BÜYÜK VERİDE MULTITHREADING İLE BENZER KAYITLARIN TESPİT EDİLMESİ

Eray Karataş

200202079

Engin Tosun

200202028

I. Özet

Bu rapor Yazılım Laboratuvarı I Dersinin 2. Projesini açıklamak ve sunumunu gerçekleştirmek amacıyla oluşturulmuştur. Bu proje Java dilinde Netbeans ortamında ve kaggle üzerinde python kullanarak geliştirilmiştir.Raporda projenin tanımı, özet, yöntem,karşılaşılan sorunlar ve çözümler,sözde kod, sonuç bölümünden oluşmaktadır. Proje aşamasında yararlanılan kaynaklar raporun son bölümünde bulunmaktadır.

II. Proje Tanimi

BU projede müşteri şikayetleri kayıtlarının tutulduğu bir veri seti içerisindeki benzer kayıtlar tespit edilecek ve tespit edilen kayıtlar masaüstü uygulamasında gösterilecektir. Multithreading kullanarak benzerlik arama süresini düşürmek amaçlanmaktadır.

Projedeki amaçlar:

1. Veri seti içerisindeki arama işlem süresini multithreading kullanılarak azaltmak. 2. Belirtilen sütun/sütunlar için her bir satırdaki kayıtların birbiriyle kelime bazlı karşılaştırılması ve aralarındaki benzerliğin tespit edilmesi. 3. Uygulama içerisinde istenen özelliklere göre kayıtları filtrelemek ve kullanıcıya göstermek. 4. Masaüstü uygulama geliştirme hakkında bilgi ve beceriye sahip olmak.

Multithreading (Çok İş Parçacıklı Çalışma):

Multithreading (çok iş parçacıklı çalışma), bir merkezi işlem biriminin (CPU) (veya çok çekirdekli bir işlemcideki tek bir çekirdeğin) aynı anda işletim sistemi tarafından desteklenen birden çok yürütme iş parçacığı sağlama yeteneğidir.
Bu tür programlamada birden çok iş parçacığı aynı anda çalışır. Çok iş parçacıklı model, sorgulamalı olay döngüsü kullanmaz. CPU zamanı boşa harcanmaz. Boşta kalma süresi minimumdur. Daha verimli programlarla sonuçlanır. Herhangi bir nedenle bir iş parçacığı duraklatıldığında, diğer iş parçacıkları normal şekilde çalışır.

Veri Seti:

Bu veri seti; finansal ürünler ve hizmetler hakkında alınan gerçek dünya şikayetlerini içermektedir. Veri seti,müşterilerin Kredi Raporları, Öğrenci Kredileri, Para Transferi vb. gibi finans sektöründeki birden fazla ürün ve hizmet hakkında yaptığı şikayetlerin farklı bilgilerini içermektedir.

Veri seti aşağıdaki kurallara uygun olacak şekilde yeniden düzenlenmelidir: Elde edilen tabloda 6 farklı sütun bulunmalıdır: Product (Ürün), Issue (Konu), Company (Şirket), State, Complaint ID, Zip Code.
Null değer içeren kayıtlar bulunmamalıdır.
Kayıtlardaki noktalama işaretleri kaldırılmalıdır.
Kayıtlardaki stop word'ler kaldırılmalıdır (nltk kütüphanesi kullanılabilir).

Benzerlik Tespiti:

Geliştirilecek projede tüm kayıtlar arasındaki benzerlik ilişkisinin incelenmesi beklenmektedir.
Bu nedenle her bir kaydın diğer bir kayıtla karşılaştırılması gerekmektedir. Karşılaştırmanın mümkün olduğunca hızlı olması için multithread kullanılmalıdır.
Benzerlik, kayıtların içerdikleri ortak kelime sayısına göre olmalıdır. Örneğin; ilk kayıt 5, ikinci kayıt 4 kelimeden oluşuyorsa ve ortak kelime sayısı 2 ise benzerlik oranı;

III. Yöntem

Projede istenen isterler için veri seti bize verilen link üzerindeki Kaggle'dan notebook açılıp python dili ile istenilen şekilde düzenlenmesi yapıldı.
Bu düzenlemede izlediğimiz adımlar şu şekildedir. Pandas dataframe'ini import etmek. Rows.csv dosyasını kaggle notebook'una yüklemek. İstenilen 6 sütunu rows.csv dosyasından çekmek. Dropna() fonksiyonu ile null değerleri temizlemek. Nltk kütüphanesini kullanarak stop wordleri temizlemek.
Son olarak import string kütüphanesi ile noktalama işaretlerini

veri setinden arındırdık. Güncel istenilen şekildeki dosyamızı dfc.tocsv('rows.csv') files.download('rows.csv') komutları ile indirdik.

Multithreading ve arayüzü yapmak için java programlama dilini kullandık.Bu kısım şu şekilde yapıldı:

Düzenlenmiş rows.csv dosyasındaki verileri bufferedreader kullanarak arraylistlere attık.Thread havuzu oluşturup bu havuzlarda birden fazla thread çalıştırılabilmesini sağladık.
Thread classlarına gidip burada verilerin tutulduğu arraylistleri kullanıp kıyaslama algoritmalarını yazdık.

