Frecuencias

Gabriela Sánchez Y.
5064

En el presente trabajo se realiza un análisis del libro *"Anne of Green Gables"* obtenido del sitio de Project Gutenberg.

1. Introducción

El análisis del libro de texto se realiza con la ayuda del lenguaje de programación R versión 4.0.2 [3], haciendo uso de tres librerías: `gutenbergr` que permite acceder al texto plano del libro y, `tidytext` y `dplyr` que permiten la descomposición del texto.

Dicho análisis se basa en el estudio de las frecuencias de las palabras y letras del texto. El primer paso para poder proceder con el estudio es obtener el texto plano del libro, lo cual es posible mediante la función `gutenberg_download`.

El procedimiento realizado para el análisis puede encontrarse en el script `t2.R` [1].

2. Letras

La descomposición del texto en caracteres se realiza con la función `unnest_tokens`. Ya que es de interés únicamente la frecuencia de las letras, se eliminan todos los caracteres que no lo son. En este caso, el único caracter que no es una letra es "|".

Para mejorar la visualización, las frecuencias se ordenan en forma decreciente. De esta manera es posible observar que las primeras tres letras más usadas en el texto son *e, t* y *a*, mientras que las menos usadas son *x, q* y *z*, tal y como se muestra en la figura 1.

3. Palabras

La descomposición en letras no dice mucho acerca del texto por lo que se procede a realizar una descomposición en palabras. Para esto, nuevamente se usa la función `unnest_tokens`.

Como segundo paso en este análisis, se realiza un filtrado: son eliminadas aquellas palabras que en inglés se conocen como *stop words* (palabras vacías), ya que no serán útiles para el estudio [2]. Son palabras muy comunes en el idioma que pueden eliminarse sin sacrificar el significado de una oración. En inglés algunos ejemplos de palabras vacías son *at, the, is, of, to*.

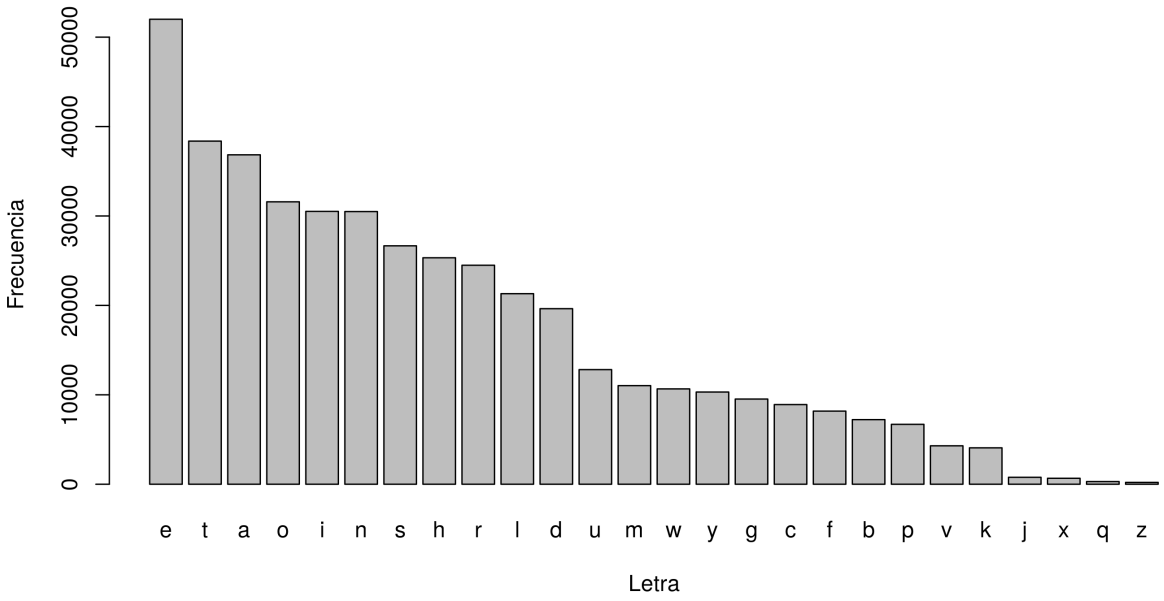


Figura 1: Gráfico de barras de la frecuencia de las letras del abecedario en el texto analizado.

Una vez hecho este filtro, se aplica otro que toma en cuenta únicamente las palabras con una frecuencia mayor a uno. En el cuadro 1 se pueden observar las primeras 10 palabras más frecuentes. Estos resultados permiten inferir que *Anne, Marilla, Diana* y *Matthew* son personajes principales en la novela, siendo *Anne* el principal ya que la frecuencia está muy por encima de los otros.

