Posgrado de Ingeniería de Sistemas

Distribuciones en oraciones de Dracula

Alberto Benavides
21 de septiembre de 2020

1. Introducción

Existen diversos métodos para determinar la autoría de libros. De entre ellos destaca el que utiliza minería de textos combinado con técnicas de agrupamientos como K–medias. Estos métodos utilizan como sus características la cantidad de palabras y comas de las oraciones presentes en las obras, entre otras. Un breve ejemplo sobre esta técnica se halla en `http://www.aicbt.com/authorship-attribution/`.

2. Recursos

El libro de *Dracula* [3] escrito por Stoker en 1897 se encuentra disponible de manera gratuita en el sitio de Project Gutenberg [4]. Éste se descarga por la librería `gutembergr` [2] de R. La separación en oraciones se realiza mediante `corpus` [1].

3. Preprocesamiento

Después de obtenido el libro mediante la función `gutenberg_download`, se analiza su contenido inicial y final con las funciones `head` y `tail`. Esto muestra la existencia de apartados introductorios y anexos que se escapan del contenido del libro en sí, por lo que se procede a remover del contenido esos elementos. Destaca aquí el uso de la función `grep` que permite identificar expresiones regulares y, en este caso, la ubicación del final del libro que coindice con las palabras “THE END”.

4. Resultados y análisis

Con el contenido del libro se pueden extraer las oraciones que contiene. Para ello primero es necesario convertir el libro obtenido en una sola cadena de texto, lo que se hace con la función `paste`, seguido del uso de la función `text_split` de la librería `corpus` que separa las oraciones de dicho texto. Se prefiere el uso de esta librería puesto que obvia abreviaturas que usan puntos para evitar separar oraciones por estos motivos. De estas oraciones se cuentan las palabras que contienen y las comas con `strsplit` para la primera y `str_count` para la última. Enseguida se grafican los histogramas de palabras y comas por oración junto a una función de distribución que se equipara a sus proporciones. Estas gráficas pueden verse en las figuras 4.1 (p. 3) y 4.2 (p. 4), respectivamente. Los procedimientos computacionales y sus códigos pueden revisarse en `https://github.com/jbenavidesv87/probabilidad/blob/master/tema3/tarea.ipynb`

La distribución de palabras por oración que sigue la obra de *Dracula* coincide con una distribución binomial negativa. Este tipo de distribuciones se generan tras obtener, en un transcurso de r_1 repeticiones, las veces que debe repetirse un experimento para que un determinado evento con probabilidad p_1 sea exitoso un total de k veces. En este caso, la distribución binomial negativa generada que equipara los resultados de la distribución del número de palabras por oración, toma los parámetros $r_1 = 90000$, $p_1 = 0.087$ y $k = 2$.

Por su parte, la distribución de comas en una oración de la obra citada coincide con una distribución geométrica que se desarrolla a partir del número de veces que debe hacerse un experimento de probabilidad p_2 (repetido r_2 veces) para que tal experimento sea exitoso una primera vez. Los parámetros que toma esta distribución para ser semejante a la del número de comas por oración en *Dracula* son $p_2 = 0.35$ y $r_2 = 10000$.

Referencias

[1] Leslie Huang. corpus: Text Corpus Analysis. `https://www.rdocumentation.org/packages/tidytext/versions/0.2.5`, 2020.

[2] David Robinson. gutenbergr. `https://cran.r-project.org/web/packages/corpus/index.html`, 2019.

[3] Bram Stoker. *Dracula*. Oxford University Press, New York, 1897.

[4] Bram Stoker. Dracula by Bram Stoker. `https://www.gutenberg.org/ebooks/345`, 2020.

