Análisis de datos de Twitter en el Día Mundial de la Salud

Xavier Vivancos García
Máster de Inteligencia de Negocio y Big Data
Análisis de datos en redes sociales

Consultora: Dra. Laia Subirats Maté
Dra. Maria Pujol Jover / Dr. Joan Melià Seguí

25 de junio de 2018

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Análisis de datos de Twitter en el Día Mundial de la Salud*
Nombre del autor:	*Xavier Vivancos García*
Nombre del consultor/a:	*Dra. Laia Subirats Maté*
Nombre del PRA:	*Dra. Maria Pujol Jover y Dr. Joan Melià Seguí*
Fecha de entrega:	06/2018
Titulación:	*Máster en Inteligencia de Negocio y Big Data*
Área del Trabajo Final:	*Análisis de datos en redes sociales*
Idioma del trabajo:	*Castellano*
Palabras clave:	*Análisis de datos, Redes Sociales, Minería de datos*

Resumen del Trabajo:

El principal objetivo del proyecto es analizar *tweets* capturados en el Día Mundial de la Salud (7 de abril) para posteriormente narrar una historia a través de los datos, utilizando Twitter como origen principal de los mismos.

La obtención de los datos se lleva a cabo mediante la *streaming* API de la herramienta R. Una vez capturados, se realizan análisis estadísticos y descriptivos, como también un análisis de sentimientos y del grafo social resultante (usuarios y *tweets* más relevantes, métricas y propiedades de la red y visualización del grafo). También se utilizan *wordclouds* para visualizar las palabras más frecuentes, y mapas para conocer desde donde los usuarios escriben los *tweets*. Varios de los análisis se repiten para *tweets* escritos en inglés, castellano y catalán, de forma que podamos avaluar posibles diferencias.

Abstract:

The main objective of the project is to analyze tweets captured on World Health Day (April 7th) to tell a story through the data, using Twitter as the main source.

The data is obtained through the streaming API of R. Once captured, we carry out statistical and descriptive analysis, as well as a sentiment analysis. We also construct and analyze the retweet network (most relevant users and tweets, metrics and properties of the network and visualization of the graph). Furthermore, we use wordclouds to display the most frequent words, and maps to visualize the most active countries on Twitter. Several of the analyzes are repeated for tweets written in English, Spanish and Catalan, so that we can assess possible differences.

Índice

Lista de figuras

Lista de tablas

1. Introducción

1.1 Contexto y justificación del Trabajo

El Día Mundial de la Salud reivindica la cobertura sanitaria universal. Es decir, se lucha para garantizar que todas las personas, en cualquier lugar, puedan tener acceso a servicios de salud esenciales y de calidad sin tener que pasar apuros económicos. Sólo hace falta leer unos pocos datos[1] de la situación mundial actual para darnos cuenta de la necesidad de un día como este:

- Por lo menos la mitad de la población mundial no puede recibir servicios de salud esenciales.
- Casi 100 millones de personas afectadas por la pobreza extrema se ven obligadas a pagar los servicios de salud de su propio bolsillo.
- Más de 800 millones de personas (casi el 12% de la población mundial) se gastan como mínimo el 10% del presupuesto familiar en gastos de salud.

El Día Mundial de la Salud trata de conseguir avances en la consecución de la cobertura sanitaria mundial haciendo partícipes y otorgando responsabilidades a diferentes colectivos, ya sean gobiernos, ministerios, partidos políticos, medios de comunicación, etc. Incluso cualquier individuo, sin necesidad de pertenecer a ninguna agrupación, puede tomar la iniciativa y aportar su granito de arena para visibilizar aún más esta problemática. En este sentido, el auge de las redes sociales ha dado la oportunidad a los usuarios de hacer oír su voz.

Durante el desarrollo de este trabajo vamos a analizar datos capturados en el Día Mundial de la Salud (7 de abril) para conocer de primera mano la concienciación y opinión de la sociedad entorno a este día. La red social utilizada para llevar a cabo la captura será Twitter. El motivo es que se trata de una plataforma que dispone de mucha información pública a escala global acerca de multitud de contenidos y temáticas. Para ello nos beneficiaremos de la *streaming* API, que nos permite capturar *tweets* en tiempo real en base a una serie de condiciones. En nuestro caso, buscaremos tweets conteniendo aquellas palabras clave o *hashtags* que se hayan utilizado de forma mayoritaria durante el día Mundial de la Salud (#WorldHealthDay o #HealthForAll, por ejemplo).

[1] *Mensajes del Día Mundial de la Salud 2018*. [en línea] Organización Mundial de la Salud. http://www.who.int/campaigns/world-health-day/2018/key-messages/es/

1.2 Objetivos del Trabajo

- Analizar el contenido de los *tweets*.
- Determinar la actitud o sentimientos de los autores de los *tweets* en relación con el Día Mundial de la Salud.
- Estudiar diferencias entre los *tweets* escritos en inglés, castellano y catalán.
- Dibujar y describir la estructura de red formada entre *tweets* y *retweets*.
- Determinar en qué países la concienciación en Twitter es mayor/ menor.

1.3 Enfoque y método seguido

1. Capturar datos de Twitter en el día Mundial de la Salud.
2. Describir y analizar estadísticamente los datos.
3. Analizar los textos contenidos en los *tweets* mediante técnicas de *text mining*.
4. Clasificar la polaridad de los *tweets*.
5. Evaluar las diferencias entre *tweets* de diferentes idiomas.
6. Obtener la red o grafo social formado por las relaciones entre *tweets* y *retweets*.
7. Extraer métricas y propiedades del grafo social, así como otros datos de interés: usuarios y *tweets* más relevantes, detección de comunidades, etc.
8. Situar los *tweets* en un mapa, en función de su procedencia geográfica.
9. Confeccionar las conclusiones

1.4 Planificación del Trabajo

N	***Actividades***	***Fecha***
1.	PEC 1: Propuesta de trabajo y planificación temporal	03/04/2018
1.1.	Decisión del área de trabajo y datos a capturar	01/04/2018
1.2.	Redacción de la propuesta de proyecto y planificación temporal	01/04/2018
2.	PEC 2: Apartados iniciales y primeros análisis	30/04/2018
2.1.	Búsqueda bibliográfica	15/04/2018
2.2.	Ampliación de los apartados introductorios	20/04/2018
2.3	Captura de datos	23/04/2018
2.4.	Primeros análisis descriptivos y estadísticos de los datos	26/04/2018
3.	PEC 3: Análisis avanzados	04/06/2018
3.1.	Extracción de la polaridad de los *tweets*	05/05/2018
3.2	Visualización y descripción del grafo social	15/05/2018
3.3.	Visualización de los *tweets* en un mapa	20/05/2018
3.4.	Redacción de resultados y conclusiones	30/05/2018
3.5	Redacción de la memoria	20/06/2018
4.	Defensa del trabajo	10/07/2018

Tabla 1.1. Planificación del Trabajo

2. Estado del arte

En la primera parte del proyecto llevamos a cabo la captura de datos de los *tweets*. Lo explicamos más en detalle en el apartado 3, pero resumiendo, utilizamos la *streaming* API de Twitter para capturar *tweets* en tiempo real durante el Día Mundial de la Salud, apoyándonos en las librerías de R ROAuth, streamR y twitteR.

Durante el trabajo también utilizamos el análisis de sentimientos[2]. Esta técnica nos permite determinar la actitud del usuario con respecto a algún tema, basándose en relaciones estadísticas y de asociación. Para ello, se catalogan individualmente las palabras de los *tweets* en función del sentimiento asociado, y se suman las frecuencias de las diferentes categorías para conocer la contribución de cada sentimiento en los *tweets*. El proceso que hemos seguido para la elaboración de las visualizaciones se puede resumir mediante la siguiente figura:

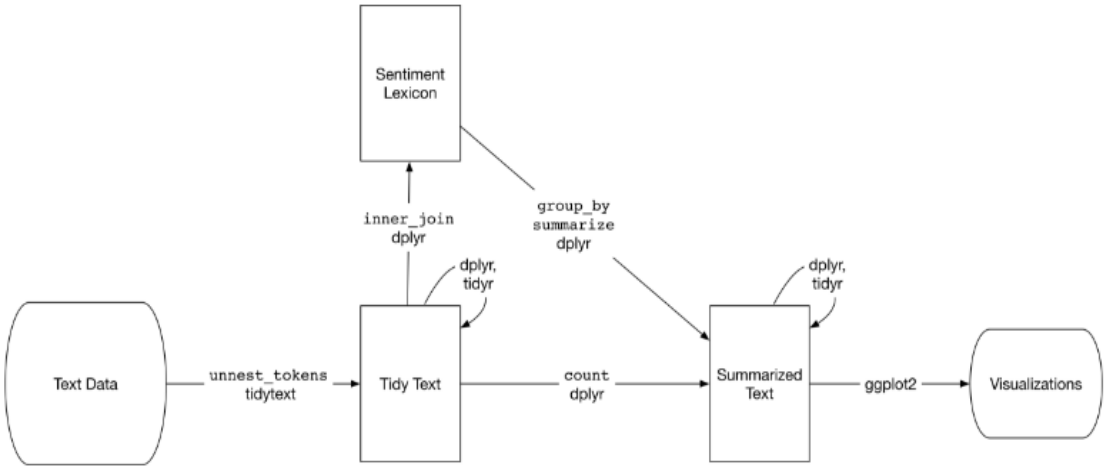

Figura 2.1. Proceso para el análisis de sentimientos

Como vemos, primero dividimos las frases en palabras o *tokens*. A continuación asociamos a cada *token* un sentimiento (una misma palabra puede tener asociados varios sentimientos a la vez) y calculamos sus frecuencias. Finalmente agrupamos por sentimiento y sumamos la contribución de cada uno.

[2] Colaboradores de Wikipedia. *Análisis de sentimiento* [en línea]. Wikipedia, La enciclopedia libre, 2018 [fecha de consulta: 19 de junio del 2018]. Disponible en https://es.wikipedia.org/w/index.php?title=An%C3%A1lisis_de_sentimiento&oldid=108714922

Las librerías que hemos utilizado en este proceso son tidytext[3] y tidyverse[4] (que incluye dplyr[5], tidyr[6] y ggplot2[7], entre otras). Durante la sección del análisis de sentimientos utilizamos diferentes métodos o también denominados lexicones que incluye la librería tidytext. Son los siguientes:

- Lexicón AFINN[8] (de Finn Årup Nielsen). Categoriza las palabras según una escala numérica del -5 al 5. Cuanto mayor es la puntuación, más positiva es la palabra.
- Lexicón Bing[9] (de Bing Liu). Clasifica las palabras en dos categorías: positivas o negativas.
- Lexicón NRC[10] (de Saif Mohammad y Peter Turney). Categoriza las palabras según emociones tales como: ira, miedo, tristeza, confianza, disgusto, sorpresa, etc.

Una vez descrita la sección del análisis de sentimientos, pasamos a explicar el proceso de elaboración de los *wordclouds* con las palabras más frecuentes. El primer paso consiste en la creación del Corpus, el cual contiene la colección de textos a analizar. Antes de contabilizar las palabras, es necesario limpiar o acondicionar el texto de los *tweets* para eliminar ruido, entendiendo por ruido aquello que es superfluo y prescindible. Este acondicionamiento incluye las siguientes transformaciones: eliminación de los signos de puntuación, espacios y números. Además, es necesario excluir ciertas palabras que no tienen ningún valor para el análisis, como pueden ser las preposiciones, artículos, pronombres, conjunciones, etc. Aunque estas palabras ayudan a conectar palabras y frases, carecen de sentido propio.

---

[3] Silge J, Robinson D (2016). "tidytext: Text Mining and Analysis Using Tidy Data Principles in R." _JOSS_, *1*(3). doi: 10.21105/joss.00037 (URL: http://doi.org/10.21105/joss.00037), <URL: http://dx.doi.org/10.21105/joss.00037>.

[4] Hadley Wickham (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse

[5] Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2018). dplyr: A Grammar of Data Manipulation. R package version 0.7.5. https://CRAN.R-project.org/package=dplyr

[6] Hadley Wickham and Lionel Henry (2018). tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions. R package version 0.8.1. https://CRAN.R-project.org/package=tidyr

[7] H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2009.

[8] Finn Årup Nielsen (2011). AFIIN lexicon. http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010

[9] Bing Liu. Bing lexicon. https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

[10] Mohammad S, Turney P (2010). NRC lexicon. http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

Una vez disponemos del texto acondicionado, es momento de crear la matriz de términos[11]. Esta matriz nos indica el número de ocurrencias de los términos en una colección de documentos. La primera columna contiene los términos, mientras que los documentos (*tweets* en nuestro caso) están colocados en la primera fila. Finalmente, ordenamos las palabras en función de su frecuencia de aparición y creamos el *wordcloud*.

Para la creación del Corpus, el acondicionamiento de los textos y la creación de la matriz de términos hemos utilizado la librería tm[12]. Puesto que el análisis se lleva a cabo en idiomas diferentes, hemos excluido las palabras redundantes para cada uno de ellos (*what* o *didn't*, por ejemplo). Para la visualización de la nube de palabras se ha hecho uso de la librería wordcloud2[13].

Otro de los aspectos importantes durante el proyecto es la geolocalización de los *tweets* y su visualización mediante un mapa. Para estas dos tareas utilizamos la librería ggmap[14]. Los *tweets* se geolocalizan utilizando el campo *location*, obteniendo las coordenadas de latitud y longitud. R únicamente nos permite 2.500 peticiones por día, así que hemos dividido la columna de las localizaciones en grupos de 2.500 para geolocalizarlas por separado.

Con respecto al grafo social de *tweets* y sus *retweets*, hemos utilizado la librería igraph[15] para su elaboración y el *software* Gephi para la visualización. Vamos a definir algunas métricas y propiedades de grafos, puesto que posteriormente las utilizaremos para categorizar la red creada entre *tweets* y *retweets*:

- Un nodo es un punto de intersección, conexión o unión de varios elementos que confluyen en el mismo lugar. En nuestro análisis, los nodos representan *tweets*.
- Los arcos o aristas enlazan los diferentes nodos. En nuestro caso, las aristas representan la existencia de un *retweet*.

