

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

Seda Nur Ekici-200201050

Begüm Erva Şahin-200201020

• Özet

Bu rapor Yazılım Laboratuvarı 1 Dersinin 2.projesini açıklamak ve sunumunu gerçekleştirmek amacıyla oluşturulmuştur. Raporda projenin tanımı, isterleri, yapım aşaması kullanılan araç ve yöntemler, kod parçacıkları vb. Bulunmaktadır. Proje aşamasında yararlanılan kaynaklar raporun son bölümünde bulunmaktadır.

• Giriş

Bu projede bizden istenen 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. bu projedeki amaç Veri seti içerisindeki arama işlem süresini multithreading kullanılarak azaltmaktır.Belirtilen sütun/sütunlar

için her bir satırdaki kayıtların birbiriyle kelime bazlı karşılaştırılması ve aralarındaki benzerliğin tespit edilmesidir. Uygulama içerisinde istenen özelliklere göre kayıtları filtrelemek ve kullanıcıya Göstermektir. Masaüstü uygulama geliştirme hakkında bilgi ve beceriye sahip olmaktır. Bu projede bizden istenenler öncelikle verilen veri setini düzenlememizdir. Daha sonrasında düzenlediğimiz bu veri setindeki tüm kayıtlar arasındaki benzerlik kontrolünün yapılmasıdır.Bu kontrol sırasında multithreading kullanmamız istenmiştir. Ayrıca her threadin çalışma zamanı ve tüm threadler için toplam çalışma zamanı bilgilerini uygulama arayüzünde göstermemiz istenmiştir. Bu benzerlikleri göstermemiz için bir arayüz tasarlamamız istenmiştir.

• İlerleyiş ve Yöntem

SENARYO 1

SENARYO 2

SENARYO 3

SENARYO 4

BENZERLİK ORANI :

SÜTÜN SEÇİNİZ :

Product

~SEVA~

	Product	Issue	Company	State	ZIP code	Complaint ID
0	checking savi...	managing...	navy federal...	fi	328XX	3238275
1	checking savi...	managing...	boeing empl...	wa	98204	3238228
2	debt collection	communic...	curo interm...	tx	751XX	3237964
3	credit reportin...	incorrect i...	ad astra rec...	la	708XX	3238479
4	checking savi...	managing...	ally financial...	az	85205	3238460
5	mortgage	closing m...	statebridge...	nj	08302	3237885
7	student loan	struggling...	student loan...	tx	773XX	3238221
8	debt collection	attempts c...	diversified c...	sc	296XX	3238545
9	credit reportin...	incorrect i...	contract call...	md	20774	3238458
10	vehicle loan le...	struggling...	ally financial...	nv	891XX	3238036
11	debt collection	false state...	miramed re...	va	23156	3237941
13	debt collection	false state...	commonwe...	tx	774XX	3237971

1. Başlamadan Önce

Bu projeye başlamadan önce proje dokümanında verilen bilgileri derleyip kullanabileceğimiz metodları araştırdık.

Kütüphaneden ve internet üzerinden yararlanabileceğimi kaynakları araştırdık. Veri seti düzenlemeye ve multithreading hakkında bilgi edindik.

2. Başlangıç

Öncelikle veri setini bizden istenen 6 farkı sütun (Product (Ürün), Issue (Konu), Company (Şirket), State, Complaint ID, Zip Code.) olacak şekilde python ile düzenledik. Veri setini düzenlerken pandas kütüphanesini kullandık. csv formatındaki belgeyi pythonda okuyup daha sonrasında düzenlediğimiz veri setindeki Null değer içeren kayıtları veri setimizden sildik.

Veri setinde bulunan noktalama işaretlerini kaldırdık ve son olarak veri setindeki kayıtlarda bulunan stop word'leri (stop word ler için nltk kütüphanesini kullandık)kaldırdık. Düzenlediğimiz veri setini csv formatında kaydettik ve javada multithreading için kullandık. Sonrasında java ile arayüz tasarımı yaptık. Arayüzü bizden istenen 4 senaryo için oluşturduk. İçerisinde hesapladığımız benzerlik oranını kullanıcıdan alabilmek için bir kutu ekledik. Hangi sütunların karşılaştırılacağını kullanıcının

seçebilmesi için bir buton ekledik.

Sonrasında kaç thread

kullanacağımızı kullanıcıdan almak için bir buton ve üç giriş ekledik. Yaptığımız benzerlikleri arayüzde göstermek için bir araç ekledik. Benzerliklerin daha hızlı hesaplanması için multithreading kullandık.

