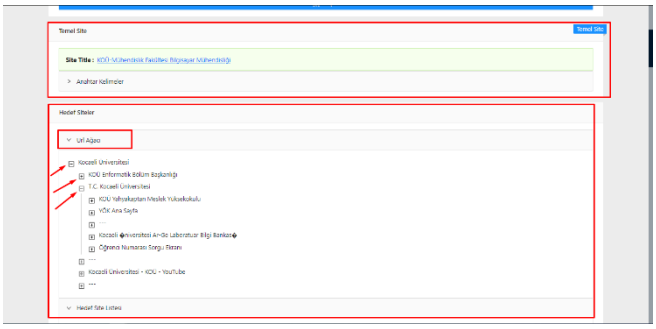
Oluşturulan sonuç ekranında üst tarafta hedeflediğiniz ana site ile ilgiler verilmektedir. Alt tarafta ise havuzunuzun içerisinde bulunan tüm sitelerin bilgileri ve benzerlik skorlaması bulunmaktadır.

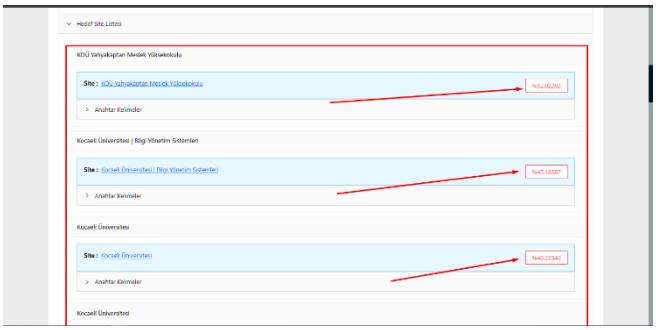


Oluşturulan ağacın solunda bulunan + ve – butonlarına tıklayarak tespit edilen alt URL'ler gösterilmektedir. (--- şeklinde gösterilen başlıklar ise sayfada başlık olmadığını belirtmektedir.)

Tespit edilen tüm sayfalar benzerlik skorlamasına göre büyükten küçüğe doğru sıralanmaktadır.







G. Semantik (Ödev Madde 5)

Sayfanın ortasında bulunan veri girişi nesnesi (input) içerisine kontrol edilmesi istenen link girilmektedir.

Ardından havuza eklenmesini istediğiniz linkleri havuza ekle butonunu kullanarak ekleyebilirsiniz.

Bir hata ile karşılaşılması durumunda ekranın sağ üst köşesinden çıkan bir açılır pencere (pop-up) ile bilgilendirme sağlanmaktadır.

Havuza eklemiş olduğunuz bir linki sağ tarafında bulunan çöp kutusu görünümündeki butona tıklayarak havuzdan kaldırabilirsiniz.

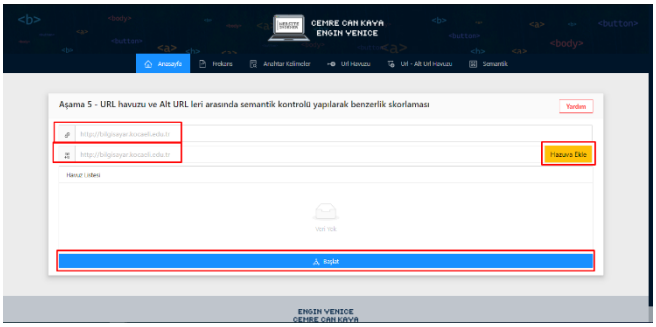
Ardından başlat butonuna tıklayarak gerekli sonuçları elde edebilirsiniz.

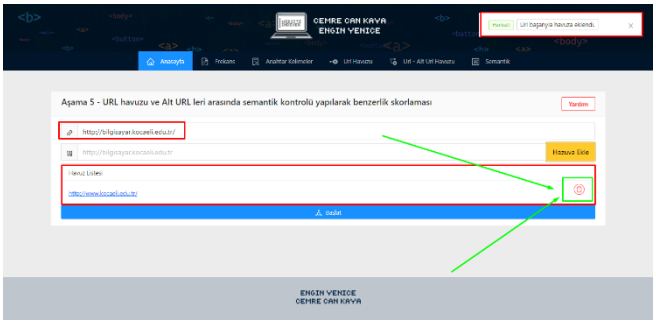
Oluşturulan sonuç ekranında üst tarafta hedeflediğiniz ana site ile ilgiler verilmektedir. Alt tarafta ise havuzunuzun içerisinde bulunan tüm sitelerin bilgileri ve benzerlik skorlaması bulunmaktadır.

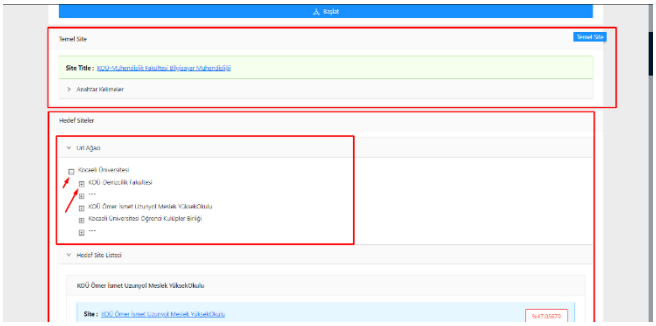
Oluşturulan ağacın solunda bulunan + ve – butonlarına tıklayarak tespit edilen alt URL'ler gösterilmektedir. (--- şeklinde gösterilen başlıklar ise sayfada başlık olmadığını belirtmektedir.)

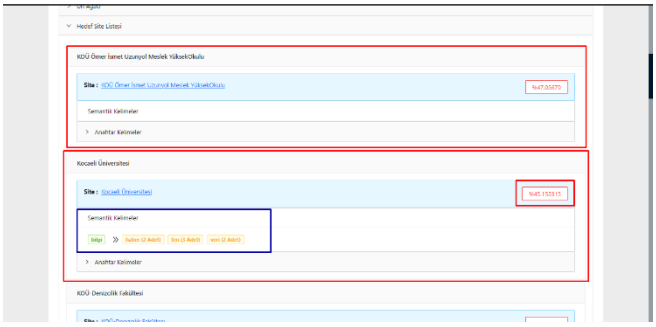
Tespit edilen tüm sayfalar benzerlik skorlamasına göre büyükten küçüğe doğru sıralanmaktadır.

Alt sitelerde bulunan semantik kelimeler ise tespit edildiği sitenin alt bilgileri arasında yer almaktadır.









V. ÇEVRİMİÇİ LİNKLER

A. *Backend (Arka Plan)*

1) *http://yazlab21.somee.com/*

B. *Frontend (Kullanıcı Arayüzü (UI))*

1) *https://siteindexer-729f6.web.app/home*

VI. KAYNAKÇA

1) *https://ng.ant.design/docs/introduce/en*
2) *http://angular.io/*
3) *https://enginyenice.com/*
4) *https://enginyenice.com/seo-ve-google-analytics/*