Veri Analizi ve Kullanıcı İlişkilendirmesi Raporu

Alperen Karacan
Öğrenci No: 210201096

I. Özet

Bu proje, JSON data seti üzerinden elde edilen kullanıcı verilerini analiz etmeyi ve bu veriler üzerinden kullanıcıların ilgi alanlarını belirlemeyi amaçlamaktadır.Proje,kullanıcıların,kullanıcı adları,adları,soyadları,bölgeleri,tweetlerini,takipçi sayılarını ve takip ettikleri hesapları içeren geniş bir veri setinden yararlanır.Hedef,bu verileri işleyerek kullanıcıların ilgi alanlarını çıkarmak ve bu bilgileri graph olarak sunmaktır.Proje,metin işleme,veri yapıları,algoritmalar ve arayüz geliştirme gibi birçok programlamayı öğrenip kullandırmayı amaçlamaktadır. Kazanım:

- Bu proje,öğrencilere veri yapıları,graflar,hash tabloları,arama algoritmaları,veri analizi gibi konularda deneyim kazanma fırsatı sunacaktır.Hem veri yapıları ve algoritmaların hem de gerçek veri analizi ve ilişkisel çıkarımların pratiğini yaparak problem çözme ve analitik düşünme becerilerimizi geliştirmeye fayda sağlayacaktır.

II. Giriş

Bu projede,kullanıcıların etkileşimlerinden yola çıkarak ilgi alanlarını tespit etmekdir.Bu tespit, veri toplama,işleme,analiz etme ve sonuçları görselleştirme adımlarından oluşur.Projenin temel amacı, kullanıcıların hangi konulara daha fazla ilgilendiklerini anlamaktır

III. Projenin Aşamaları

A. 1- Veri Kaynağı

Projede 50 bin kullanıcının verilerini içeren bir JSON dosyasını veri kaynağı olarak kullandım fakat 50 bin kullanıcın verilerini işlemek çok uzun sürdüğü için ne kadar kullanıcının verilerinin alınacağını fonksiyon yazarak esnek bıraktım.

B. 2- Kullanıcı

Kullanıcı bilgileri User sınıfı aracılığıyla temsil edilir.Veri depolama ve erişim için özelleştirilmiş HashTable ve Hashmap gibi veri yapıları kullanılır.

C. 3- Analiz

Kullanıcıların tweetleri önemsiz sayılan kelimeler çıkartılarak ve kelimeleri köklerine ayırarak analiz edilir.

D. 4-İlgi Alanına Göre Eşleme

Analiz sonucunda elde edilen veriler,önceden belirlenen ilgi alanlarına göre kategorize edilir.Bu kısım kullanıcıların hangi konulara daha fazla ilgi gösterdiğini ortaya çıkarır.

- Kullanıcı ilgi alanlarını hash tablolarında tutulmalıdır.Her bir ilgi alanı için bir hash tablosu oluşturulmalı ve ilgi alanını anahtar olarak kullanarak bu ilgi alanını paylaşan kullanıcıların listesini değer olarak saklanmalıdır.
- Kullanıcıların takip ettikleri,takipçileri ve tweet içerikleri arasında benzerlikleri tespit ederek,ortak ilgi alanlarına sahip kullanıcıları eşleştirilmelidir.
- Verilen bir ilgi alanı arama algoritmaları (hashing vb.) kullanılarak bulunmalıdır.Ayrıca bu ilgi alanını paylaşan kullanıcıların listesi çıkarılmalıdır
- Kullanıcıların ortak ilgi alanlarını belirlemek için hash tabloları ve karşılaştırma algoritmaları (arama, sıralama algoritmaları vb.) kullanılmalıdır.
- Hash tabloları ve arama algoritmaları kullanılarak elde edilen benzer ilgi alanlarına sahip kullanıcılar arasındaki ilişkiler gösterilmelidir

– Kullanıcıların ilgi alanlarına dayalı detaylı analiz raporları oluşturularak metin dosyasına kaydedilmelidir
– Belirli bölge (konum) ve dil için trend olan hashtagler veya konular listelenmelidir.