Oluşturduğumuz threadlerin çalışma sürelerini currentTimeMillis() yardımıyla hesapladık. Toplam çalışma sürelerini de currentTimeMillis() yardımıyla hesapladık.

Arka planda ise düzenlenmiş veri setini bufferreader ile

okuyup.Parçalama işlemi yaptık ve bu parçalanmış satırları iki boyutlu dizi yani matris olarak kaydettik .

Daha sonra benzerlik için bir fonksiyon oluşturup benzerlik tanımlamalarını yaptık.

İstenen senaryolar için satırları kendi arasında karşılaştırıp karşılaştırılan satırlar arasında hangisi büyükse onu paydaya ortakları ne ise onu paya yazıp bir yüzde hesabı yaptık.

3. İLERLEYİŞ

Bizden istenen benzerlik oranlarını bulmak için öncelikle düzenlediğimiz veri setini bir csv dosyasına aktardık. Bu csv dosyasını BufferedReader() yardımıyla okuduk. Dosyayı satır satır okuyarak iki boyutlu diziye aktardık. Her bir satırı virgül yardımıyla parçalayıp iki tane for döngüsüyle her bir değeri elde ediyoruz.

Sonrasında bizden istenen dört tane senaryo için benzerlik hesaplamalarımızı yapıyoruz. Dört tane senaryo için switch case döngüsü yardımıyla case :1 içinde senaryo biri hesapladık, case :2 içinde senaryo ikiyi hesapladık, case :3 içinde senaryo üçü hesapladık ve son olarak da case : 4 içinde senaryo dördü hesapladık.

Senaryo bir için case : 1 içinde bir for döngüsü tanımladık. Bu for döngüsü düzenlediğimiz veri seti boyutu kadar dönüyor ve kayıt1 içine değerleri satır satır kaydediyor. Bir tane daha for döngüsü yardımıyla satırların içerisindeki kelimeleri kayıt2 içerisinde kayıt ediyoruz. Daha sonrasında ise benzerlikBul() fonksiyonu ile karşılaştırma yapıyoruz. Senaryo iki için case :2 içinde bir for döngüsü tanımladık. Bu for döngüsü düzenlediğimiz veri seti boyutu kadar dönüyor ve product2, issue2, company2 içine değerleri satır satır kaydediyor. B1 içinde bizden istenen benzerliği atıp if döngüsü yardımıyla issue2 ve company2 yi ekrana yazdırırız. Senaryo üç için case : 3 içinde bir for döngüsü tanımladık. Bu for döngüsü düzenlediğimiz veri seti boyutu kadar dönüyor ve kayıt1 içinde aktarıyor. Bizden istenen Complaint Id ile karşılaştırma yapıp eğer istenen Complaint Id ye eşit ise iki tane for döngüsü yardımıyla karşılaştırmaları yapıp benzerlik oranını hesaplıyoruz ve bu benzerlik oranı 50 den fazlaysa issue leri ekrana yazduyoruz. Senaryo dört için case : 4 içinde bir for döngüsü tanımladık. Bu for döngüsü

düzenlediğimiz veri seti boyutu kadar dönüyor ve kayıt1 içinde aktarıyor. İki tane for döngüsü yardımıyla karşılaştırmaları yapıp benzerlik oranını hesaplıyoruz ve bu benzerlik oranı yüzde 80 den fazla olan Issue leri yazdırıyoruz.

• Arayüz fotoğrafları

A	B	C	D	E	F	G	H
	Product	Issue	Company	State	ZIP code	Complaint	ID
0	checking simanaging	navy federfl			328XX		3238275
1	checking simanaging	boeing em wa			98204		3238228
2	debt colleccommunic	curo interr tx			751XX		3237964
3	credit repcincorrect	irad astra rela			708XX		3238479
4	checking simanaging	ally financi az			85205		3238460
5	mortgage closing	mostatebridg nj			8302		3237885
7	student lostruggling	student loitx			773XX		3238221
8	debt colleccattempts	c diversified sc			296XX		3238545
9	credit repcincorrect	ir contract camd			20774		3238458
10	vehicle loa struggling	ally financi nv			891XX		3238036
11	debt collecfalse	stateimiram ed r va			23156		3237941
13	debt collecfalse	stateicommonw tx			774XX		3237971
14	credit repcproblem	c rally financi tx			774XX		3237837
15	debt collecttook	threa navy feder fl			346XX		3237928
16	debt colleccommunic	mercantileny			14609		3238682
17	debt colleccattempts	chcfs healthaz			852XX		3237088
19	checking simanaging	ally financico			81302		3237160
20	debt collecfalse	stateidiversified nj			087XX		3236983
21	debt collecfalse	stateially financi mi			48185		3237404
22	debt collecwritten	no weltman w oh			432XX		3236857
23	debt collectthreaten	ecashcall in ca			944XX		3236737
24	debt colleccattempts	c ability recnc			28216		3237833
25	credit repcincorrect	ir nissan mottx			797XX		3237655
26	credit repcimproper	t first advan ut			84401		3237100
27	debt collecwritten	no midwest fi il			60025		3237425
28	debt colleccattempts	chcfs healthfl			33056		3236953

