BÜYÜK VERİDE MULTI-THREADING İLE BENZER KAYITLARIN TESPİT EDİLMESİ

Berk Sunduri

Bilgisayar Mühendisliği Bölümü

Kocaeli Üniversitesi

berksunduri@gmail.com

1-)Projenin Tanımı

Bu kısım sadece projenin açıklamasını okuyup edindiğim ön bilgiye göre yazılmıştır.

Bize verilen pdf dosyasında projenin amacını öğrendim. Verilen isterleri dikkatlice okudum

Projede bizden istenen şeyin multi-threading kullanarak bize sunulan bir dosyaya benzerlik testi yapıp istenilen değerleri görmemizi sağlayacak bir arayüz oluşturmamız olduğunu anladım.

1.1-)Problem Tanımı

Bu kısımda bizden yapmamız istenilenler içermektedir.

Birinci adımda bizden veri setinin hocaların isteği üzerine yeniden düzenlememiz istenmiştir.

İkinci ister olarak düzenlenmiş veri setinde ki kayıtlar arasında benzerlik kontrolü yapmamız istenmiştir. Buna ek olarak kaç threadle çalışmamız gerektiği uygulama arayüzünden seçilmesi beklenmektedir.

Üçüncü aşamada ise bizden Her threadin çalışma zamanını ve tüm threadler için toplam çalışma zaman bilgilerinin uygulama arayüzünde göstermemiz istenmiştir.

Dördüncü ister olarak bizden istenilen sütün ya da sütünlar arasında ki girilen benzerlik oranı ve üzerinde benzerliğe sahip kayıtları masaüstü uygulamamızda göstermemiz istendi.

Beşinci ve son ister olarak uygulama için basit bir arayüz oluşturmamız istenmiştir. Bu arayüzde benzerlik oranının seçilebileceği bir araç, benzerliklerinin araştırılması istenen sütun veya sütünların seçilebileceği bir araç, kaç tane thread kullanılacağının seçilebileceği bir araç, Her bir threadin çalışma zamanını gösteren bir araç ve sonuçların açıkça ekranda gösterilebileceği bir araç yapmamız istenmiştir.

2-)Yapılan Araştırmalar ve Karşılaşılan Sıkıntılar

Bu kısım proje öncesi ve sonrası araştırmaları ve de projenin yapım aşamasındaki sıkıntıları ve çözümlerini içermektedir.

İlk karşılaştığım sorun hangi programlama dilini kullanacağım olduğuydu. Bunun için Python dilini kullanmada karar kıldım. Veri işlemek için uygun bir dil olacağını düşündüm fakat threading kısmı beni hayal kırıklığına uğrattı.

Daha sonrasında dosyada sütünlar arası benzerlik testi yapmak için farklı algoritmalar denedim. Fakat bir çok algoritmanın ya implemantasyonu çok zor ya da imkansız ya da testi yapması hem çok uzun sürüyor hem de testler her zaman tutarlı olmuyordu. Sonunda kendi yazdığım basit "Brute Comparing" denilebilecek bir algoritmayı kullanmakta karar kıldım. İşlem sırasında stringleri sürekli bölmek zorunda kaldığım için algoritma yavaş çalışıyor fakat tek doğru sonuç alabildiğim algoritma bu oldu.

2.1-Proje Sırasında Yararlanılan Teknolojiler

Projeyi Python dili kullanarak PyCharm IDE'sinde yazdım.

Pythonun çeşitli kütüphanelerinden yararlandım.

3-)Tasarım

3.1-Akış Diyagramı

Kısım ektedir.(1)

4-)Genel Yapı

4.1-Kullanıcı Kısmı

Program çalıştığında karşımıza oluşturduğum User Interface çıkmaktadır. Kullanıcı bu kısımda yapmak istediği işlemleri istediği şekilde gerçekleştirebilir.

Kullanıcı istediği değerleri girip Search butonuna bastığı anda program istenilen işlemleri yaparak kullanıcının karşısına bir dataframe sunmaktadır.

4.2-Kod Kısmı

Kod kısmına baktığımızda ise kod fonksiyonlar yazılarak geliştirilmiştir. En başta dosyaların yüklenmesi, dosyaların üzerinde istenilen işlemlerin yapılması ve hemen altında da karşılaştırma fonksiyonu bulunmaktadır.

Bundan sonra programın çalışması ve yazılabilirliğini kolaylaştırmak için tam beş tane ana fonksiyon ayrı olarakta arayüzün oluşturulmasında on bir tane fonksiyon bulunmaktadır.

Kullandığım fonksiyonlar csv dosyasını düzenlemek ve sonrasında istek üzerine bağlı olarak farklı karşılaştırma algoritmalarını içerir.

Kullandığum kütüphaneler raporun 5. bölümünde belirtilmiştir.

5-)Kütüphaneler

Bu kısımda projeye import ettiğim kütüphaneler bulunmakta:

- pandas
- csv
- numpy
- re
- time
- nltk
- tkinter

7-)Referanslar

1-)Tkinter Treeview

”www.plus2net.com”

2-) CSV File Operations

”stackhowto.com “

3-) Stack Over Flow
”https://www.stackoverflow.com”

4-)Pandas Documentation

“pandas.pydata.org”

5-)Data Analysis and Visual Presentation

“datacarpentry.org”

6-)Corey Schafer-Pyhton Threading Tutorial

7-)Multithreading in Pyhton

”www.geeksforgeeks.com”