E. 5- Graf Üzerinde Analiz

Kullanıcıdan alınan veri ile ilgili kişinin takip ettikleri kullanıcılar ile grafi çizilir.Aşağıda örnek bir graf bulunmaktadır.

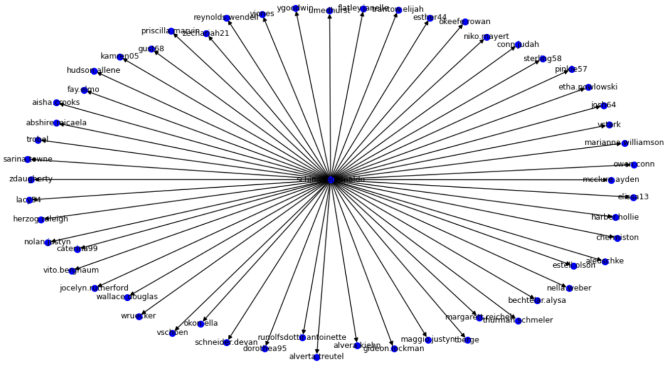

Fig. 1. Graf Örneği

IV. DENEYSEL SONUÇLAR

A. Veri Seti ve İşlenmesi

Projede kullanılan veri seti,JSON dan elde edilen 50.000 kullanıcıya ait verileri içermektedir.Bu veriler,kullanıcı bilgileri,tweet içerikleri,takipçi sayıları,takip ettikleri hesaplar ve dil,bölge gibi bilgilerdir.Veri setini daha yönetilebilir ve daha hızlı olması için boyutu indirgeyere ilk 100 kullanıcıya kadar kısıtladım,Bu kısıtı kendiniz esnek bir şekilde değiştirebilirsiniz fonksiyonunu yazdım.Bu işlem,Analiz süresini ve kaynak kullanımını olabildiğince az kullanmak için önemlidir.

B. Tweet Analizi ve Kelime Frekansı

Kullanıcıların tweetlerinin analizi,Her tweetteki kelimeleri Türkçe köklerine indirgeyerek ve önceden belirlediğim önemsiz kelimeleri hariç tutarak gerçekleştirilmiştir.Bu işlem,doğal dil işleme tekniklerini kullanarak metinden anlamlı bilgiler çıkarmayı hedefler.Analiz sonucunda,her kullanıcının tweetlerinde en sık geçen kelime ve bu kelimelerin frekansları tespit edilmiştir.Bu sonuçlar, kullanıcıların hangi konular üzerinde daha fazla konuştuğunu ve hangi kelimelerin onlar için önemli olduğunu belirlemek için kullanılır.Kullanılan algoritma bubble sort dur.Merge ve pythonun yerleşik sıralama algoritmalarıyla da denedim aralarında bariz perfonmans farkları olmamasından dolayı bubble sort ile yaptım sizler diğer bahsettiğim algoritmaları da koda kolayca entegre ederek yapabilirsiniz.

```
Hangi kullanıcının analizini almak istersiniz? (1 - 100, Çıkmak ve kullanıcı graphını açmak için -1): 3

Kullanıcı Adı: monserrat82
Adı: Freda Abbott IV
Takipçi Sayısı: 100
Takip Ettiği Kişi Sayısı: 84
Dil: sl
Bölge: IQ

Kullanıcının İlgi Alanları ve İlgili Kelimeler:
İlgili Kategori: Siyasi, İlgili Kelime: savaş,Kelimenin Frekansı: 3

En Çok Kullanılan İlk 5 Kelime:
yer: 5
başladı: 4
yap: 4
savaş: 3
köy: 3
```

Fig. 2. Tweet Analizi ve Frekans Örneği

C. İlgi Alanlarını Kategorize Etme

Elde edilen kelime frekansları,önceden belirlediğim ilgi alanları ile eşleştirilerek her kullanıcının ilgi alanları belirlenmiştir.İlgi alanlarınıda kendim txt dosyalarına yazdığım ilgi alanı ile ilgili kelimeler bularak txt den kelimeleri okuyarak frekans sayımını yaptım.Bu süreç,her kelimenin hangi ilgi alanına ait olduğunu belirlemek için anahtar kelime listelerini kullanır.Örneğin,"dünya" kelimesi sıkça geçiyorsa, kullanıcının "Astronomi" ile ilgileniyor olduğu sonucuna varılabilir.