Continuando con las palabras más frecuentes, en la figura 2 se muestra un gráfico de barras que muestra la frecuencia de las 20 palabras siguientes en frecuencia a las del cuadro 1. Pueden observarse otros nombres como *Gilbert y Jane* y, lo que parecen ser apellidos *Lynde* y *Barry*, por lo que podríamos decir que son personajes secundarios en la novela.

Una persona que ya ha leído el libro sabe que *Barry* es el apellido de *Diana*. Esto advierte

Cuadro 1: Palabras más comunes y su frecuencia.

Palabra	Frecuencia
anne	1107
marilla	797
diana	386
matthew	339
time	178
girl	170
school	152
miss	148
home	144
white	142

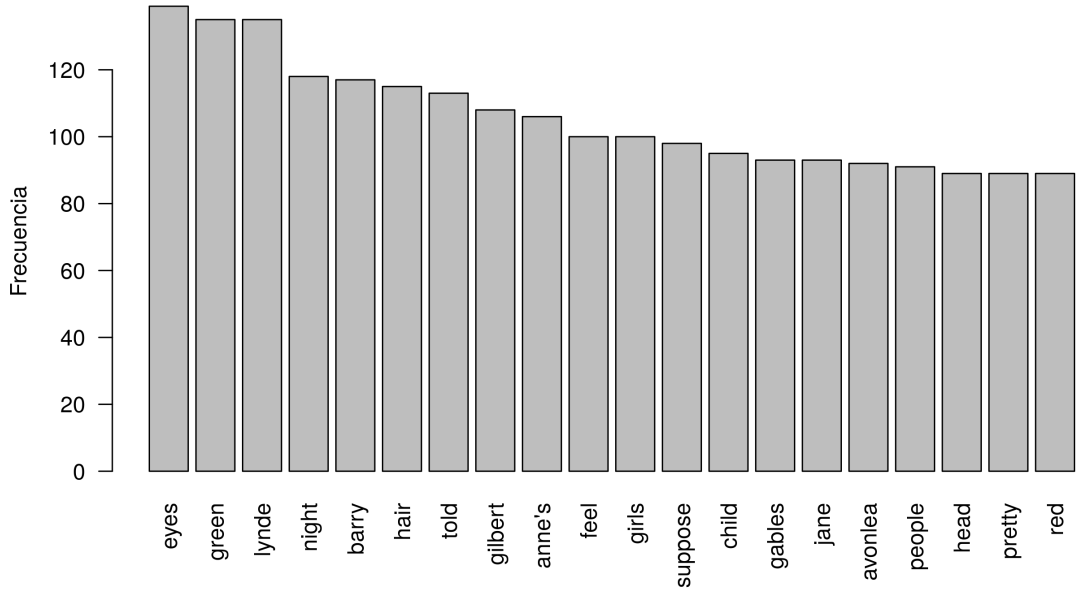


Figura 2: Gráfico de barras de las palabras más frecuentes en el texto, una vez realizados dos filtros.

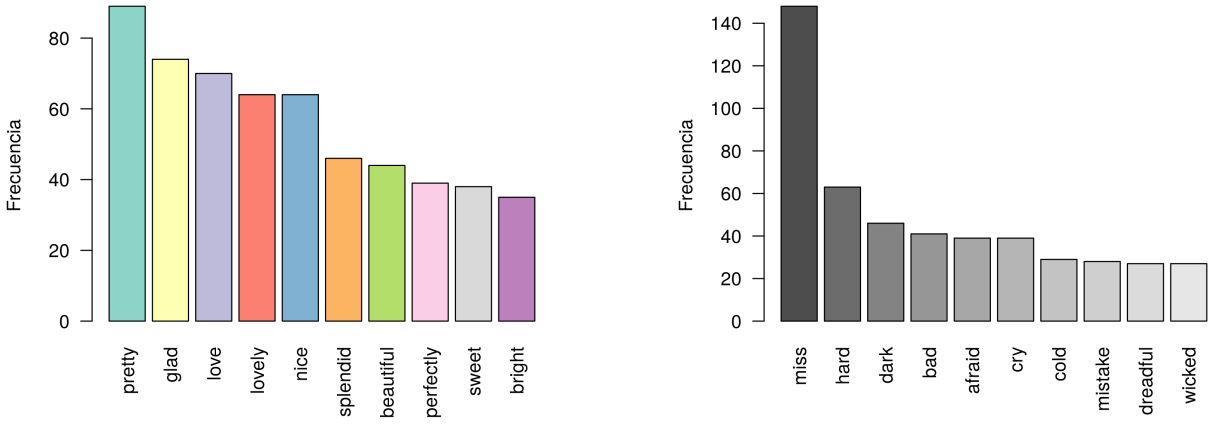
que un análisis de palabras "sueltas" no permite inferir mucho acerca del contenido del texto, una mejor opción es considerar la frecuencia en que aparecen dos o más palabras juntas.