---

[11] Wikipedia contributors. (2018, May 14). Document-term matrix. In *Wikipedia, The Free Encyclopedia*. Retrieved 19:27, June 22, 2018, from https://en.wikipedia.org/w/index.php?title=Document-term_matrix&oldid=841158065

[12] Ingo Feinerer and Kurt Hornik (2017). tm: Text Mining Package. R package version 0.7-3. https://CRAN.R-project.org/package=tm

13 Dawei Lang (NA). wordcloud2: Create Word Cloud by htmlWidget. R package version 0.2.0. https://github.com/lchiffon/wordcloud2

14 D. Kahle and H. Wickham. ggmap: Spatial Visualization with ggplot2. The R Journal, 5(1), 144-161. URL http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf

15 Csardi G, Nepusz T: The igraph software package for complex network research, InterJournal, Complex Systems 1695. 2006. http://igraph.org

- Un grafo dirigido es un tipo de grafo en el cual las aristas tienen un sentido definido.
- Un grafo denso es un grafo en el que el número de aristas es cercano al número máximo de aristas. Lo opuesto, un grafo con sólo algunas aristas, es un grafo disperso.
- El coeficiente de agrupamiento o transitividad se calcula como la media de los valores del coeficiente de agrupamiento de cada nodo. Esta métrica calcula la fracción de posibles triángulos que existen en la red.
- La reciprocidad en un grafo dirigido mide la tendencia a tener conexiones mutuas entre los nodos.

Una vez explicados todos los conceptos y técnicas que se utilizan durante el proyecto, podemos pasar a explicar la captura de datos en el siguiente apartado.

3. Captura de datos

3.1 Preparación del entorno

Para poder empezar con la captura de datos, previamente tenemos que conseguir unas credenciales adecuadas para acceder a la API de Twitter. Los pasos a seguir son los siguientes,

1. Disponer de una cuenta de Twitter. En mi caso ya tengo la siguiente: @Xavier91vg
2. Registrar una aplicación en Twitter en https://dev.twitter.com/apps.
3. Acceder a la pestaña "Key and Access Tokens" y obtener las credenciales.
4. Instalar los paquetes en R que permiten interactuar con Twitter.
5. Registrar las credenciales de la aplicación e iniciar el proceso de autenticación.
6. Ejecutar una consulta sencilla para validar que las capturas funcionan correctamente.

3.2 Herramienta y librerías

Con respecto a la herramienta utilizada, el proceso de captura se lleva a cabo mediante RStudio y desde la cuenta @Xavier91vg. Las librerías empleadas son las siguientes,

- ROAuth[16]. Una vez finalizado el proceso de validación y registro de las credenciales, esta librería permite almacenarlas en un objeto OAuth. De esta forma, las próximas veces que trabajemos con estas credenciales, sólo deberemos cargar estos datos en lugar de repetir todo el proceso.
- streamR[17]. Proporciona una serie de funciones que permiten a los usuarios de R acceder a datos de Twitter.
- twitterR[18]. Facilita una interfaz para la API de Twitter.

[16] Jeff Gentry and Duncan Temple Lang (2015). ROAuth: R Interface For OAuth. R package version 0.9.6. https://CRAN.R-project.org/package=ROAuth

[17] Pablo Barbera (2014). streamR: Access to Twitter Streaming API vida R. R package version 0.2.1. https://CRAN.R-project.org/package=streamR

[18] Jeff Gentry (2015). twitteR: R Based Twitter Client. R package version 1.1.9. https://CRAN.R-project.org/package=twitteR

3.3 Ejecución de la captura

Para la obtención de los datos utilizamos la *streaming* API[19], que permite conectar y filtrar los *tweets* que se están publicando en tiempo real. Es decir, se deja el proceso funcionando durante un cierto período de tiempo, y se recogen todos los *tweets* que se han publicado durante este período. La captura se ha iniciado a las 15:17 en el Día Mundial de la Salud (7 de abril), finalizando a las 18:17 y durando un total de tres horas.

En cuanto a las palabras clave o *hashtags*, hemos especificado los siguientes:

- #WorldHealthDay
- #DiaMundialDeLaSalud
- #DiaMundialDeLaSalut
- #SalutPerATothom
- #HealthForAll
- #WHD2018

Hemos tenido en cuenta varios idiomas (castellano, catalán e inglés) para poder evaluar las diferencias de los *tweets* en función de la lengua. Una vez la captura finaliza, los resultados son guardados en un archivo JSON ubicado en nuestro directorio de trabajo. El fichero contiene un total de 42.838 *tweets* listos para ser analizados.

3.4 Diccionario de datos

A continuación, mostramos los campos[20] que contiene el archivo con los *tweets* capturados, incluyendo una pequeña explicación del significado de cada uno.

Campo	Tipo	Explicación
created_at	String	Momento de la publicación del *tweet*
id	Int64	Identificador único del *tweet*
id_str	String	Identificador único del *tweet*
text	String	Contenido del *tweet* codificado en UTF-8
source	String	Aplicación desde donde se envía el *tweet*
truncated	Boolean	Indica si el valor del parámetro text se ha truncado
in_reply_to_status_id	Int64	Identificador del *tweet* original
in_reply_to_status_id_str	String	Identificador del *tweet* original
in_reply_to_user_id	Int64	Identificador del autor del *tweet* original
in_reply_to_user_id_str	String	Identificador del autor del *tweet* original
in_reply_to_screen_name	String	Nombre del autor del *tweet* original
user	User object	Lista de atributos del autor del *tweet*

[19] *Twitter streaming API*. [en línia] Twitter Developer Documentation. https://dev.twitter.com/streaming/overview

[20] *Field Guide*. [online] Twitter Developer Documentation. https://dev.twitter.com/overview/api/tweets

coordinates	Coordinates	Coordenadas de longitud y latitud
place	Places	Información geográfica
quoted_status_id	Int64	Identificador del *tweet* citado
quoted_status_id_str	String	Identificador del *tweet* citado
is_quote_status	Boolean	Indica si un *tweet* cita a otro *tweet*
quoted_status	Tweet	Contiene el objeto *tweet* original
retweeted_status	Tweet	Información relativa al *tweet* original
quote_count	Integer	Número de citas del *tweet*
reply_count	Int	Número de respuestas al *tweet*
retweet_count	Int	Número de r*etweets*
favorite_count	Integer	Número de *likes* del *tweet*
entities	Entities	Metadatos e información contextual adicional
extended_entities	Extended Entities	Metadatos e información contextual adicional
favorited	Boolean	Indica si el *tweet* ha recibido un *like* de su autor
retweeted	Boolean	Indica si el *tweet* ha recibido un *retweet* de su autor
possibly_sensitive	Boolean	Indica si el *tweet* contiene un enlace
lang	String	Idioma del *tweet*

Tabla 3.1. Estructura de un *tweet*

4. Análisis de datos

4.1 Análisis estadístico y descriptivo

Podemos iniciar el análisis con un gráfico indicando el número de apariciones de cada *hashtag* en los *tweets* capturados.

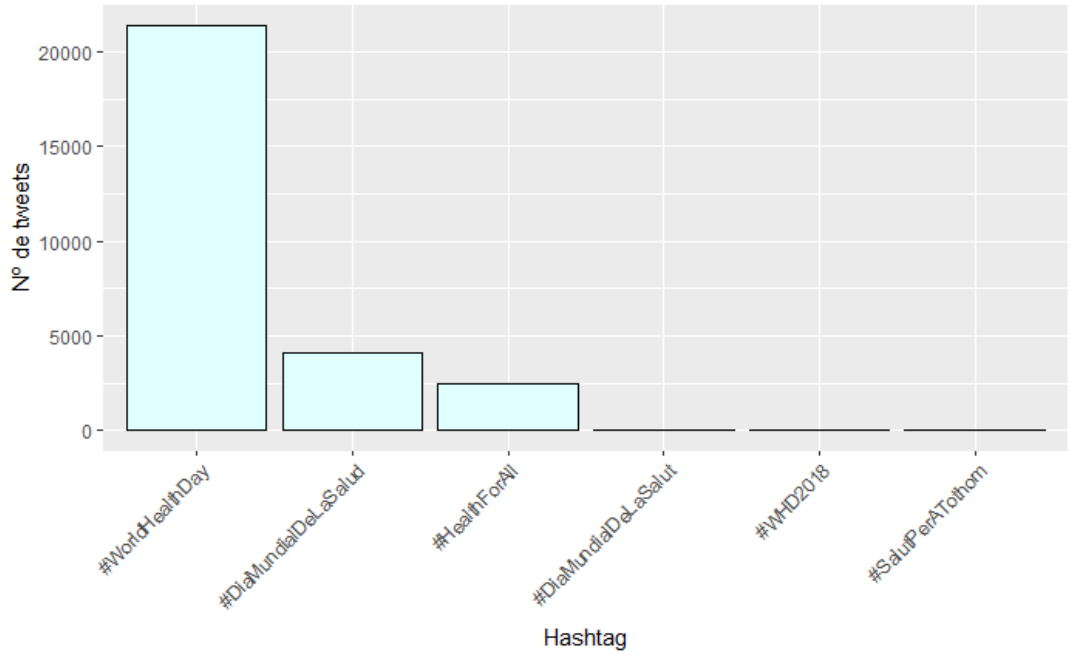

Figura 4.1. Frecuencia de los *hashtags*

El *hashtag* que aparece de forma mayoritaria es #WorldHealthDay, seguido de #DiaMundialDeLaSalud y #HealthForAll. Los demás han resultado ser minoritarios, obteniendo menos de un centenar de apariciones en cada caso.

Podemos también visualizar el número de *tweets* a lo largo de la captura, para conocer en qué instantes se registra más/menos actividad.

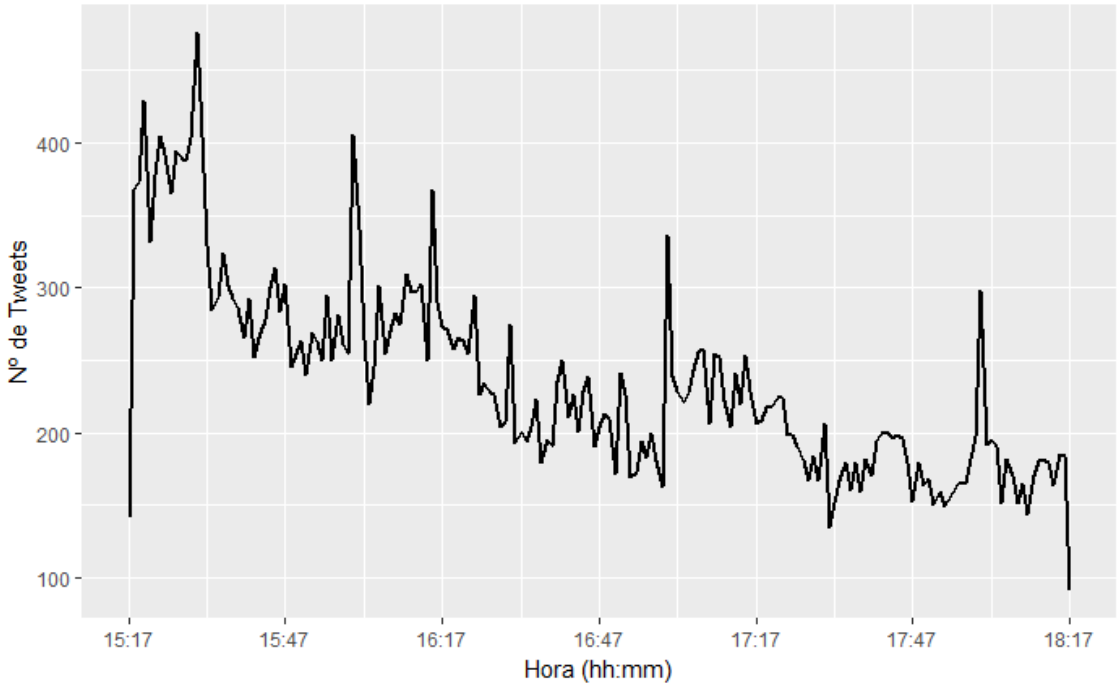

Figura 4.2. Número de *tweets* por minuto

Vemos dos mínimos pronunciados en el inicio y final de la captura. Se contabilizan menos tweets en el comienzo porque la captura empieza exactamente a las 15:17:37, de forma que no se han recogido los *tweets* generados durante los 36 primeros segundos de ese minuto. Lo mismo sucede con el final: la captura se detiene a las 18:17:35, perdiéndose los *tweets* originados durante el resto de ese minuto.

El momento exacto con más *tweets* se produce a las 15:30, con un máximo de 476. Por el contrario, a las 17:31 únicamente se capturan 135 *tweets* (sin tener en cuenta el instante final que hemos comentado en el párrafo anterior). Podemos observar una tendencia descendente en el número de *tweets* escritos a medida que pasan las horas. Es posible que la mayor actividad se produzca durante las primeras horas del día, y ésta se vaya reduciendo a lo largo de la jornada.

En cuanto al idioma de los *tweets*, los más utilizados han sido los siguientes:

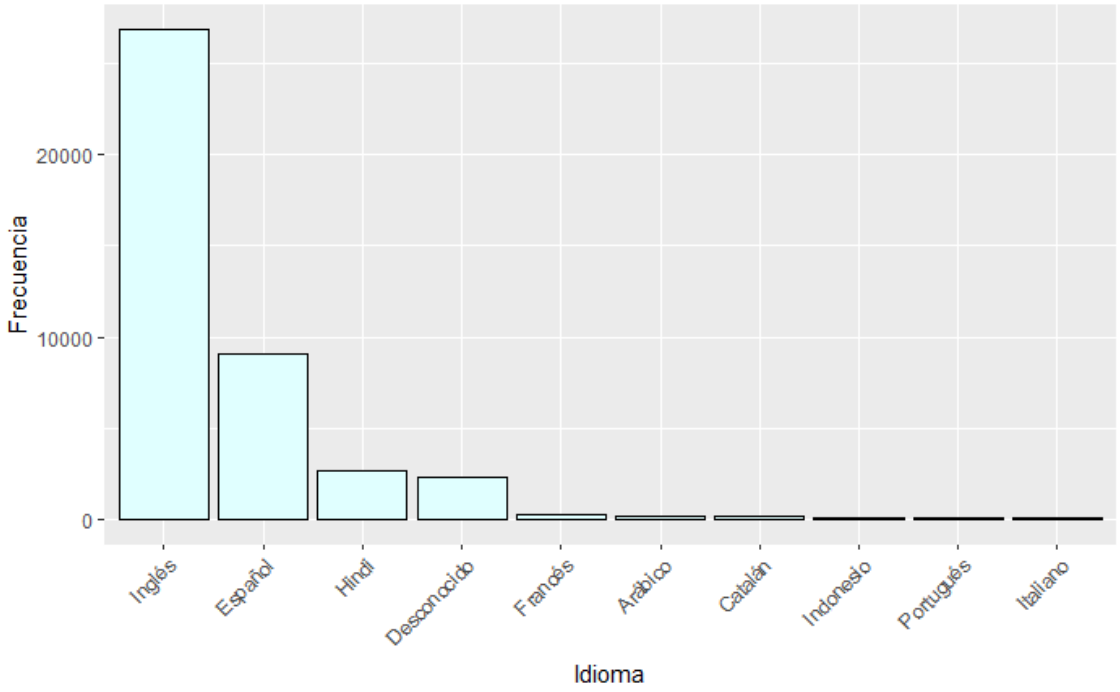

Figura 4.3. Número de *tweets* según el idioma

Como es de esperar, el inglés es el lenguaje que más predomina en los *tweets* capturados, seguido del español. El hecho de haber utilizado palabras clave en estos idiomas ha facilitado la obtención de un gran número de *tweets*, con excepción quizás del catalán.

