Text Mining aplicado a una novela de Charles Dickens

Máster Universitario en Modelización y Análisis de Datos Económicos (MUMADE)

Autores: Bermann, M.A. & Pérez, R.S. [Grupo D]

17 abril, 2022

Índice

Resumen

El uso de técnicas analíticas avanzadas para su aplicación a distintos ámbitos está empezando a ser una continua constante. Una de las técnicas que está proliferando en los últimos años es el análisis de discursos, mensajes, publicaciones en redes sociales, etc. Para ello, se utilizan las llamadas técnicas de minería de texto (text mining). Este creciente interés es el que se intentará abordar en este trabajo. Así, en este caso, se va a leer y analizar una famosa novela, pero la única diferencia es que, esta vez, lo hará el programa R por "nosotros". De esta forma, se ha escogido la novela **Oliver Twist**, una obra del exponente de la novela social, Charles Dickens, el cual recoge en dicha obra un libro con una trama y un conjunto de ambientes que, como se verá, se caracterizan por la decadencia, las cuestiones trágicas, el pesimismo, etc. Además, se verán cambios de entorno según el capítulo analizado, con lo que se podrá ver, momento a momento, la evolución del guión gracias a a las técnicas de *text mining* mencionadas.

I. Procesamiento del texto

En este primer capítulo se detalla cómo se ha obtenido y procesado el texto.

En primer lugar se ha **descargado e importado**, al entorno de trabajo, la novela, a través del Project Gutenberg, donde *Oliver Twist* tiene la numeración 730.

Pueden verse las primeras líneas de la obra extraída a continuación:

```
# A tibble: 6 x 2
  gutenberg_id text
         <int> <chr>
1          730 "Oliver Twist"
2          730 ""
3          730 "OR"
4          730 "THE PARISH BOY'S PROGRESS"
5          730 ""
6          730 "by Charles Dickens"
```

En segundo lugar, se ha ajustado el marco de análisis de la novela a partir del inicio del Capítulo I.

```
# A tibble: 6 x 1
  texto
  <chr>
1 "CHAPTER I. TREATS OF THE PLACE WHERE OLIVER TWIST WAS BORN AND OF THE CIRCUM~
2 ""
3 "Among other public buildings in a certain town, which for many reasons it wi~
4 "For a long time after it was ushered into this world of sorrow and trouble, ~
5 "Although I am not disposed to maintain that the being born in a workhouse, i~
6 "As Oliver gave this first proof of the free and proper action of his lungs, ~
```

En tercer lugar, se han eliminado las filas en blanco, así como los cambios de renglones vacíos.

```
# A tibble: 6 x 1
  texto
  <chr>
1 "CHAPTER I. TREATS OF THE PLACE WHERE OLIVER TWIST WAS BORN AND OF THE CIRCUM~
2 ""
3 "Among other public buildings in a certain town, which for many reasons it wi~
4 "For a long time after it was ushered into this world of sorrow and trouble, ~
5 "Although I am not disposed to maintain that the being born in a workhouse, i~
6 "As Oliver gave this first proof of the free and proper action of his lungs, ~
```

```
# A tibble: 6 x 1
  texto
  <chr>
1 CHAPTER I. TREATS OF THE PLACE WHERE OLIVER TWIST WAS BORN AND OF THE CIRCUMS~
2 Among other public buildings in a certain town, which for many reasons it wil~
3 For a long time after it was ushered into this world of sorrow and trouble, b~
4 Although I am not disposed to maintain that the being born in a workhouse, is~
5 As Oliver gave this first proof of the free and proper action of his lungs, t~
6 The surgeon had been sitting with his face turned towards the fire: giving th~
```