Düzenlemiş veri setinin son hali yularıdaki resimde gösterilmiştir. İstenen stopwordler kaldırılmış, belirtilen satırlar silinmiş ve tüm noktalama işaretleri kaldırılmıştır. Son hali csv olarak yazdırıldı.

~SEVA~

SENARYO 1SENARYO 2SENARYO 3SENARYO 4* * *

BENZERLİK ORANI :

10

SÜTUN SEÇİNİZ :

Product

THREAD SAYISI :

1

☐ 2☐ 4☐ 10

GÖSTER

No	Kayıt1	Kayıt2	Benzerlik Oranı
1	managing account	managing account	0.0
2	managing account	managing account	0.0
3	managing account	managing account	0.0
4	managing account	problem lender com...	0.0
5	managing account	managing loan lease	0.0
6	managing account	managing account	0.0
7	managing account	managing account	0.0
8	managing account	managing account	0.0
9	managing account	managing account	0.0
10	managing account	managing account	0.0
11	managing account	problem lender com...	0.0
12	managing account	managing loan lease	0.0
13	managing account	managing account	0.0
14	managing account	managing account	0.0
15	managing account	managing account	0.0
16	communication tactics	communication tactics	0.0
17	communication tactics	communication tactics	0.0
18	communication tactics	communication tactics	0.0
19	communication tactics	communication tactics	0.0
20	communication tactics	communication tactics	0.0
21	communication tactics	communication tactics	0.0
22	incorrect information ...	incorrect information ...	0.0
23	incorrect information ...	threatened contact s...	0.0
24	incorrect information ...	incorrect information ...	0.0
25	incorrect information ...	improper use report	0.0
26	incorrect information ...	incorrect information ...	0.0
27	incorrect information ...	incorrect information ...	0.0
28	incorrect information ...	incorrect information ...	0.0
29	incorrect information ...	incorrect information ...	0.0
30	incorrect information ...	incorrect information ...	0.0
31	incorrect information ...	incorrect information ...	0.0
32	incorrect information ...	incorrect information ...	0.0
33	incorrect information ...	incorrect information ...	0.0

• Son Söz

Bu proje bize çok fazla bilgi birikimi sağladı ve farklı bir bakış açısı kazandırdı. Projeyi yaparken çok fazla araştırma yapıp eksiklerimizi kısa zamanda tamamladık ve projede bizden istenenleri elimizden gelen en iyi şekilde yapmaya çalıştık. Bir veri seti düzenlemeyi öğrendik..tread ve multithread kavramlarını öğrendik. Bir masaüstü uygulamasının nasıl oluşturulacağı işleneceğine dair detaylı bir bilgi sahibi olduk. Projede python ve java dillerini eş zamanlı olarak kullandık.

~SEVA~

SENARYO 1SENARYO 2SENARYO 3SENARYO 4

BENZERLİK ORANI :

SÜTUN SEÇİNİZ :

Product

THREAD SAYISI :

☐ 2☐ 4☐ 10

GÖSTER

Product	Issue	Company	State	Zip code	Complaint ID
---------	-------	---------	-------	----------	--------------

• kaynakça

*<https://www.youtube.com/watch?v=bv6dGZTpXXw>

*<https://www.youtube.com/watch?v=cG2AoJ5TKLY>

*<https://www.youtube.com/watch?v=xv-1ax50BKM>

*<https://www.kaggle.com/code/mert34/python-le-temel-veri-analizi>

*<https://gelecegiyazanlar.turkcell.com.tr/blog/pandas-ile-veri-analizi>

*<https://www.yusufsezer.com.tr/java-thread/>

*<https://devnot.com/2021/thread-nedir-detayli-bir-thread-incelemesi/>

*https://www.tutorialspoint.com/java/lang/system_currenttimemillis.htm

*<https://www.btkakademi.gov.tr/port al/course/player/deliver/java-ile-programlamaya-giris-9617>

