EDA - Deep Dive

México

Duplicados en conjunto

Por [title, seendate, domain]

Total --> 84 --> 0.09% del total (89,362)

Análisis de títulos

Por caracteres

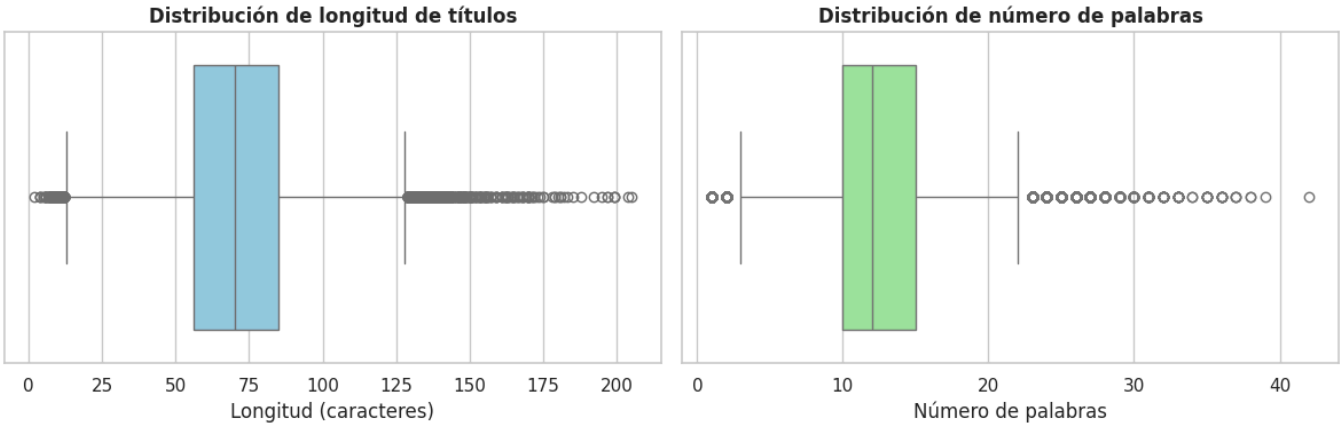
Métrica	Valor
Total	89,348
Media	70.51
Desviación estándar	23.18
Mínimo	2
25%	56
50%	70
75%	85
Máximo	205

Por palabras

Métrica	Valor
Total	89,348
Media	12.60
Desviación estándar	4.50
Mínimo	1
25%	10
50%	12
75%	15