En cuarto lugar, se han agrupado los párrafos por capítulos y se ha *tokenizado* por palabras, es decir, se ha tomado cada palabra como unidad de significado.

```
# A tibble: 50 x 1
   texto
   <chr>
 1 CHAPTER I. TREATS OF THE PLACE WHERE OLIVER TWIST WAS BORN AND OF THE CIRCUM~
 2 CHAPTER II. TREATS OF OLIVER TWIST'S GROWTH, EDUCATION, AND BOARD
 3 CHAPTER III. RELATES HOW OLIVER TWIST WAS VERY NEAR GETTING A PLACE WHICH WO~
 4 CHAPTER IV. OLIVER, BEING OFFERED ANOTHER PLACE, MAKES HIS FIRST ENTRY INTO ~
 5 CHAPTER V. OLIVER MINGLES WITH NEW ASSOCIATES. GOING TO A FUNERAL FOR THE FI~
 6 CHAPTER VI. OLIVER, BEING GOADED BY THE TAUNTS OF NOAH, ROUSES INTO ACTION, ~
 7 CHAPTER VII. OLIVER CONTINUES REFRACTORY
 8 CHAPTER VIII. OLIVER WALKS TO LONDON. HE ENCOUNTERS ON THE ROAD A STRANGE SO~
 9 CHAPTER IX. CONTAINING FURTHER PARTICULARS CONCERNING THE PLEASANT OLD GENTL~
10 CHAPTER X. OLIVER BECOMES BETTER ACQUAINTED WITH THE CHARACTERS OF HIS NEW A~
# ... with 40 more rows
```

```
# A tibble: 6 x 1
  texto
  <chr>
1 CHAPTER I. TREATS OF THE PLACE WHERE OLIVER TWIST WAS BORN AND OF THE CIRCUMS~
2 Among other public buildings in a certain town, which for many reasons it wil~
3 For a long time after it was ushered into this world of sorrow and trouble, b~
4 Although I am not disposed to maintain that the being born in a workhouse, is~
5 As Oliver gave this first proof of the free and proper action of his lungs, t~
6 The surgeon had been sitting with his face turned towards the fire: giving th~
```

```
# A tibble: 6 x 2
  texto                                                          capitulo
  <chr>                                                          <chr>
1 Among other public buildings in a certain town, which for many reaso~ CHAPTER~
2 For a long time after it was ushered into this world of sorrow and t~ CHAPTER~
3 Although I am not disposed to maintain that the being born in a work~ CHAPTER~
4 As Oliver gave this first proof of the free and proper action of his~ CHAPTER~
```

```
5 The surgeon had been sitting with his face turned towards the fire: ~ CHAPTER~
6 "Oh, you must not talk about dying yet."                              CHAPTER~
```

```
# A tibble: 6 x 2
  capitulo                                                              word
  <chr>                                                                 <chr>
1 CHAPTER I. TREATS OF THE PLACE WHERE OLIVER TWIST WAS BORN AND OF THE C~ publ~
2 CHAPTER I. TREATS OF THE PLACE WHERE OLIVER TWIST WAS BORN AND OF THE C~ buil~
3 CHAPTER I. TREATS OF THE PLACE WHERE OLIVER TWIST WAS BORN AND OF THE C~ town
4 CHAPTER I. TREATS OF THE PLACE WHERE OLIVER TWIST WAS BORN AND OF THE C~ reas~
5 CHAPTER I. TREATS OF THE PLACE WHERE OLIVER TWIST WAS BORN AND OF THE C~ prud~
6 CHAPTER I. TREATS OF THE PLACE WHERE OLIVER TWIST WAS BORN AND OF THE C~ refr~
```

II. Análisis de sentimientos

En este segundo capítulo, se va a realizar un análisis *token* por *token* (en este caso, palabra por palabra) de la novela, determinando si estos corresponden a **sentimientos positivos o negativos**, a partir de la colección de palabras de *Bing*.