Bu algoritmalarda verilerin benzerlik oranlarını hesaplayıp yeni oluşturduğumuz arraylistlere attık.Bu işlemlerden sonra thread havuzunu kapatıpbu zamana kadar ki her bir threadin ve toplam threadlerin çalışma süresini hesapladık.

Arayüzü çalıştırıp,arayüzün bulunduğu classdaki ilgili constructora benzerlik oranı ve bu orana ait verileri attık.
Arayüzde buton,label,combobox, jtable gibi yapılar oluşturup bu yapılarda verilere ve thread sürelerine ait istenen sonuçları gösterdik.

Projede Yapılması istenen isterler hakkında

Projede yapılması gerekenler 1. Verilen veri seti istenen şekilde yeniden düzenlenmelidir.6 farklı sütun bulunmalıdır: Product (Ürün), Issue (Konu), Company (Şirket), State, Complaint ID, Zip Code.

2. Düzenlenmiş veri setindeki kayıtlar arasında benzerlik kontrolü yapılmalıdır. Kontrol sırasında mutlaka multithreading kullanılmalıdır.
Multithreading için kullanılacak thread sayısı uygulama arayüzünden girilmelidir.

3. Her thread'in çalışma zamanı ve tüm thread'ler için toplam çalışma zamanı bilgileri uygulama arayüzünde gösterilmelidir.

4. İstenilen sütun ya da sütunlar arasındaki girilen benzerlik oranı (threshold) ve üzerinde benzerliğe sahip kayıtlar masaüstü uygulamasında gösterilmelidir.

5. Uygulamanızı sunmak üzere basit bir arayüz geliştirmeniz istenmektedir. Arayüz Özellikleri:

• Benzerlik oranının (Threshold değeri) seçilebileceği / girilebileceği bir araç,
• Benzerliklerinin araştırılması istenen sütun veya sütunların seçilebileceği bir araç,
• Kaç tane thread kullanılacağının seçilebileceği / girilebileceği bir araç, • Her bir thread'in çalışma zamanını ve toplam çalışma zamanını gösteren araçlar
• Sonuçların açıkça ekranda gösterilebileceği bir araç.

IV. Sözde Kod

1-BAŞLA

2-VERİLERİN TUTULACAĞI "ArrayList" LERİ OLUŞTUR

3-CSV DOSYASINDAN VERİLERİ OKUYUP "ArrayList" LERE AT

4-VERİLERİ SÜTUNLAR(Product,Issue gibi)ŞEKLİNDE AYIRIP HER BİR SÜTUNUN BİR "ArrayList" TE TUTULDUĞU YENİ "ArrayList" LERE AT

5-KULLANİCİDAN THREAD SAYİSİ VE AYNI ANDA ÇALIŞACAK THREAD SAYISI AL

6-KULLANİCİDAN ALİNAN THREAD SAYİSİ VE AYNİ ANDA ÇALIŞACAK THREAD SAYİSİNA GÖRE THREAD HAVUZU OLUŞTUR

7-THREADLERİN OLDUĞU CLASSLARA GİT

8-KIYASLAMA ALGORİTMALARININ BULUNDUĞU "run" METHODUNU ÇALIŞTIR

9-"ArrayList" LERDE TUTULAN VERİLERİN HER BİRİYLE BENZERLİK ORANLARINI BUL

10-BULUNAN BENZERLİK ORANINA SAHİP VERİLERİ,VERİLERİN SIRALARINI VE BENZERLİK ORANLARINI YENİ "ArrayList" LERE AT

11-THREAD HAVUZUNUN ÇALIŞMASINI DURDUR

12-THREADLERİN ÇALIŞMA SÜRESİNİ HESAPLA

13-THREADLERİN TOPLAM ÇALIŞMA SÜRESİNİ HESAPLA

14-ARAYÜZÜN BULUNDUĞU CLASSDAKİ İLGİLİ "Constructor" A BENZERLİK ORANI VE BU ORANA AİT VERİLERİN TUTULDUĞU "ArrayList" LERİ AT

15-ARAYÜZÜ ÇALIŞTIR

16-ARAYÜZE BUTON,LABEL,COMBOBOX,JTABLE EKLE

17-BUTON,COMBOBOX LARIN ”ActionPerformed” METHODLARINI EKLE

18-”ActionPerformed” METHODLARINA YAPILACAK İŞLEMLERİ EKLE

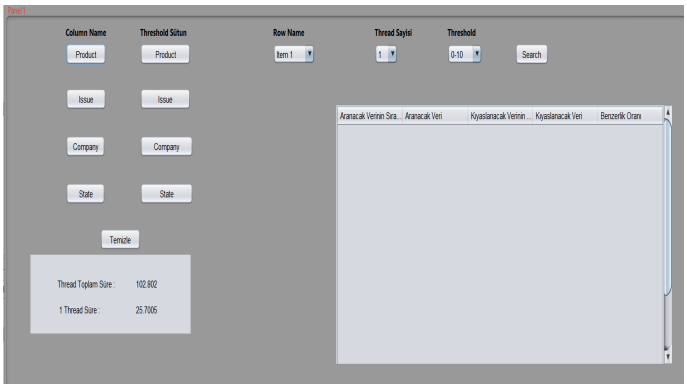
19-1 THREADİN ÇALIŞMA SÜRESİ VE THREADLERİN TOPLAM ÇALIŞMA SÜRELERİNİ LABEL'A YAZDIR

