KOCAELİ ÜNİVERSİTESİ
MÜHENDİSLİK FAKÜLTESİ
BİLGİSAYAR MÜHENDİSLİĞİ

YAZILIM LABORATUVARI-1
PROJE -2

200202113
Yunus KARAMAN

BÜYÜK VERİDE MULTITHREADING İLE BENZER KAYITLARIN TESPİT EDİLMESİ

1) Projenin Özeti:

Bu proje de python yazılım dili kullanılmıştır. Jupyter notebook üzerinde masaüstü projesi geliştirilmiştir

Problemin Tanımı:

Proje için gerekli olan bir veri seti vardır. Daha önceden hazır verilen bu veri setindeki ürünler ve konular üzerinde multithreading kullanarak bir benzerlik tespiti yapılmak istenmektedir. Ürünler kolonunda müşteri şikayetleri vardır. Benzer kayıtlar tespit edilecek ve tespit edilen kayıtlar masaüstü uygulamasında gösterilecektir.

Proje amacı:

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. Uygulama içerisinde istenen özelliklere göre kayıtları filtrelemek ve kullanıcıya göstermek.

2.GİRİŞ

Proje python yazılım dili ile jpyter notebook üzerinde yazılmaktadır. Projede hazır veri seti kullanılmaktadır. Hazır veri setinideki boşlukları noktalama işaretleri ve gereksiz kelimeler kaldırılmaktadır. Yapılan temizleme işlemi üzerinden Veri seti içerisindeki ürünler ve konular üzerinde benzerlik tespiti yapılmaktadır. Veri setindeki her bir kayıt için ayrı ayrı veri benzerlik tespiti yapılmaktadır. Veri seti içerisindeki ürünler ve konular kısmındaki benzerlik tespiti ürünler sutunu ürünlerle konular sutunu konular ile benzerlik tespiti yapılır. Karşılaştırılcak olan kayıtlar alınır içerisindeki kelimeler karşılaştırılır ve benzer olan kelimler tespit edilir. Tespit edilen kelimelerin sayısı karşılaştırılan kayıtlar arasındaki kelime sayısı fazla olan kayıdın kelime sayısına

bölünür ve buradan yüzde kaç benzer olduğuna dair bir sonuç çıkarılır. Uygulama da 4 farklı benzerlik tespiti yapılmaktadır. Kullanıcıdan bir yüzde istenir ve girilen yüzde ile tüm ürünler üzerinden benzerlik oranları gösterikmektedir ve Tüm konular üzerinden benzerlik oranları gösterilmektedir. Kullanıcı veri setindeki ürünler için bir Id ve yüzde girerek Id ye ait olan kayıtın diğer tüm ürünler ile karşılaştırılır ve benzerlik oranları tespit edilir. Girilen Id ait olan benzerlik oranları ile tespit edilen kayıtlar üzerinden kullanıcı yüzde girerek bu kayıtlar ile konuları arasında benzerlik oranları tespit edilmektedir. Gösterilen benzerlik oranları için multithread kullanılmaktadır. 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. Ürünler için yapılan benzerlik tespiti için kulanılan multiThread ile işlemler daha kısa sürede yapılmaktadır.

