TWITTER API İLE BÜYÜK VERİ İŞLEMLERİ VE VERİNİN GÖRSELLEŞTİRİLMESİ

ENES ÇAVUŞ

İçerik

- Giriş
- Kullanılan Teknolojiler
- Veri kaynağı İşlemleri – Twitter API
- Big Data İşlemleri
- Görselleştirme ve Analiz
- Sonuç

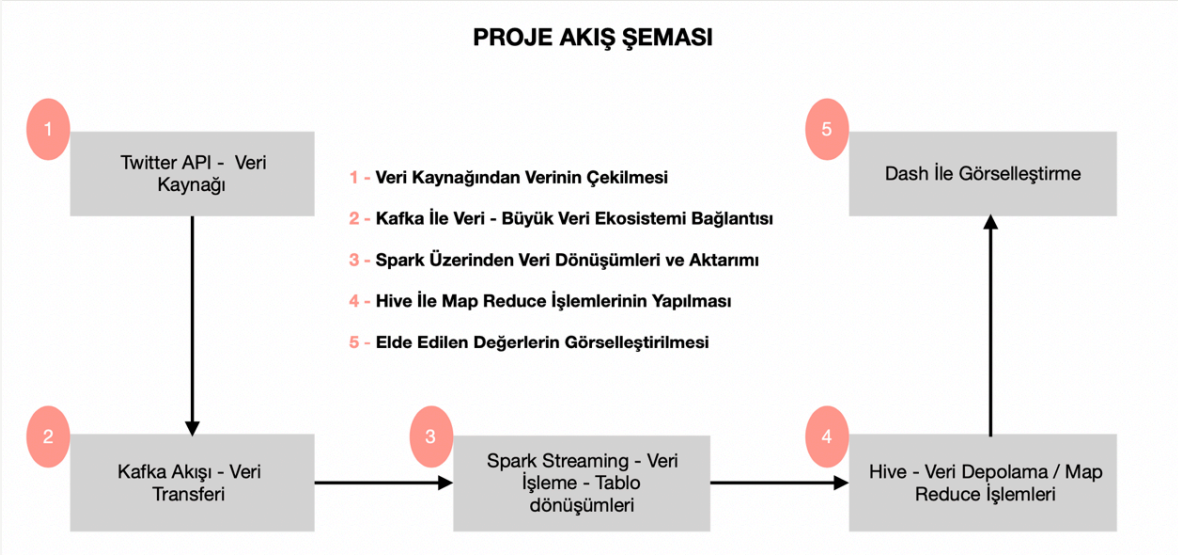
Giriş

- Bu çalışmada, veri kaynağı olarak seçilen sosyal medya ağı Twitter üzerinden toplanan verilerin, büyük veri teknolojileri kullanılarak işlemlere tutulması, verinin ayrıştırılıp dönüştürülerek son haline getirilmesi ve ardından görselleştirmeler yapılarak veriden çıkarım ve analizlerde bulunulması hedeflenmiştir.

- Projede üç temel adım bulunmaktadır. Birincisi veri kaynağının belirlenmesi ve verinin toplanması. İkincisi, elde edilen verinin büyük veri ekosistemine bağlanması ve işlemlerden geçirilerek büyük veri teknolojilerinden yararlanılması. Son adımda ise elde edilen sonuç verilerinin ve değerlerinin görselleştirilmesi ile gösterge paneli elde edilmesidir. Bu sayede hiç işlenmemiş ham bir veriden bilgi elde edilmesi amaçlanmıştır.

Giriş – Proje Akış Şeması

- Projenin akış şeması yandaki gibidir.
- Veriler veri kaynağından toplanır ve Kafka ile büyük veri ekosistemine bağlanır.
- Kafka üzerinden Spark Streaming kütüphanesine aktarılır ve veride gerekli dönüşümler gerçekleştirilir.
- Bu işlemler tamamlandıktan sonra Hive üzerinde veri tabanı işlemleri yapılır ve veriler depolanmış olur.
- Sonraki aşamalarda verilerin görselleştirilmesi ve analizi bulunmaktadır.



