Bilgisayar Mühendisliği Bölümü

Yazılım Laboratuvarı II 2020-2021 Bahar Dönemi Proje I

Web İndeksleme Uygulaması

190201105 Alperen İleri-170201093 İlayda Dişiaçık

I.Proje Tanımı ve Aşamaları

I-I.Tanım

"Web İndeksleme Uygulaması" adlı proje verilen bir URL'deki web sayfa içeriğine göre diğer birden fazla web sayfasını benzerlik bakımından indeksleyip sıralanması sağlayan web tabanlı bir uygulamadır. Bu proje beş isterden oluşmaktadır. İlk ister için web sitesinde URL girilecek bir alan oluşturulacak ve oraya girilen URL'nin metnindeki her kelime için frekans hesaplama yapılacaktır. İkinci ister için web sitesindeki alana girilen URL metninde geçen kelimelerden en önemlileri belirlenerek anahtar kelimeler çıkartılacak. Üçüncü ister için girilen iki adet URL için benzerlik skoru hesaplanması yapılacak ve birinci URL'nin tüm anahtar kelimeler ve birinci URL içeriğinde frekansı, bu iki URL için benzerlik skoru, ikinci URL'de geçen tüm anahtar kelimeler ve bunların ikinci URL içeriğindeki frekansları yazdırılacak. Dördüncü ister için girilen bir URL'nin içeriği ile web site kümesindeki her bir web sayfasının içeriklerinin benzerlik skorları ayrı ayrı hesaplanacaktır. Tüm alt URL'ler dikkate alınacak ve en son web siteleri benzerlik skorlarına göre sıralanacaktır. Son isterde ise verilen URL içerisindeki anahtar kelimelerle alakalı kelimeler bulunacaktır.

I-II.Aşamalar

Projede bize gerekli olacak toplam kelime, en çok kullanılan kelime, benzer kelime frekansları gibi listeler ve beş adet URL için değişkenler tanımlandı. URL'nin metin içeriğinde geçen kelimeleri almak için BeautifulSoap kütüphanesinden yararlanarak kazma işlemi yapıldı ve p tagine sahip kelimler listede biriktirildi. strip() fonksiyonu ile metnin içeriği kelime kelime ayrıldı ve noktalama işaretleri çıkartılıp büyük harfler küçük harf haline getirildi. Bu kelimeler içerisinden en çok kullanılan n tane kelime alındı ve bu aşamalar her bir URL için tekrarlandı. Kelimeler alındıktan sonra intersection() fonksiyonu ile URL'ler arasındaki ortak kelimeler değişkenlerde tutuldu. Tutulan kelimelerin toplam frekans hesabı yapıldı. Kelime frekanslarının len() fonksiyonu ile bulduğumuz kelimelerin sayısına bölünmesiyle benzerlik skoru oluşturuldu.

Web sitesinin giriş ekranına beş adet URL için alan ve hesaplama işlemine geçilmesi için submit butonu koyuldu. asama1, asama2, asama3 ve asama4 olmak üzere ayrı sayfalarda sonuçlar yazdırıldı.

II.Temel Bilgiler ve Yapılan Araştırmalar

II.I-Proje Sırasında Yararlanılan Teknolojiler

Proje, Python programlama dili kullanılarak PyCharm geliştirme ortamında oluşturuldu. Projeyi geliştirirken, Python dilinin bize sunduğu kütüphane ve fonksiyonlardan yararlanıldı. Bunlara; json, collections request gibi yapılar örnek verilebilirken BeautifulSoup, Flask gibi frameworkler ile de desteklenildi.

II.II-Yapılan Araştırmalar

Proje için web sitelerinden veri çıkarmayı ve bu dağıtık haldeki verileri daha düzgün şekilde sunmayı sağlayan web scraping araştırıldı.

Bunu gerçekleştirmemizi sağlayacak olan Python'un BeautifulSoup kütüphanesinin kullanımı araştırıldı ve makaleler incelendi. Karşılaştırma ve benzerlik gibi işlemleri gerçekleştirecek Python fonksiyonları araştırıldı ve örnekler incelendi. Python ile web tarafında kullanılabilecek bir framework olan Flask araştırıldı.

