Assignment 9 - Twitter

Pratik Agrawal (804861)

15/01/2020

##Setup

```
library(ggpubr)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: magrittr
```

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------
---------- tidyverse 1.3.0 --
## v tibble 2.1.3    v dplyr   0.8.3
## v tidyr  1.0.0    v stringr 1.4.0
## v readr  1.3.1    v forcats 0.4.0
## v purrr  0.3.3
```

```
## -- Conflicts -----------------------------------------------------------------
## x tidyr::extract()masks magrittr::extract()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x purrr::set_names() masks magrittr::set_names()
```

```
# Open Libraries
library(rtweet)
```

```
##
## Attaching package: 'rtweet'
```

```
## The following object is masked from 'package:purrr':
##
##     flatten
```

```
# Speficy Authentification Token's provided in your Twitter App

create_token(
  app = "Sentiment_ndtv",






)
```

```
## <Token>
## <oauth_endpoint>
##  request:  https://api.twitter.com/oauth/request_token
##  authorize: https://api.twitter.com/oauth/authenticate
##  access:   https://api.twitter.com/oauth/access_token
## <oauth_app> Sentiment_ndtv
##   key:    N7mHCm2vQ12lBsN0Q26aL6kNc
##   secret: <hidden>
## <credentials> oauth_token, oauth_token_secret
## ---
```

Task 1

1. What is the main research question? Briefly explain the essence of the paper in 1-3 sentences.

The main research question the paper is how do one automatically quntify and scale the customer needs for innovtion and business growth?

The paper suggests one approch to answer above question using machine learning techniques and socialmedia data. The authors suggest techniques such as surveys, interviews etc are not automated and not scalable enough to extract insights about customer needs. They also demostrate the approch using e-mobility twitter data.

2. Replicate the search performed by the authors on page 8 (section “Data Acquisition & Labeling”). I.e., using the Twitter API, collect every instance (tweet), excluding retweets, which contains at least one item of predefined keyword list:

• e-tankstelle, eauto, elektroauto, elektrofahrzeug, elektromobilitaet, elektro-mobilität, ladesaeule, ladesäule

• ecar, electric mobility, EV vehicle, e-mobility, emobility

• bmw i3, egolf, eup, fortwo electric drive, miev, nissan leaf, opel ampera, peugeot ion, renault zoe, tesla model s

```r
list_of_tweets <- search_tweets2(
  c("e-tankstelle OR eauto OR elektroauto OR elektrofahrzeug OR elektromobilitaet
    OR elektromobilität OR ladesaeule OR ladesäule",
    "ecar OR \"electric mobility\" OR \"EV vehicle\" OR e-mobility OR emobility",
    " \"bmw i3\" OR egolf OR eup OR \"for two electric drive\" OR miev
    OR  \"nissan leaf\" OR \"opel ampera\" OR \"peugeot ion\" OR \"renault zoe \"
    OR \"tesla model s\" "),
  type = "recent", include_rts = FALSE,
geocode = NULL, max_id = NULL, parse = TRUE, token = NULL,
retryonratelimit = TRUE, verbose = TRUE,lang = "en"
)
dim(list_of_tweets)
```

```
## [1] 2738  91
```

```r
tail(list_of_tweets$text)
```

```
## [1] "Our rebate has been extended through March 31, 2020! Make your next car 100% Elect
## [2] "I spent a day with a Tesla Model S this past week. It was great. Supercharger charg
## [3] "@Lemonosity_ renault zoe! read mad good reviews on it, obvs u pay little to no road
## [4] "Well another day another new bmw i3 driver at the MK hub and a new subscriber:) hin
## [5] "@shortword @The_real_colin @Tweetermeyer coincidentally, the Egolf is the only curr
## [6] "@Carpervert @FullyChargedShw P.s. eGolf driver here"
```

```r
write_as_csv(list_of_tweets, "EV.csv", prepend_ids = TRUE, na = "", fileEncoding = "UTF-8"
```

3. Reflect on the search query. If you were to conduct similar research now, can the keyword list be extended? If so, what words would you add?

Yes, The keyword list can surely be extended. I would add other car companies in the list which are working on their own versions of electric cars. for example, Diamler and other US companies.

4. Based on the language information of Twitter, filter for tweets by language by creating three subsamples: German, English, and other-language tweets. How many cases are there in each subsample? Save the subsamples as a comma-separated value file.

German

