WEB İNDEKSLEME UYGULAMASI

Şevki Karagöl
Bilgisayar Mühendisliği Bölümü
Kocaeli Üniversitesi
170201009

Mustafa Yiğit
Bilgisayar Mühendisliği Bölümü
Kocaeli Üniversitesi
180201108

ÖZET- Bu projede, projeyi yapan kişiler için web tabanlı uygulamaların çalışma mantığını anlaması, bunun ardından web indeksleme yöntemleri hakkında bilgi edinilmesi ve web tabanlı uygulama yazma becerisinin geliştirilmesi amaçlanmaktadır. Bu amaç doğrultusunda öğrencilerden, bir web sitesinin frontend ve backend kısmının geliştirilmesi ve belli isterler doğrultusunda bu web sitesinin bir amaca hizmet etmesi istenmiştir.

Bu amaçlar ve isterler doğrultusunda Python (Flask), HTML,CSS (Bootstrap) ve internet üzerinden veri çekmek için çeşitli kütüphaneler kullanılarak verilen URL'ler üzerinden proje isterlerini uygulayan bir web sitesi geliştirilmiş, test çalışmaları yapılmış ve kullanıma hazır hale getirilmiştir. Web sitesi çevrimiçi yayına alınmamış ve "local" olarak çalışmaktadır.

Anahtar kelimeler-

Web geliştirme, frontend, backend, flask, bootstrap, indeksleme, veri, semantik analiz, benzerlik skorlaması, frekans

I.GİRİŞ

Bu projede web sitesi üzerinden kullanıcıdan URL'ler alınmış, ardından bu URL'lerdeki metin içerikleri kullanılarak web sitelerinde geçen kelimeler ve her kelimeye ait frekans değerleri bulunmuştur. Bu frekans değerleri ve birkaç parametreye bağlı olarak anahtar kelime listesinin çıkarılması, çıkarılan anahtar kelimelere uygulanan bir formül yardımıyla benzerlik skoru bulunması işlemleri yapılmıştır. Ardından sitenin alt dallarının tespit edilmesiyle site yapısı ortaya çıkarılmış ve bu yapıdaki her URL'e aynı analizler uygulanmıştır. Son olarak sitede aynı içerik içerisinde kullanılmış olan eş anlamlı sözcüklerin semantik analiz yardımıyla saptanması işlemi yapılmıştır. Kullanıcıların tüm bu işlemleri yapabilmesi için bir web sitesi tasarlanmıştır. İşlemler aşama aşama kullanıcının anlayacağı şekle getirilerek bilgi olarak bu web sitesinde sunulmuştur.

Bu projede Python web framework'ü olan Flask, Bootstrap, internetten veri çekmek için BeautifulSoup ve Request kütüphanelerinin bir arada kullanımına yönelik bir çalışma gerçekleştirilmiştir. Aynı zamanda öğrencilerin, proje isterlerlerinin çözümününe yönelik araştırdığı algoritmalar IDE aracılığıyla bilgisayar ortamına aktarılmıştır.

II.TEMEL BİLGİLER

Bu proje Python framework'ü Flask ile geliştirilmiş olup, geliştirme ortamı olarak "Visual Studio Code" kullanılmıştır. İlk etapta proje için bir yol haritası çıkarılarak ön hazırlık sürecine girilmiştir. Bu aşamada projenin isterlerine yönelik araştırmalar gerçekleştirilmesi adına grup içerisinde bir iş bölümü yapılmış olup elde edilen veriler doğrultusunda projenin ana hatları ortaya çıkarılmış ve büyük ölçüde karşılaşılabilecek problemler saptanıp çözümlendirilmeye çalışıldıktan sonra IDE ortamında projenin ilk adımları atılmıştır.

Yapılan ön hazırlık sürecinde web sitesi üzerinden URL'in alınması, alınan URL'in sahip olduğu metin içeriğinin işlenmesi, işlenen içerik üzerinden proje isterlerinin nasıl gerçekleştirilebileceği gibi problemler üzerinde durulmuştur. Bu konulara ve problemlere yönelik gerekli araştırmalar yapıldıktan sonra projeye şekil verme aşamasına gidilmiştir.

Proje ön hazırlık süreciyle birlikte yaklaşık on günlük bir süreçte tamamlanmıştır.

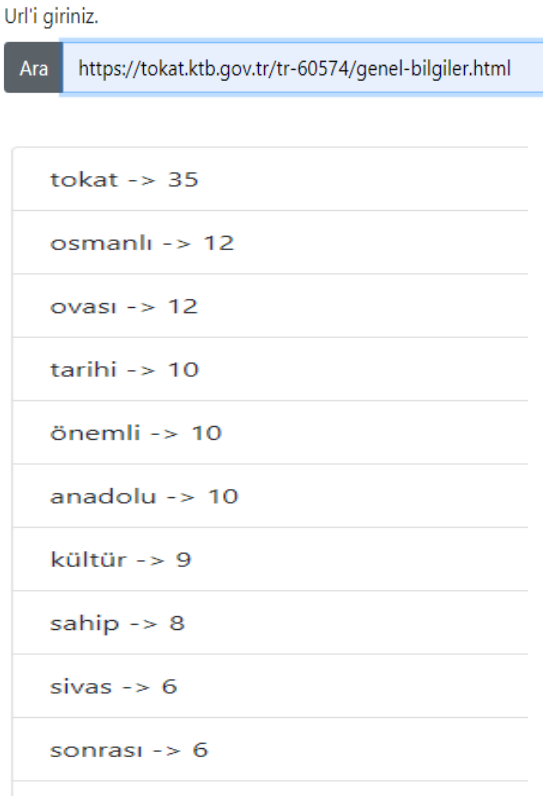
Not: İsterlerin tamamı eksiksiz bir şekilde tamamlanmıştır.