III.Genel Yapı

III.I-Kullanıcı Kısmı

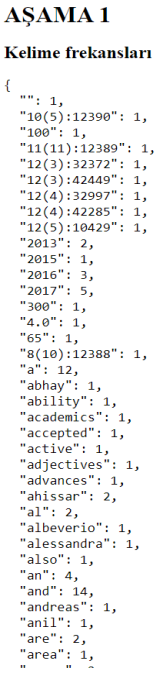
Kullanıcı programı çalıştırdığında bir web arayüzü oluşacaktır. Bu arayüzde URL girilebilmesi için beş adet boş alan, işlemleri gerçekleştirmesi için de bir submit butonu bulunmaktadır. URL'leri doldurmadan submit butonuna tıklandığında hata verecektir. URL'ler girilip submit butonuna tıklandığında tüm maddeler için tek tek hesaplama yapılacak ve sonuçlar ayrı sayfalar üzerinde gösterilecektir.

III-II-Kod Kısmı

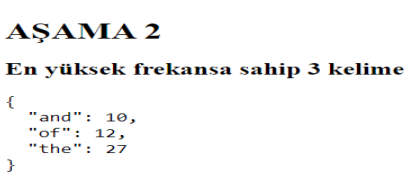
Programa toplam 290 kod satırdan oluşmaktadır.

IV.Ekran Çıktıları

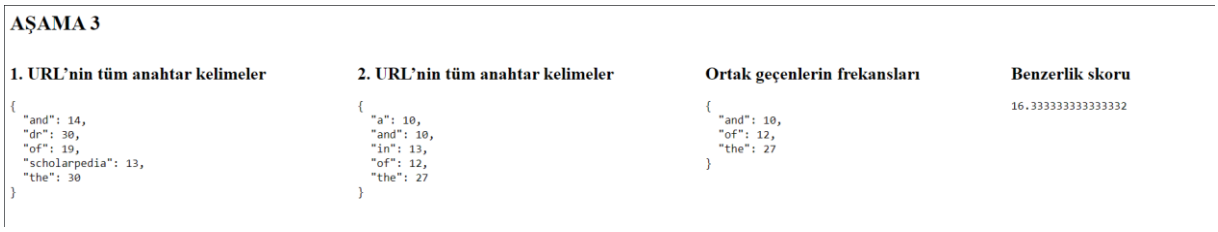
URL'lerin alındığı giriş ekranı



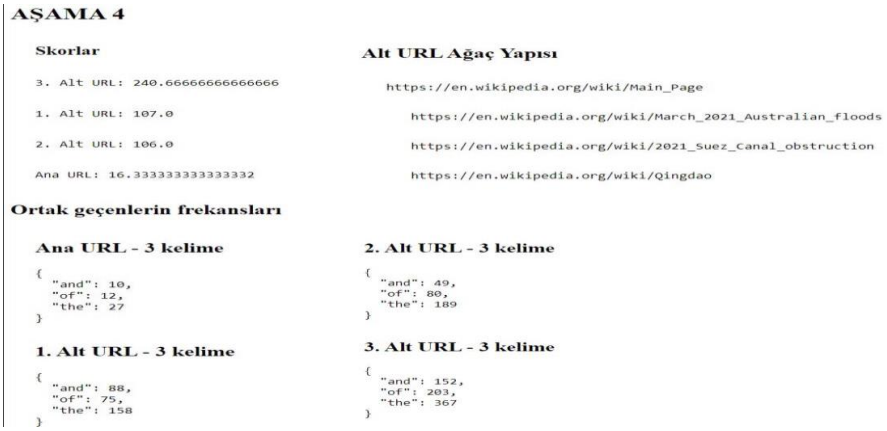
1.maddenin ekran çıktısı



2.maddenin ekran çıktısı



3.maddenin ekran çıktısı



4.maddenin ekran çıktısı

V.Kaynakça

https://docs.python.org/3/library/json.html

https://www.w3schools.com/python/python_json.asp

https://www.crummy.com/software/BeautifulSoup/bs4/doc/

https://docs.python.org/3/library/collections.html

https://www.udemy.com/course/python-dersleri/

https://www.tutorialspoint.com/flask/index.htm

https://medium.com/@awesome_nyn/python-ile-flask-microframework-kullanarak-nas%C4%B1l-web-projesi-olu%C5%9Fturulur-fbca456e7c71