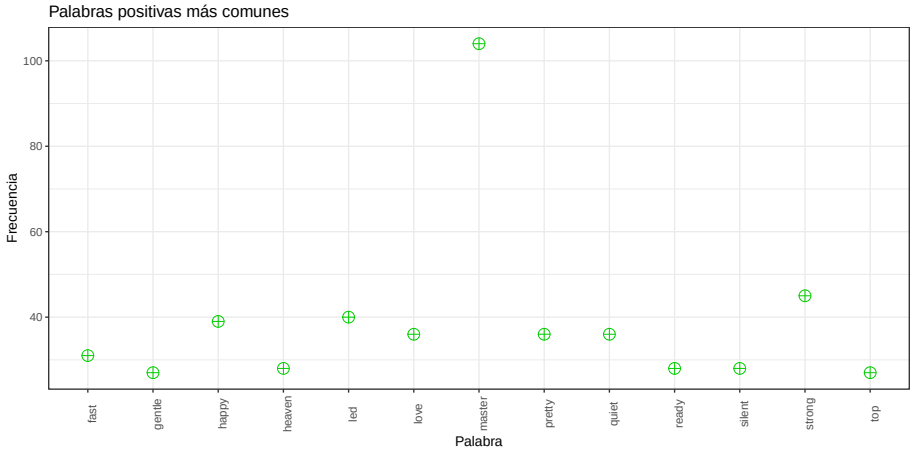
Se ha decidido utilizar el repositorio de palabras *Bing* ya que el resto ofrecido por R categorizan las palabras en varios grupos, por lo que no son tan sencillos de representar e identificar. En este sentido, cabe decir que se ha encontrado muy interesante la fuente *"afinn"*, pero al tener un tercio de las palabras del repositorio de *Bing* se ha decidido descartar.

Para realizar todo lo comentado se han importado las palabras del repositorio *Bing* y se han cruzado los datos con las palabras que contiene la novela de Dickens, identificando así qué palabras se consideran "positivas" y cuáles "negativas" en dicha novela.

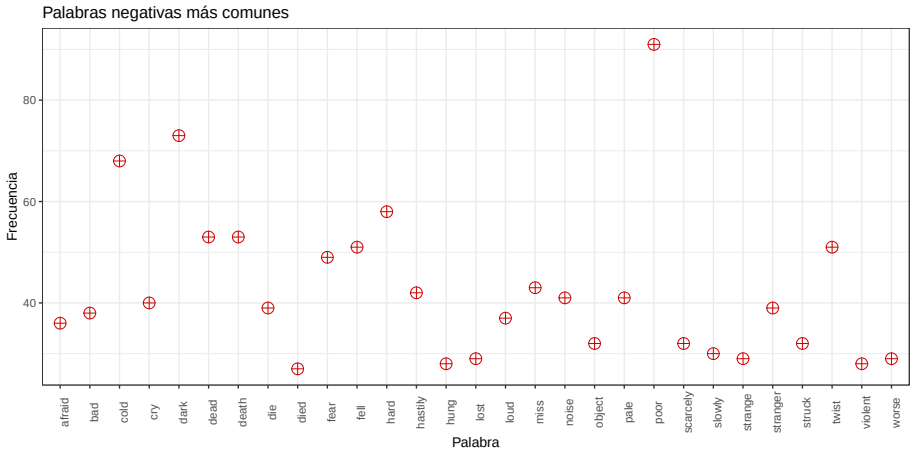
```
# A tibble: 6 x 2
  word       sentiment
  <chr>      <chr>
1 2-faces    negative
2 abnormal   negative
3 abolish    negative
4 abominable negative
5 abominably negative
6 abominate  negative
```

II.I. Análisis de sentimientos sobre la obra en conjunto

Así, en primer lugar, se puede obtener un gráfico de las **palabras "positivas"** más frecuentas en la novela de Dickens. Llama la atención que la palabra más frecuente es *master*, la cual, aunque se clasifique como positiva, en el contexto de la obra puede verse más como un signo de sumisión, y por lo tanto, negativa. También se observan palabras como *happy* o *love*, que son eminentemente positivas.



En cuanto a las **palabras "negativas"**, si bien se descarta *Twist*, por ser el apellido de Oliver, el protagonista, estas palabras tienen una frecuencia acumulada mucho mayor que las palabras positivas, siendo palabras unívocas y que van en la línea de la trama de la novela. *Oscuro*, *pobre*, *frío* o *muerto* solo pueden ser interpretadas de un modo, y es en ese ambiente de decadencia que describe continuamente la novela donde las injusticias, el maltrato y la pobreza son una constante en la vida del pequeño Oliver. Destaca especialmente también las palabras *muerte* y *muerto*, dos palabras muy negativas que se reptien, ambas, 53 veces.