20-KULLANICIN ARAYÜZÜ NASIL KULLANACAĞINA BAĞLI OLARAK ”ActionPerformed” METHODLARINDAN ”addRowToJTable” METHODLARINA GİT VE JTABLE A VERİ YAZDIR

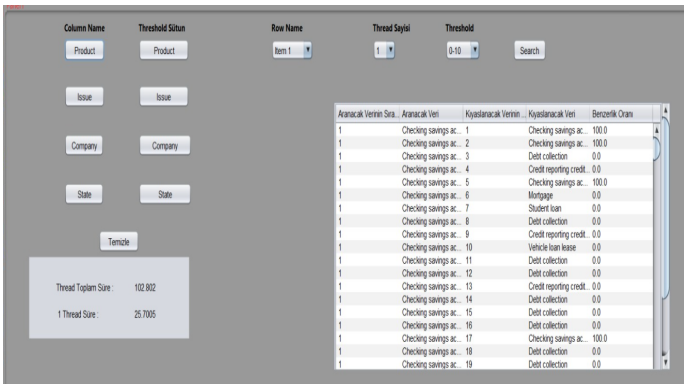
21-ÇIKIŞ

V. Sonuç

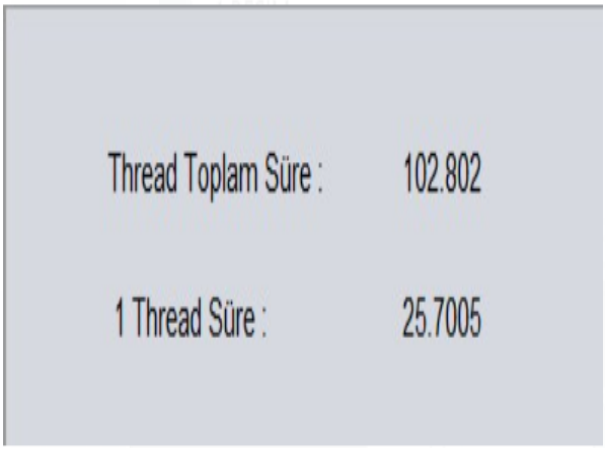
Masaüstü uygulama arayüzü:



Listeleme:



Thread süreleri:




```
Company Sutunundaki 10.ci kayit ile 9523.ci kayit arasindaki benzerlik orani:%25.0
Company Sutunundaki 10.ci kayit ile 9524.ci kayit arasindaki benzerlik orani:%33.3
Company Sutunundaki 10.ci kayit ile 9525.ci kayit arasindaki benzerlik orani:%0.0
Company Sutunundaki 10.ci kayit ile 9526.ci kayit arasindaki benzerlik orani:%33.3
Company Sutunundaki 10.ci kayit ile 9527.ci kayit arasindaki benzerlik orani:%0.0
Company Sutunundaki 10.ci kayit ile 9528.ci kayit arasindaki benzerlik orani:%33.3
Company Sutunundaki 10.ci kayit ile 9529.ci kayit arasindaki benzerlik orani:%0.0
Company Sutunundaki 10.ci kayit ile 9530.ci kayit arasindaki benzerlik orani:%25.0
Company Sutunundaki 10.ci kayit ile 9531.ci kayit arasindaki benzerlik orani:%0.0
Company Sutunundaki 10.ci kayit ile 9532.ci kayit arasindaki benzerlik orani:%33.3
Company Sutunundaki 10.ci kayit ile 9533.ci kayit arasindaki benzerlik orani:%0.0
Company Sutunundaki 10.ci kayit ile 9534.ci kayit arasindaki benzerlik orani:%0.0
Company Sutunundaki 10.ci kayit ile 9535.ci kayit arasindaki benzerlik orani:%25.0
```

```
Thread 1 isini bitirdi

Threadlerin toplam calisma suresi:117.896 sn

Threadlere gelene kadar ki calisma suresi:9.192 sn

Threadlerin uyku suresi: 2 sn

Tek bir threadin calisma suresi:29.474
```

VI. Kaynakça

https://www.udemy.com/course/sifirdan-ileri-seviyeye-komple-java-gelistirici-kursu/

https://ufukuzun.wordpress.com/2015/04/05/javada-multithreading-bolum-5-thread-havuzlari-thread-pools/

https://www.w3schools.com/java/java$_t hreads.asp$

https://www.yusufsezer.com.tr/java-thread/

https://docs.oracle.com/javase/tutorial/uiswing/

https://www.w3schools.com/python/pandas/default.asp

https://docs.python.org/3/library/string.html Genel Sorunlar için; -stackoverflow.com -theprogrammershangout.com LaTeX Raporu hazırlamak için gerekli ekipman ve bilgiler; -www.overleaf.com