D. Kullanıcı İlişkileri ve Graf

Projede,kullanıcıların birbirleriyle olan ilişkilerini görselleştirmek için graf çizdirdim.Bu graf, kullanıcıların kimleri takip ettiğini gösterir.Tkinter arayüzü kullanılarak oluşturulan grafikte, kullanıcılar düğümler olarak,takip ilişkileri ise kenarlar olarak temsil edilir.Bu görselleştirme, kullanıcıların sosyal medya üzerindeki etkileşim dinamiklerini anlamak için önemli bir araçtır.Ayrıca txt dosyasına da benzer ilgi alanları olan kullanıcıları kaydettim.

Fig. 3. Benzer İlgi Alanlarının Yazıldığı Txt Örneği

E. Perfonmans ve Optimizasyon

Kodlama sürecinde,çeşitli sıralama algoritmaları (bubble sort, merge sort ve Python'un yerleşik sort fonksiyonu) larını kullandım performans açısından karşılaştırdım aralında farklar vardı en hızlısı yerleşik sort du fakat hazır veri yapısı olduğu için kullanmadım,bubble sort ile merge sort arasında da çok fark olmadığı için bubble sortu tercih ettim.Bu karşılaştırma,en uygun algoritmayı seçmeme yardımcı oldu.

KAYNAKÇA

https://stackoverflow.com
https://chat.openai.com/chat
https://www.youtube.com

V.
PROJENIN ÖRNEK GÖRSELLERI

```
Hangi kullanıcının analizini almak istersiniz? (1 - 100, Çıkmak ve kullanıcı graphını açmak için -1): 55

Kullanıcı Adı: goldner.yazmin
Adı: Elva Schmeler DDS
Takipçi Sayısı: 135
Takip Ettiği Kişi Sayısı: 28
Dil: nl
Bölge: PT

Kullanıcının İlgi Alanları ve İlgili Kelimeler:
İlgili Kategori: Astronomi, İlgili Kelime: dünya,Kelimenin Frekansı: 4

En Çok Kullanılan İlk 5 Kelime:
ara: 5
dan: 4
dünya: 4
yer: 4
taraf: 4
```