Kullanılan Teknolojiler

- Python ve Kütüphaneleri
  - Kafka-Python
  - PySpark
  - PyHive
  - Pandas
  - Flask
  - Dash Plotly
  - Tweepy
- Apache Kafka
  - Veri Aktarımı - Bağlantılar
- Apache Hadoop
  - Depolama – Veri İşleme
- Apache Spark
  - Veri dönüştürme – Veri Tabanına yazma
- Apache Hive
  - Veri tabanı ve depolama yönetimi

Veri Kaynağı İşlemleri - Twitter API

- Twitter API, Twitter tarafından sağlanan bir sosyal medya API'ıdır.
- Birçok farklı türde veri çeşitleri, farklı erişim yöntemleri ve farklı kullanım amaçları bulunmaktadır.
- Bu çalışmada hedeflenen, tweet içerikleri ve trend verileri kullanılarak sayısal değerlerin görselleştirilmesi ve analiz edilmesidir.
- Yandaki görselde tweet verisine ait örnek bir içerik bulunmaktadır.


```
{"created_at":"Mon Mar 01 08:31:54 +0000 2021",
"id":                    "id_str":                    ",
"text":"RT
"source":"\u003ca href=\"http:\/\/twitter.com\/download\/android\" rel=\"nofollow\"\u003e
"truncated":false,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply
"user":{"id"                    ,"id_str":                    ","name"          ,"screen_
},"geo":null,"coordinates":null,"place":null,"contributors":null,
    "retweeted_status":{"created_at":"Mon Mar 01 08:29:43 +0000 2021",
    "text":"                                                                    #thailand #
    "truncated":false,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_r
    "retweet_count":1,
    "favorite_count":1,
    "entities":{"hashtags":[{"text":"thailand","indices":[63,72]},{"text":"coronavirus","
},"is_quote_status":false,"quote_count":0,"reply_count":0,
"retweet_count":0,
"favorite_count":0,
"entities":{"hashtags":[{"text":"thailand","indices":[81,90]},{"text":"coronavirus","ind
,"extended_entities":{"media":[{"id"                    ,"id_str":                    ","
,"favorited":false,"retweeted":false,"possibly_sensitive":false,"filter_level":"low","lan
```

Big Data İşlemleri - Kafka

- Verinin büyük veri teknolojileri ile kullanılabilmesi için veri girişi olarak kullanılacak teknoloji Kafka'dır.
- Kafka ile paralel işlemler yürütülebilir ve veri akışı hızlı bir şekilde gerçekleştirilebilir.
- Kafka, çeşitli portlar üzerinde ve belirlenen Kafka Topic'leri üzerinde verileri taşır.
- Yandaki görselde örnek Kafka sunucusu çalıştırma ve örnek topic oluşturma görselleri verilmiştir.

```
bin/zookeeper-server-start.sh config/zookeeper.properties
bin/kafka-server-start.sh config/server.properties
```

```
bin/kafka-topics.sh --create --bootstrap-server localhost:9092 \
--replication-factor 1 --partitions 1 --topic tweettablo
```

Big Data İşlemleri – Spark / Hive

- Bu adımda Spark üzerinde veri dönüşümleri yapılmaktadır.
- Kafka üzerinden iletilen veriler JSON veri formatındadır dolayısıyla bu format üzerinde dönüşümler yapılması gerekmektedir.
- Hedeflenen içerikler yandaki görseldeki gibidir.
- Bazı durumlarda sayısal değerler hesaplanmalı, bazı durumlarda kategorik veriler hesaplanmalıdır.
- Yandaki tablo gerçek zamanlı olarak Hive veri depolama tarafına aktarılmaktadır.

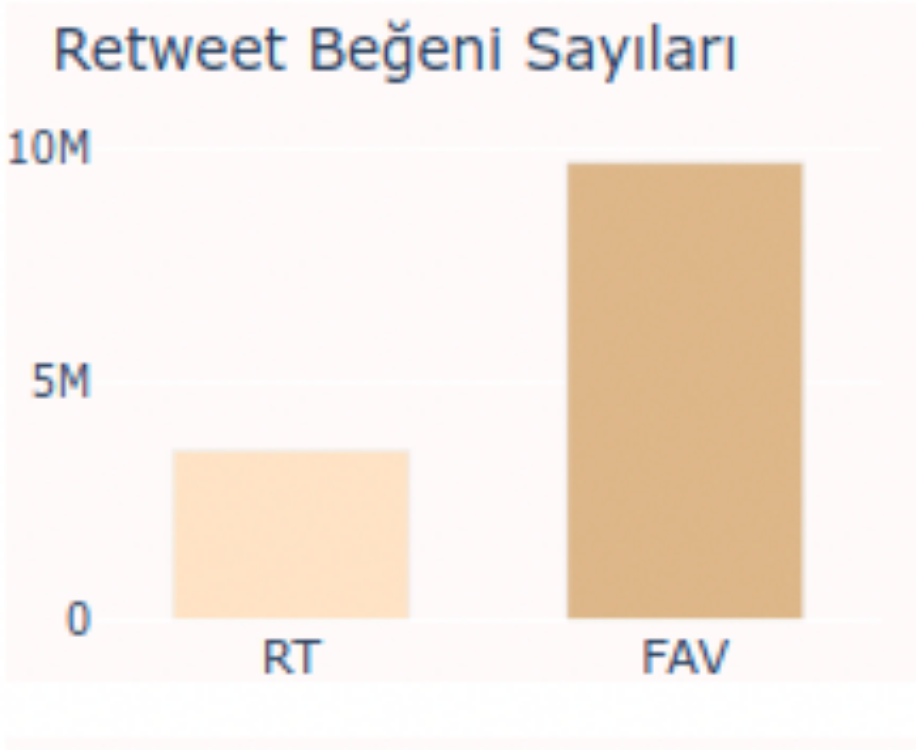
tweetlen	rtcount	fvcount	isrt	location	device
121	0,20	324	Yes	England	Andorid
54	0	2	No	None	WebApp
138	1753	5629	Yes	India	Andorid
140	187	2001	Yes	Spain	iPhone
100	0	0	No	London,UK	Other
113	96	500	Yes	USA	Android
121	3	12	Yes	India	iPhone

