Minería de Texto

Fabiola Vázquez

15 de septiembre de 2020

Resumen

El objetivo de esta tarea es analizar un libro te texto, mediante las palabras, letras, personajes y sentimientos que se mencionan. Se utilizan gráficas de barras y figuras, como los word clouds, para representar mejor, de una forma visual, estos resultados.

---

1. Introducción

La minería de textos extrae información útil e interesante de textos, como lo pueden ser de páginas web, documentos, artículos. En este caso, se realizó un estudio de la novela *Alice's Adventures in Wonderland* [1] con el software estadístico R versión 4.0.2 [2] en el IDE **R Studio** [3]. Se analizan las palabras, letras más frecuentes del libro, así como también, los sentimientos que se mencionan en él.

2. Análisis del texto

La novela se extrajo de la librería electrónica Project Gutenberg, de la siguiente manera.

```
library(gutenbergr)
alice <- gutenberg_download(c(11))
```

Una vez cargado, procedemos a crear dos listas que contengan los caracteres y las palabras de nuestro libro, respectivamente.

```
letras = alice %>% unnest_tokens(chars, text, "characters")
palabras = alice %>% unnest_tokens(word, text, "words")
```

2.1. Análisis de caracteres

De nuestra lista de letras, no solo aparecen letras, si no también números, en el cuadro 2 aparecen los números cero y tres con frecuencia uno. Para el análisis, quitamos esos valores no deseados y procedemos a trabajar únicamente con las letras.
En la figura 1 podemos apreciar que la letra que tiene menor frecuencia es la letra **z** y la de mayor frecuencia es la **e**. El libro contiene 107 721 letras, de las cuales 13 576 son la letra **e**, eso quiere decir que son aproximadamente el 13 % de las letras.

Cuadro 1: Fragmento de la tabla de frecuencias de los caracteres que aparecen en el libro.

	Caracter	Frecuencia
1	0	1
2	3	1
3	a	8 791
4	b	1 475
5	c	2 399
6	d	4 931

Cuadro 2: Fragmento de la tabla de frecuencias de las palabras que aparecen en el libro.

	Palabra	Frecuencia
1	the	1 644
2	and	872
3	to	729
4	a	632
5	she	541
6	it	530
7	of	514
8	said	462

Figura 1: Gráfica de barras de las frecuencias de caracteres.

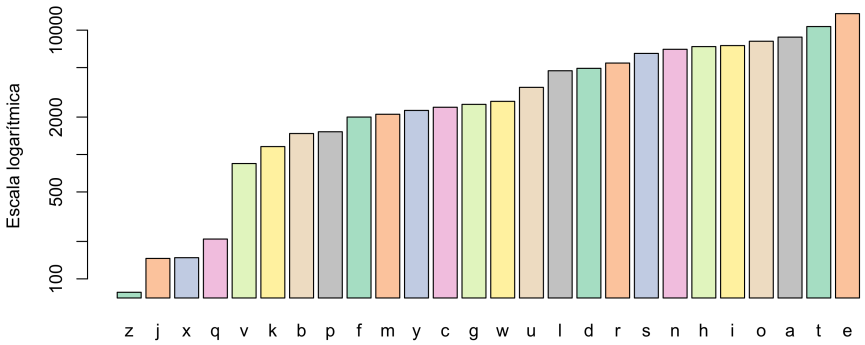


Figura 2: Palabras más frecuentes en el libro.