Métrica	Valor
Máximo	42

Gráficas

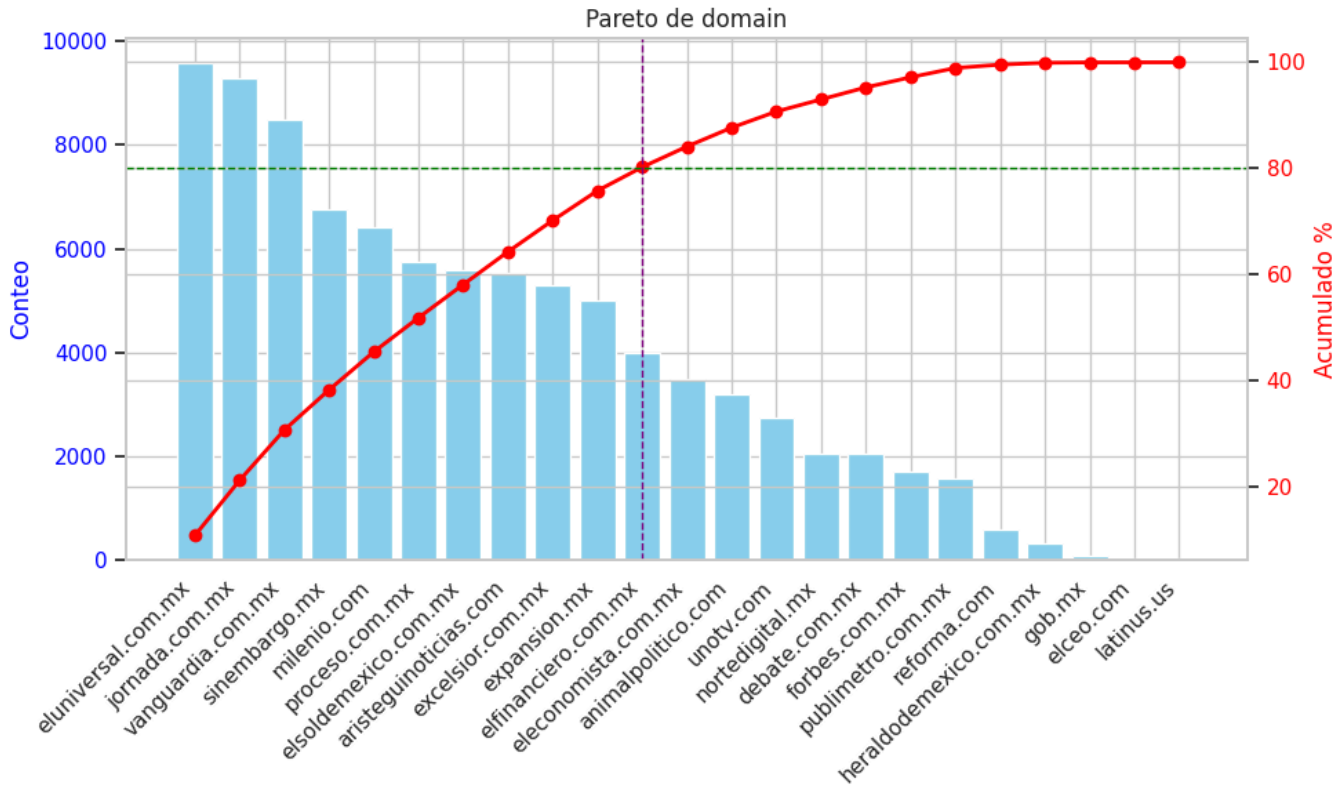


Los títulos largos, por lo que se puede ver son malas extracciones de scrapping, por ejemplo, el más largo se ve como

```
Stori lanza herramienta que permite al ahorro obtener el 15 % de GAT Nominal ¿ Cómo generar más rendimientos en tus ahorros ? Stori Cuenta+ te ayuda , el aliado para incrementar rendimientos en tus ahorros
```

Quizás funcionaria mantener los títulos a 1 std mayor y menor por palabras

Análisis de dominios



Análisis de n-gramas

- Antes de cualquier limpieza de texto, los **2-gramas** más comunes son **uniones de stopwords**, por ejemplo, "de la". Sin embargo, también se encuentran los siguientes boilerplates:
  - La jornada - Top 1
  - El financiero - Top 10
  - Forbes méxico
  - la inflación
- A partir de **3-gramas**, el conteo se repetición se va por debajo de las 700 apariciones, teniendo menos uniones de stopwords:
  - sol de méxico
  - el sol de
  - el sol de méxico
  - tasa de interés
  - precio del dólar
  - de ciudad juárez
  - norte de ciudad juárez
  - la jornada el
  - la jornada la
  - deportes gossip columnas

- noticias deportes gossip
- de méxico noticas

US

Duplicados en conjunto

```
Columnas evaluadas: ['title', 'seendate', 'domain']
Total de duplicados: 284
Proporción: 0.26% del total (107,432 filas)
```

Análisis de títulos

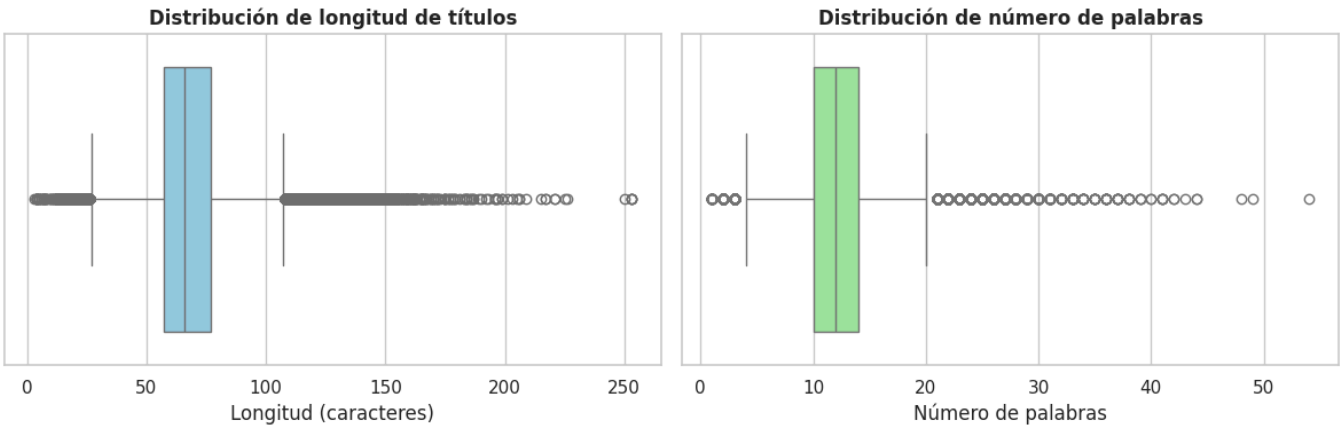
Por caracteres

```
count 107,420
** mean 67.69 **
** std 18.65 **
min 3
25% 57
50% 66
75% 77
max 253
```

