LIMPIANDO Y ORGANIZANDO TU DATA DE TW

CORTO Y DULCE

Pregunta de Investigación

- No lo vamos a cubrir
- ¡Importantísimo!
- Nuestra ventaja comparativa

Data

- ~~Obtener data (TW)~~
- Limpiar data
- Organizar la data
- Explorar la data

Crear la red

- Introducción a igraph
- V(net) y E(net) en igraph

Características de la red

- Betweenness
- Centrality
- Paths

Detección de comunidades

- Greedy algorithms / Randomwalk
- Características de las comunidades.

Visualizar la red

- Layout
  - Force-directed, random.
- Centro / Periferia
- Baja y alta reslución

Análisis de la red I

- Activiación de la red
- La paradoja de la amistad
- Rising stars / rising tides

Análisis de la red II

- Disonancias y resonancias cognitivas
- Viralidad
- Análisis de texto
- Experimentos

¿QUÉ INFORMACIÓN TENEMOS?

(LA MEJOR DE TODAS… INFORMACIÓN GRATIS)

QUÉ BUSCAMOS Y CÓMO:

- Hay varias formas de analizar una red. La más directa, es basándonos en un evento de nuestro interés. Por ejemplo, si quiero saber qué pasa en las redes durante las elecciones de Colombia, buscaría:

```
28  search_tweets(q = 'colombia OR eleccion',
29                n = 100,
30                include_rts = T,
31                retryonratelimit = TRUE)
```

- Por lo general (y más adelante veremos exactamente porqué), la mejor manera de construir una red es utilizando términos generales que no tengan un sesgo inherente.

QUÉ BUSCAMOS Y CÓMO:

- Si para el mismo tema de las elecciones colombianas busco el hashtag:

```
28  search_tweets(q = 'duquepresidente',
29                n = 100,
30                include_rts = T,
31                retryonratelimit = TRUE)
```

- Lo más probable es que obtengan más tuits de gente que votará por Iván Duque (y no tendré información sobre quienes no van a votar por Duque).
- **Nota**: no te preocupes si en tu búsqueda obtienes tuits que *no* están relacionados con tu tema de interés. Más adelante veremos cómo solucionamos esto.
- **La política general es:** *mejor que sobre a que falte (data).*

RECOLECTANDO LA DATA

- Comenzamos buscando un término de interés:

```
duque_tw <- search_tweets(q = 'duque', # Hagamos sobre Iván Duque
                          n = 10000000, # Vamos a necesitar muchos tuits
                          include_rts = T, # Si quiere que incluya RTs
                          retryonratelimit = TRUE) # Sin esto, no funciona
```

- Necesitamos *mucha* data. Mientras más mejor. Recolectar un millón de tuits puede tomar cerca de 12 horas. La cantidad total de tuits dependerá de la frecuencia con la que apareció en TW en los últimos 6 a 9 días.
- Si estamos interesados en un evento que está comenzando pero que se está desarrollando (e.g., un paro indígena), entonces lo ideal es correr el código cada dos o tres días, hasta que termine el evento.

¿QUÉ INFORMACIÓN TENEMOS?