Si se refleja en una tabla la información anterior, se podrá ver cómo la frecuencia de palabras negativas repetidas es mucho mayor que la frecuencia de palabras positivas. En total, hay más de dos veces palabras negativas (1239) que positivas (505), aun contando y teniendo en cuenta *master* como palabra postiva.

word	sentiment	n
master	positive	104
strong	positive	45
led	positive	40
happy	positive	39
love	positive	36
pretty	positive	36
quiet	positive	36
fast	positive	31
heaven	positive	28
ready	positive	28

word	sentiment	n
poor	negative	91
dark	negative	73
cold	negative	68
hard	negative	58
dead	negative	53
death	negative	53
fell	negative	51
twist	negative	51
fear	negative	49
miss	negative	43

II.II. Análisis de sentimientos por capítulos

Una vez realizado el análisis de los sentimientos de la obra en su conjunto, teniendo en cuenta la evolución de los escenarios y de lo que transcurre a lo largo de dicha obra, puede ser de interés ver **cómo se diferencian los distintos sentimientos a lo largo de los capítulos**. Para ello, se han agrupado las palabras positivas y negativas por capítulos.

Así, en primer lugar, si se analizan las **palabras positivas**, se observa cómo hay dos secciones claves en las que hay más palabras positivas. Primero, en los capítulos XII y XIV hay una mayor frecuencia de palabras positivas. Véase únicamente cómo se llama, por ejemplo, el capítulo XII. La segunda sección de palabras positivas frecuentes se da justo antes de finalizar en los capítulos XXXII, XXXIII, XXXIV y XXXVI. Sólo hace falta fijarse en el capítulo XXXII, donde Oliver Twist emprende diferentes amistades y aumenta su felicidad y bienestar respecto a anteriores situaciones.

```
# A tibble: 50 x 3
```

```
# Groups:   capitulo [50]
   capitulo                                                  sentiment     n
   <chr>                                                     <chr>     <int>
 1 CHAPTER I. TREATS OF THE PLACE WHERE OLIVER TWIST WAS BORN A~ positive     26
 2 CHAPTER II. TREATS OF OLIVER TWIST'S GROWTH, EDUCATION, AND ~ positive     86
 3 CHAPTER III. RELATES HOW OLIVER TWIST WAS VERY NEAR GETTING ~ positive     65
 4 CHAPTER IV. OLIVER, BEING OFFERED ANOTHER PLACE, MAKES HIS F~ positive     42
 5 CHAPTER IX. CONTAINING FURTHER PARTICULARS CONCERNING THE PL~ positive     62
 6 CHAPTER L. THE PURSUIT AND ESCAPE                             positive     43
 7 CHAPTER V. OLIVER MINGLES WITH NEW ASSOCIATES. GOING TO A FU~ positive     70
 8 CHAPTER VI. OLIVER, BEING GOADED BY THE TAUNTS OF NOAH, ROUS~ positive     41
 9 CHAPTER VII. OLIVER CONTINUES REFRACTORY                      positive     33
10 CHAPTER VIII. OLIVER WALKS TO LONDON. HE ENCOUNTERS ON THE R~ positive     55
# ... with 40 more rows
```

En segundo lugar, si se analizan las **palabras negativas**, en primer lugar se observan que sus frecuencias son mucho mayores. Parece paradójico que la mayor concentración de estos sentimientos negativos se dan al final de la obra, donde casi se alcanzan en uno de los capítulos las 200 palabras negativas.

```
# A tibble: 50 x 3
# Groups:   capitulo [50]
   capitulo                                                  sentiment     n
   <chr>                                                     <chr>     <int>
 1 CHAPTER I. TREATS OF THE PLACE WHERE OLIVER TWIST WAS BORN A~ negative     43
 2 CHAPTER II. TREATS OF OLIVER TWIST'S GROWTH, EDUCATION, AND ~ negative    106
 3 CHAPTER III. RELATES HOW OLIVER TWIST WAS VERY NEAR GETTING ~ negative    117
 4 CHAPTER IV. OLIVER, BEING OFFERED ANOTHER PLACE, MAKES HIS F~ negative     71
 5 CHAPTER IX. CONTAINING FURTHER PARTICULARS CONCERNING THE PL~ negative     46
 6 CHAPTER L. THE PURSUIT AND ESCAPE                             negative    165
 7 CHAPTER V. OLIVER MINGLES WITH NEW ASSOCIATES. GOING TO A FU~ negative    165
 8 CHAPTER VI. OLIVER, BEING GOADED BY THE TAUNTS OF NOAH, ROUS~ negative     89
 9 CHAPTER VII. OLIVER CONTINUES REFRACTORY                      negative    103
10 CHAPTER VIII. OLIVER WALKS TO LONDON. HE ENCOUNTERS ON THE R~ negative     99
# ... with 40 more rows
```