Por palabras

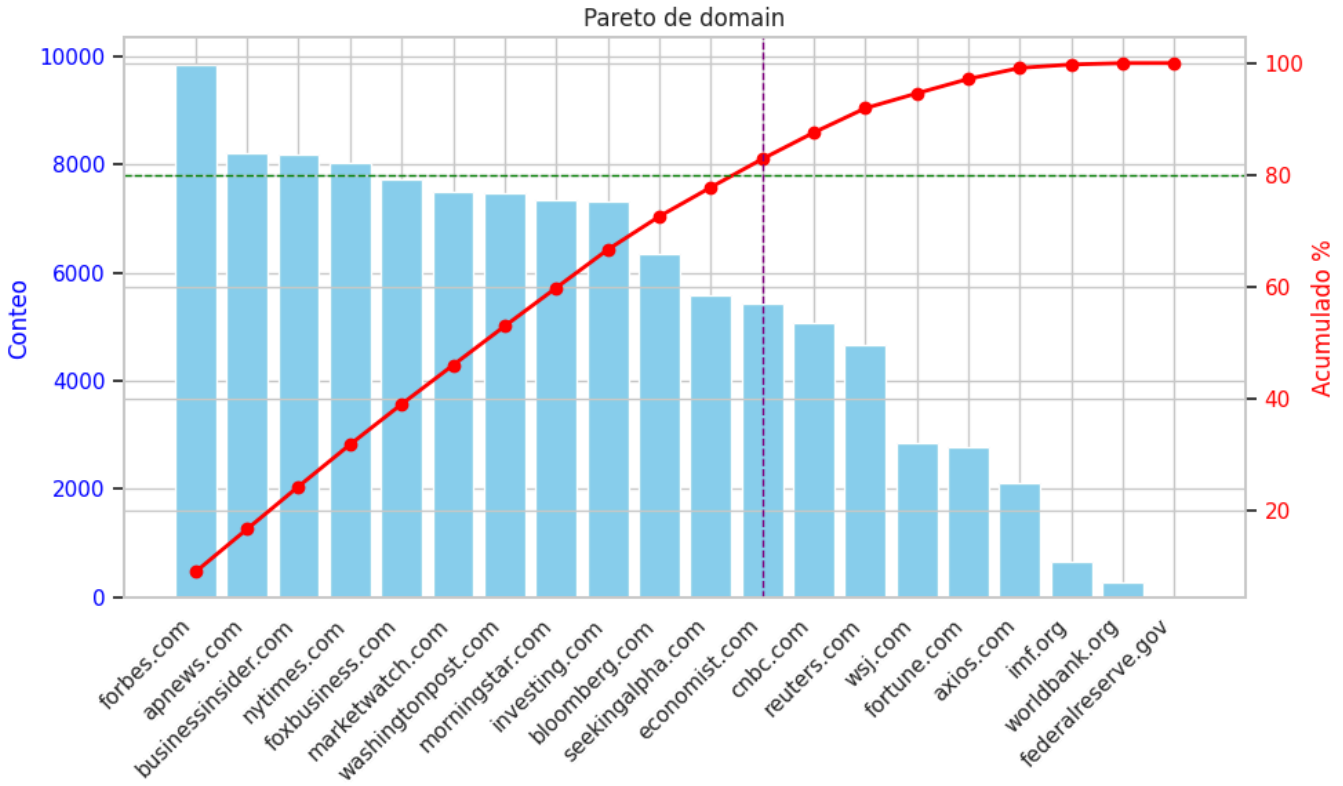
```
count 107,420
** mean 12.09 **
** std 3.72 **
min 1
25% 10
50% 12
75% 14
max 54
```

Gráficas




```
Argentina : First Review under the Stand – By Arrangement ; Inflation
Consultation ; Financing Assurances Review ; and Request for Rephasing ,
Augmentation , Waivers of Nonobservance and Applicability of Performance
Criteria , and Modification of Per...
```