Görselleştirme ve Analiz

- Görselleştirme kısmında Dash Plotly python kütüphanesi kullanılmıştır.
- Bu kütüphane, web tabanlı olacak şekilde çeşitli arayüz ve grafikler ile entegre şekilde çalışmaktadır.
- Gerçek zamanlı veri görselleştirmelerinde tercih edilmektedir.
- Hive ile Dash arasında bir işlem daha yapılmaktadır, Hadoop dağıtık veri işleme ve depolama sisteminde bulunan verilere ait hesaplamaların bilgisayarın lokal hafızasına geçriilmesi işlemidir.
- Bu ara adımda PyHive kullanılarak Hive üzerindeki Map Reduce işlemleri gerçekleştirilecek ve son bilgiler lokal disk üzerinde python veri yapılarında saklanacaktır.
- Daha sonra bu verilere Dash kodları ile erilişecek ve veri tiplerine uygun grafikler ile görselleştirme adımları gerçekleştirilecektir.

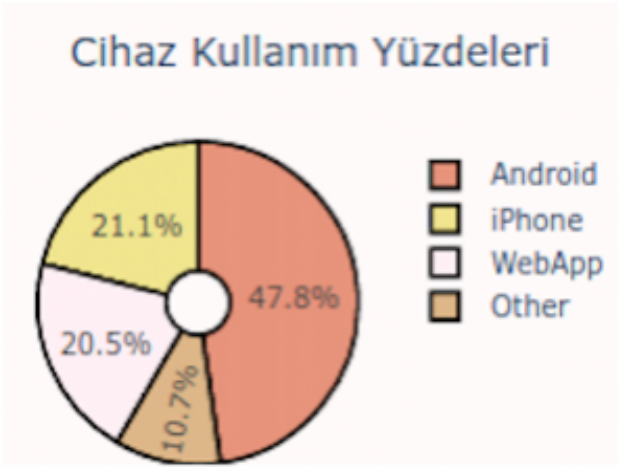
Görselleştirme ve Analiz / Bar Grafiği

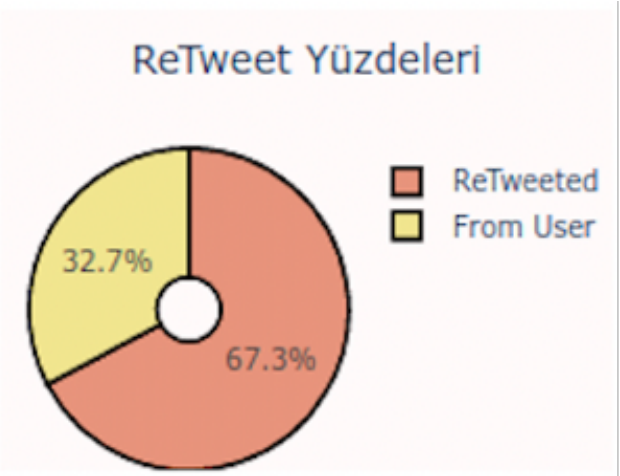
- Sayısal değerlerin görselleştirilmesindeki en popüler grafik tiplerinden olan bar grafiği, reTweet ve Beğeni sayıları gibi veriler için uygundur.
- Bu grafik ile konunun ne kadar çok konuşulduğu ve ilgi gördüğü gözlemlenebilmektedir.
- Aynı zamanda her bir grafik gerçek zamanlı olarak veya günlük olarak güncellenebilir. Bu çalışmada son 7 güne ait güncellemeler ve haftalık veriler üzerinden görselleştirmeler yapılmıştır