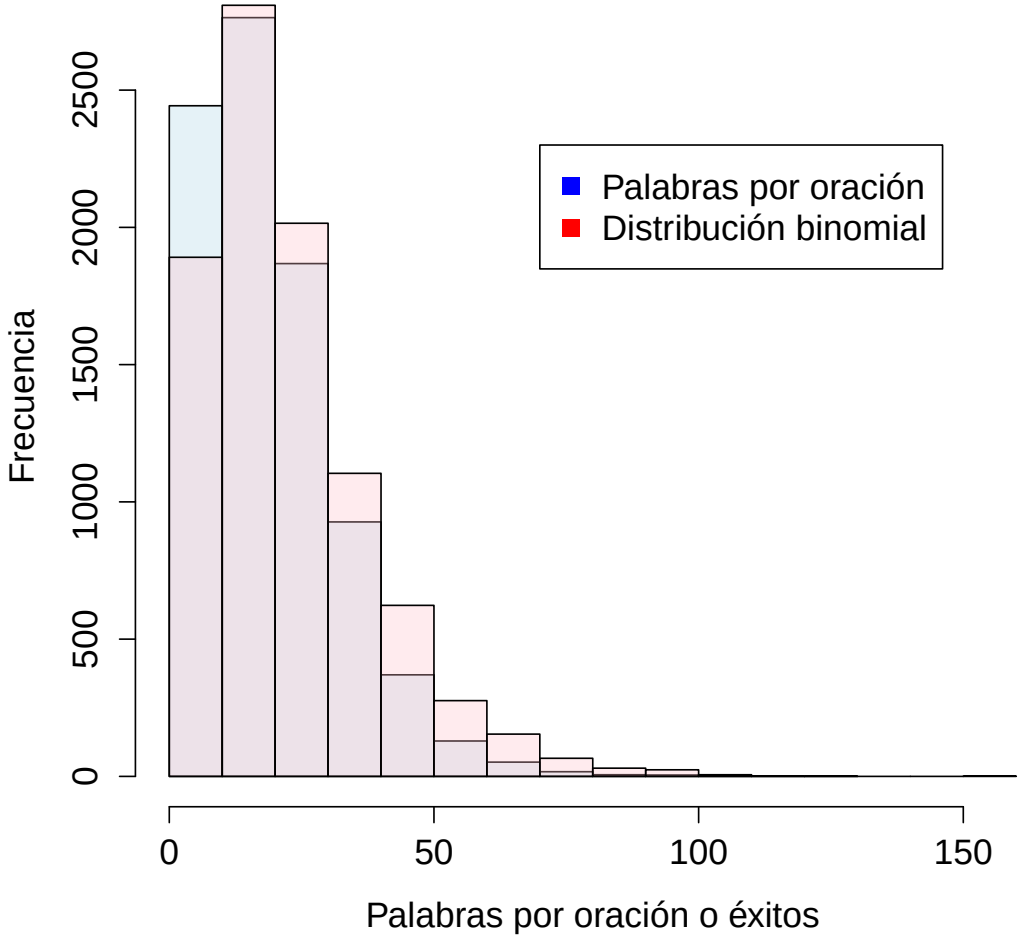

Figura 4.1: Histogramas de palabras por oración de *Dracula* (azul) y función de distribución binomial negativa (rojo).

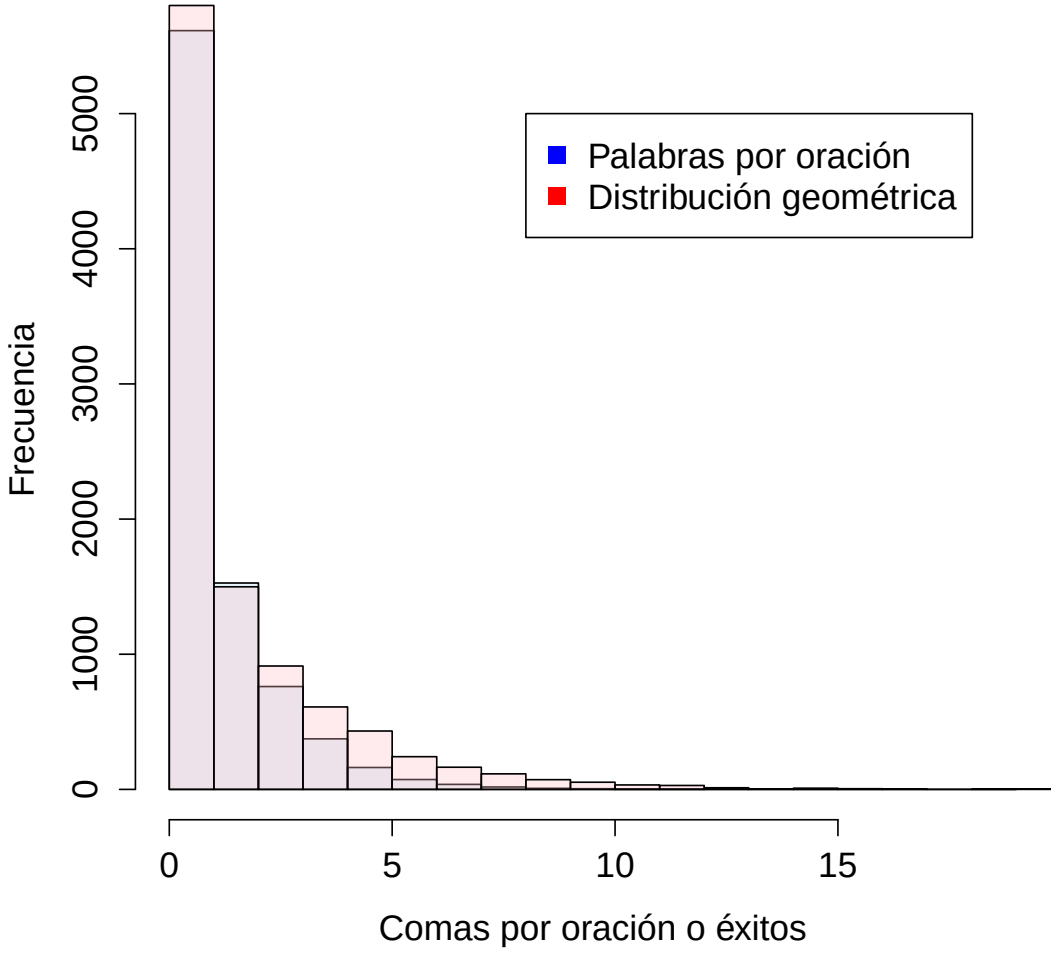

Figura 4.2: Histogramas de comas por oración de *Dracula* (azul) y función de distribución geométrica (rojo).