Vamos a seguir con un análisis de la longitud de los *tweets*, contabilizando tanto el número de caracteres como el número de palabras en función del idioma.

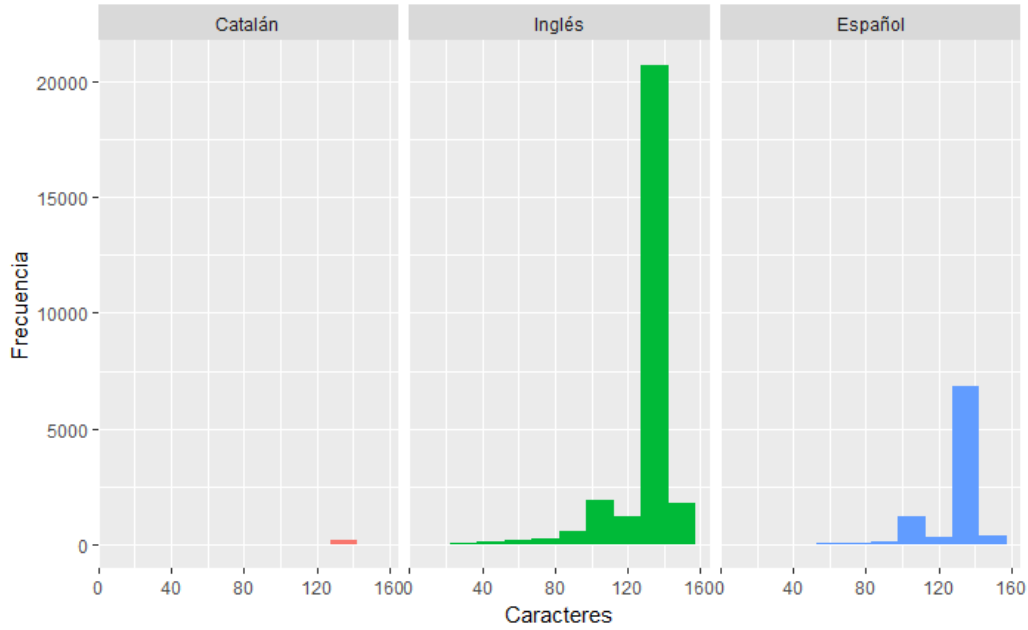

Figura 4.4. Histogramas del número de caracteres en función del idioma

Puede verse de forma muy visual como el número de *tweets* capturados en catalán es muy reducido. Parece que la distribución se concentra mayormente entre los 120 y 160 caracteres. Hasta un total de 22.519 *tweets* tienen una longitud de 140 caracteres, por eso obtenemos tanta concentración en esa franja.

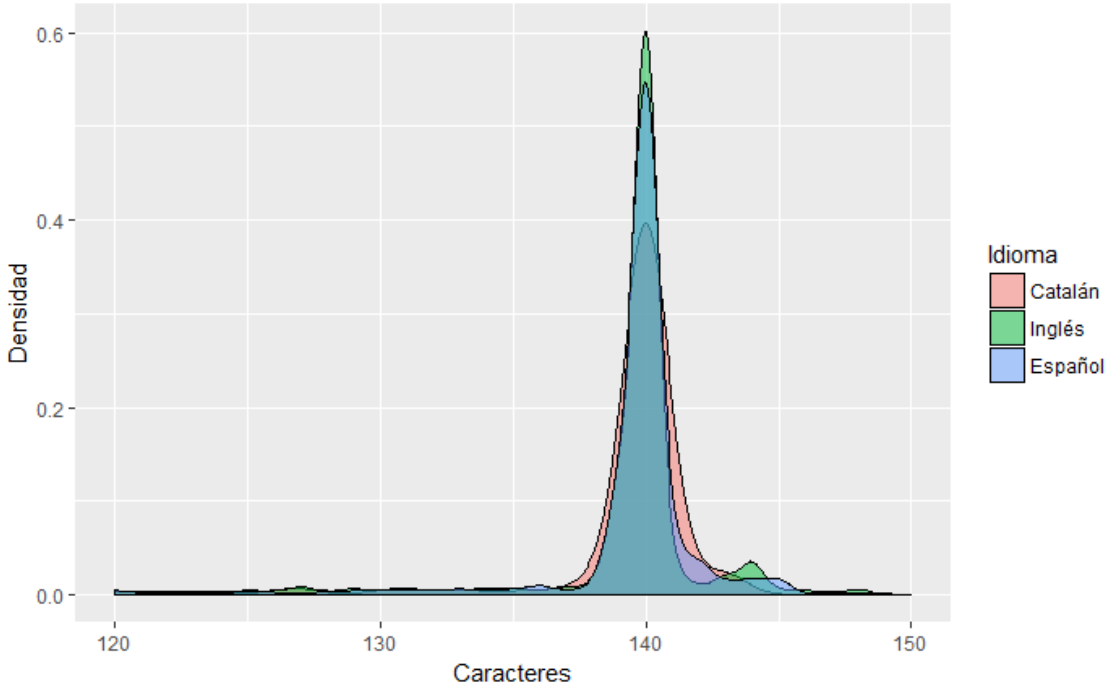

Figura 4.5. Gráfico de densidad del número de caracteres en función del idioma

Con respecto a la distribución del número de palabras, obtenemos lo siguiente:

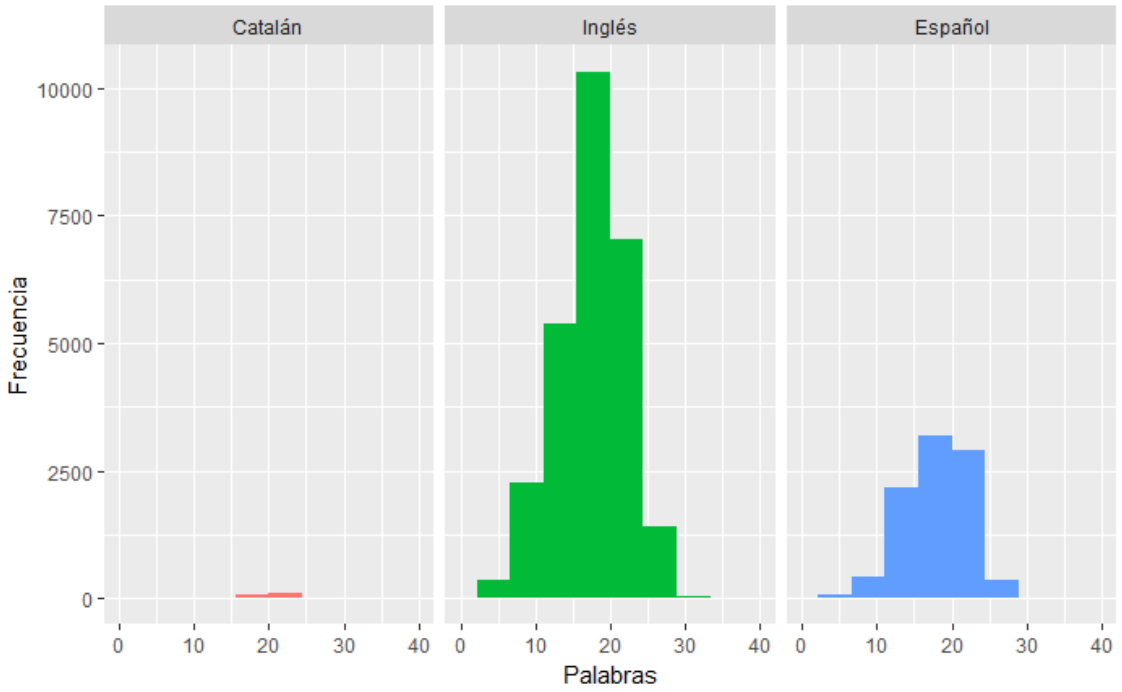

Figura 4.6. Histogramas del número de palabras en función del idioma

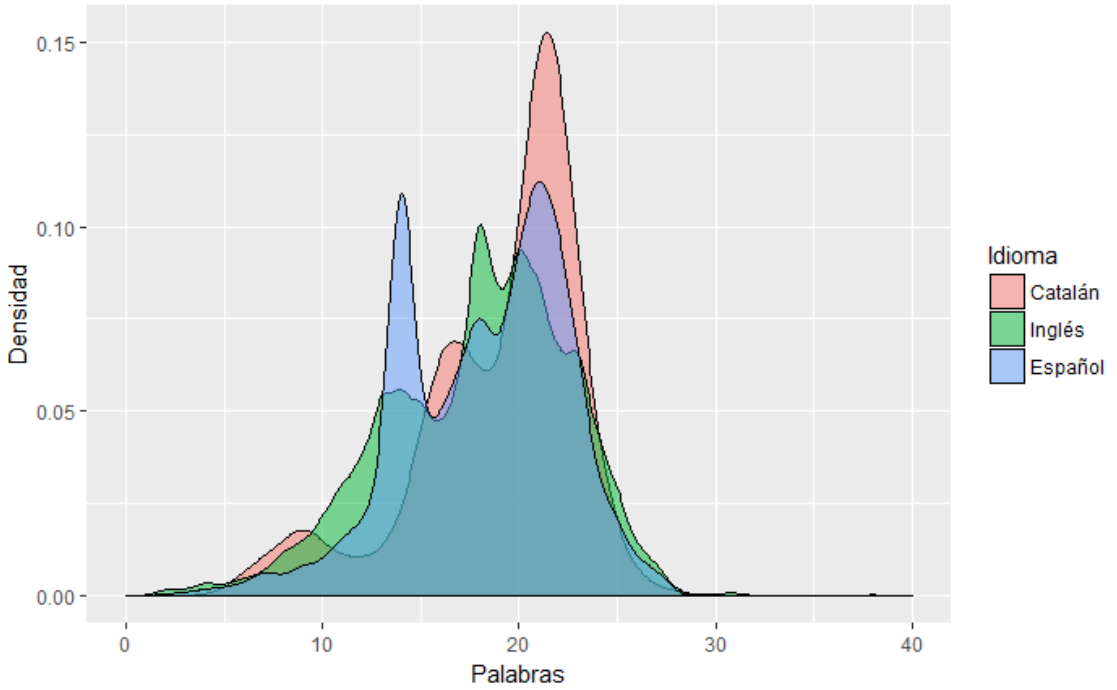

Figura 4.7. Gráfico de densidad del número de palabras en función del idioma

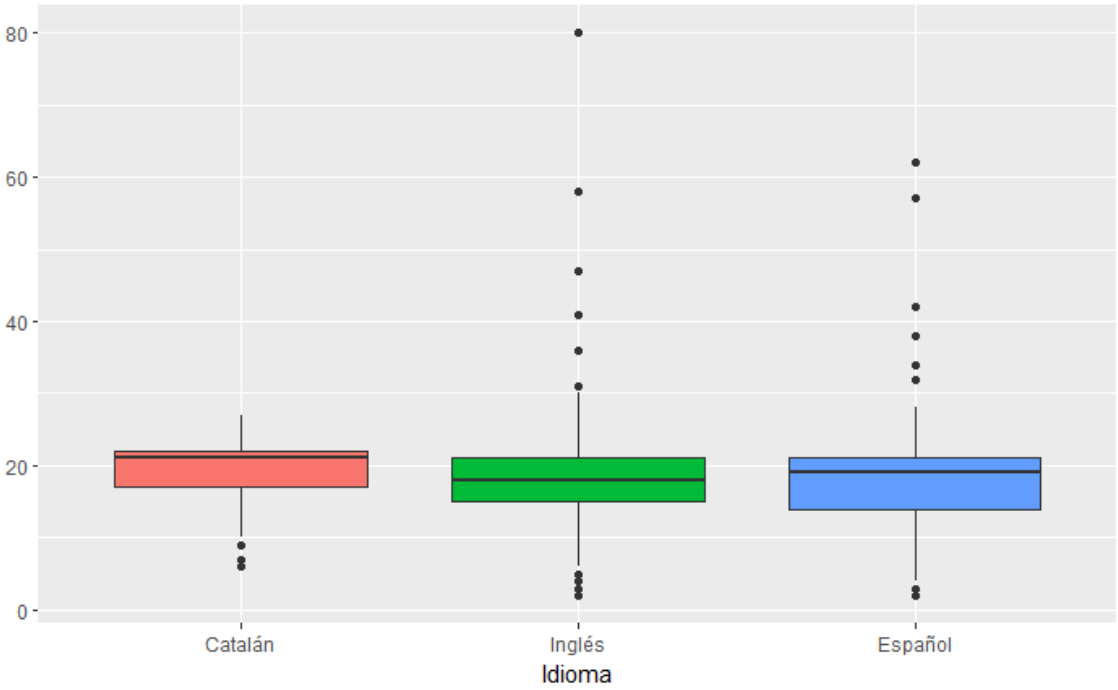

Figura 4.8. *Boxplots* del número de palabras en función del idioma

En el caso de los caracteres no hemos elaborado el gráfico de los *boxplots* porque las cajas son muy estrechas y la visualización es muy mala.

Podemos también visualizar las distribuciones de los siguientes atributos numéricos de los autores de los *tweets* capturados, localizables dentro del campo *user*:

- *friends_count*. Número de usuarios que el autor del *tweet* sigue.
- *followers_count*. Número de usuarios que siguen al autor del *tweet*.
- *favourites_count*. Número de *likes* que ha dado el autor del *tweet*.
- *statuses_count*. Número de *tweets* del autor del *tweet* (incluyendo *retweets*).