```r
list_of_tweets_de <- search_tweets2(
  c("e-tankstelle OR eauto OR elektroauto OR elektrofahrzeug
    OR elektromobilitaet OR elektromobilität OR ladesaeule
    OR ladesäule","ecar OR \"electric mobility\" OR \"EV vehicle\"
    OR e-mobility OR emobility","\"bmw i3\" OR egolf OR eup
    OR \"fortwo electric drive\" OR miev OR \"nissan leaf\"
    OR  \"opel ampera\" OR \"peugeot ion\" OR \"renault zoe \"
    OR \"tesla model s\" "),
  type = "recent", include_rts = FALSE,
geocode = NULL, max_id = NULL, parse = TRUE, token = NULL,
retryonratelimit = TRUE, verbose = TRUE, lang = "de"
)
```

```
## retry on rate limit...
## waiting about 6 minutes...
```

```r
dim(list_of_tweets_de)
```

```
## [1] 2696 91
```

```r
tail(list_of_tweets_de$text)
```

```
## [1] "Auch das Laden beim Arbeitgeber soll steuerlich bevorteilt werden (geldwerter Vort
## [2] "Kaufdeal! BMW i3 120 Ah bei Märtin [11.381 Euro brutto Ersparnis] | Verbrauch: • S
## [3] "@Solokrieger @UdyrSux @GenderP1 @zeitonline Das einzige Bauteil, wo man Seltene Er
Verbrenner"
## [4] "CES 2020: Mehr Fotos vom BMW i3 Urban Suite Concept https://t.co/PWNZ4JnTiD"
## [5] "Mein zweiter Tag auf der #CES2020 beginnt damit das ich den #BMW i3 Urban Suite vo
## [6] "@YogicCEO Das war mein Beitrag vor etwa 1 Jahr: Leider werden gute #Elektroautos systematis
```

```
write_as_csv(list_of_tweets_de, "EV_DE.csv", prepend_ids = TRUE, na = "", fileEncoding = "
```

French

```
list_of_tweets_fr <- search_tweets2(
  c("e-tankstelle OR eauto OR elektroauto OR elektrofahrzeug OR elektromobilitaet
  OR elektromobilität OR ladesaeule OR ladesäule","ecar OR \"electric mobility\"
  OR  \"EV vehicle\"
    OR e-mobility OR emobility","\"bmw i3\" OR egolf OR eup
  OR \"fortwo electric drive\" OR miev OR \"nissan leaf\"
  OR  \"opel ampera\" OR \"peugeot ion\" OR \"renault zoe \" OR \"tesla model s\" "),
  type = "recent", include_rts = FALSE,
geocode = NULL, max_id = NULL, parse = TRUE, token = NULL,
retryonratelimit = TRUE, verbose = TRUE, lang = "fr"
)

dim(list_of_tweets_fr)
```

```
## [1] 363 91
```

```
tail(list_of_tweets_fr$text)
```

```
## [1] "@TheTeslaShow Tesla Model S - Cybertruck edition?"
## [2] "2014 beu légui mangui twitter mousoumeu def lou melni wayé nak wolof njaay néna lou
## [3] "@yethiooo @Macky_Sall Walahi\nT le pire Moy Etat bi yonam nekouci... \nIls ont bana
## [4] "#CES2020 de Las Vegas : BMW courtise les VTC américains avec une petite électrique
## [5] "Après 1,000 km parcourus en 3 semaines, je confirme que cette petite BMW I3 60aH es
## [6] "CES 2020 : interconnecter votre Renault et les objets connectés de votre maison, c
```

```
write_as_csv(list_of_tweets_fr, "EV_fr.csv", prepend_ids = TRUE, na = "", fileEncoding = "
```

Samples language wise,

```
cat("English: " ,nrow(list_of_tweets))
```

```
## English: 2738
```

```
cat("German: ",nrow(list_of_tweets_de))
```

```
## German: 2696
```

```
cat("French: ",nrow(list_of_tweets_fr))
```

```
## French: 363
```

5. Look at the results (check the first ten rows of the English subsample). Do you note any issues with your data? What clean-up steps associated with social media data would you recommend?

Tweets can be cleaned by,

1. Removing usernames i.e @usernames

2. Removing Special Characters i.e. Emojis

3. Removing URLs

Task 2

Find the 1000 most recent tweets by Katy Perry (https://twitter.com/katyperry), Kim Kardashian West (https://twitter.com/KimKardashian), and Ariana Grande (https://twitter.com/ArianaGrande). Save the sample as a comma-separated value file.