3.YÖNTEM

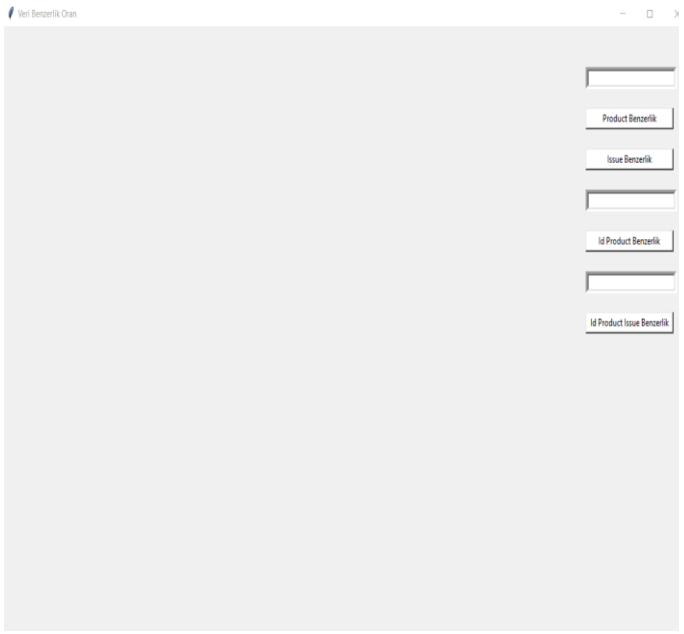
Bu proje de bizden istenilen multithread kullanılarak verilen veri setindeki ürünler ve konular olmak benzerlik tespiti yapılmak istenmektedir. Bu proje python yazılım dili ile yazılmaktadır. Verilen hazır veri seti kullanılmaktadır. Hazır veri tipindeki fazla olan sutunları pandas kütüphanesi ile temizleme işlemleri yapılmaktadır. Temizlenen sutunlar üzerinden çıkarılan veri sitesinde noktalama işaretleri ve gereksiz kelimeler string kütühanesi nltk kütüphanesi kullanılmaktadır. Veri seti içerisindeki dil ingilizcedir. Kullanılan nltk kütühanesi içerisinde olan noktalama işaretleri ve ingilizce gereksiz kelimeler (stopwords) hazır olarak çekilmektedir ve bu işlemleri kullanarak temiz veri seti ortaya çıkarılmaktadır. Çıkan temiz veri seti benzerlik tespiti için hazır hale getirilmektedir. Temiz veri setindeki ürünlerin ismi aynı olanları bir liste içerisinde toplanmaktadır ve aynı olan ürünlerin kayıt sayıları da bir listede toplanmaktadır. Bu işlemler konular için de yapılmaktadır. Yapılan işlemler sonucun da var olan listeler ve temiz veri seti bulunmaktadır. Benzerlik_orani() adındaki method Benzerlik için elimizde bunun ürün isimleri ve ürün sayıları üzerinden her bir kaydı karşılaştırmak için bir döngü içirisine girmektedir. Her iki kayıt için kullanılan ürün metnini kelimeler bölmek için bir nltk kütühanesi kullanırız nltk kütüphanesi sayesinde girilen metin liste halinde bize geri döner. Kelimeler den oluşan bu iki liste kelimeleri karşılaştırılır. Karşılaştırılan kelimlerin benzer olanlarını sayarız bu toplam benzer kelimeleri karşılaştırdığımız iki kayıt içerisindeki en uzun kelime sayısı olana böleriz ve çıkan sonucu yüz ile çarpıp iki kayıt arasındaki benzerlik yüzdesini bulmuş oluruz. Projede bizden her hangi bir kaydın diğer tüm kayıtlar arasındaki benzerlik oranları da istenmektedir. Bunu için bir veri setinde olan ürün Id üzerinden kayıt seçilir. Seçilen kayıt için ürünler ve konular için oluşturduğumuz benzerlik_orani() işlemlerini yapılmaktadır. İstenilen benzer kayıtlar üzerinden bizden konuların benzerlik oranı tespiti yapılmak istenirse ürünlere ait olan tüm konular çekilmektedir. Oluşturulan veri üzerinden tekrar bir yüzde girilip konular üzerinde benzerlik_oran() methodu kullanarak benzerlik tespiti yapılamaktadır. Oluşturulan bu methodları masaüstü programı olarak kullanmak için bir

arayüz tasaralanmaktadır. Arayüz için python tkinter kütüphanesi kullanılmaktadır. Arayüzde veri setindeki karşılaştırdığımız verileri gösterbilmek için bir tablo oluşturulmaktadır. Ürün benzerlik tespiti, konular benzerlik tespiti, id ürün benzerlik tespiti ve seçilen ürünleerin konulara göre benzerlik tespiti olmak toplamda 4 buton bunlunmaktadır. Istenilen yüzdeler için 2 tane Entry vardır ve kaç tane thread kullanıldığını gösteren bir label kulanılmaktadır.

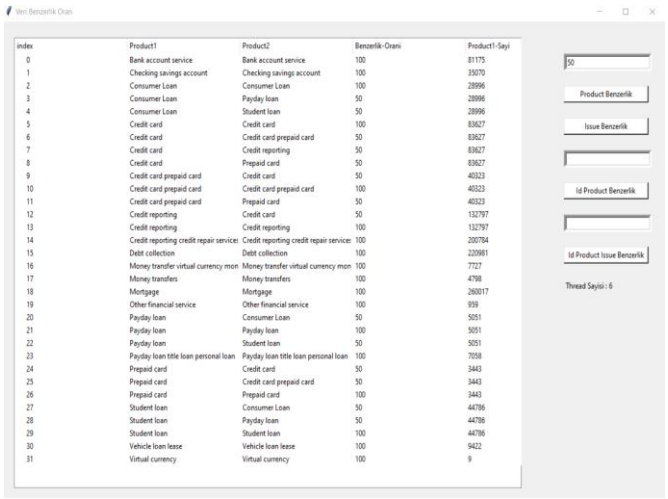
4. DENEYSEL SONUÇLAR

Program Çıktısı:

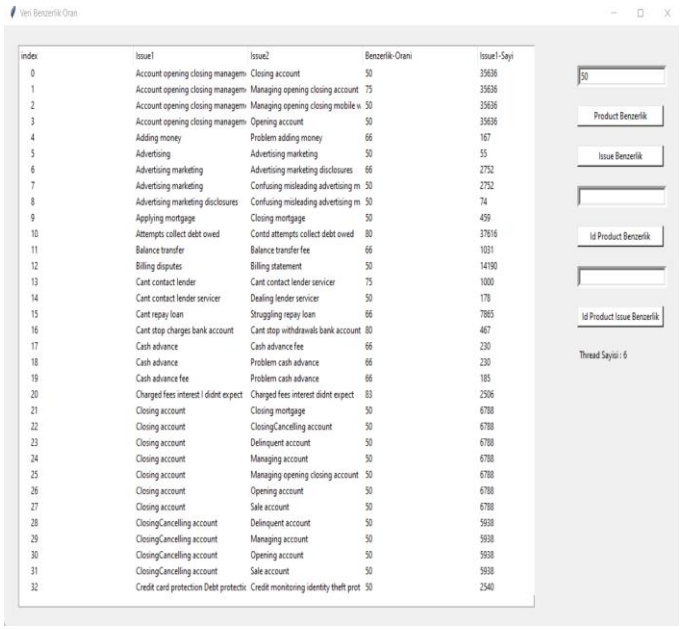
anasayfa



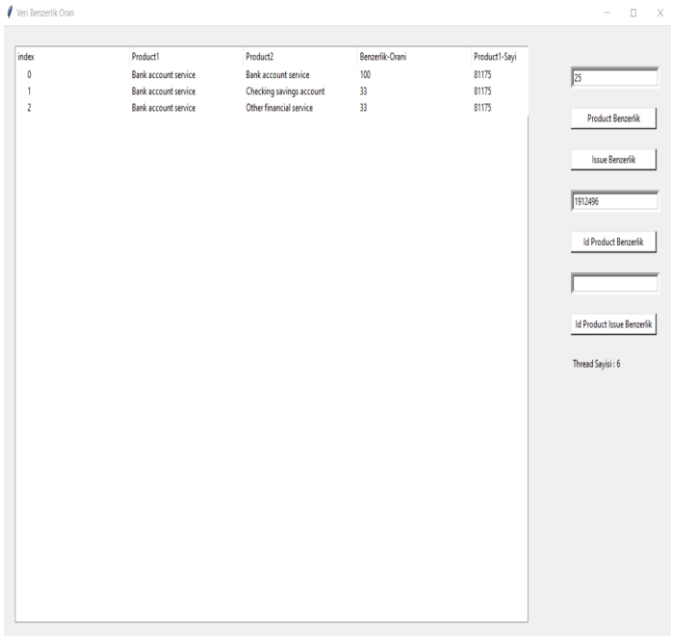
Ürünlerin benzerlik tespiti



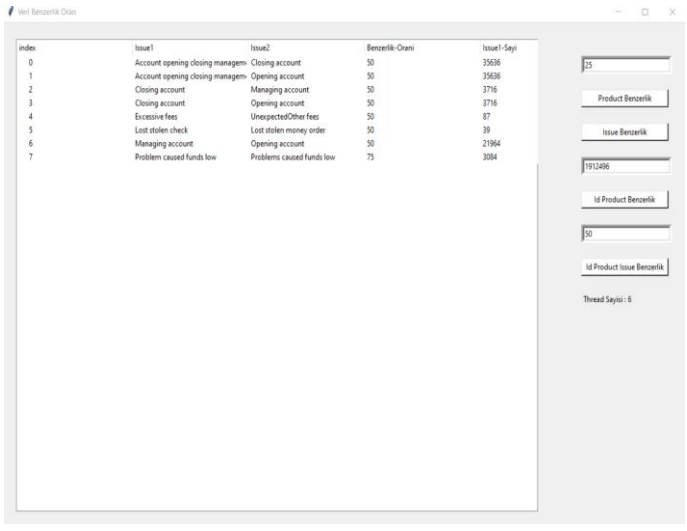
Konuların benzerlik tespiti



Id girilen ürünün benzerlik tespiti



Id girilen ürünün benzerlik tespiti yapılan kayıtların konularına göre benzerlik tespiti



5. SONUÇ

Proje multiTreading ile ürünlerin benzerlik oranları tespit edilmektedir. Projede ürünlerin ve konuların birbiri ile benzerlik oranları tespit edilmektedir. Seçilen ürünün diğer tüm kayıtlar ile benzerlik oranları tespit edilir ve edilen tüm benzer kayıtlar içerisinden konularına göre bir benzerlik yüzdesi vererek konularına göre benzerlik oranları tespit edilmektedir.

6.KAYNAKÇA

Proje geliştirilirken ve araştırma aşamasında faydalanılan kaynaklar;

Web Site

https://www.geeksforgeeks.org/multithreading-python-set-1/

https://www.nltk.org/

https://pandas.pydata.org/

https://realpython.com/python-gui-tkinter/