KOCAELİ ÜNİVERSİTESİ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ
YAZILIM LAB. I - 2. Proje

PROJE TESLİM TARİHİ: 20.11.2022

BÜYÜK VERİDE MULTITHREADING İLE BENZER KAYITLARIN TESPİT EDİLMESİ

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.

Amaç:

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.

Programlama Dili: C#, Java, Python, vb.

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.

Daha fazla bilgi için aşağıdaki linklere bakabilirsiniz:

- https://mertmekatronik.com/thread-ve-multithread-nedir
- https://www.tutorialspoint.com/operating_system/os_multi_threading.htm
- https://www.javatpoint.com/multithreading-in-java
- https://totalview.io/blog/multithreading-multithreaded-applications#:~:text=Multithreading%20is%20a%20model%20of,to%20their%20own%20CPU%20core.
- https://www.geeksforgeeks.org/multithreading-python-set-1/

Veri seti:

- Kullanılacak veri setine aşağıdaki linkten ulaşabilirsiniz:

https://www.kaggle.com/datasets/selener/consumer-complaint-database

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;

$$\frac{2\ (benzer\ kelime\ sayısı)}{5\ (uzun\ olan\ kaydın\ kelime\ sayısı)} = \%40'tır.$$

Kayıtlar arasında olabilecek benzerlik oranları için örnek tablo aşağıda verilmiştir:

Kayıt 1	**Kayıt 2**	**Benzerlik Oranı**
Debt collection	Debt collection	%100
Debt collection	Mortgage	%0
Managing loan lease	Problems end loan lease	%25
Managing loan lease	Struggling pay loan	%33.3
Credit reporting credit repair services personal consumer reports	Payday loan title loan personal loan	%12.5

Credit reporting credit repair services personal consumer reports	Credit card prepaid card	%25
Closing account	Managing account	%50
Closing account	Problem lender company charging account	%20

İsterler:

1. Verilen veri seti istenen şekilde yeniden düzenlenmelidir.
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. Bu arayüz aşağıdaki isterleri içermelidir:
    - 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.
6. Uygulamada aşağıdaki ve benzer senaryolardan elde edilen sonuçlar ekranda gösterilmelidir:
    - Senaryo 1: Ürün (Product) sütununda %60 ve üzeri benzer olan kayıtları ekranda gösteriniz.
    - Senaryo 2: Aynı ürünler (Product) için % 70 ve üzeri benzerlikteki konuları (Issue) içeren Şirket (Company) isimlerini ekranda gösteriniz.
    - Senaryo 3: 'Complaint Id' = 3198084 olan şikayet kaydı için % 50 ve üzeri benzerlikteki konuları (issue) içeren kayıtları ekranda gösteriniz.
    - Senaryo 4: 5 Thread ile Konular(Issue) sütununda %80 ve üzeri benzer olan kayıtları ekranda gösteriniz

Proje Teslimi:

- Rapor IEEE formatında (önceki yıllarda verilen formatta) 4 sayfa, akış diyagramı veya yalancı kod içeren, özet, giriş, yöntem, deneysel sonuçlar, sonuç ve kaynakça bölümünden oluşmalıdır. Raporda kullanılan algoritma açıklanmalı ve algoritmanın kaba kodu yazılmalıdır.
- Dersin takibi projenin teslimi dahil edestek.kocaeli.edu.tr sistemi üzerinden yapılacaktır. edestek.kocaeli.edu.tr sitesinde belirtilen tarihten sonra getirilen projeler kabul edilmeyecektir.
- Proje ile ilgili sorular edestek.kocaeli.edu.tr sitesindeki forum üzerinden Arş. Gör. Ayşe Gül EKER veya Arş. Gör. Gamze KORKMAZ ERDEM'e sorulabilir.
- Demo sırasında algoritma, geliştirdiğiniz kodun çeşitli kısımlarının ne amaçla yazıldığı ve geliştirme ortamı hakkında sorular sorulabilir. Kullandığınız herhangi bir satır kodu açıklamanız istenebilir.