Görselleştirme ve Analiz / Pie/Dilim Grafiği

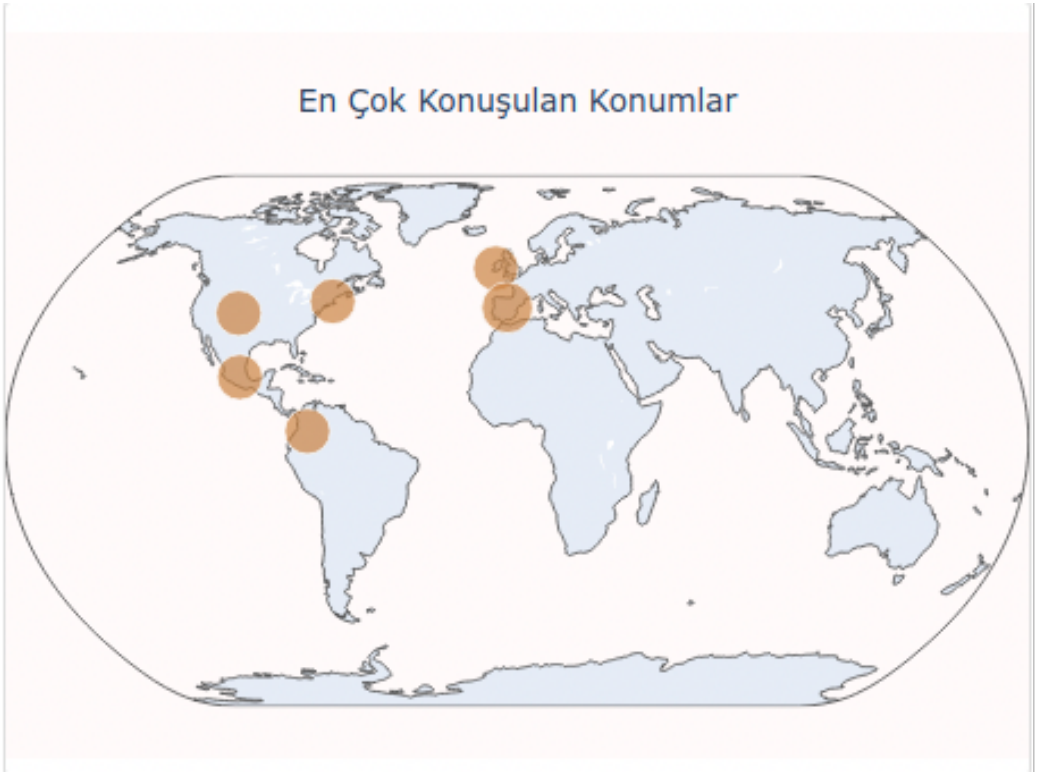
- Bu grafikler yüzde veya oran içeren veriler için kullanıma uygundur.
- Bu çalışmada ReTweet durumu ve kullanılan cihazlara ait veriler bu grafikler ile görselleştirmeye uygun görülmüştür.
- Aşağıdaki görseller bahsedilen iki yüzde dilimli grafiğe ait örnek çıktılardır.