```r
## compare account activity
tmls <- get_timeline(
c("KimKardashian", "katyperry","ArianaGrande"),
n = 1000
)
```

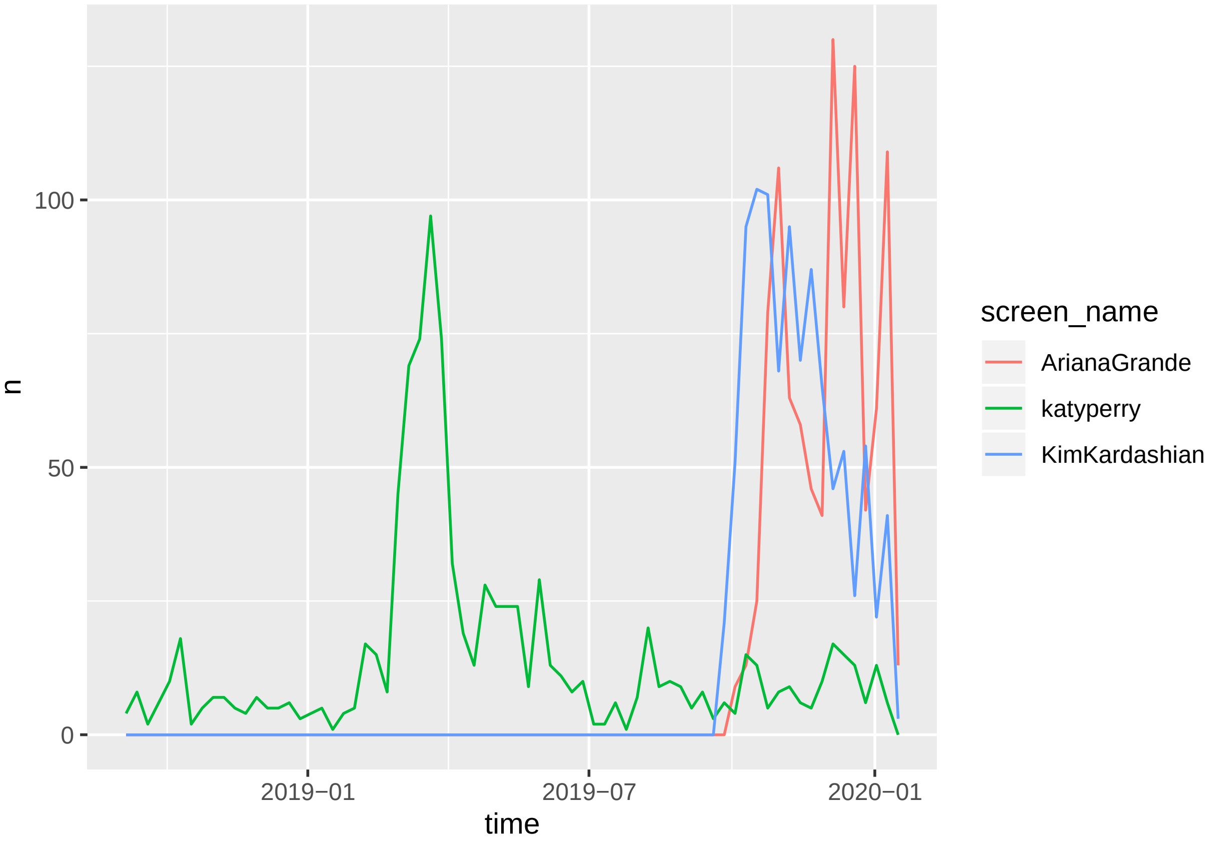
```r
head(tmls$text)
```

```
## [1] "The @skims Essential Bodysuit Collection Available in sizes XXS - 5X, the bodysuit
## [2] "The @skims Essential Bodysuit Collection is coming soon! I’ve been wearing these su
## [3] ".@kimkardashian wears shades Pink Aura, Rose Dust and Cosmic Brown from the Night S
## [4] ".@KKWBeauty's Celestial Skies collection is inspired by soft sunsets and shimmery
## [5] "I’ve had the new @kkwbeauty Bronze Heaven palette in my purse for months - The gol
https://t.co/bkD1E1Zgl5"
## [6] "I love the new @kkwbeauty Sepia Sunset Eyeshadow Palette’s warm matte and metallic
```

```r
#save the Tweets in CSV file
write_as_csv(tmls, "celebs.csv", prepend_ids = TRUE, na = "", fileEncoding = "UTF-8")
```

Visualize the tweet frequency of the above celebrities by Weeks. Who posts most often? Who posts least often?