III. Nube de palabras

En este tercer apartado, se realizará una **nube de palabras** para representar las palabras más comunes a lo largo del libro. Así, no sólo se verán las palabras más frecuentes, sino que, a través del tamaño, se verá su importancia de un solo vistazo. Esto se realizará a cabo dos veces: en primer lugar para todo el vocabulario de la novela, y en segundo lugar con las palabras que constituyen sentimientos positivos y negativos.

Así, respecto al primer planteamiento, si se realiza una nube de palabras para el conjunto del vocabulario de la novela, se verá cómo en dicha nube destacan personajes (Oliver Twist, Bumble, Fagin, un judío -y no judíos en plural, con lo que es posible que un personaje sea caracterizado así en vez de una comunidad-, un doctor, Noah, etc.), elementos que dan una visión de la ambientación (*boy*, *girl*, *gentleman*, *sir*, *child*, *fire*, *hat*) que es de la vida infantil de la Inglaterra de los años 1830. Por último, hay palabras que bien podrían estar en cualquier libro (*door*, *head*, *eyes*, *time*, *hand*, *day*, *night*).

En segundo lugar, se acota la nube de palabras a las palabras categorizadas en el análisis de sentimientos como positivas y negativas, de manera que se sabrá que tendrán un significado relevante para el análisis aquí realizado. De esta forma, se observa que, de igual manera que se vio gracias a la tabla en el anterior apartado, *master* tiene una frecuencia muy alta. Dentro del resto de palabras positivas, se ven algunas que tienen un deje de tristeza, como *heaven*, *quiet*, *silent* o incluso *fine* (se debe recordar que también significa multa). Las palabras negativas que más aparecen son *poor*, *cold*, *dark* y *dead*. Aparece *twist*, pero al ser el apellido

del protagonista, no se debe tomar en cuenta como palabra negativa.

IV. Frecuencia de palabras

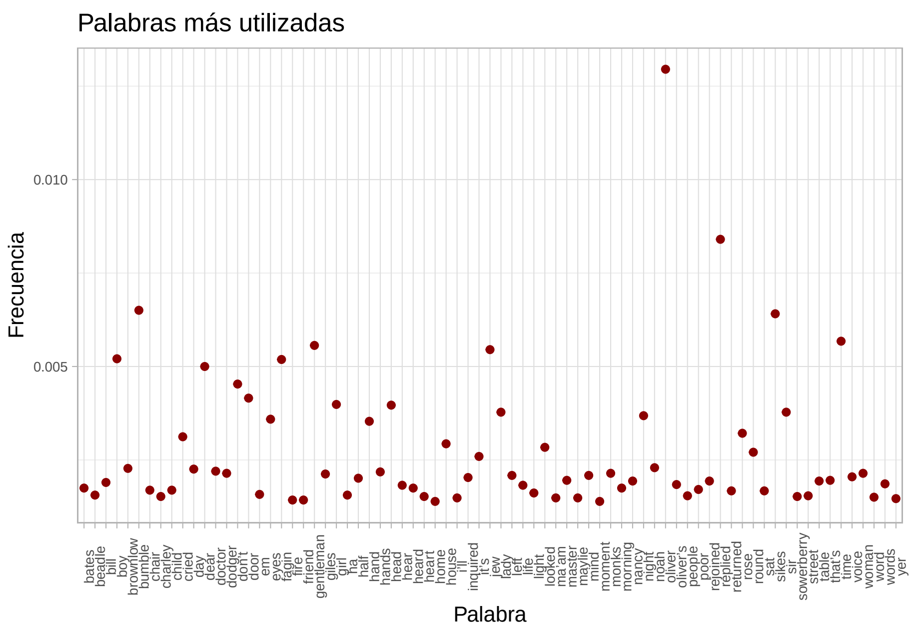
En este cuarto capítulo se va a explorar qué palabras aparecen más veces en el libro, a través de una tabla y de un gráfico. Ya que es la misma información obtenida en la primera nube de palabras graficada en el anterior capítulo, se verán muchas palabras repetidas. Sin embargo, al estar en un formato menos gráfico, pero más completo y preciso, en el que se verán cuántas veces se repite exactamente cada palabra, se podrá ver con más detalle las características del texto.