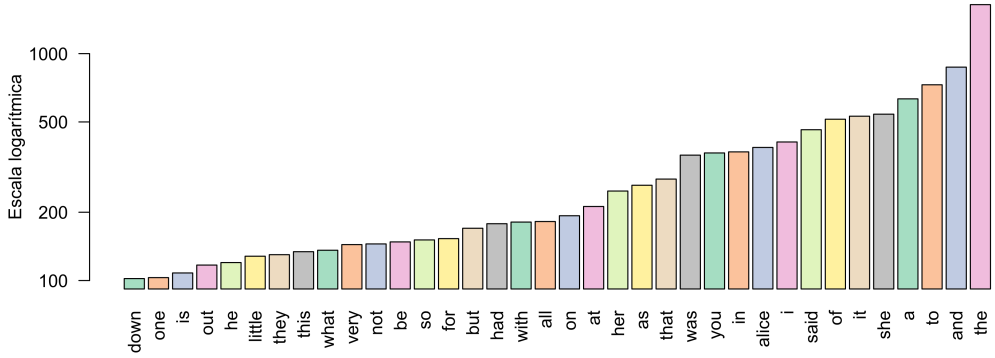
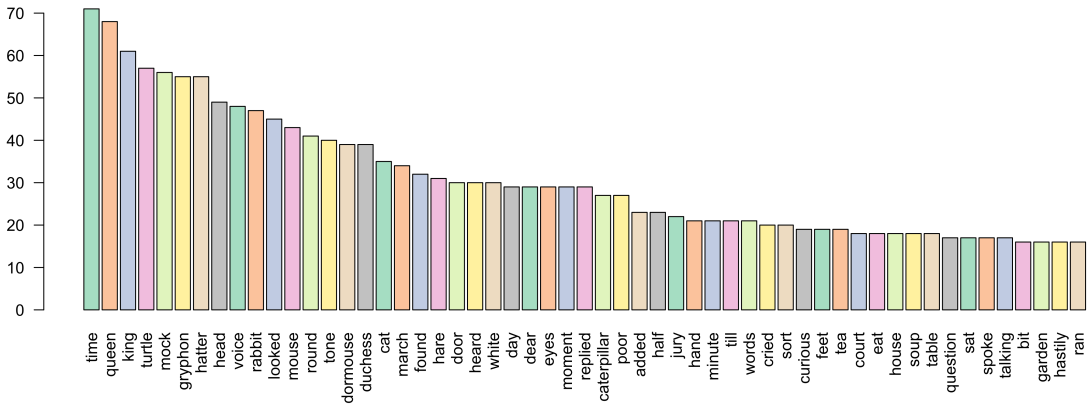


Figura 3: Palabras más frecuentes en el libro omitiendo pronombres y artículos.



2.2. Análisis de palabras

Se realizó una tabla de frecuencias de las palabras que aparecen en el libro y se obtuvo la figura 2, en donde apreciamos que aparecen palabras del tipo artículo, pronombres, etcétera. Por ejemplo, la palabra *the* tiene una frecuencia de 1 644. Para una mejor apreciación, quitamos dichas palabras, el gráfico correspondiente aparece en la figura 3. Otra manera de visualizar las palabras más frecuentes del libro, es usando un *word cloud*, donde, las palabras de mayor tamaño son aquellas que tienen una frecuencia mayor en el texto, como lo podemos apreciar en la figura 4.

2.3. Personajes

El libro cuenta con una gran variedad de personajes. En este análisis consideramos solo aquellos que son más relevantes en la trama. La figura 5 muestra que el nombre del personaje que se menciona más durante la trama del libro es **Alice**. Los nombres de los demás personajes se mencionan de manera similar entre sí, pero con menor frecuencia que la del personaje principal.

Figura 4: Wordcloud con las palabras más frecuentes en el libro.

Figura 5: Personajes más frecuentes en el libro.

