TWITTER VERİ ANALİZİ ÇALIŞMASI

Ali “Şabobey” ŞABAHAT

KONU: #TFF ETİKETİ İLE YAPILMIŞ TWEETLERİN İNDİRİLMESİ VE ANALİZİNİN YAPILMASI

ÖZET: 21 GÜN BOYUNCA TWITTER PLATFORMU ÜZERİNDEN #TFF ETİKETİ İLE YAPILAN PAYLAŞIMLARIN GÜNLÜK OLARAK TWITTER GELİŞTİRİCİ HESABI ARACILIĞI İLE İNDİRİLMESİ VE RSTUDİO VE GEREKLİ KÜTÜPHANELERCE ANALİZİNİN YAPILIP RAPORLANMASI.

VERİ MADENCİLİĞİ SÜRECİ AŞAMALARI:

1.AŞAMA: PROBLEMİN TANIMLANMASI

2.AŞAMA: VERİYİ ANLAMA

3.AŞAMA: VERİNİN HAZIRLANMASI

4.AŞAMA: MODELLEME

5.AŞAMA: DEĞERLENDİRME

6.AŞAMA: YAYILIM

#TFF ETİKETİ İLE YAPILMIŞ TWEETLERİN ANALİZİNİN AŞAMALARI VE SUNULMASI:

1. TWITTER GELİŞTİRİCİ HESABININ ALINMASI:

Öncelikle bir Twitter kullanıcı hesabına sahip olunması gerekir.
Ardından “developer.twitter.com” adresinden Twitter Geliştirici Hesabı için başvuruda bulunulması gerekir. Başvuru yapılırken neden bu hesaba ihtiyacımız olduğu ayrıntılı ve doğru şekilde anlatılmalıdır yoksa hesap onaylanmaz ya da tekrar tekrar başvuruda bulunulması gerekir. Twitter geliştirici hesabı onaylandıktan sonra izin verilen miktarda veri çekilebilir.

Veri çekebilmek için bize verilen gerekli özel anahtar ve şifreler kullanılır ve bu bilgiler paylaşılmamalıdır:

```
options(httr_oauth_cache=T)

consumer_key <- [redacted]
consumer_secret <- [redacted]
access_token <- [redacted]
access_secret <- [redacted]

setup_twitter_oauth(consumer_key, consumer_secret, access_token, access_secret)
```

2. RSTUDIO PROGRAMI İLE TWEETLERİN ÇEKİLMESİ VE ANALİZ EDİLMESİ:

RStudio programında gerekli kütüphaneler indirildikten sonra Twitter Geliştirici Hesabımızdan elde ettiğimiz anahtar ve şifreleri kullanarak veri çekildi ardından 21 gün boyunca (genel olarak saat 22.30-00.00 arasında çekildi) çekilen veriler hergün indirildi ve bu veriler bir araya getirilerek toplu şekilde analiz edildi:

-Gerekli Kütüphaneler:

İHTİYACIMIZ OLAN KÜTÜPHANELERİN:

YÜKLENMESİ

```
install.packages("twitteR")
install.packages("ROAuth")
install.packages("openssl")
install.packages("httpuv")
install.packages("tm")
install.packages("readxl")
install.packages("readr")
install.packages("stringi")
install.packages("stringr")
install.packages("writexl")
install.packages("tidytext")
install.packages("dplyr")
install.packages("ggplot2")
install.packages("wordcloud")
install.packages("RColorBrewer")
install.packages("syuzhet")
install.packages("lubridate")
install.packages("scales")
install.packages("reshape2")
```

ÇAĞIRILMASI

```
library(twitteR)
library(ROAuth)
library(openssl)
library(httpuv)
library(stringi)
library(stringr)
library(tm)
library(readxl)
library(readr)
library(writexl)
library(tidytext)
library(dplyr)
library(ggplot2)
library(wordcloud)
library(RColorBrewer)
library(syuzhet)
library(lubridate)
library(scales)
library(reshape2)
```