Análisis de dominios



Análisis de n-gramas

```
================================================================
                              TOP 20 3-5 GRAMAS EN 'TITLE'
================================================================
                           ngram  count
               by investing com   1381
            the washington post   1050
                 new york times    824
                   the new york    815
              the new york times    802
        earnings call transcript    517
           results earnings call    418
results earnings call transcript    418
              stock market today    414
        european midday briefing    364
                the stock market    352
          north american morning    347
 north american morning briefing    347
       american morning briefing    347
                     of the day    330
           emea morning briefing    312
                     news of the    241
            news highlights top    238
                news of the day    238
                   need to know    225

----------------------------------------------------------------
EJEMPLOS DE N-GRAMAS MÁS FRECUENTES (3)
----------------------------------------------------------------
by investing com: 1381 veces
the washington post: 1050 veces
new york times: 824 veces
****************************************************************
```

```
============================================================
                              TOP 20 2-5 GRAMAS EN 'TITLE'
============================================================
              ngram  count
         by reuters   2632
              of the   1731
              in the   1584
      investing com   1462
       by investing   1384
   by investing com   1381
           new york   1363
       stock market   1341
           covid 19   1143
    washington post   1072
     the washington   1057
the washington post   1050
             the new   1043
              how to   1022
        wall street    997
             for the    979
              on the    958
            ahead of    941
          york times    824
     new york times    824

------------------------------------------------------------
EJEMPLOS DE N-GRAMAS MÁS FRECUENTES (3)
------------------------------------------------------------
by reuters: 2632 veces
of the: 1731 veces
in the: 1584 veces
************************************************************
```

Canadá

Duplicados en conjunto

```
Columnas evaluadas: ['title', 'seendate', 'domain']
Total de duplicados: 16
Proporción: 0.05% del total (29,774 filas)
```

Análisis de titulares

Por caracteres

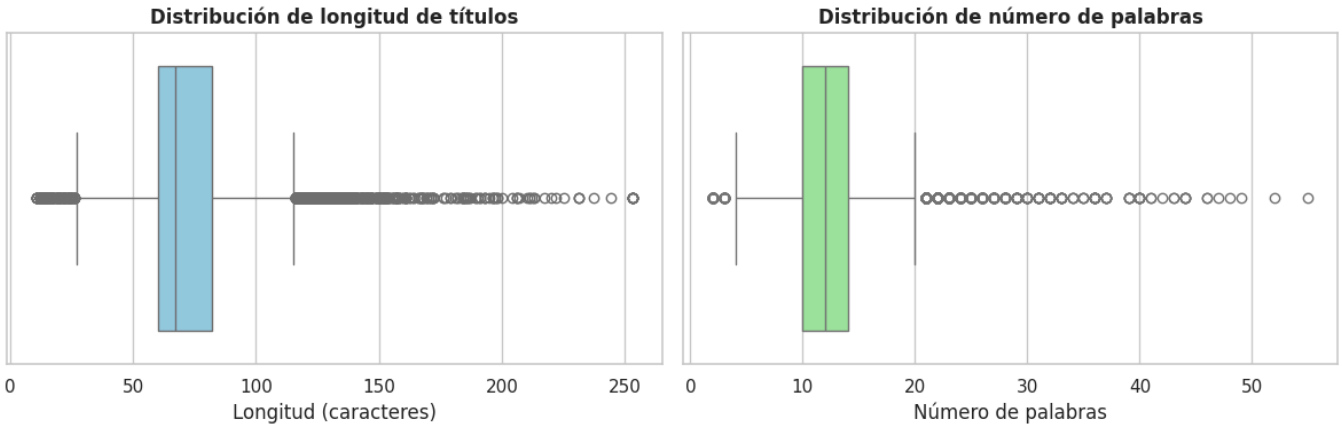
```
count 29,774
mean 71.748875
std 19.094297
min 11
```

```
25% 60
50% 67
75% 82
max 253
```