Görselleştirme ve Analiz / Harita Grafiği

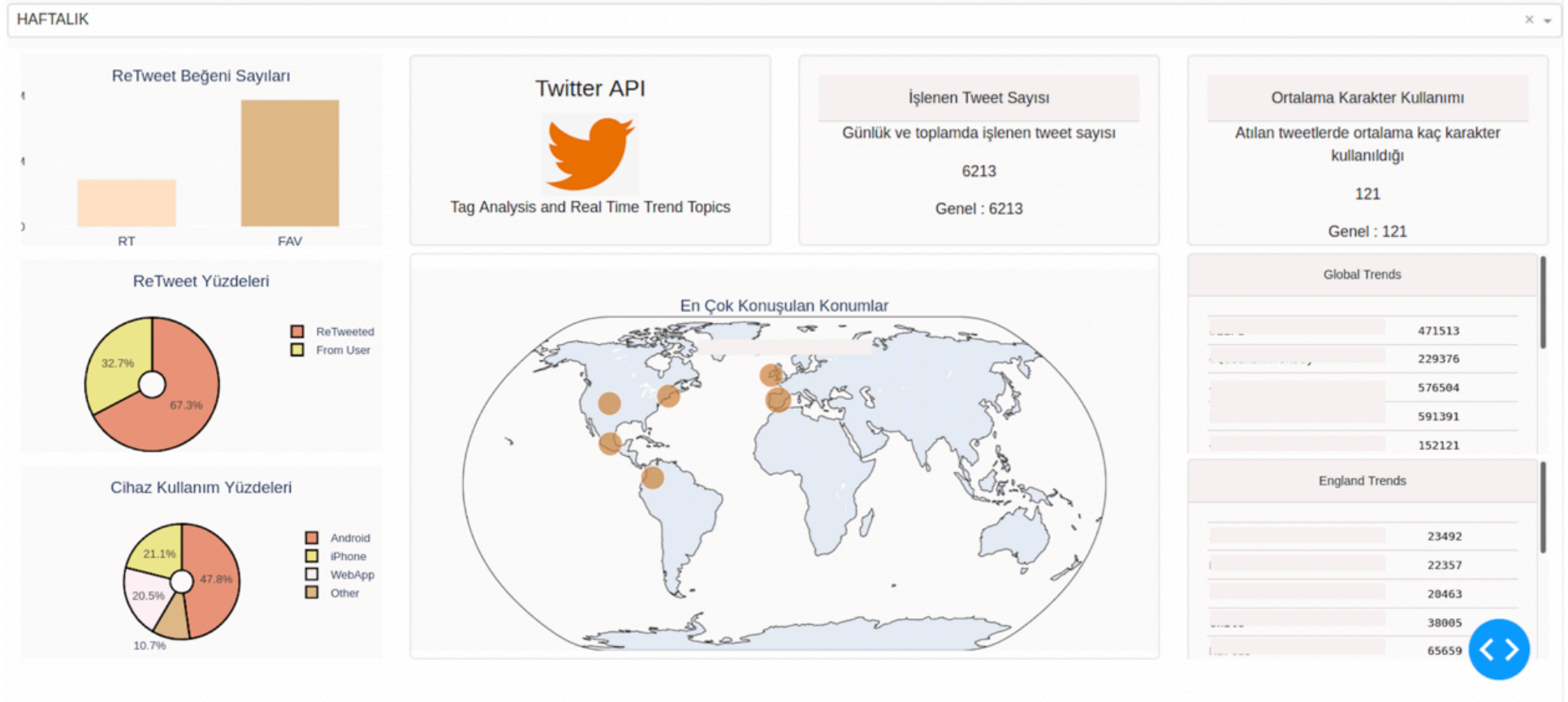
- Harita grafikleri özellikle konum bazlı veriler üzerinde sıklıkla kullanılmaktadır.
- Bu projede, çoğunlukla konum bilgileri elde edilmektedir.
- Elde edilen konum bilgilerinden en çok tekrar edilenler göz önünde bulundurularak, incelenen konu başlığının (ör: covid-19 , big-data) dünya çapında hangi konumlarda gündemde olduğu ve çok konuşulduğu gözlemlenebilmektedir.
- Yandaki görselde örnek bir harita görseli bulunmaktadır.



Görselleştirme ve Analiz / Bootstrap Kartları

- Dash birçok farklı web tabanlı kütüphaneler ile birlikte entegre çalışabilmektedir.
- Bootstrap hem ölçeklendirme hem de html elementeleri ile kolay kullanım sağlayan bir aracı kütüphanedir.
- Çalışmada incelenen verilerin daha fazla detayını görselleştirmek amacı ile aşağıdaki gibi birkaç kart görselleştirmesi yapılmıştır.
- Aynı zamanda bu kartlar üzerine tablolar yerleştirilerek, daha önceden bahsedilen trend verilerinin içeriği de gerçek zamanlı olarak arayüzde görselleştirilmektedir.

Genel Arayüz Çıktısı



SONUÇ

- Projede, veri kaynağı üzerinden elde edilebilecek sayısal ve kategorik veri tiplerinin görselleştirilmesi hedeflenmiş ve gerçekleştirilmiştir.
- Çalışma süreci boyunca birçok farklı teknoloji birbiri ile bağlanmış, entegre halde çalıştırılmıştır. Bu sayede gerçek zamanlı bir büyük veri akışı / pipeline / veri boru hatları tasarlanmış ve gerçekleştirilmiştir.
- Veri tipleri, veri dönüşümleri, veri depolaması, veri görselleştirmesi ve veri analizi gibi ana başlıklarda işlemler yürütülmüş, sonuç olarak bir arayüz elde edilmiştir.
- Çalışma sonunda Büyük Veri, Veri Akışı, API, Görselleştirme ve Analiz alanlarında tecrübe edinilmiş, proje tamamlanmıştır.

TWITTER API İLE BÜYÜK VERİ İŞLEMLERİ VE VERİNİN GÖRSELLEŞTİRİLMESİ

ENES ÇAVUŞ