Antes de proceder a analizar conjuntos de palabras, se examinan las palabras positivas y negativas más comunes de acuerdo al léxico `bing`, que es posible obtener a través de la función `get_sentiments`. La figura 3 muestra las 10 palabras más comunes positivas (figura 3a) y negativas (figura 3b).

En el gráfico de barras mostrado en la figura 3b se puede apreciar que la palabra negativa con más frecuencia en el libro es *miss*. Sin embargo, aquí se señala un problema. Se sabe que tanto en español como en inglés existen palabras homónimas, es decir, palabras cuya pronunciación o escritura es igual o similar pero que tienen diferente significado. En este caso, personas que ya han leído el libro, pueden advertir que el personaje principal se refiere a su profesora como *Miss Stacy*, otro personaje relevante en la novela. por lo que no se garantiza que *miss* en su connotación "negativa" realmente tenga una frecuencia más alta que las otras palabras mostradas en el gráfico.

Por último, se analiza la frecuencia en que aparecen las secuencias de dos palabras, también llamado bigrama. El análisis de los brigramas se realiza con los datos que previamente filtraron las palabras vacías, ya que si se omite ese paso, los resultados obtenidos muestran serán los que se muestran en el cuadro 2 y dicha información no aporta sobre el contenido del libro.

En la figura 4 se observan los bigramas más comunes en el texto filtrado. Resalta el nombre de *Green Gables*, lugar donde se desarrolla la mayoría de la trama de la novela y nombres de personajes como *Miss Stacy, Gilbert Blythe, Anne Shirley, Rubby Gillis,* entre otros.



(a) Palabras positivas

(b) Palabras negativas

Figura 3: Gráfico de barras de las palabras positivas y negativas más frecuentes.

Cuadro 2: Bigramas más frecuentes en el texto sin filtro.

Bigrama	Frecuencia
in the	413
to be	325
of the	308
it was	273
to the	240
and i	198
on the	191
i don't	174
going to	165
was a	164

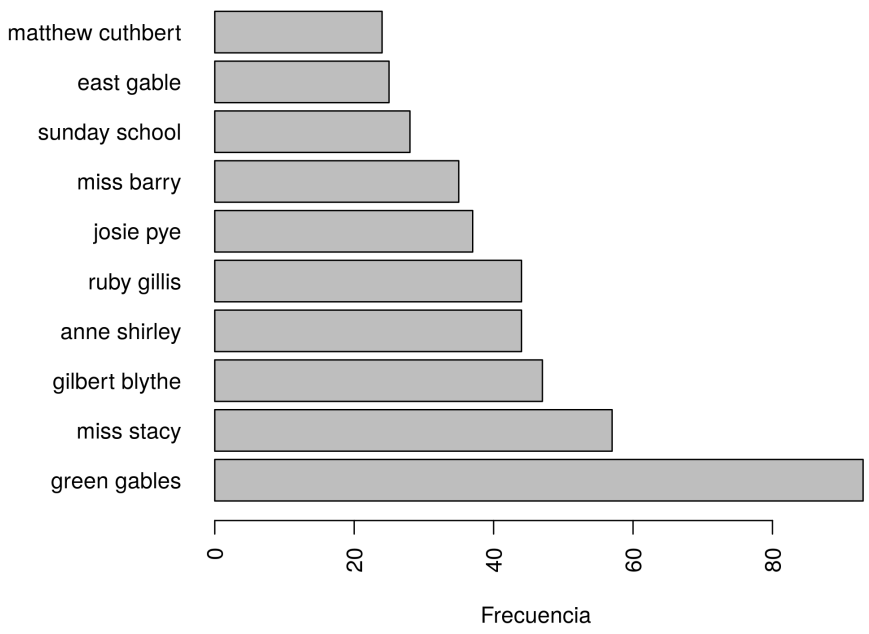


Figura 4: Bigramas con mayor frecuencia en el texto filtrado.

Referencias

[1] Gabriela Sánchez Y. Modelos Probabilistas Aplicados. https://github.com/Saphira3000/MPA.

[2] Julia Silge and David Robinson. Text Mining with R. https://www.tidytextmining.com/tidytext.html.

[3] The R Foundation. The R Project for Statistical Computing. https://www.r-project.org/.