Por palabras

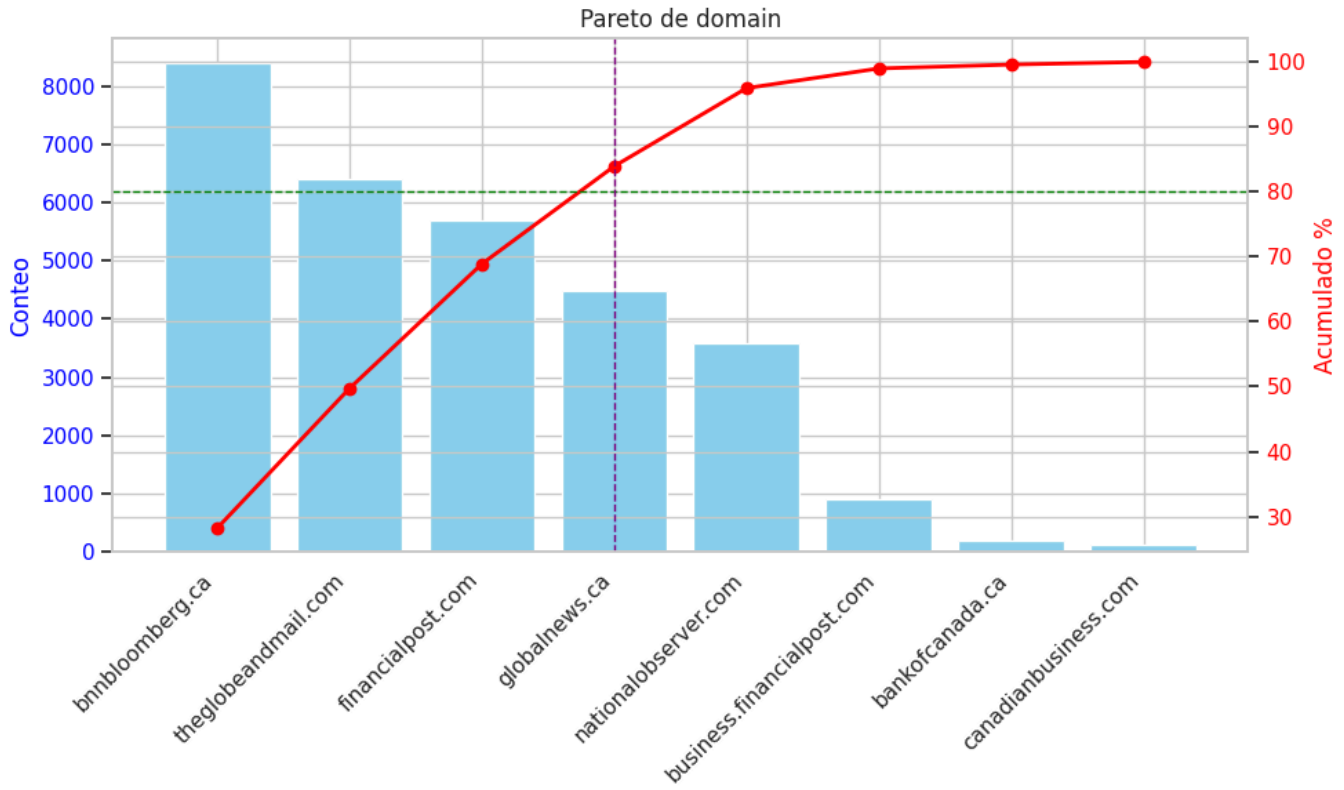
```
count 29774.00000
mean 12.54027
std 3.71308
min 2.00000
25% 10.00000
50% 12.00000
75% 14.00000
max 55.00000
```

Gráficas




```
Letters to the editor : Feb . 1 : Perhaps Conservatives should worry not
only if they can get elected without a viable climate platform , but whether
they can even get a date . How appealing are the Conservative leadership
contenders ? Plus other le...
```

Análisis de dominios



Análisis de n-gramas

```
============================================================
                          TOP 20 2-5 GRAMAS EN 'TITLE'
============================================================
                                ngram  count
                            of canada    546
                              bank of    525
                             covid 19    483
                        bank of canada    466
                       canada national    393
                     national observer    382
              canada national observer    374
                                in the    324
                         news analysis    311
        canada national observer news    311
               observer news analysis    311
                         observer news    311
       national observer news analysis    311
                national observer news    311
canada national observer news analysis    311
                                of the    296
                               for the    270
                            the globe    259
                               to the    253
                               on the    248


============================================================
                          TOP 20 3-5 GRAMAS EN 'TITLE'
============================================================
                                ngram  count
                        bank of canada    466
              canada national observer    374
                observer news analysis    311
                national observer news    311
         canada national observer news    311
       national observer news analysis    311
canada national observer news analysis    311
                          globe and mail    232
                     the globe and mail    230
                          the globe and    230
                      the daily chase      90
                letters to the editor      71
                       letters to the      71
                        to the editor      71
                     before the bell       65
                   stock market today      65
              upgrades and downgrades      64
       analyst upgrades and downgrades     64
                           the bank of     64
                 analyst upgrades and      64
```

China

Duplicados en conjunto

```
Columnas evaluadas: ['title', 'seendate', 'domain']
Total de duplicados: 12
Proporción: 0.07% del total (16,794 filas)
```