Fig. 4. Konsol Örneği

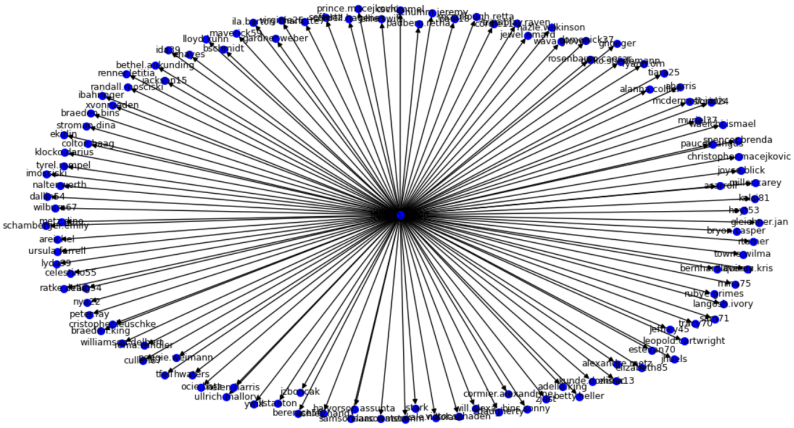

Fig. 5. Graf Örnek

Fig. 6. Benzer txt Örnek

VI.
Projenin UML Diyagrami ve Akış Şemasi

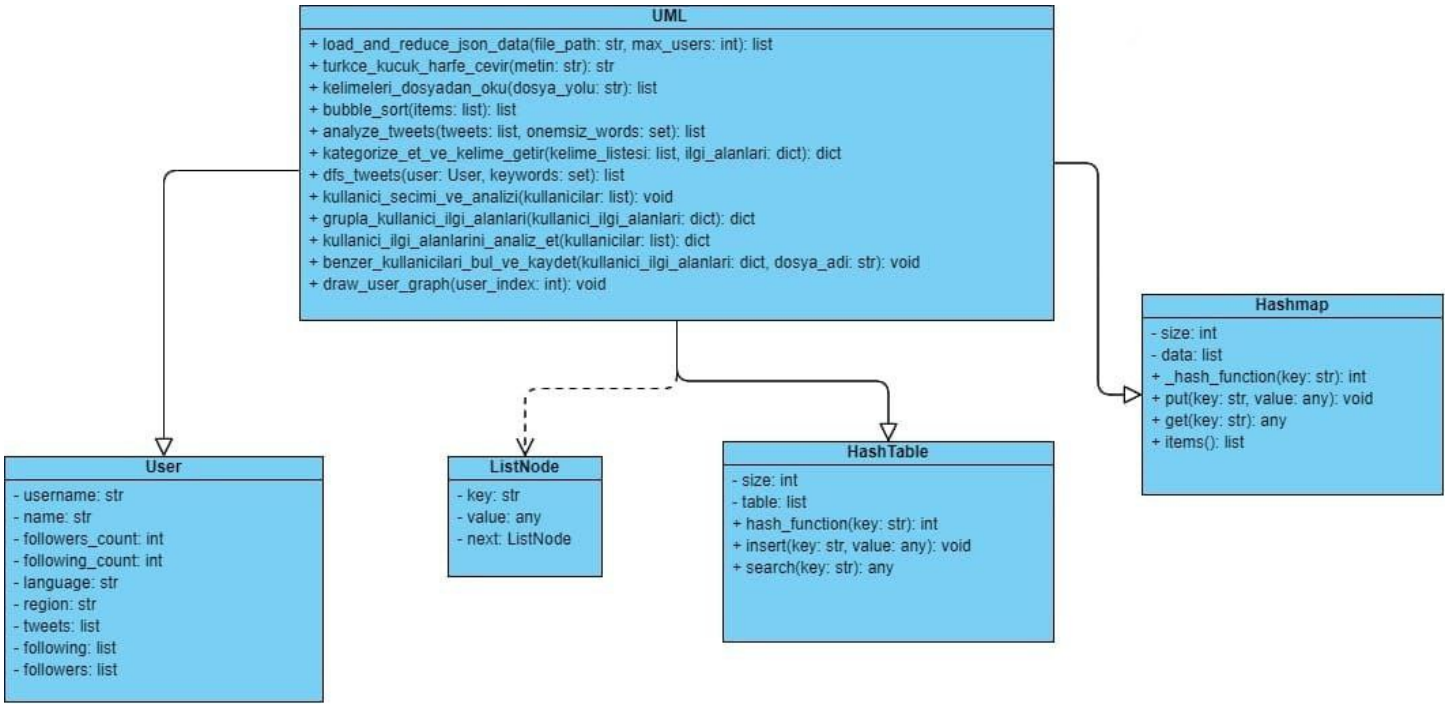

Fig. 7. UML Diyagramı

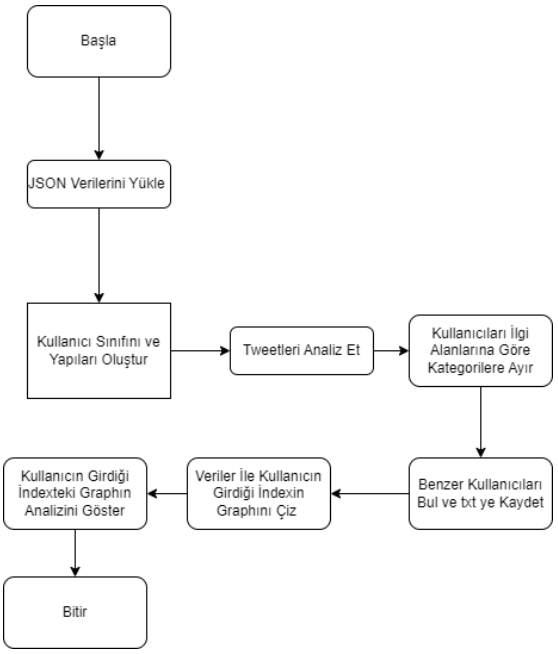

Fig. 8. Akış Diyagramı