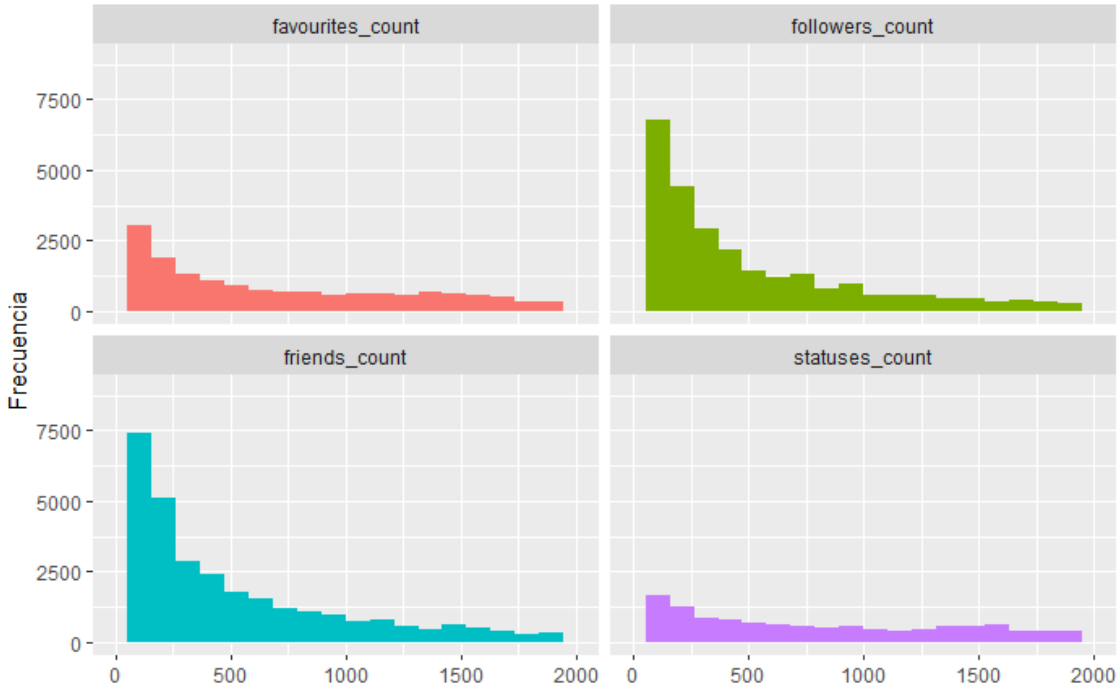

Figura 4.9. Histogramas de algunos atributos de los usuarios

Vemos como la tendencia con los cuatro atributos es parecida, aunque en algunos casos es más pronunciada que en otros. Para valores bajos, la concentración de usuarios es muy alta. Por ejemplo, la gran mayoría de usuarios tienen un número de seguidores comprendido entre 0 y 500. En el caso del número de *tweets* y favoritos, las distribuciones están más esparcidas. Eso significa que a la hora de escribir *tweets* (y hacer *retweets*) podemos encontrar usuarios de todo tipo: que escriben poco, bastante, mucho, etc. Lo mismo sucede con los favoritos.

Podemos visualizar la relación existente entre el número de seguidores y el número de amigos de los usuarios.

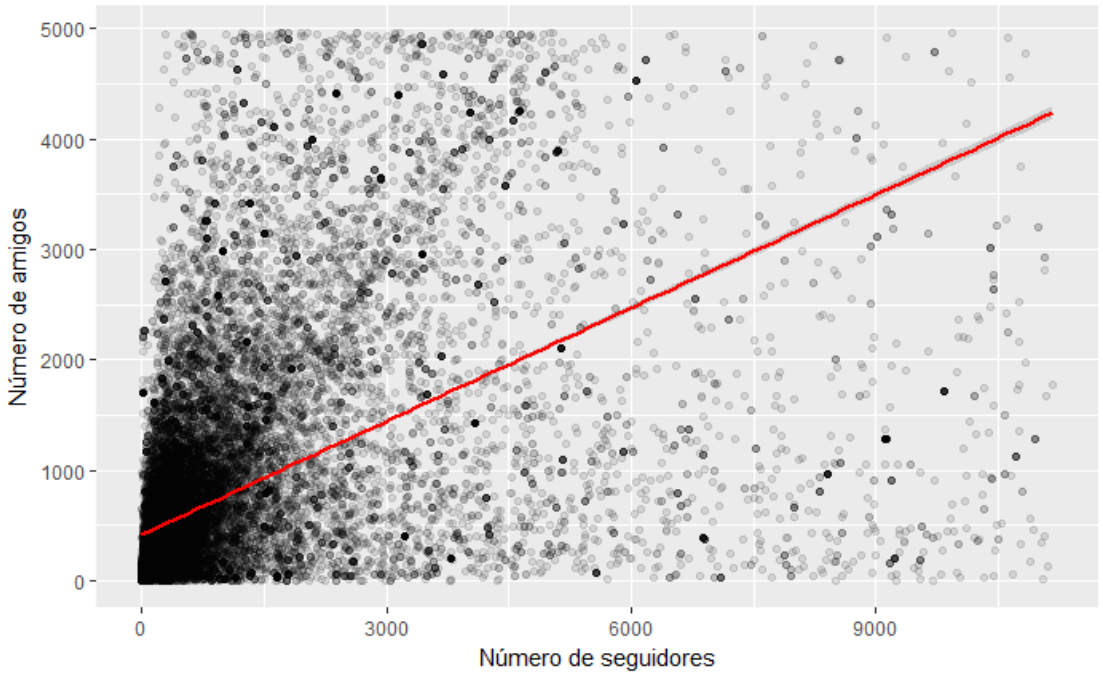

Figura 4.10. Correlación entre número de amigos y seguidores

Como vemos, la mayor parte de los puntos se encuentran concentrados en la esquina inferior izquierda. Es decir, hay una minoría de usuarios con una gran cantidad de seguidores y/o amigos, mientras que la inmensa mayoría tienen muy pocos.

Para acabar con este apartado, vamos a visualizar las palabras más frecuentes dentro del campo de descripción del perfil de Twitter de los usuarios, de forma que podamos extraer el máximo conocimiento sobre ellos: intereses, aficiones, inquietudes, etc.

Figura 4.11. *Wordcloud* de la descripción de los usuarios

Los usuarios que han escrito *tweets* durante la captura están comprometidos socialmente y realmente involucrados en cuestiones de salud. Como curiosidades, parece que las aficiones más frecuentes entre los usuarios son la música y viajar. También encontramos bastantes estudiantes y creyentes.

Las 20 palabras más repetidas en la descripción de los perfiles han sido las siguientes:

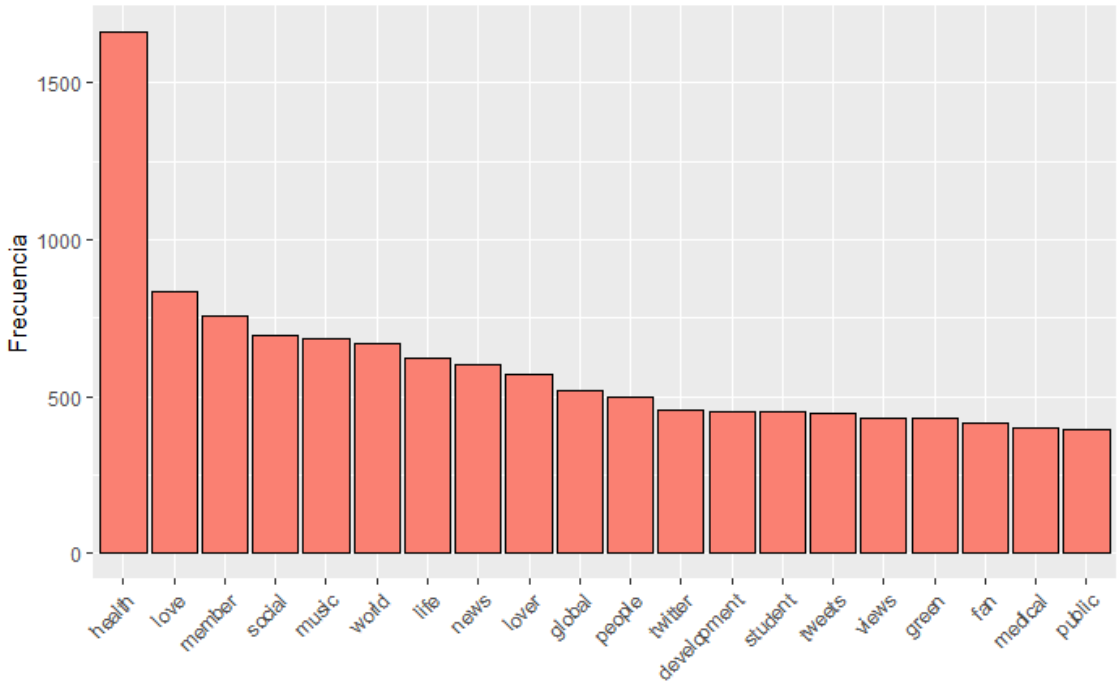

Figura 4.12. Palabras más frecuentes de la descripción de los usuarios

4.2 Palabras más repetidas

Para proseguir nuestro análisis, vamos a descubrir y mostrar cuáles son los términos más comunes que aparecen en los textos de los *tweets* capturados en función del idioma. Para ello utilizaremos una representación visual denominada nube de palabras o *wordcloud*, en inglés. Este tipo de visualización muestra las palabras más frecuentes de un texto. Su principal particularidad es que los términos aparecen con diferentes tamaños, en función de sus frecuencias de aparición. Empezamos con los *tweets* escritos en inglés,