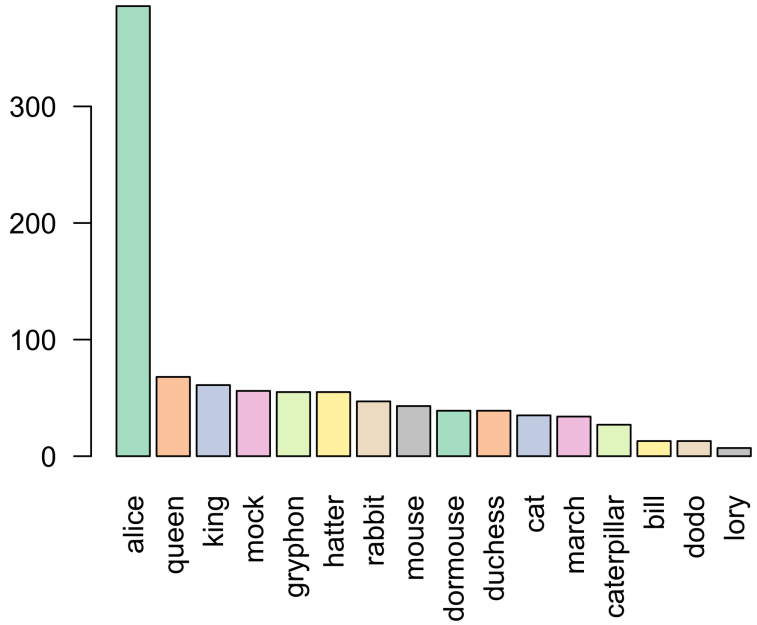
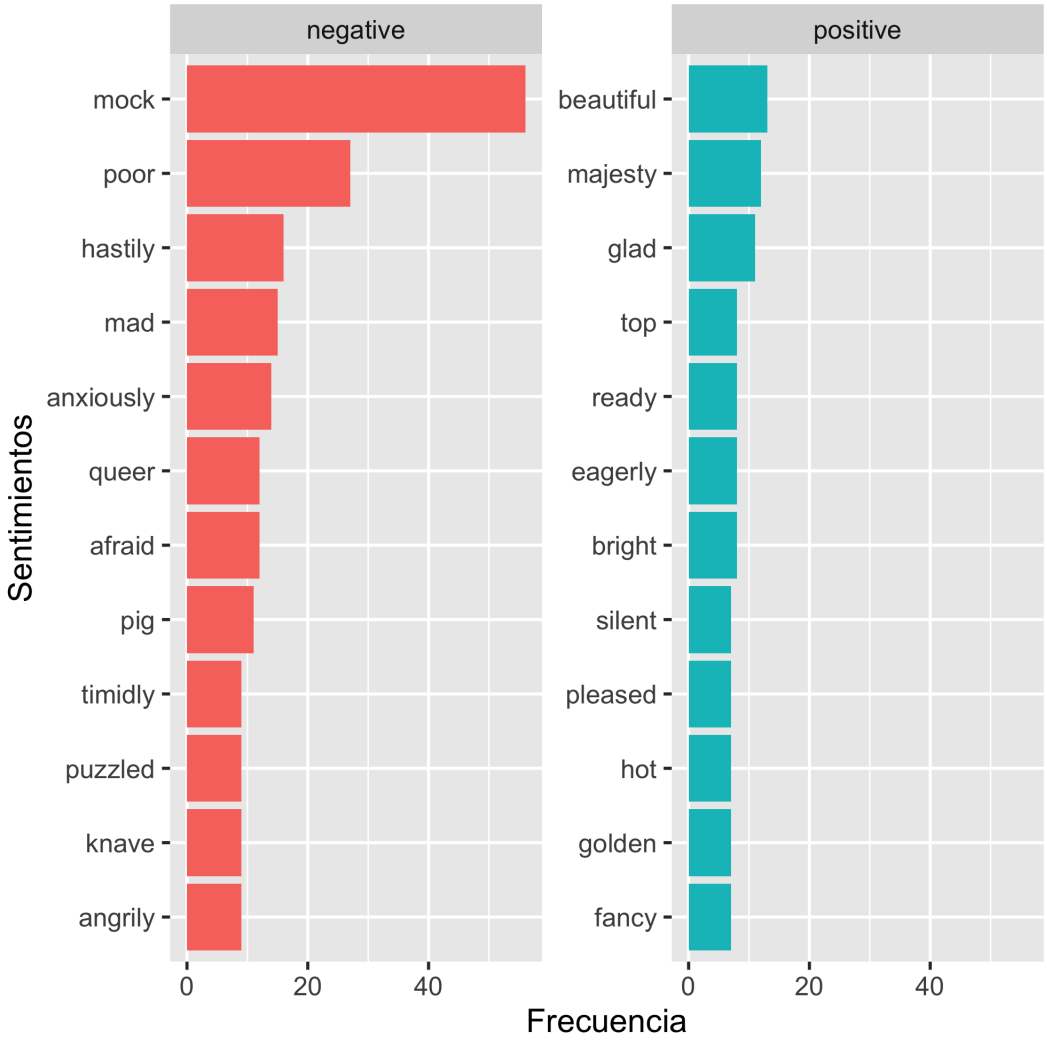


Figura 6: Comparación de sentimientos positivos y negativos.



2.4. Análisis de los sentimientos

Una novela puede transmitir diversos sentimientos, dependiendo del tipo de novela que es. Para este análisis es necesario una librería que nos permite acceder a diversos diccionarios que contienen palabras asociadas a ciertos sentimientos, o clasificadas como sentimientos positivos o negativos. En la figura 6 podemos apreciar una comparación entre los sentimientos negativos y positivos que se encuentran en el libro. Como vemos, el lado izquierdo, referente a los negativos, tiene palabras más frecuentes que el apartado derecho, es decir que se habla más de sentimientos negativos.

3. Conclusiones

Este tipo de análisis de textos es muy útil ya que te ayuda a obtener información importante del texto que se analiza. Por ejemplo, las palabras más frecuentes en el libro (omitiendo pronombres, artículos, conectores) nos sirven para darnos una idea sobre la trama del texto.

Figura 7: Wordcloud comparativo entre sentimientos negativos y positivos.

Referencias

[1] Lewis Carroll. *Alice's Adventures in Wonderland*. Macmillan, England, UK, 1865.

[2] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

[3] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA, 2020.