-Veri Çekme Aşaması:

```
#Twitterdan veri cekme asamasi icin gereken ozel bilgiler ve kodlama kısmı

options(httr_oauth_cache=T)

consumer_key <- "[redacted]"
consumer_secret <- "[redacted]"
access_token <- "[redacted]"
access_secret <- "[redacted]"

setup_twitter_oauth(consumer_key, consumer_secret, access_token, access_secret)

tweets <- searchTwitteR("#tff", n=1000,  locale = "tr_TR")

tweets.df <- twListToDF(tweets)

#-------------------------------------
```

Gerekli bilgiler girilip çalıştırıldığında "console" ekranında doğrulama uyarısı görülür:

```
[1] "Using direct authentication"
```

Daha sonra:

tweets <- searchTwitteR("#tff", n=1000, locale= "tr_TR") komutu ile veri çekilir.

tweets.df <- twListToDF(tweets) komutu ile çekilen veriler tweets.df olarak tabloya dönüştürülür.

-Verilerin günlük kaydedilmesi:

Verilerin 21 gün boyunca toplanıp bir arada analiz edilebilmesi için write_xlsx komutuyla tweets.df tablosu hedef dosyaya excel formatında kaydedildi.

```
#Gunluk cekilen tweetlerin indirilmesi

write_xlsx(tweets.df,
           "D:\\DERS\\21-22 guz\\veri madenciligi\\VeriCekmeTwitter\\veriler\\gun1.xlsx")
```

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	text	favorited	oriteCo	replyToSN	created	truncated	eplyToSI	id	eplyToUI	tusSour	reenNar	weetCo	sRetwee	etweete	longitude	latitude
2	https://t.co/F3j5GJMpBo	0	0		###	1		1472286222684939	<a href="	ErkanCeli		0	0	0		
3	uygulayan Dünya d...	0	0		###	0		1472285436404019	<a href="	veysel_u		6	1	0		
4	@SuleymanRodop Hakemler o	0	0	SuleymanRod	###	0	1472274	1472285	2146680	<a href="	srm4460		0	0	0	

-Verilerin bir araya getirilip birleştirilmesi ve RStudio'ya yüklenmesi:

21 gün boyunca belli saatlerde düzenli olarak çekilmiş günlük verilerin tutulduğu excel dosyaları kontrol edilerek tekrara düşen tweetler tarihlerine dikkat edilerek temizlendi, birleştirildi, veri setimizde ihtiyacımız olmayan diğer kısımlar silindi ve sadece tweetlerin yazı içeriği ile yüklendi. 21 Gün boyunca tekrarlanan #tff etiketi veri çekme işlemi sonucu 4009 adet tweet elde edildi.

Veri setindeki en eski tweetin paylaşılma tarihi: "2021-12-09 23:12:57 UTC", en yeni tweetin paylaşılma tarihi ise: "2022-01-07 21:23:13 UTC".

```
setwd("D:/DERS/21-22 güz/veri madenciliği/VeriCekmeTwitter/Odev")

tweets = read.csv("tweets.csv", header=T, sep = ";")
```

	text
1	Hakemler o kadar Trabzonspor'u kulluyor ki FENERBAHÇE-GALATASARAY ve BEŞİKTAŞIN başkanları bence bu hakeml...
2	@Besiktas Yazıklar olsun #TFF #tffistifa !!!!
3	@Ibrhm_Lothbrok @alifuatsekmen @mackolik İyi top oynuyorsunuz. Adamalarin sutlari yok buraya kadar tamam. Ha...
4	RT @hknozcinarBJK: Çok güzel bir tiyatro izliyoruz ve galiba sonu beli yazık verin kupayı bitsin bu iş #tff #TSvYMS İ...
5	TrabzonFutbolFederasyonu?? #TFF #Trabzon #MHK @TFF_Org https://t.co/ztJSO4v82Z
6	@Fenereditor @CanerErkin @TFF_Org Caner birdaha ceza alırsa alacağı olan o bir maçı hesabında düş #tff
7	RT @hknozcinarBJK: Çok güzel bir tiyatro izliyoruz ve galiba sonu beli yazık verin kupayı bitsin bu iş #tff #TSvYMS İ...
8	Çok güzel bir tiyatro izliyoruz ve galiba sonu beli yazık verin kupayı bitsin bu iş #tff #TSvYMS İyi Beşiktaşlıyım... #t...
9	RT @besiktas190344: #prenaltı #tff #TSvYMS AYIP &AYIP https://t.co/fbRCl1aTv8
10	#prenaltı #tff #TSvYMS AYIP &AYIP https://t.co/fbRCl1aTv8