Figura 4.13. *Wordcloud* de los *tweets* en inglés

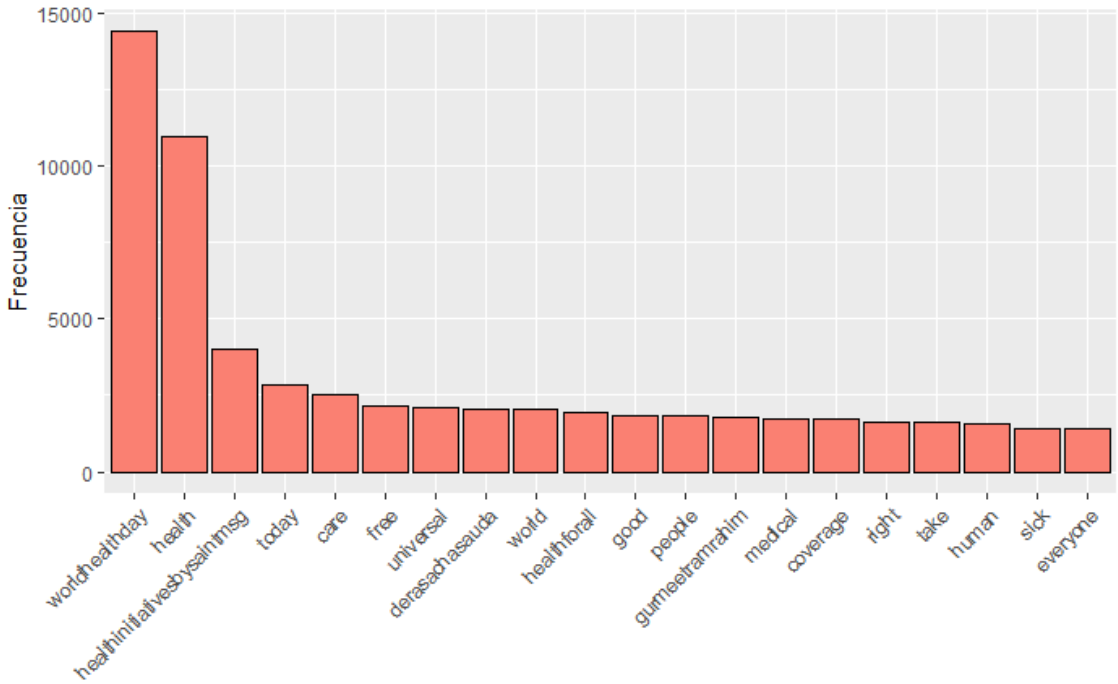

Figura 4.14. Palabras más frecuentes de los *tweets* en inglés

Veamos ahora las palabras más frecuentes en castellano,

Figura 4.15. *Wordcloud* de los *tweets* en castellano

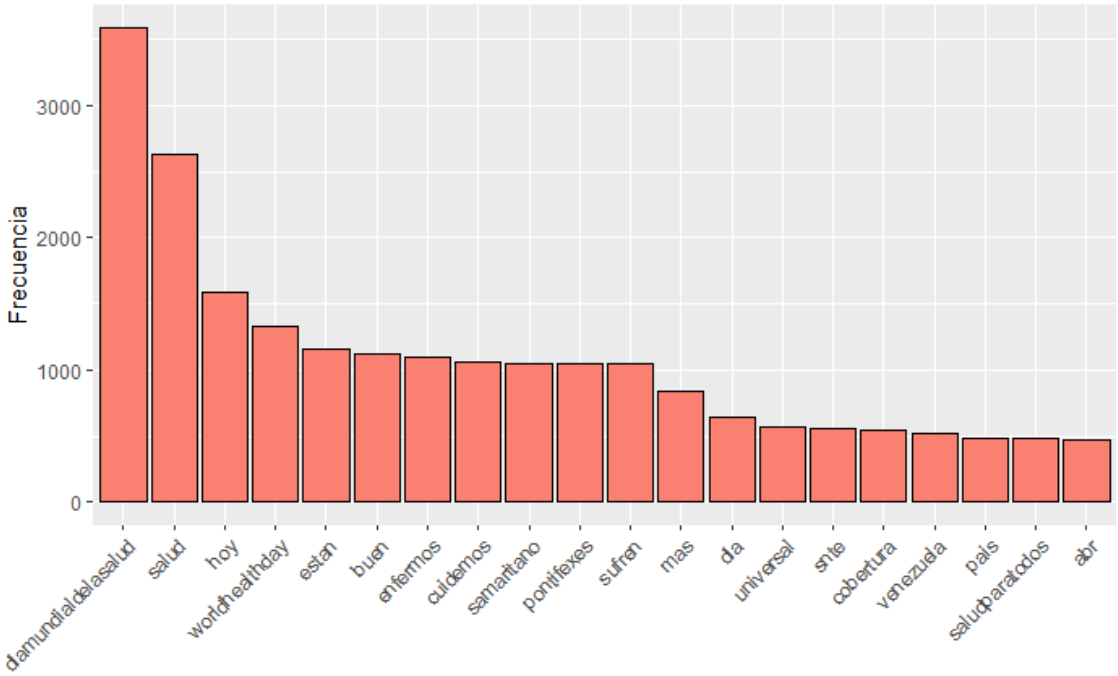

Figura 4.16. Palabras más frecuentes de los *tweets* en castellano

Finalmente repetimos el mismo proceso para los *tweets* en catalán,

Figura 4.17. *Wordcloud* de los *tweets* en catalán

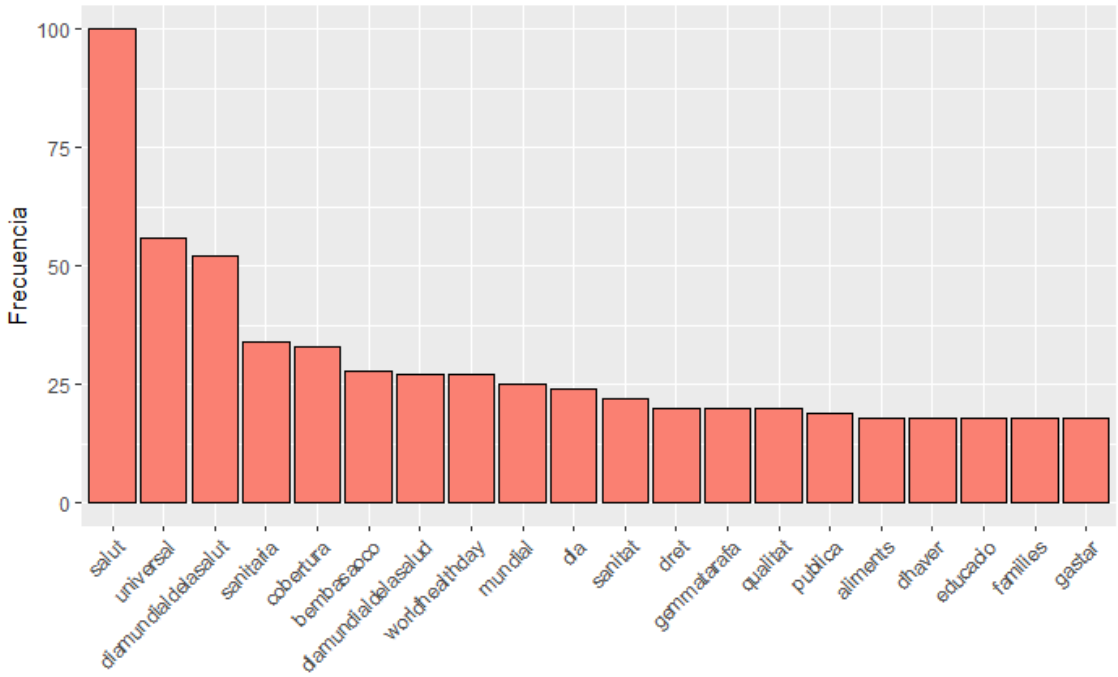

Figura 4.18. Palabras más frecuentes de los *tweets* en catalán

4.3 Análisis de sentimientos

A continuación vamos a clasificar la polaridad de los textos mediante el análisis de sentimiento o también denominado minería de opinión. En el primer gráfico vemos las palabras categorizadas en dos grupos, según su connotación: positivas o negativas (método Bing).

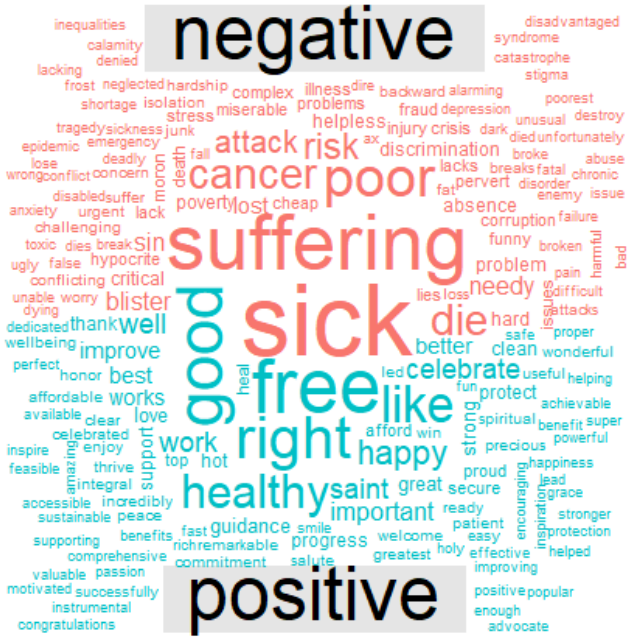

Figura 4.19. Palabras positivas y negativas

Es posible elaborar una clasificación más compleja que sea capaz de categorizar las palabras según sentimientos más concretos (método NRC), tales como: ira, miedo, tristeza, confianza, disgusto, etc. En el siguiente gráfico de barras vemos la frecuencia de cada sentimiento en las palabras de los *tweets* capturados.

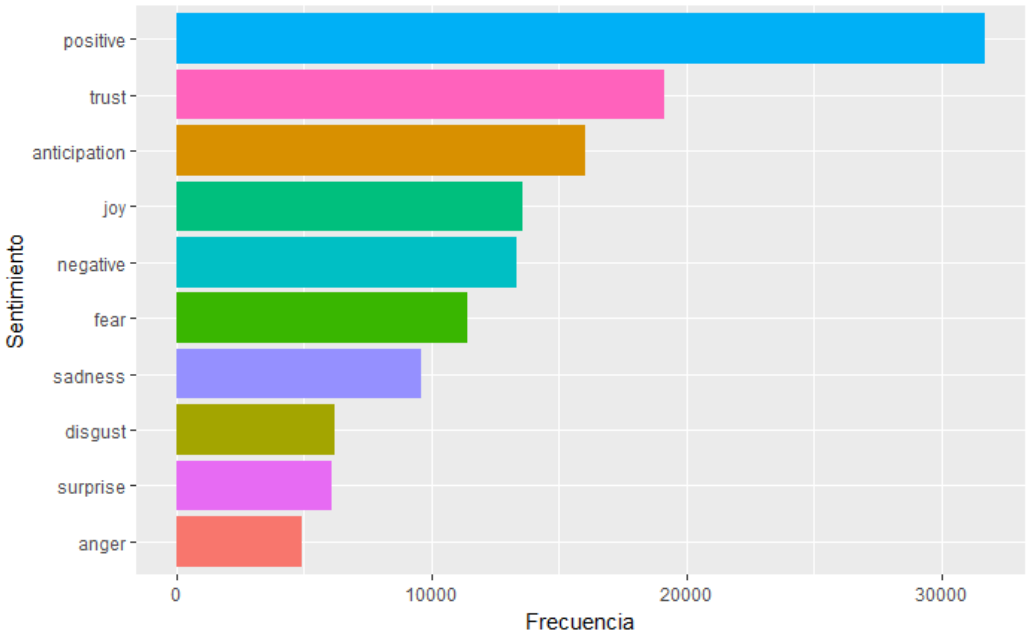

Figura 4.20. Análisis de sentimientos

Las conclusiones que podemos extraer de este gráfico son muy positivas. La mayoría de las palabras que los autores utilizan en sus *tweets* durante el Día Mundial de la Salud están asociadas con sentimientos positivos: confianza, esperanza y alegría. Por otro lado, los sentimientos más tóxicos aparecen con mucha menos frecuencia. Me parece un éxito que el sentimiento de enfado haya aparecido en último lugar, no teniendo cabida en un día como este. En definitiva, parece que el Día Mundial de la Salud genera conciencia en los usuarios de Twitter, quienes esperan y confían que algún día todas las personas, en cualquier lugar, puedan tener acceso a servicios de salud esenciales y de calidad sin tener que pasar apuros económicos.

Podemos consultar también las palabras más repetidas para cada sentimiento:

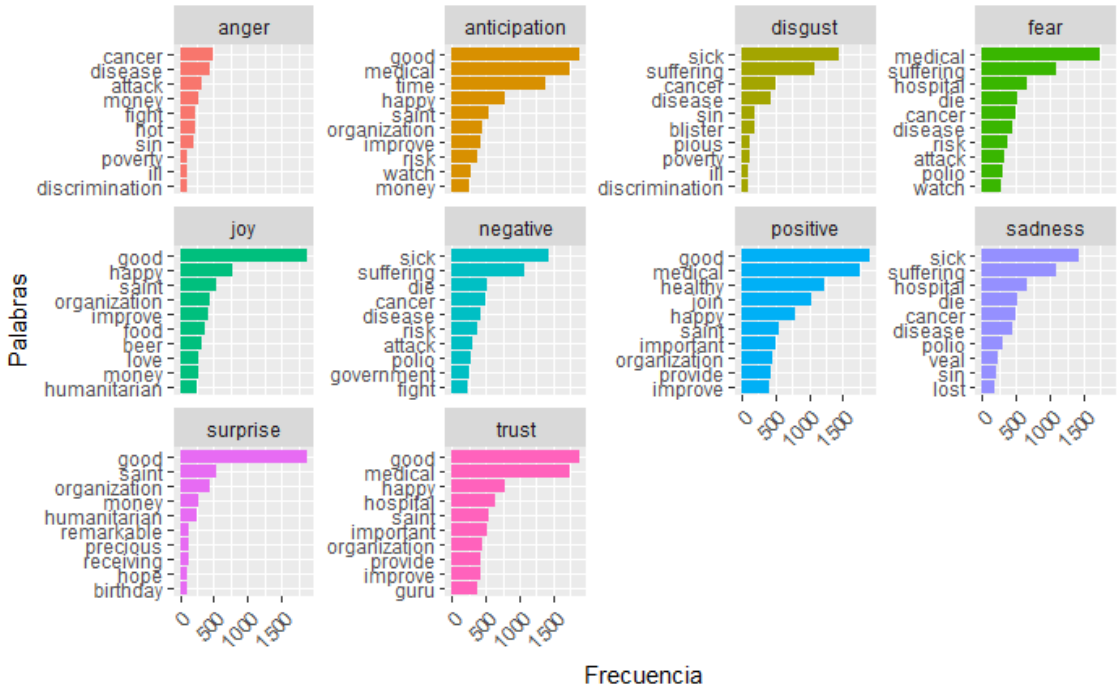

Figura 4.21. Palabras más repetidas para cada sentimiento

Por último, es posible clasificar las palabras según una escala numérica entre -5 y 5 (método AFINN). Cuanto mayor es la puntuación, más positiva es la palabra. A continuación podemos ver las palabras más positivas y negativas de los *tweets* extraídos según el sistema AFINN (independientemente de su frecuencia de aparición).

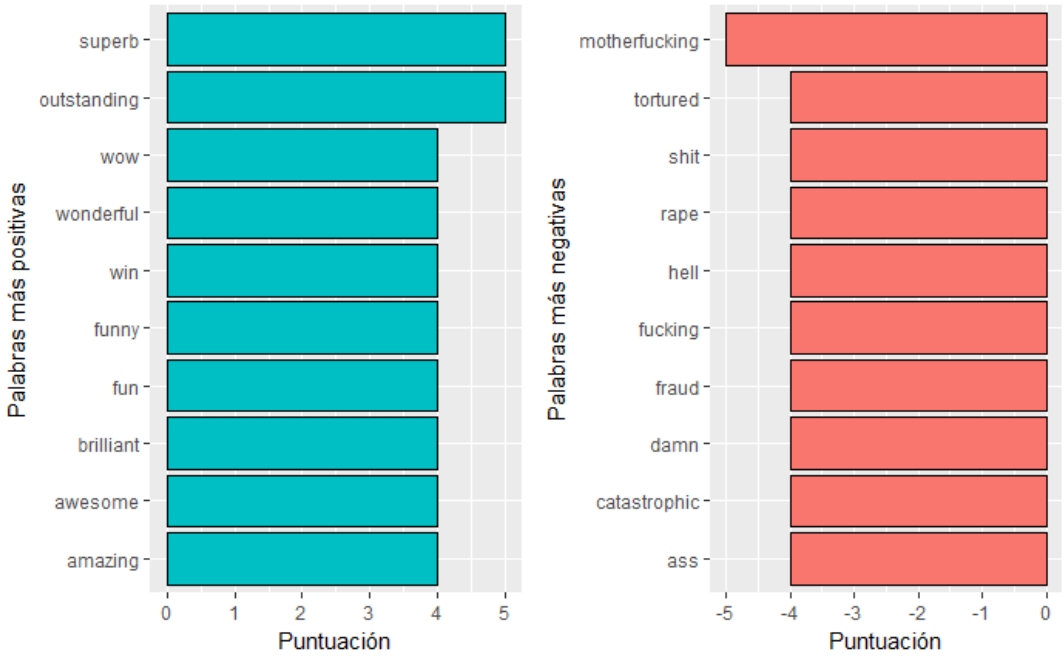

Figura 4.22. Palabras más positivas y negativas

Si también tenemos en cuenta la frecuencia de aparición, las palabras más relevantes del análisis son las siguientes:

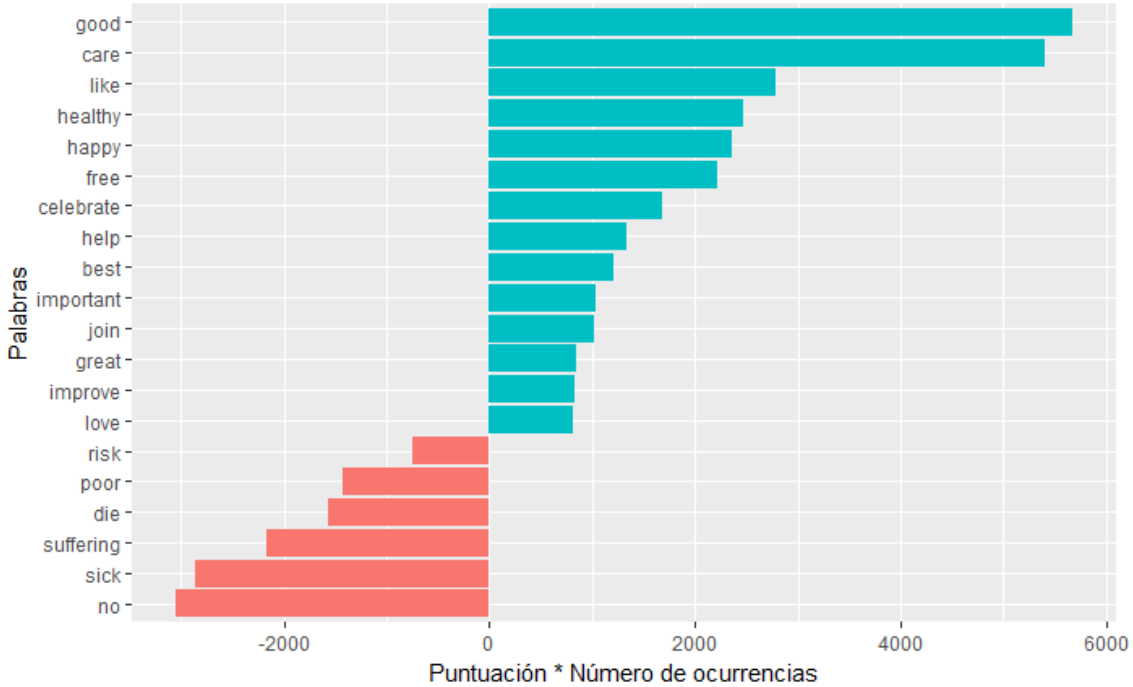

Figura 4.23. Palabras más relevantes del análisis de sentimiento

Para calcular la contribución de cada palabra en este último gráfico, hemos multiplicado su puntuación asociada (entre -5 y 5) por el número de ocurrencias de la palabra. Finalmente nos hemos quedado con las que tienen una mayor contribución, en valor absoluto.