Análisis de titulares

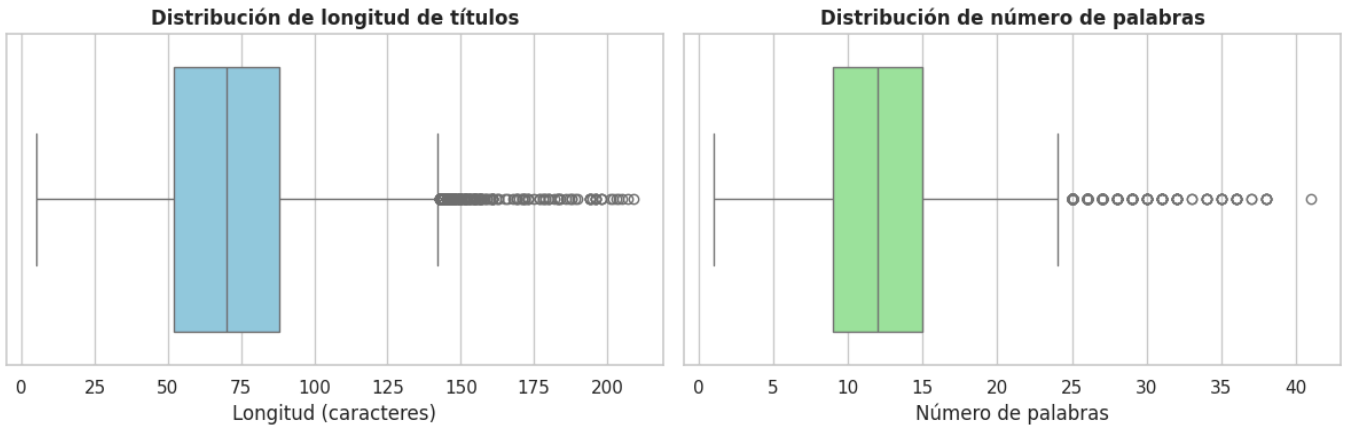
Por caracteres

```
count 16794.000000
mean 72.540669
std 26.823323
min 5.000000
25% 52.000000
50% 70.000000
75% 88.000000
max 209.000000
```

Por palabras

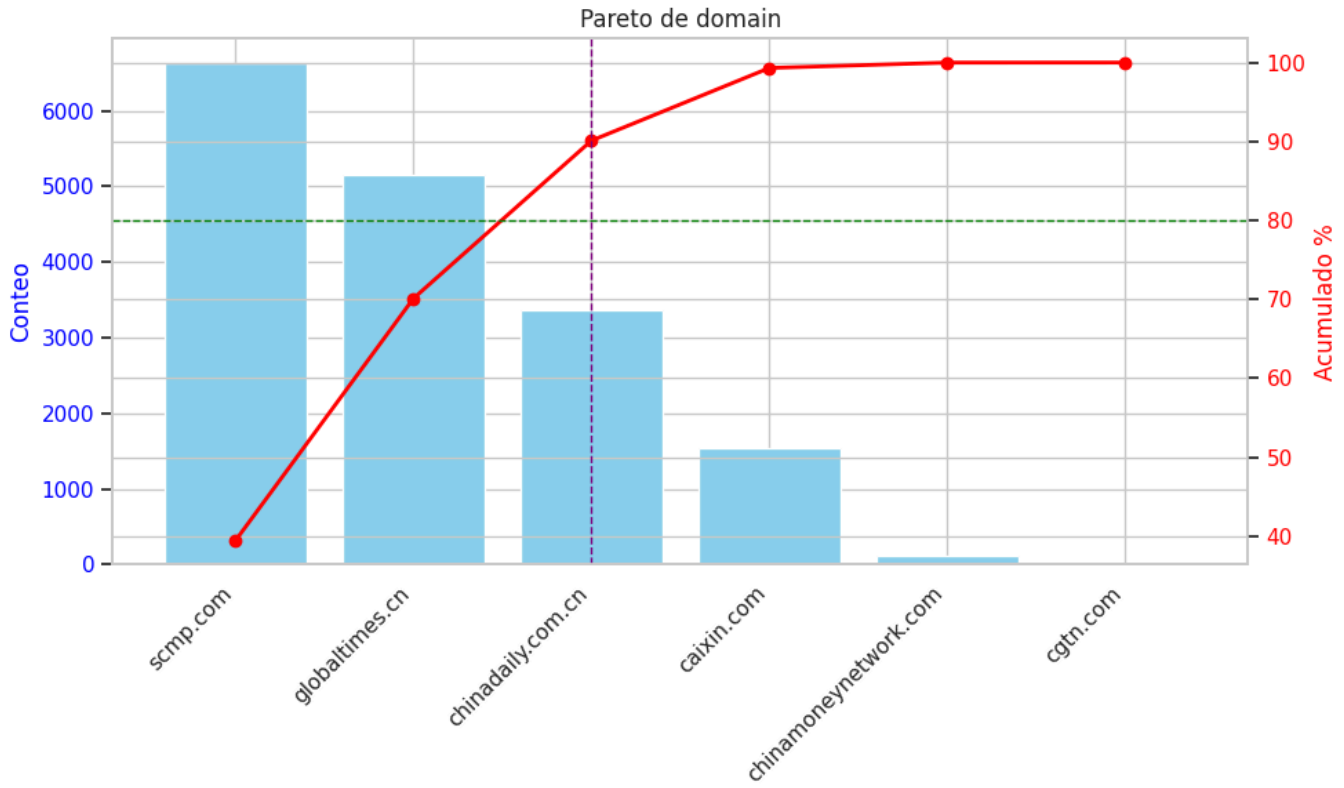
```
count 16794.000000
mean 12.153626
std 5.014964
min 1.000000
25% 9.000000
50% 12.000000
75% 15.000000
max 41.000000
```