III.YÖNTEM

Bu projede izlenilen yol aşağıda anlatılmıştır:

Aşama 1:

İlk olarak kullanıcıdan URL alınmıştır. Alınan URL bir fonksiyona gönderilerek sayfada geçen tüm kelimelerin frekansının bulunması işlemi yapılmıştır. Kullanıcıya sayfada geçen kelimeler frekanslarıyla beraber "büyükten küçüğe sıralanarak" listelenmiştir.



Aşama 2-3:

İkinci aşama üçüncü aşamanın ön hazırlığı niteliğinde olduğu için bu iki aşama web sitesinde tek bir sayfada sunulmuştur.

Birinci aşamadaki kelime frekansları bulma fonksiyonu ve “title” etiketi üzerinden kelime bulma fonksiyonu kullanılarak bir anahtar kelime listesi bulma algoritması tasarlanmıştır. Ve tüm anahtar kelimeler alt alta listelenmiştir.

Üçüncü aşamada ise listelenilen bu anahtar kelimelerden faydalanılarak benzerlik skorlaması yapılmıştır. Benzerlik skorlamasında kullanılan formül şu şekildedir:

a = ortak anahtar kelimelerin toplam frekans değeri

b = tüm anahtar kelimelerin toplam frekans değeri

x = benzerlik skoru

x = (a / b) * 100

Aşama 4-5:

Beşinci aşama dördüncü aşamanın devamı niteliğinde olduğu için bu iki aşama web sitesinde tek bir sayfada sunulmuştur.

Dördüncü aşamada girilen URL ve URL kümesi arasında kümenin tüm birinci ve ikinci alt URL’leri de hesaba katılarak farklı bir benzerlik skorlaması bulma algoritması tasarlanmıştır.

Beşinci aşamada, dördüncü aşamada kullanılan algoritmaya ek olarak incelemesi yapılan URL içeriğinde geçen eş anlamlı kelimelerin bulunmasını sağlayan algoritma kullanılmıştır. Bu kelimelerin "alakalı anahtar kelimeleri" anahtar kelimelerin yanlarına yazılmıştır. Benzerlik skorlaması büyükten küçüğe sıralanmıştır.

1. yapay -> 132
2. zekâ -> 88
3. aı -> 33
4. nedir -> 33
5. oracle -> 33
6. türkiye -> 33
7. veri ---> bilgi -> 32
8. iş ---> işlem -> 19
9. makine -> 17

Dördüncü ve beşinci aşamaya dair genel çıktı ek olarak sunulmuştur. ***ek1**

IV.KABA KOD

1)Program çalıştı.

2)Kullanıcı "Aşama 1" sekmesine gitti ve URL girişi yaparak "Ara" butonuna bastı.

3)"Aşama 1" için çıktı görüntülendi.

4)Kullanıcı "Aşama 2|3" sekmesine gitti ve iki adet URL girişi yaparak "Ara" butonuna bastı.

5)URL'ler alınarak gerekli fonksiyonlara gönderildi.

6)Fonksiyonlardan döndürülen anahtar kelimelerin listesi, benzerlik skoruyla beraber web sitesinde gösterildi.

7)Kullanıcı "Aşama 4|5" sekmesine gitti. Bir adet URL ve bir adet URL kümesi girdi (URL kümesi alt alta girilmelidir, aksi takdirde hata ile karşılaşılır.).

8) Girilen URL'ler alt URL'leriyle birlikte "ağaç yapısı" şeklinde gösterildi.

9)Her sitenin anahtar kelimeleri ve eğer sitede varsa eş anlamlı kelimeleri yan yana gösterildi.

10)Alt URL'ler de kullanılarak bulunan benzerlik skoru gösterildi.

V.REFERANSLAR

[1]https://www.youtube.com/channel/UCZNZj3mkdCGJfCoKyl4bSYQ

[2]https://flask.palletsprojects.com/en/1.1.x/#user-s-guide

[3]https://getbootstrap.com/docs/4.0/getting-started/introduction/

[4] https://www.geeksforgeeks.org/

[5] https://www.moradam.com/

[6] https://es-anlamli.gen.sx/

[7] https://stackoverflow.com/

[8] https://www.w3resource.com/

[9]https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[10]http://repository.bilkent.edu.tr/handle/11693/23211

***ek1**

Ağaç Yapısı

- https://tr.wikipedia.org/wiki/Yapay_zek%C3%A2

100.0

1. yapay -> 75
2. zekâ -> 34
3. değiştir -> 20
4. insan ---> in -> 20
5. kaynağı -> 19
6. vikipedi -> 18

- https://www.oracle.com/tr/artificial-intelligence/what-is-ai/

54.56

1. yapay -> 132
2. zekâ -> 88
3. ai -> 33
4. nedir -> 33
5. oracle -> 33
6. türkiye -> 33
7. veri ---> bilgi -> 32
8. iş ---> işlem -> 19
9. makine -> 17

- https://www.sozcu.com.tr/

0.0

1. dünya -> 23
2. ekonomi -> 12
3. son -> 11
4. keşfet -> 10
5. gündem -> 10
6. gerçek ---> doğru -> 6
7. sözcü -> 5
8. gazetesi -> 5
9. türkiyenin -> 5
10. tek -> 5

- https://www.sozcu.com.tr/hayatim

1. burcu -> 12
2. haberler -> 6
3. bin -> 5
4. hayat ---> yaşam -> 5
5. gün -> 4
6. magazin -> 3
7. kültür -> 3
8. sanat -> 3
9. haberleri -> 3
10. şık -> 3

- https://www.sozcu.com.tr/hayatim/magazin-haberleri

• • •