Para acabar con la sección del análisis de sentimiento, podemos estudiar la frecuencia de cada sentimiento para cada minuto durante la captura. De esta forma podemos analizar en qué instantes se registra una mayor actividad de un determinado sentimiento, e intentar examinar el porqué. Hemos utilizado un gradiente para representar la frecuencia de cada sentimiento: cuanto más intenso es el rojo, más frecuencia.

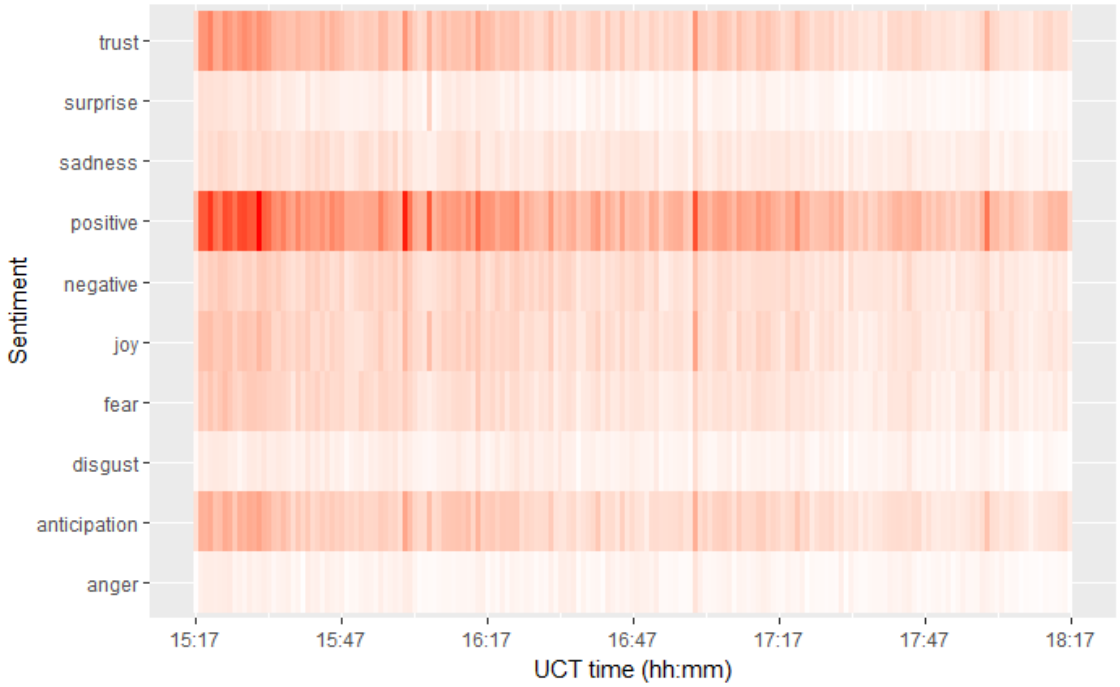

Figura 4.24. Frecuencia de cada sentimiento a lo largo de la captura

Vemos como las palabras asociadas a sentimientos positivos son las más fácilmente visibles, sobre todo durante la primera media hora de la captura. Si comparamos con uno de los primeros gráficos que hemos elaborado (Figura 2. Número de *tweets* por minuto), las zonas más rojizas coinciden con los momentos de mayor actividad en cuanto a la publicación de *tweets* (17:32, 16:02 y 17:02 aproximadamente).

También es curioso ver como la evolución de las frecuencias a lo largo de la captura es prácticamente la misma, aunque en algunos sentimientos se evidencia más que en otros. Si estudiamos la correlación entre estas variables, podemos ver como están fuertemente relacionadas:

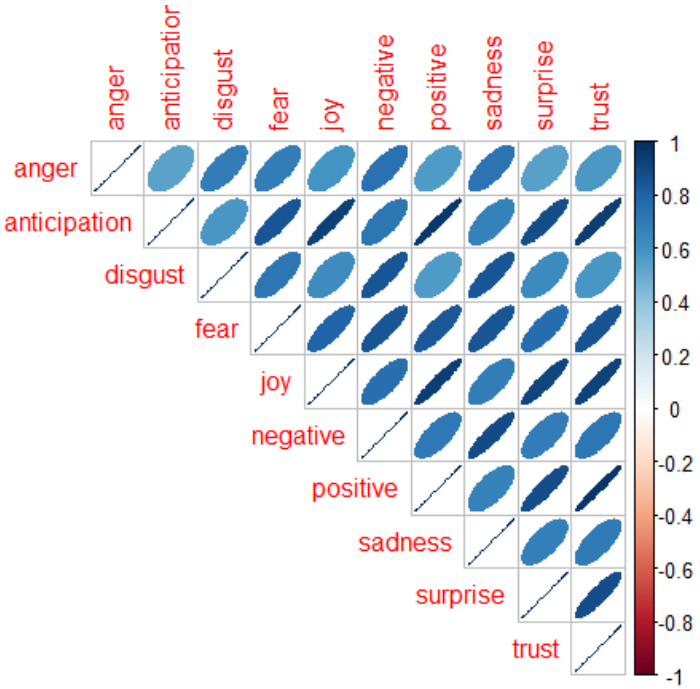

Figura 4.25. Correlaciones entre frecuencias de cada sentimiento

Esto puede ser porque una misma palabra puede estar asociada con múltiples sentimientos, provocando que las frecuencias de varios sentimientos crezcan a la vez. Además, cuantos más *tweets* se publican por minuto, más aumentan (en mayor o menor medida) las frecuencias de todos los sentimientos.

4.4 Información geográfica

Empezamos con un gráfico de las zonas horarias desde donde más escriben los usuarios:

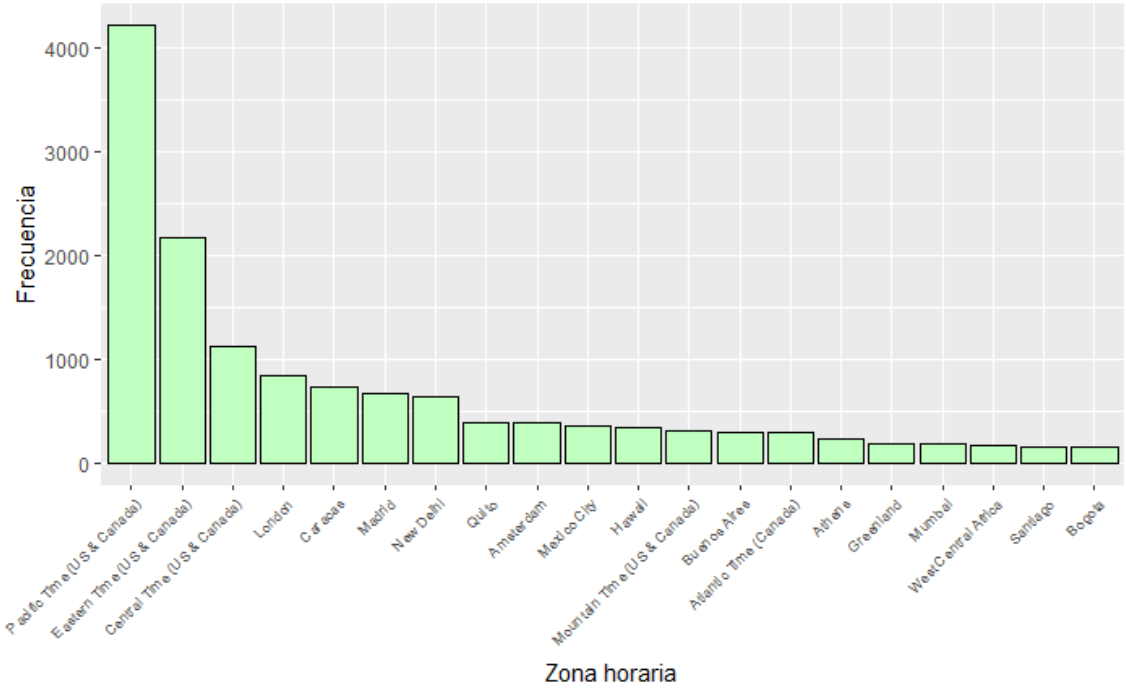

Figura 4.26. Zonas horarias más frecuentes

También podemos situar los *tweets* capturados en un mapa. No podemos hacerlo mediante el campo *coordinates* de los *tweets* capturados, puesto que todos sus valores son NULL. En su lugar, vamos a geolocalizar el campo *location* de los usuarios. El resultado es el siguiente:

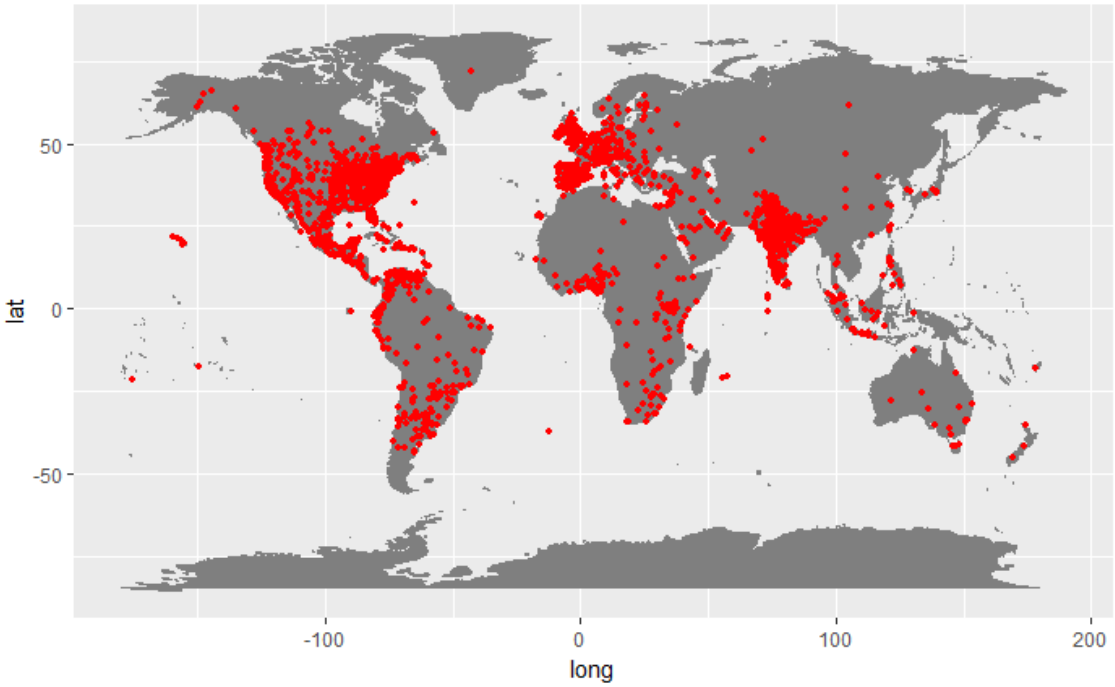

Figura 4.27. *Tweets* geolocalizados

No todos los *tweets* son geolocalizables, por eso no hay tantos puntos rojos como *tweets* capturados. El siguiente enlace http://rpubs.com/xvivancos/tfm contiene un mapa interactivo donde es posible hacer *zoom* para analizar con más detalle las zonas desde donde se han escrito más *tweets*.

4.5 Estructura de la red

En este apartado vamos a analizar la red que se forma con los *tweets* y *retweets* de la captura. Es decir, el grafo resultante tendrá como nodos los *tweets* implicados y como arcos la existencia de un *retweet* entre dos *tweets*. De esta forma podremos identificar los usuarios y *tweets* más relevantes. Para su elaboración vamos a utilizar el *software* Gephi, puesto que para una red tan grande no creo que pueda visualizarse correctamente en R.

En la red tenemos etiquetados los cinco nodos con más grados de entrada con el nombre del usuario cuyos *tweets* han tenido más *retweets*.