```r
## group by screen name and plot each time series
ts_plot(dplyr::group_by(tmls, screen_name), "weeks")
```



Ariana Grande Posts more often , Katy Perry Posts less often.

Task 3

Consider three important German (or a country of your origin) political figures of your choice. Briefly reason your choice (1-2 sentences). Similar to exercise 2, find 2000 most recent tweets by these three prominent political figures of your choice. Save the sample as a comma-separated value file.

I am exploring Tweets from Indian Politicians.

Narendra Modi is a prime minister of india, Rahul Gandhi and Arvind Kejriwal are most popular opposition leaders.

Narendra Modi has 52.5 Million followers in twitter, which is more than half of total population of germany.

other two leaders have combined more than 28 Million twitter followers.

```
# NAMO: 52.5 Mil, RAGA: 11.8 MIL, KEjri:16.3
## compare account activity
tmls_politics <- get_timeline(
c("narendramodi", "RahulGandhi","ArvindKejriwal"),
n = 3000
)
dim(tmls_politics)
```

```
## [1] 8993 90
```

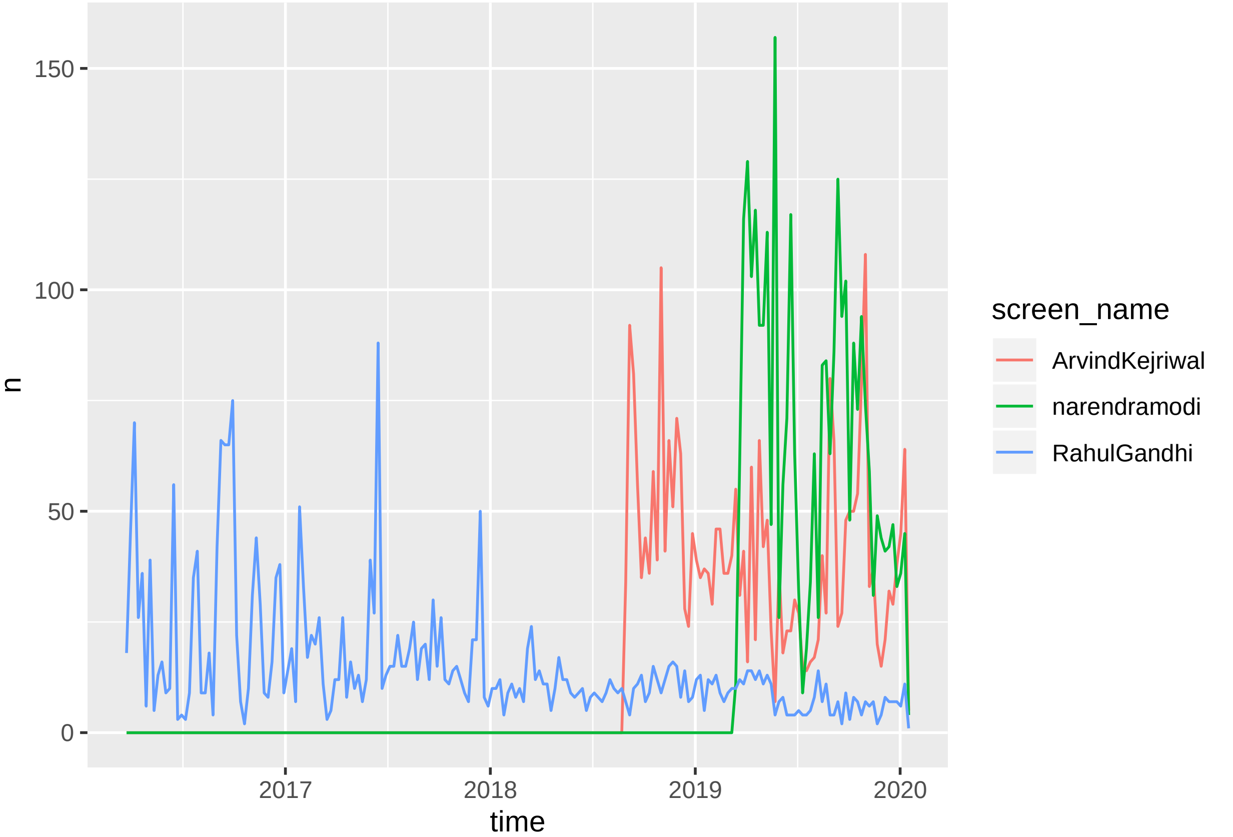
```
head(tmls_politics$screen_name)
```