Gráficas



```
Skincare brand Allies of Skin was inspired by a serious accident — now , it
drawing attention worldwide with its groundbreaking products using growth
factor serum , Singaporean founder Nicholas Travis explains
```

Análisis de dominios



Análisis de n-gramas

```
============================================================
                              TOP 20 2-5 GRAMAS EN 'TITLE'
============================================================
                    ngram  count
                hong kong   1901
            global times    461
             south china    400
           china morning    360
      china morning post    360
            morning post    360
south china morning post    360
     south china morning    360
        hong kong stocks    300
             kong stocks    300
                  com cn    279
          chinadaily com    279
       chinadaily com cn    279
                covid 19    268
                 财新网 财新网    216
                in china    200
                 in hong    189
            in hong kong    189
                gt voice    186
  global times editorial    176


============================================================
                              TOP 20 3-5 GRAMAS EN 'TITLE'
============================================================
                     ngram  count
south china morning post    360
      china morning post    360
     south china morning    360
        hong kong stocks    300
       chinadaily com cn    279
            in hong kong    189
  global times editorial    176
   national security law     99
         hong kong police     79
      hong kong protests     74
            of hong kong     53
           for hong kong     51
   china daily editorial     49
    world chinadaily com     48
 world chinadaily com cn     48
       china tech digest     45
     hong kong property     35
           latest on the     35
     latest on the novel     35
            on the novel     35
```

Brasil

Duplicados en conjunto

```
Columnas evaluadas: ['title', 'seendate', 'domain']
Total de duplicados: 122
Proporción: 0.21% del total (59,089 filas)
```

Análisis de titulares

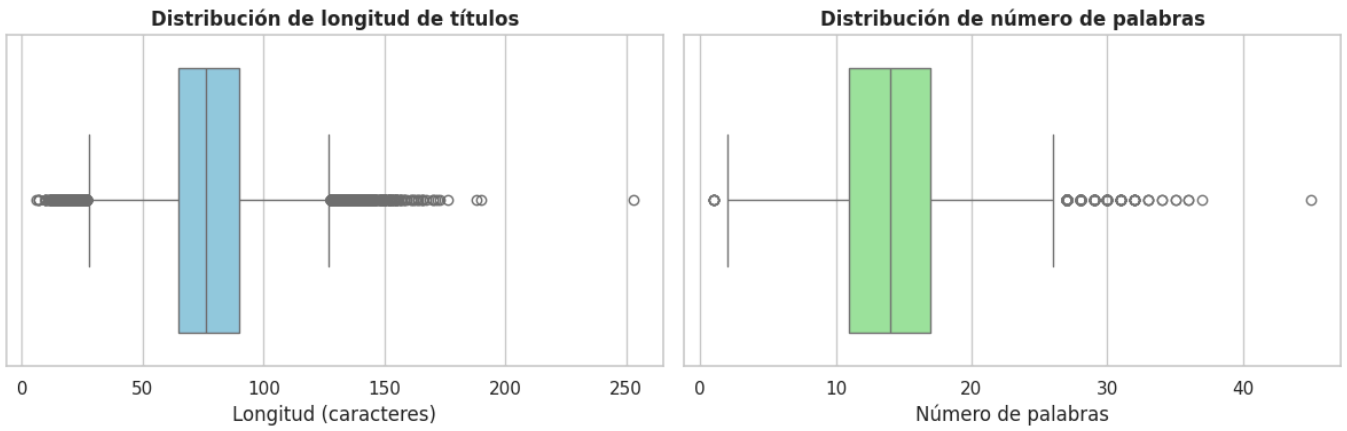
Por caracteres

```
count 59089.000000
mean 77.509841
std 20.195909
min 6.000000
25% 65.000000
50% 76.000000
75% 90.000000
max 253.000000
```