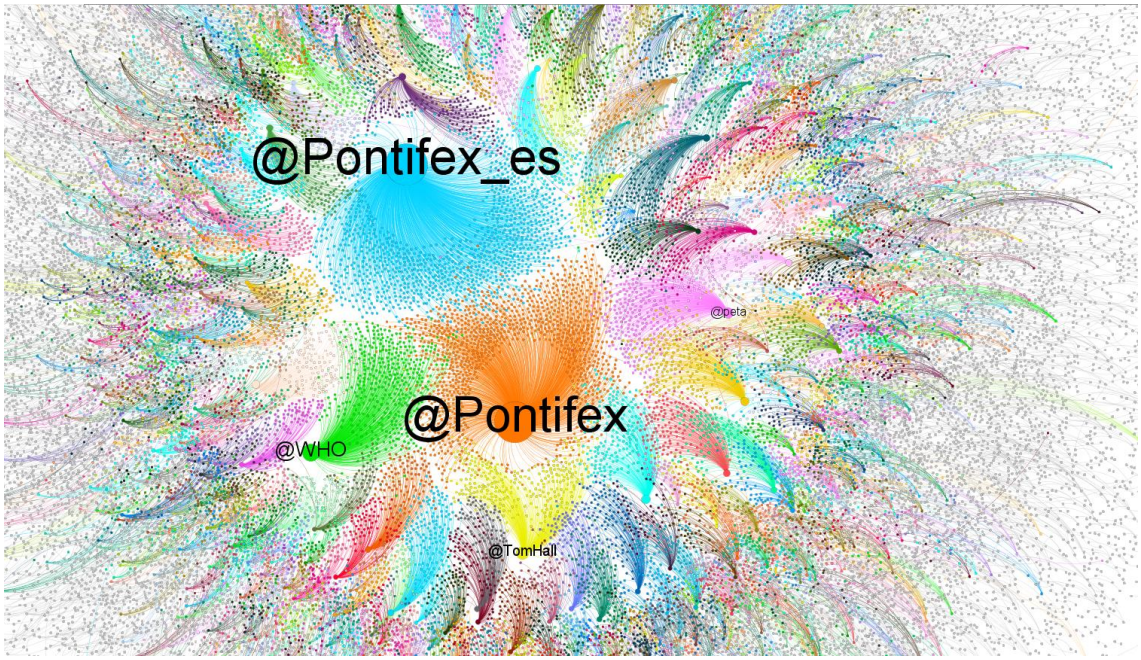

Figura 4.28. Grafo de *tweets* y *retweets*

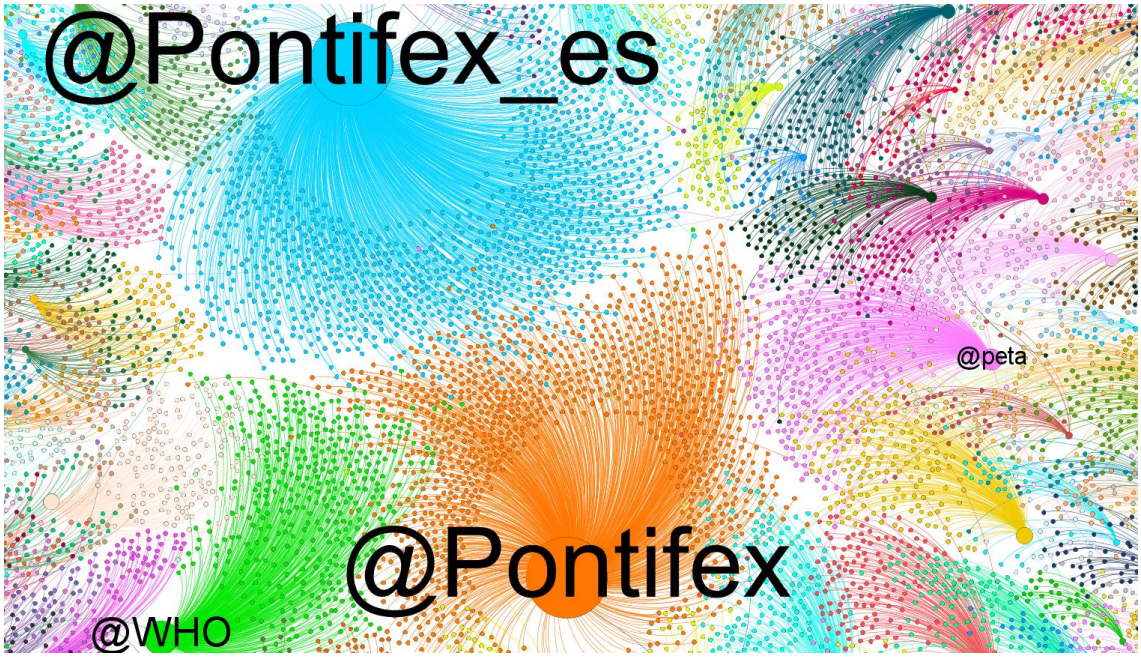

Figura 4.29. Grafo de *tweets* y *retweets* (*zoom*)

Podemos caracterizar el grafo con una serie de propiedades y métricas:

- El grafo tiene 48.024 nodos y 33.905 arcos.
- El grafo es dirigido, pero no bipartito.
- Se trata de un grafo disperso, al obtener una densidad de arcos (*edge density*) cercana a cero: 1.45e-05.
- Tanto la transitividad como la reciprocidad es 0.

La distribución de los grados de entrada es la siguiente:

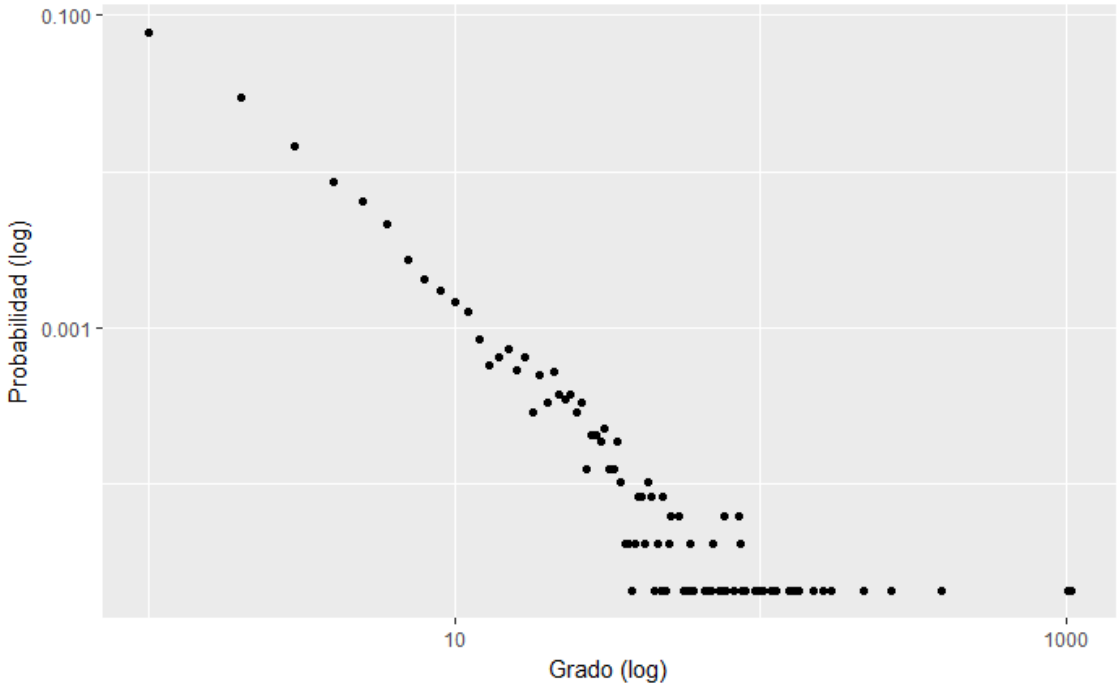

Figura 4.30. Distribución de los grados de entrada

En la figura podemos ver que existe una gran cantidad de nodos con grados relativamente bajos, mientras que sólo existen unos pocos usuarios con un grado muy superior a los demás. Estos nodos son llamados *hubs*, ya que concentran una gran cantidad de relaciones a su alrededor. Podemos ver como el histograma de grados sigue, más o menos, una línea recta descendiente. Cuando esto ocurre, estamos hablando de una red que cumple con la ley de la potencia.

Veamos el texto del *tweet* más compartido durante la captura, perteneciente al Papa Francisco.

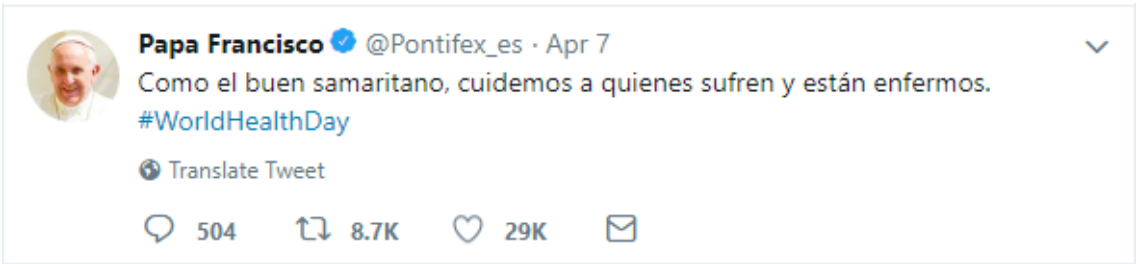

Figura 4.31. *Tweet* más relevante

Los textos de los siguientes cuatro *tweets* con más nodos de entrada a continuación:

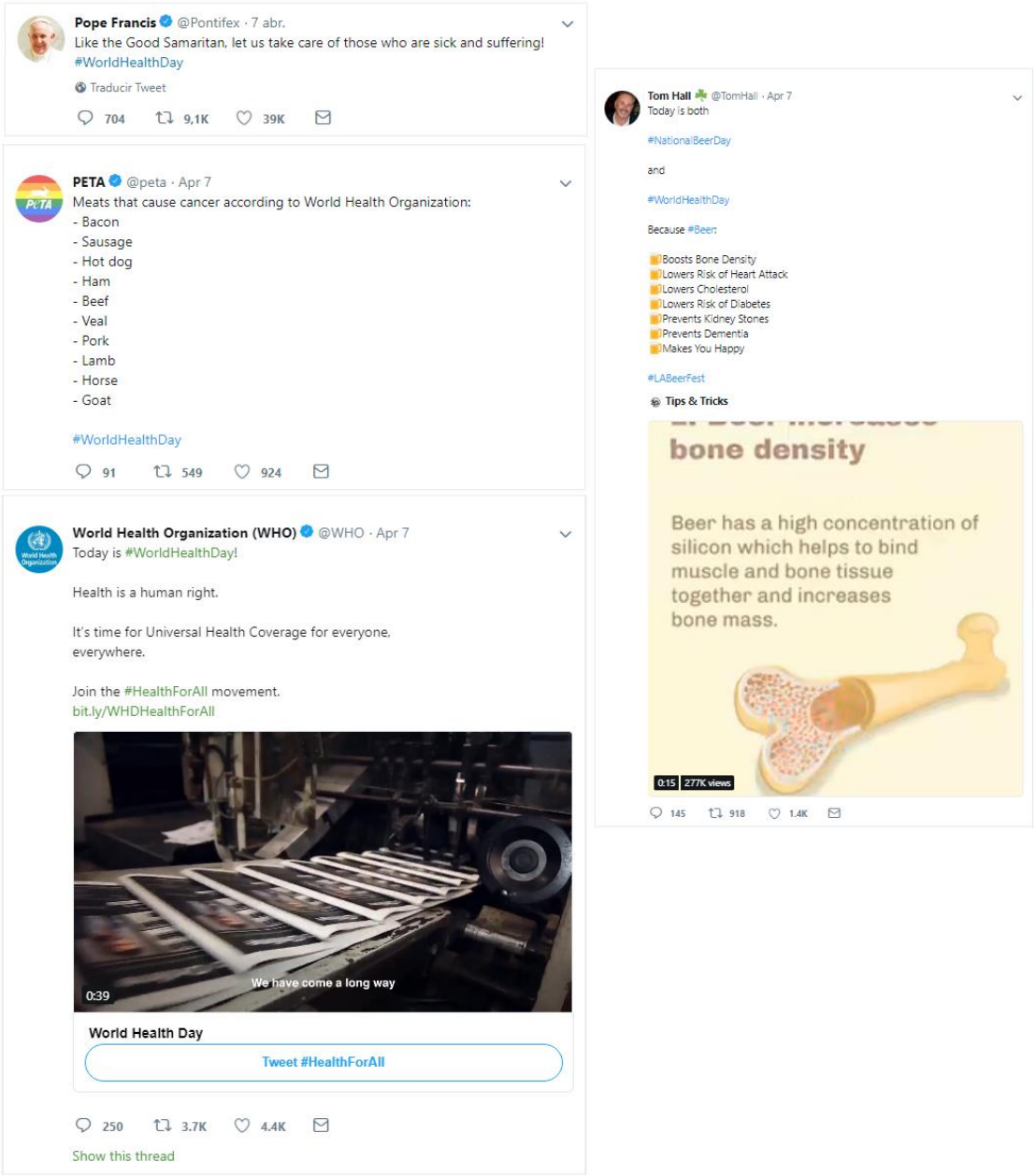

Figura 4.32. *Tweets* más relevantes

Vemos como la religión cobra mucha importancia entre los usuarios que han participado en Twitter durante el Día Mundial de la Salud, puesto que los dos *tweets* más relevantes pertenecen a las cuentas en español e inglés del Papa Francisco. Otros dos corresponden a las organizaciones de PETA (*People for the Ethical Treatment of Animals*) y WHO (*World Health Organization).* El *tweet* restante nos informa de los beneficios para la salud que aporta la cerveza, y que por eso el Día Nacional de la Cerveza 2018 coincide con el Día Mundial de la Salud.