De esta forma, se observa que la palabra más frecuente es el nombre de pila del protagonista, que se menciona casi 700 veces. Llaman la atención palabras de estilo que se repiten mucho, como *replied*, *time* o incluso *dear*. También son muy frecuentes nombres de personajes, como Fagin, Sikes o Bumble. Hay que tener en cuento que se ha escapado alguna palabra vacía como *don't* o *it's* que, al ser contracciones, no aparecen en la base de *stop words* utilizada.

También se puede navegar en las palabras que eran demasiado raras para aparecer en la nube de texto, pero son suficienteme frecuentes como para tener un impacto en el texto, como los verbos usados (*spoke*, *stood*, *brittles*, *stopped*, *appeared*, *fell*, *cried*) en el texto, que también dan una idea de lo que pasa en el libro, con acciones físicas como caídas y paradas, y acciones más literarias como hablar, aparecer o llorar.

```
# A tibble: 9,495 x 2
   word          n
   <chr>     <int>
 1 oliver      689
 2 replied     447
 3 bumble      346
 4 sikes       341
 5 time        302
 6 gentleman   296
 7 jew         290
 8 boy         277
 9 fagin       276
10 dear        266
# ... with 9,485 more rows
```

Si todo lo expuesto se representa gráficamente, esto sirve para ver mejor las frecuencias de cada palabra, ya que en el resto de representaciones se ha visto la frecuencia en valores absolutos. Desde luego, al ser una obra extensa, no hay palabras tan comunes. Si fuese, por ejemplo, un cuento, se verían frecuencias relativas más altas.


Palabras más utilizadas
Frecuencia
0.010
0.005
Palabra

V. Conclusiones

Las técnicas de análisis de datos están en continuo desarrollo en plena sociedad de la información. En este sentido, las técnicas de minería de texto (*text mining*) aparecen como una herramienta que permite analizar de forma más cualitativa ámbitos que, con técnicas exclusivamente orientadas a datos numéricos no son capaces de considerar.

De esta forma, en este trabajo, se ha escogido la obra *Oliver Twist*, una novela de *Charles Dickens* enmarcada en plena Revolución Industrial en una Inglaterra donde, dicha obra, escenifica una ambientación de decadencia y de pesimismo por las consecuencias de dicha época. En este sentido, dicha obra se ha importado del *Project Gutenberg*, y con las técnicas de minería de texto se ha procesado el texto y se ha podido ver que, Oliver Twist, es una obra que a lo largo de su duración predominan los sentimientos tristes, y se ambienta en la época en la que se escribió y los nombres de sus personajes.

También se ha podido ver que relata una historia que ocurre en la parte baja de la sociedad, con un personaje que representa las minorías religiosas y que, en su desenlace, hay un aumento de palabras positivas y sobre la amistad. Gracias a ello, se deduce que tiene un carácter social con un final feliz, una historia que sirvió para acercar una realidad a aquellas personas que tenían la suerte de estar alejada de ella. Esto entra en la coherencia de la línea literaria de Dickens, un autor entregado a las causas sociales en cada una de sus obras, y que incluso traspasaría su propia dimensión filantrópica.

Sin embargo, el análisis realizado en este trabajo también supone perder partes fundamentales del texto: cuáles son los hechos por los que pasa el protagonista (que efectivamente está presente a lo largo de la novela), las enseñanzas que nos aporta el final feliz, cómo actúa cada personaje y, sobre todo, la lectura placentera de uno de los grandes clásicos victorianos de Dickens.

Referencias

En esta sección se incluyen las referencias bibliográficas utilizadas para el desarrollo del proyecto.