Es mucha… pero no todo nos sirve.

```
> colnames(duque_tw)
 [1] "user_id"                "status_id"              "created_at"
 [4] "screen_name"            "text"                   "source"
 [7] "display_text_width"     "reply_to_status_id"     "reply_to_user_id"
[10] "reply_to_screen_name"   "is_quote"               "is_retweet"
[13] "favorite_count"         "retweet_count"          "quote_count"
[16] "reply_count"            "hashtags"               "symbols"
[19] "urls_url"               "urls_t.co"              "urls_expanded_url"
[22] "media_url"              "media_t.co"             "media_expanded_url"
[25] "media_type"             "ext_media_url"          "ext_media_t.co"
[28] "ext_media_expanded_url" "ext_media_type"         "mentions_user_id"
[31] "mentions_screen_name"   "lang"                   "quoted_status_id"
[34] "quoted_text"            "quoted_created_at"      "quoted_source"
[37] "quoted_favorite_count"  "quoted_retweet_count"   "quoted_user_id"
[40] "quoted_screen_name"     "quoted_name"            "quoted_followers_count"
[43] "quoted_friends_count"   "quoted_statuses_count"  "quoted_location"
[46] "quoted_description"     "quoted_verified"        "retweet_status_id"
[49] "retweet_text"           "retweet_created_at"     "retweet_source"
[52] "retweet_favorite_count" "retweet_retweet_count"  "retweet_user_id"
[55] "retweet_screen_name"    "retweet_name"           "retweet_followers_count"
[58] "retweet_friends_count"  "retweet_statuses_count" "retweet_location"
[61] "retweet_description"    "retweet_verified"       "place_url"
[64] "place_name"             "place_full_name"        "place_type"
[67] "country"                "country_code"           "geo_coords"
[70] "coords_coords"          "bbox_coords"            "status_url"
[73] "name"                   "location"               "description"
[76] "url"                    "protected"              "followers_count"
[79] "friends_count"          "listed_count"           "statuses_count"
[82] "favourites_count"       "account_created_at"     "verified"
[85] "profile_url"            "profile_expanded_url"   "account_lang"
[88] "profile_banner_url"     "profile_background_url" "profile_image_url"
```

¿QUÉ INFORMACIÓN NOS SIRVE?

Información sobre el usuario que publicó el tuit y el texto de tuit.

Información sobre el usuario que publicó el tuit original (si el tuit fue retuiteado).

Información sobre el tuit.

```
> duque_tw <- duque_tw[,c(1:7,78:84,48:62,13:17,19)] # Cambio un poco el orden pero no es nece
sario
> colnames(duque_tw)
 [1] "user_id"                "status_id"              "created_at"
 [4] "screen_name"            "text"                   "source"
 [7] "display_text_width"     "followers_count"        "friends_count"
[10] "listed_count"           "statuses_count"         "favourites_count"
[13] "account_created_at"     "verified"               "retweet_status_id"
[16] "retweet_text"           "retweet_created_at"     "retweet_source"
[19] "retweet_favorite_count" "retweet_retweet_count"  "retweet_user_id"
[22] "retweet_screen_name"    "retweet_name"           "retweet_followers_count"
[25] "retweet_friends_count"  "retweet_statuses_count" "retweet_location"
[28] "retweet_description"    "retweet_verified"       "favorite_count"
[31] "retweet_count"          "quote_count"            "reply_count"
[34] "hashtags"               "urls_url"
```

ELIMINAMOS TUITS DUPLICADOS

- A veces, el API manda tuits duplicados. Estos los eliminamos.
- **IMPORTANTE**: eliminamos los tuis duplicados por el API, *no* los tuits duplicados porque usuarios les dieron RT.

```
> duque_tw <- duque_tw[!duplicated(duque_tw$status_id),]
```

¡LISTO!

(Guarden y seguimos)

Pregunta de Investigación

- No lo vamos a cubrir
- ¡Importantísimo!
- Nuestra ventaja comparativa

Data

- ~~Obtener data (TW)~~
- ~~Limpiar data~~
- ~~Organizar la data~~
- Explorar la data

Crear la red

- Introducción a igraph
- V(net) y E(net) en igraph

Características de la red

- Betweenness
- Centrality
- Paths

Detección de comunidades

- Greedy algorithms / Randomwalk
- Características de las comunidades.

Visualizar la red

- Layout
  - Force-directed, random.
- Centro / Periferia
- Baja y alta resolución

Análisis de la red I

- Activiación de la red
- La paradoja de la amistad
- Rising stars / rising tides

Análisis de la red II

- Disonancias y resonancias cognitivas
- Viralidad
- Análisis de texto
- Experimentos