Si repetimos el análisis para los *tweets* capturados en castellano, los más relevantes son los siguientes:

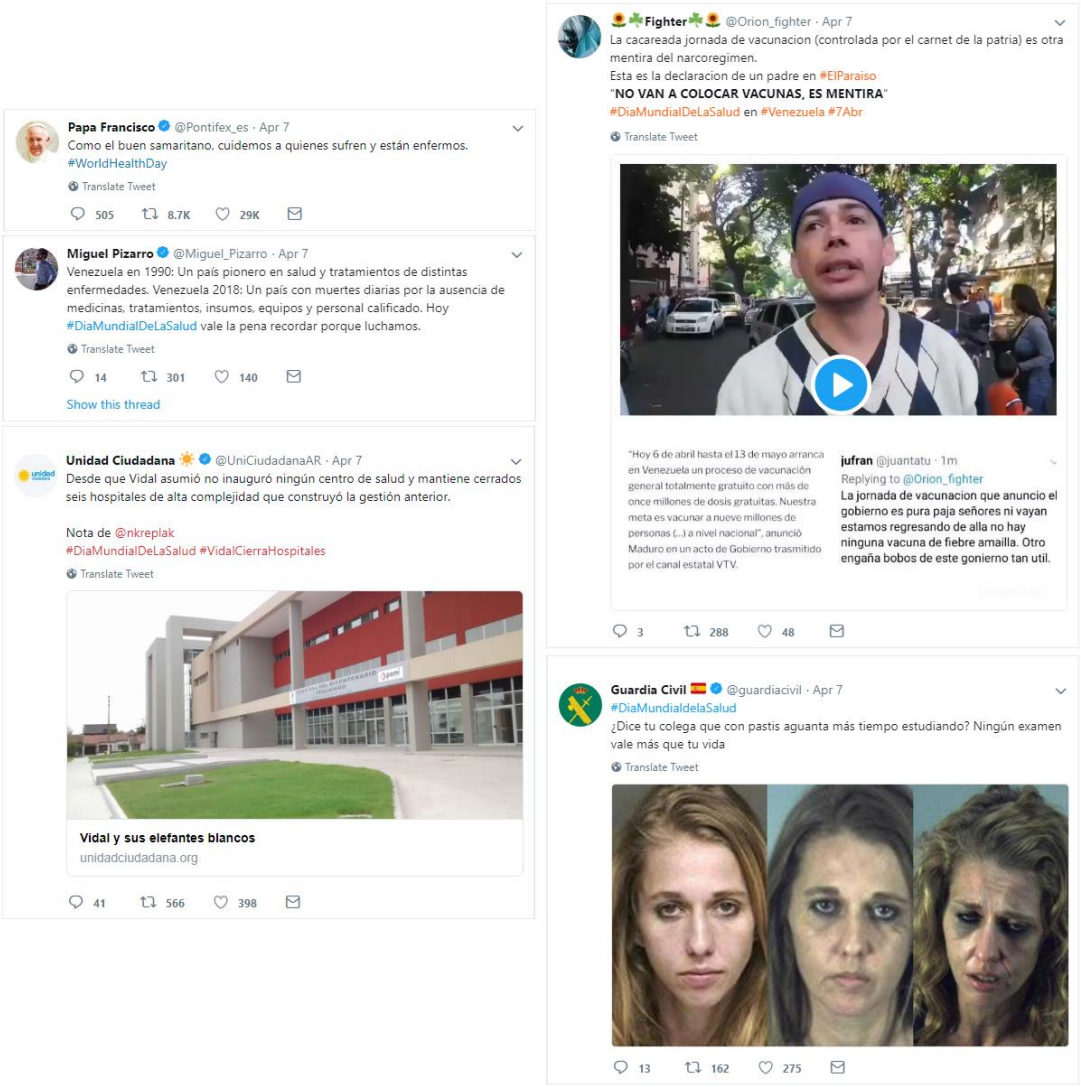

Figura 4.33. *Tweets* en castellano más relevantes

De nuevo, el *tweet* en castellano más relevante es el del Papa Francisco. Sin embargo, vemos como hasta dos *tweets* de los cinco con más *retweets* están directamente relacionados con la crítica situación política en Venezuela y sus consecuencias en el sistema sanitario del país. Los *tweets* denuncian la falta de medicinas, tratamientos, personal calificado, etc. y ponen en duda la jornada de vacunación. Otro *tweet* hace referencia al mandato de María Eugenia Vidal (gobernadora de la provincia de Buenos Aires), indicando que ha sido incapaz de inaugurar ningún centro de salud, manteniendo cerrados seis hospitales que se construyeron en la gestión anterior.

Parece que los *tweets* escritos en castellano tienen un aire más reivindicativo y de crítica hacia los gobernantes de ciertos países, que son incapaces de ofrecer un sistema sanitario de calidad.

Finalmente, los textos de los *tweets* en catalán con más *retweets* son los siguientes:

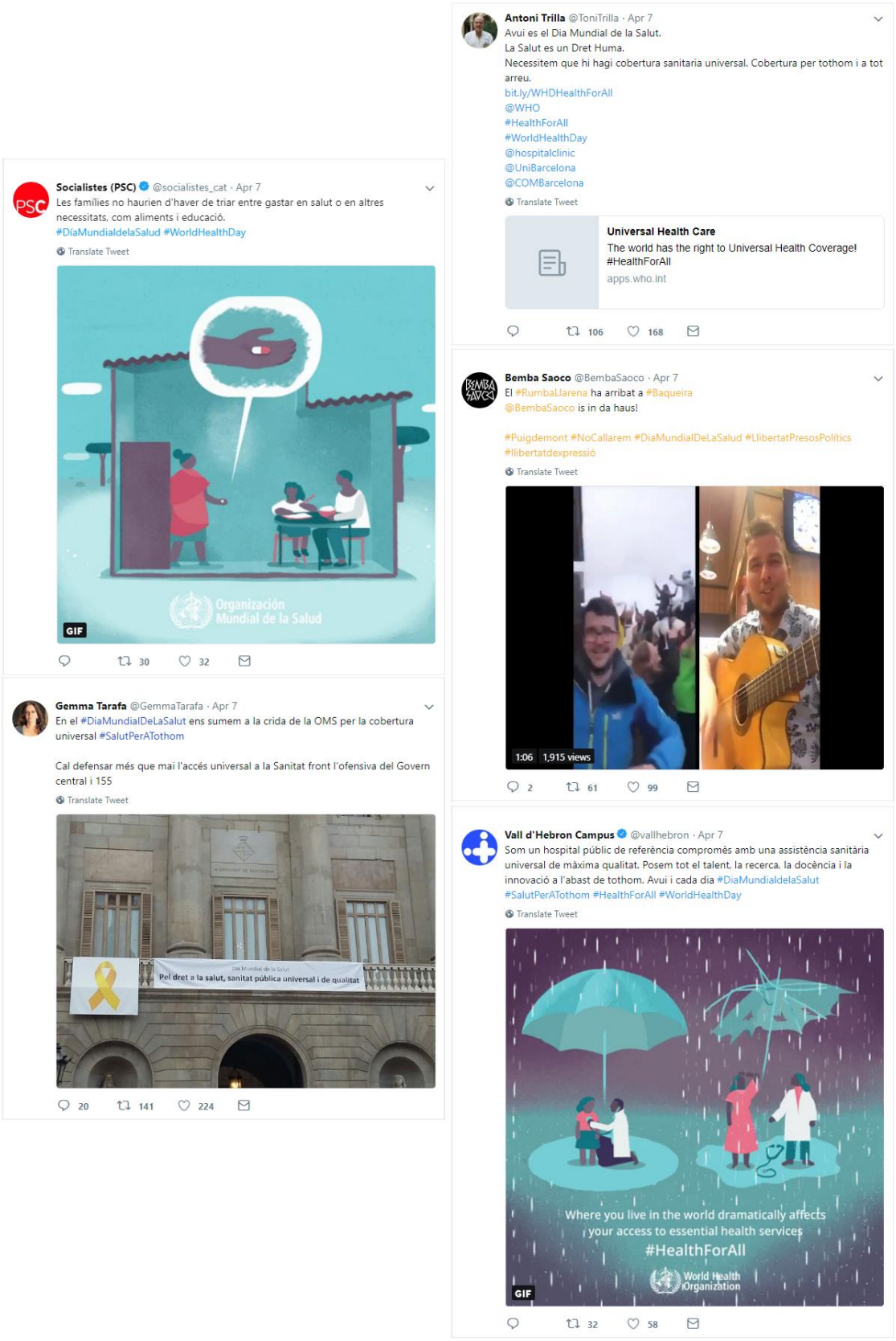

Figura 4.34. *Tweets* en catalán más relevantes

El *tweet* en catalán con más *retweets* corresponde al Partido de los Socialistas de Cataluña (PSC). En general, los *tweets* indican sobre la importancia de disponer de una cobertura sanitaria universal. También pueden verse ligeramente otros temas de la actualidad política de Cataluña, como son la encarcelación de los políticos catalanes y la aplicación del artículo 155.

5. Conclusiones

Una vez finalizado el análisis de los *tweets* capturados durante el Día Mundial de la Salud, es momento de recapitular y ver si se han cumplido los objetivos propuestos, así como extraer las principales conclusiones y lecciones aprendidas durante el desarrollo del proyecto. En cuanto a los objetivos que planteamos al inicio, hemos conseguido cumplir todos:

- ✓ Analizar el contenido de los *tweets*.
- ✓ Determinar la actitud o sentimientos de los autores de los *tweets* en relación con el Día Mundial de la Salud.
- ✓ Estudiar diferencias entre los *tweets* escritos en inglés, castellano y catalán.
- ✓ Dibujar y describir la estructura de red formada entre *tweets* y *retweets*.
- ✓ Determinar en qué países la concienciación en Twitter es mayor/ menor.

Las principales conclusiones que obtenemos a partir del análisis de los *tweets* son las siguientes:

- En cuanto al número de *tweets* por minuto durante la captura, se observa una tendencia descendente a medida que pasan las horas. Es posible que la mayor actividad se produzca durante las primeras horas del día, y ésta se vaya reduciendo a lo largo de la jornada.
- El idioma más utilizado y con mucha diferencia con respecto a los demás es el inglés.
- El número de palabras y caracteres es muy similar para *tweets* escritos en inglés, castellano y catalán. No se encuentran diferencias significativas en los gráficos de densidad y *boxplots*. Lo único que varía son los histogramas, puesto que los *tweets* en inglés son mucho más frecuentes.
- De los usuarios que han participado en la captura, la gran mayoría tiene "pocos" amigos y seguidores en Twitter. En el caso del número de *tweets* y favoritos, las distribuciones no están tan concentradas para valores bajos.
- Los usuarios que han escrito *tweets* durante la captura están comprometidos socialmente y realmente involucrados en cuestiones de salud, y así lo demuestran algunas de las palabras más frecuentes dentro del campo de descripción del perfil de Twitter (*health* ha sido la palabra más repetida en sus descripciones).
- Las palabras más repetidas son parecidas para los tres idiomas, y se observan términos religiosos con bastante frecuencia (pontífice o samaritano, por ejemplo).
- La mayoría de las palabras que los autores utilizan en sus *tweets* durante el Día Mundial de la Salud están asociadas con sentimientos positivos: confianza, esperanza y alegría. Por otro lado, los sentimientos más tóxicos aparecen con mucha menos frecuencia.
- En el mapa con los *tweets* geolocalizados encontramos mucha actividad desde el oeste de Europa, sobre todo en España y Reino Unido, y en menor medida en otros países como Francia, Italia, Alemania, Bélgica,

Holanda, etc. En el continente asiático destaca con diferencia la India (mucha concentración en Nueva Delhi), mientras que Nigeria es el país africano desde donde más *tweets* se han escrito. Pasando a América, se registran muchísimos *tweets* desde Estados Unidos (Nueva York, Washington, Los Angeles, Philadelphia, etc.). En América Central y del Sur destacan México, Venezuela, Argentina, Colombia, Ecuador, entre otros.

- Se confirma la importancia de la religión para muchos usuarios de Twitter en el Día Mundial de la Salud cuando dos *tweets* del Papa Francisco (uno en castellano y el otro en inglés) son los que tienen más *retweets* del análisis.
- Tanto en los *tweets* escritos en castellano como en catalán pueden observarse críticas y reivindicaciones derivadas de la situación política actual. Venezuela y Argentina son protagonistas entre los *tweets* más relevantes en castellano, debido al pobre sistema sanitario y a la gestión de María Eugenia Vidal en Buenos Aires. Por el otro lado, en los *tweets* en catalán aparecen críticas hacia el 155 y los presos políticos.

Finalmente, algunos puntos no se han podido llevar a cabo y sería interesante tenerlos en cuenta para trabajos futuros:

- Análisis de sentimientos para los tres idiomas: inglés, castellano y catalán. La librería tidytext funciona únicamente con el inglés. Hubiera sido muy interesante comparar la contribución de cada sentimiento para diferentes idiomas.
- Análisis de la red más detallado. A parte de los *retweets*, me hubiera gustado conocer los usuarios con más *replies* y menciones, como también utilizar más métricas y propiedades para definir el grafo resultante. También me hubiera gustado visualizar los grafos para los *tweets* en castellano y en catalán, pero por falta de tiempo hice únicamente el grafo incluyendo todos los *tweets*.
- Creo que el proceso de depuración de los textos de los *tweets* puede mejorarse, puesto que en los *wordclouds* aparece algún que otro término raro que podría ser resultado de algo incorrecto durante la fase de limpieza.

6. Bibliografía

[1] *Mensajes del Día Mundial de la Salud 2018*. [en línea] Organización Mundial de la Salud. http://www.who.int/campaigns/world-health-day/2018/key-messages/es/

[2] Colaboradores de Wikipedia. *Análisis de sentimiento* [en línea]. Wikipedia, La enciclopedia libre, 2018 [fecha de consulta: 19 de junio del 2018]. Disponible en https://es.wikipedia.org/w/index.php?title=An%C3%A1lisis_de_sentimiento&oldid=108714922

[3] Silge J, Robinson D (2016). "tidytext: Text Mining and Analysis Using Tidy Data Principles in R." _JOSS_, *1*(3). doi: 10.21105/joss.00037 (URL: http://doi.org/10.21105/joss.00037), <URL: http://dx.doi.org/10.21105/joss.00037>.

[4] Hadley Wickham (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse

[5] Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2018). dplyr: A Grammar of Data Manipulation. R package version 0.7.5. https://CRAN.R-project.org/package=dplyr

[6] Hadley Wickham and Lionel Henry (2018). tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions. R package version 0.8.1. https://CRAN.R-project.org/package=tidyr

[7] H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2009.

[8] Finn Årup Nielsen (2011). AFIIN lexicon. http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010

[9] Bing Liu. Bing lexicon. https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

[10] Mohammad S, Turney P (2010). NRC lexicon. http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

[11] Wikipedia contributors. (2018, May 14). Document-term matrix. In *Wikipedia, The Free Encyclopedia*. Retrieved 19:27, June 22, 2018, from https://en.wikipedia.org/w/index.php?title=Document-term_matrix&oldid=841158065

[12] Ingo Feinerer and Kurt Hornik (2017). tm: Text Mining Package. R package version 0.7-3. https://CRAN.R-project.org/package=tm

[13] Dawei Lang (NA). wordcloud2: Create Word Cloud by htmlWidget. R package version 0.2.0. https://github.com/lchiffon/wordcloud2

[14] D. Kahle and H. Wickham. ggmap: Spatial Visualization with ggplot2. The R Journal, 5(1), 144-161. URL http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf

[15] Csardi G, Nepusz T: The igraph software package for complex network research, InterJournal, Complex Systems 1695. 2006. http://igraph.org

[16] Jeff Gentry and Duncan Temple Lang (2015). ROAuth: R Interface For OAuth. R package version 0.9.6. https://CRAN.R-project.org/package=ROAuth

[17] Pablo Barbera (2014). streamR: Access to Twitter Streaming API vida R. R package version 0.2.1. https://CRAN.R-project.org/package=streamR

[18] Jeff Gentry (2015). twitteR: R Based Twitter Client. R package version 1.1.9. https://CRAN.R-project.org/package=twitteR

[19] *Twitter streaming API*. [en línia] Twitter Developer Documentation. https://dev.twitter.com/streaming/overview

[20] *Field Guide*. [online] Twitter Developer Documentation. https://dev.twitter.com/overview/api/tweets