Por palabras

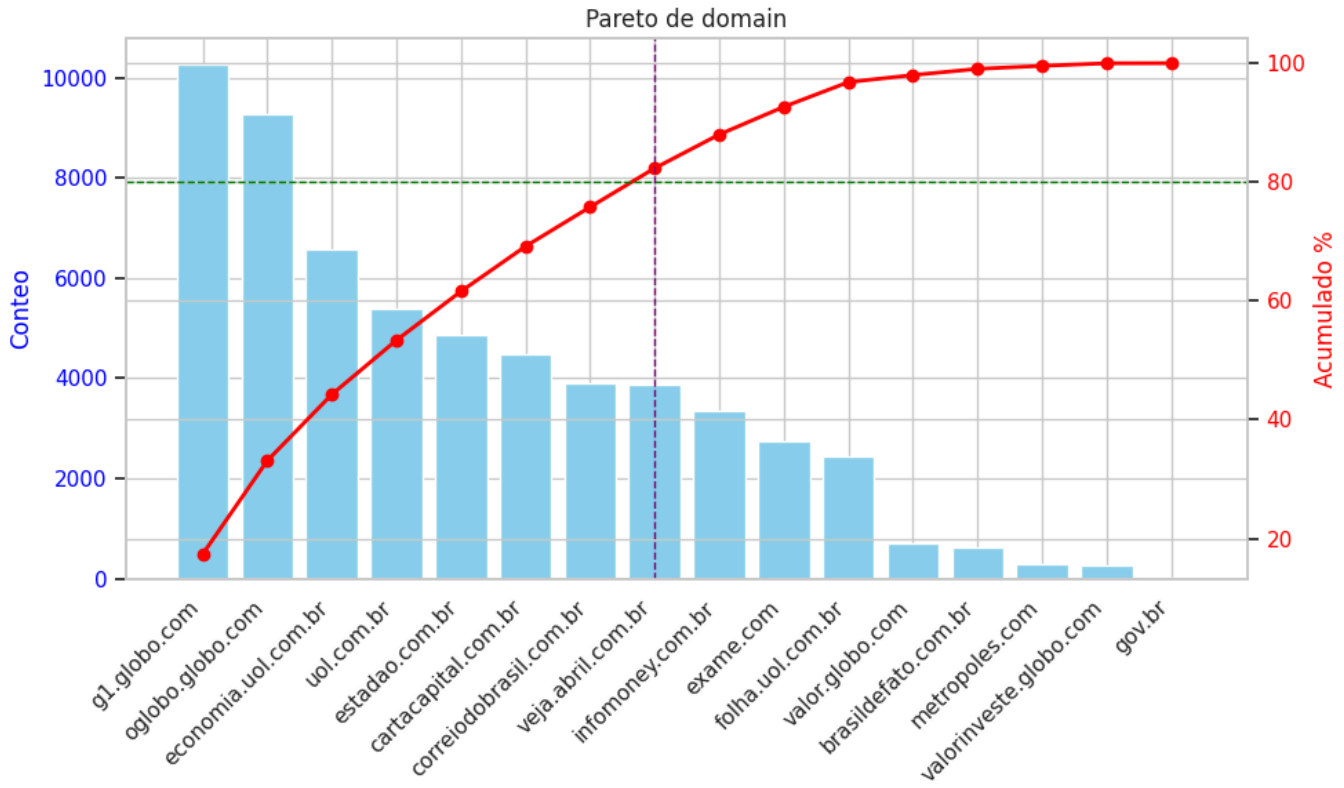
```
count 59089.000000
mean 14.097023
std 4.106178
min 1.000000
25% 11.000000
50% 14.000000
75% 17.000000
max 45.000000
```

Gráficas



```
A campanha do democrata Joe Biden reagiu à declarao do presidente Donald
Trump , após o republicano anunciar que irá pedir à Suprema Corte que a
contagem dos votos da eleio seja paralisada . Ultrajante , sem precedentes e
incorreto , disse a chefe d...
```

Análisis de dominios



Análisis de n-gramas

```
================================================================
                                TOP 20 3-5 GRAMAS EN 'TITLE'
================================================================
              ngram  count
  correio do brasil   1385
      taxa de juros    306
   vagas de emprego    202
       em são paulo    190
   do banco central    158
    contra covid 19    155
   imposto de renda    148
       de são paulo    131
        diz que não    130
lavagem de dinheiro    129
     rio de janeiro    121
  bolsonaro diz que    119
  guerra na ucrânia    111
ministério da saúde    101
       da lava jato    100
  pela primeira vez     99
         de mais de     99
      opera em alta     99
    por suspeita de     94
     anos de prisão     93

================================================================
                                TOP 20 2-5 GRAMAS EN ''
================================================================
             ngram  count
        do brasil   1783
          diz que   1560
       correio do   1385
correio do brasil   1385
          nos eua    982
        no brasil    887
          mais de    868
     de bolsonaro    864
          dos eua    853
          por que    812
         de juros    723
         covid 19    604
          de lula    566
         de trump    520
       do governo    487
            de sp    467
           do rio    446
     2025 mercado    443
    presidente do    435
    banco central    425
```