```
## [1] "narendramodi" "narendramodi" "narendramodi" "narendramodi" "narendramodi"
## [6] "narendramodi"
```

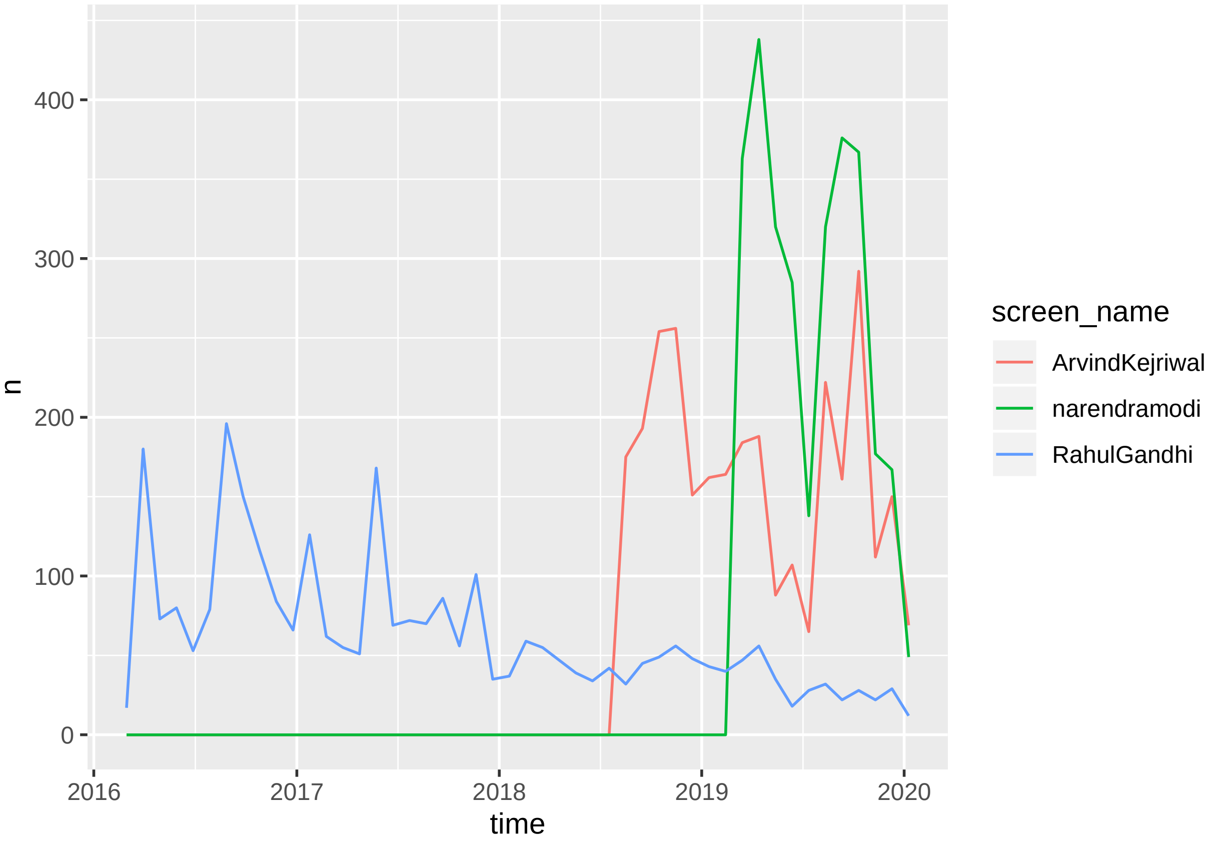
```
#save the Tweets in CSV file
write_as_csv(tmls_politics, "Politicians.csv", prepend_ids = TRUE, na = "", fileEncoding =
```

Compare the account activity for these politicians (visualize the frequency of tweets).Who posts most often? Who posts least often?

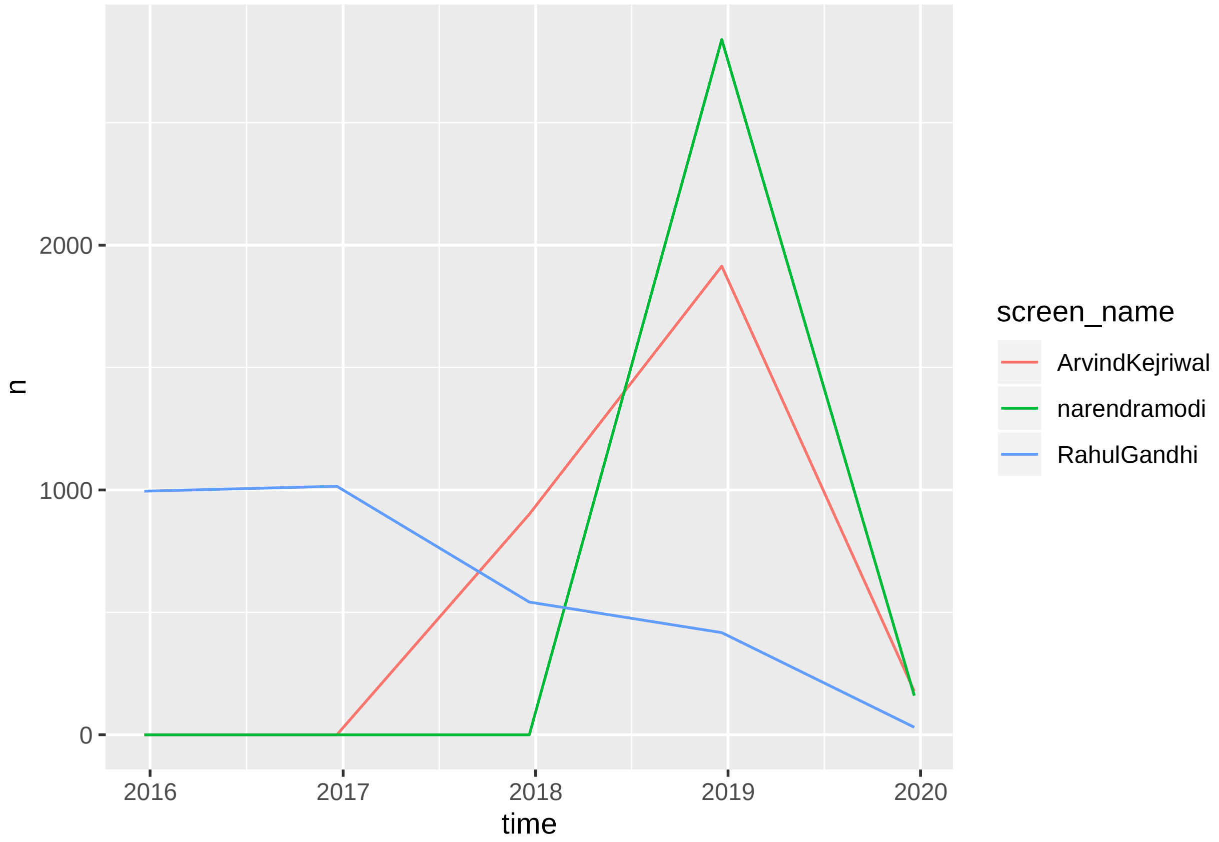
```
## group by screen name and plot each time series
ts_plot(dplyr::group_by(tmls_politics, screen_name), "weeks")
```



```
# group by screen name and plot each time series - months
ts_plot(dplyr::group_by(tmls_politics, screen_name), "months")
```



```
# group by screen name and plot each time series - years
ts_plot(dplyr::group_by(tmls_politics, screen_name), "years")
```



Narendra Modi posts more often, Rahul Gandhi posts less often.