Dickens, C. (1837). Oliver Twist. Project Gutenberg. https://www.gutenberg.org/ebooks/730.

El Cronovisor (2017). Charles Dickens, genio de la crítica social. Episodio 4. https://open.spotify.com/episode/3lLBJHZGY3verRFV6ptEbS?si=9_xGOeGWTiW4ijZuIexBeg&utm_sourc
link

Gutiérrez, M.J. (2022). Text mining con R. Aprendizaje estadístico y otras técnicas avanzadas. Máster Universitario en Modelización y Análisis de Datos Económicos. Universidad de Castilla-La Mancha.

Silge, J. & Robinson D. (2022). Text Mining with R. https://www.tidytextmining.com/tidytext.html.

Anexos

Anexo 1. Datos de la sesión

En esta sección se recogen los datos de la sesión utilizada para elaborar este informe. Es fundamental observar la versión de R, así como las versiones de los paquetes bajo los cuales se ha ejecutado el código o *script*.

```
R version 4.1.1 (2021-08-10)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 19043)

Matrix products: default

locale:
[1] LC_COLLATE=Spanish_Spain.1252  LC_CTYPE=Spanish_Spain.1252
[3] LC_MONETARY=Spanish_Spain.1252 LC_NUMERIC=C
[5] LC_TIME=Spanish_Spain.1252

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] reshape2_1.4.4     wordcloud_2.6      RColorBrewer_1.1-3 knitr_1.38
 [5] textdata_0.4.1     scales_1.2.0       ggraph_2.0.5       igraph_1.3.0
 [9] tm_0.7-8           NLP_0.2-1          tokenizers_0.2.1   gutenbergr_0.2.1
[13] forcats_0.5.1      stringr_1.4.0      purrr_0.3.4        readr_2.1.2
[17] tibble_3.1.6       ggplot2_3.3.5      tidyverse_1.3.1    tidytext_0.3.2
[21] dplyr_1.0.8        tidyr_1.2.0

loaded via a namespace (and not attached):
 [1] fs_1.5.2           bit64_4.0.5        lubridate_1.8.0    httr_1.4.2
 [5] SnowballC_0.7.0    tools_4.1.1        backports_1.4.1    utf8_1.2.2
 [9] R6_2.5.1           DBI_1.1.2          colorspace_2.0-3   withr_2.5.0
[13] tidyselect_1.1.2   gridExtra_2.3      curl_4.3.2         bit_4.0.4
[17] compiler_4.1.1     cli_3.2.0          rvest_1.0.2        xml2_1.3.3
[21] labeling_0.4.2     triebeard_0.3.0    slam_0.1-50        digest_0.6.29
[25] rmarkdown_2.13     pkgconfig_2.0.3    htmltools_0.5.2    highr_0.9
[29] dbplyr_2.1.1       fastmap_1.1.0      rlang_1.0.2        readxl_1.4.0
[33] rstudioapi_0.13    farver_2.1.0       generics_0.1.2     jsonlite_1.8.0
[37] vroom_1.5.7        magrittr_2.0.3     Matrix_1.3-4       Rcpp_1.0.8.3
[41] munsell_0.5.0      fansi_1.0.3        viridis_0.6.2      lifecycle_1.0.1
[45] stringi_1.7.6      yaml_2.3.5         MASS_7.3-54        plyr_1.8.7
[49] grid_4.1.1         parallel_4.1.1     ggrepel_0.9.1      crayon_1.5.1
[53] lattice_0.20-45    graphlayouts_0.8.0 haven_2.4.3        hms_1.1.1
```

```
[57] pillar_1.7.0       reprex_2.0.1       glue_1.6.2         evaluate_0.15
[61] modelr_0.1.8       urltools_1.7.3     vctrs_0.4.0        tzdb_0.3.0
[65] tweenr_1.0.2       cellranger_1.1.0   gtable_0.3.0       polyclip_1.10-0
[69] assertthat_0.2.1   xfun_0.30          ggforce_0.3.3      broom_0.8.0
[73] tidygraph_1.2.1    janeaustenr_0.1.5  viridisLite_0.4.0  ellipsis_0.3.2
```