-Yüklenen verilerin temizlenmesi ve hazırlanması:

Tweetlerin analiz edilebilmesi için gereksiz ifadelerden temizlenmesi gerekiyor, örneğin: RT, @ gibi ifadeler ve noktalama işaretleri gibi.

tweet_clean adında yeni bir tablo oluşturarak verilerimizi temizliyoruz:

```
#verileri temizleme ve analize hazirlama asamalari

tweet_clean <- tweets

tweet_clean$text <- stri_enc_toutf8(tweet_clean$text)

#RT ifadelerinin kaldirilmasi
tweet_clean$text <- ifelse(str_sub(tweet_clean$text,1,2) == "RT",
                           substring(tweet_clean$text,3),
                           tweet_clean$text)

#URL linklerinin temizlenmesi
tweet_clean$text <- str_replace_all(tweet_clean$text, "http[^[:space:]]*", "")

#Hashtag "#" ve "@" isaretlerinin kaldirilmasi
tweet_clean$text <- str_replace_all(tweet_clean$text, "#\\S+", "")
tweet_clean$text <- str_replace_all(tweet_clean$text, "@\\S+", "")

#Noktalama isaretlerinin temizlenmesi
tweet_clean$text <- str_replace_all(tweet_clean$text, "[[:punct:][:blank:]]+", " ")

#Tum harfleri kucuk harfe cevirme
tweet_clean$text  <- str_to_lower(tweet_clean$text, "tr")

#Rakamlarin temizlenmesi
tweet_clean$text <- removeNumbers(tweet_clean$text)

#ASCII formatina uymayan karakterlerin temizlenmesi
tweet_clean$text <- str_replace_all(tweet_clean$text, "[<].*[>]", "")
tweet_clean$text <- gsub("\uFFFD", "", tweet_clean$text, fixed =  TRUE)
tweet_clean$text <- gsub("\n", "", tweet_clean$text, fixed =  TRUE)

#Alfabetik olmayan karakterlerin temizlenmesi
tweet_clean$text <- str_replace_all(tweet_clean$text, "[^[:alnum:]]"," " )
#bosluk temizleme
tweet_clean$text <- stripWhitespace(tweet_clean$text)
```

	text
1	hakemler trabzonspor kulluyor fenerbahçe galatasa...
2	yazıklar
3	iyi top oynuyorsunuz adamalarin sutlari buraya ta...
4	güzel tiyatro izliyoruz galiba sonu beli yazık verin k...
5	trabzonfutbolfederasyonu
6	caner birdaha ceza alırsa alacağı maçı hesabında d...

-Stopwords (Durak-Etkisiz Kelimeler) Belirlenmesi ve Çıkartılması:

Analiz için etkisi olmayacak veri setini şişiren bağlaçlar, tek harf, sayılar gibi ifadeler belirlenir ve temizlenir:

```
#stopwords(durak kelimeler) belirlenmesi ve temizlenmesi
liste=c(stopwords("en"),"tag","rt","retweet","çok","daha","az","bir",
        "iki","üç","dört","beş","altı","yedi","sekiz","dokuz",
        "ile","ve","veya","ama","biraz","fakat","biz","siz",
        "onlar","ben","sen","o","mi","ya","en","kadar","artık",
        "bundan","bu","şu","dan","da","de","neden","ne","niye",
        "niçin","nasıl","için","nin","nun","un","in","ki","ilk",
        "ettiği","yaptığı","önce","sonra","öyle","böyle","şöyle",
        "devam","mı","amp","son","yine","bile","hafta","gün",
        "ay","yıl","yok","sayısı","sayı","say","den","gibi",
        "yapılan","edilen","tüm","bugün","yarın","ye","olsun",
        "değil","olarak","u","edenlere","yerde","edenler","ı",
        "e", "a","olan","alan","kalan","t","th","tf","d","süper",
        "super","savaş","savas","baris","barış","kaç","teşekkür",
        "teşekkürler","karşı","yana","yan","ora","bura",
        "olan","oldu","olmuş","bul","bulun","ın","in")
tweet_clean$text = removeWords(tweet_clean$text,liste)
```

