1. Describir los metadatos que ya tenemos

Número de columnas por periódico, por año.

Cantidad de columnas

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Diario | 2018 | 2019 | 2020 | Total |
| El Espectador | 3634 | 3048 | 3837 | 10519 |
| El Tiempo | 709 | 618 | 674 | 2001 |
| Semana | 822 | 0 | 0 | 822 |
| Total | 5165 | 3666 | 4511 | 13342 |

Cantidad de autores únicos

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Diario | 2018 | 2019 | 2020 | Total |
| El Espectador | 173 | 214 | 185 | 278 |
| El Tiempo | 150 | 138 | 141 | 245 |
| Semana | 67 | 0 | 0 | 67 |
| Total | 383 | 350 | 324 | 577 |

Palabras totales en la columna texto: 8376957.

Cantidad promedio de palabras por columna: 628.

Mínimo de palabras en una columna: 128.

Máximo de palabras en una columna: 5223.

Número de palabras diferentes en todos los textos: 336020.

1. Sobre los autores

¿Cuántos son? ¿Cuántas columnas por año escribe cada autor? ¿Proporción de hombres y mujeres? ¿Quiénes?

1. Depurar el corpus

Crear por diario y año que contengan

Nivel 1: Metadatos y cada columna en un archivo de .txt númerado

Ejemplo: Carpeta general y dentro tres carpetas 2018, 2019 y 2020.y en cada carpeta\_año (702 .txt) (625 .txt) (674 .txt) 🡪 nombrado T\_2018\_001 … / E\_2018\_001… / S\_2018\_001…

Nivel 2. Depurar los .txt de mayúsculas y signos de puntuación

Nivel 3. Depurar los .txt de las *stop words*

1. Describir el corpus