-Verilerin kelimelere bölünmesi:

Verilerimizi kelime şeklinde parçalıyoruz:

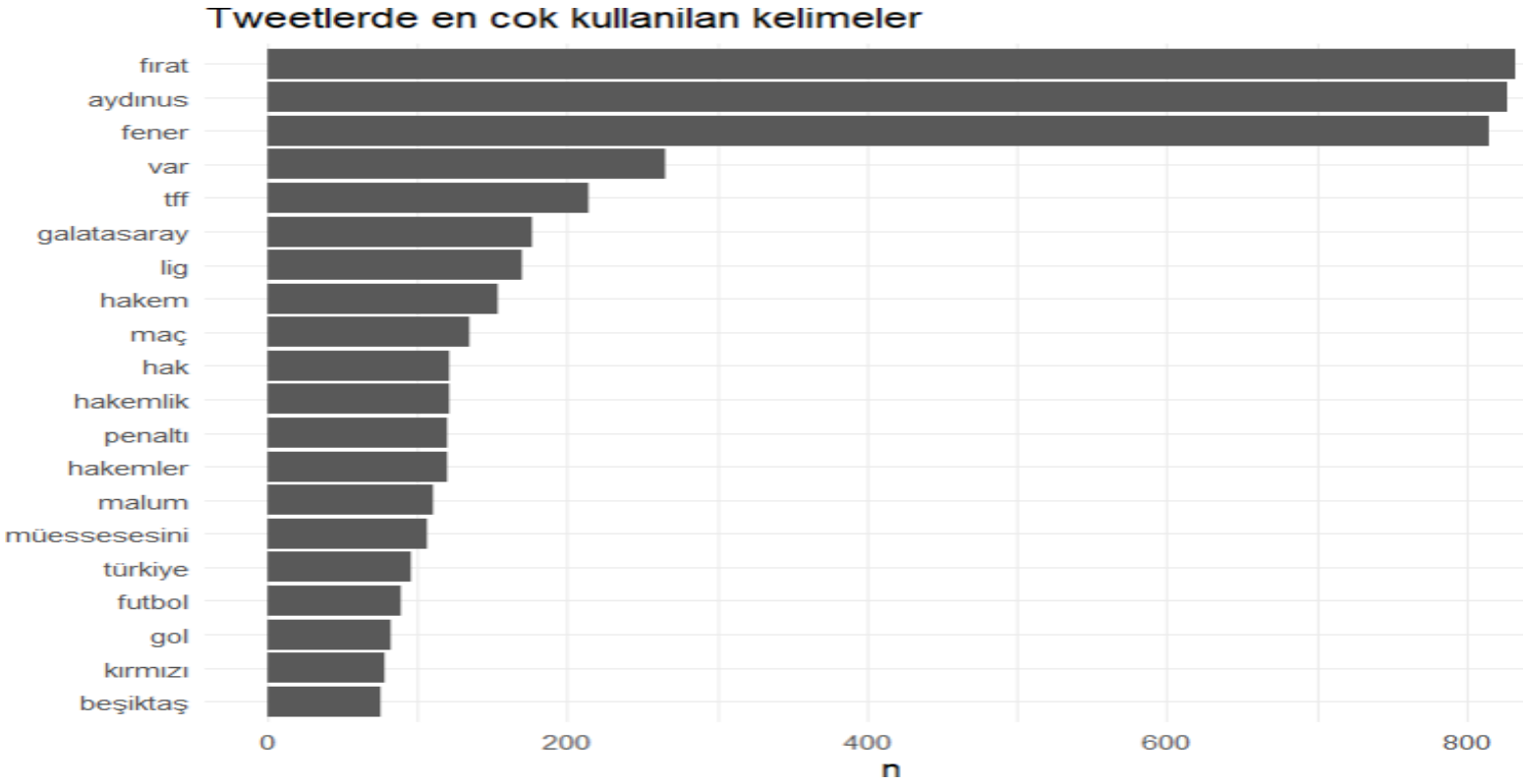
```
tidy_tweets <- tweet_clean %>% select(text) %>%
  mutate(linenumber = row_number()) %>% unnest_tokens(word, text)

tidy_tweets <- tidy_tweets
```

	linenumber	word
1	1	hakemler
2	1	trabzonspor
3	1	kulluyor
4	1	fenerbahçe
5	1	galatasaray
6	1	beşiktaşın
7	1	başkanları
8	1	bence
9	1	hakemlere
10	1	inat

-Verilerin Analizinin Yapılması

En çok tekrar eden (75 ve üzeri kez) kelimelerin plot grafiği:

```
tidy_tweets %>%
  count(word, sort = TRUE) %>%
  filter(n > 75) %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(word, n)) +
  geom_col() +
  xlab(NULL) +
  coord_flip() + theme_minimal() +
  ggtitle("Tweetlerde en cok kullanilan kelimeler")
```



En çok tekrar eden 50 kelime ve frekanları:

```
#en cok tekrar eden 50 kelime ve frekanslari
options(max.print=100)
tidy_tweets %>%count(word, sort = TRUE)
```

```
> options(max.print=100)
> tidy_tweets %>%count(word, sort = TRUE)
                word   n
1              fırat 833
2            aydınus 827
3              fener 815
4                var 265
5                tff 214
6         galatasaray 177
7                lig 170
:                  :   :
49           fonlar  45
50             ligi  45
```

Kelime bulutu oluşturulması:

En çok tekrar eden 80 adet kelime ile kelime bulutu oluşturuldu:

```
#kelime bulutunun olusturulmasi asamasi
pal <- brewer.pal(8, "Dark2")
pal <- pal[-(1:2)]

tidy_tweets %>%
  count(word) %>%
  with(wordcloud(word, n, max.words = 80, colors = pal))
```

En çok tekrar eden 3 kelime “fırat”,”aydınus” ve “fener” en büyük şekilde bulutta yer kaplamakta.

-Veri Analizinin Sonucunun Değerlendirilmesi:

21 Günlük tweet çekimi sonucu hazırlanmış veri setinde en çok tekrar eden kelime: 833 defa tekrar eden “fırat” kelimesi olmuştur, bu kelimeyi 827 defa ile “aydınus”, 815 defa ile “fener” kelimesi takip etmiştir, #tff etiketiyle paylaşılmış verilerden oluşturulmuş bu veri setinde hakkında en çok tweet atılan konu “Hakem Fırat Aydınus” olmuştur. Ardından gelen “fener” yani kastedilen Fenerbahçe kelimesi ile hakem Fırat Aydınus’un söz konusu bir Fenerbahçe maçını yönettiği çıkarımını yapabiliriz.

154 defa “hakem”, 121 defa “hakemlik” ve 120 defa “hakemler” kelimesi tekrar etmiş, TFF’ye bağlı lig hakemlerinden 833 defa “fırat” şeklinde adı tekrar eden Hakem Fırat Aydınus da düşünüldüğünde #tff etiketi ile toplanan verilerde en çok hakkında paylaşım yapılmış olan konu TFF Hakemleridir.

Sonuç olarak: #tff etiketi ile yapılmış olan yani Türkiye Futbol Federasyonu hakkında atılmış olan tweetler çoğunlukla Twitter kullanıcılarının o günkü maçın üzerlerinde bırakmış olduğu etki, federasyona olan bakış açıları, hakemler hakkındaki görüşleri ve her türlü eleştirilerini içermektedir.