

Contar el número de palabras que hay en un libro con PowerShell

Leyendo por Internet nadie se pone de acuerdo en el número de palabras que hay en El Quijote, con el siguiente código vamos a comprobar:

- Número de palabras que hay en total aunque se repitan
- Número de palabras que hay sin que se repitan
- Palabras más frecuentes

Para contabilizar una palabra, hemos quitado signos de puntuación y las palabras que contienen una única letra (primer script) y las que son de dos letras o menos (segundo script).

Primer script quitando las palabras de una letra

[crayon-6a9d836140f17978477515/]

```

1 cd ~
2
3 # Quitar signos de puntuación y similar
4 $patron = '[^a-zA-Zá-úÁ-Úäöü\']'
5
6 $total = ""
7 $total = foreach($linea in gc .\el_quijote.txt -Encoding UTF8)
8 {
9     $partir = $linea.split(" ") -replace $patron, ''
10    foreach($palabras in $partir)
11    {
12        # Palabra que tengan más de un carácter
13        if($palabras.Length -gt 1)
14        {
15            $palabras
16        }
17    }
18 }
19
20 # Contar todas las palabras que tengan más de un carácter
21 write-Host "Total de palabras que tienen más de un carácter: " $total.Count
22
23 # Contar las palabras agrupadas
24 write-Host "Total de palabras agrupadas:" ($total | Group-Object | Sort-Object Name -Descending -CaseSensitive).Count
25
26 # 10 palabras más frecuentes
27 write-Host "5 palabras más frecuentes"
28 $total | Group-Object | Sort-Object count -Descending -CaseSensitive | select -First 5

```

```

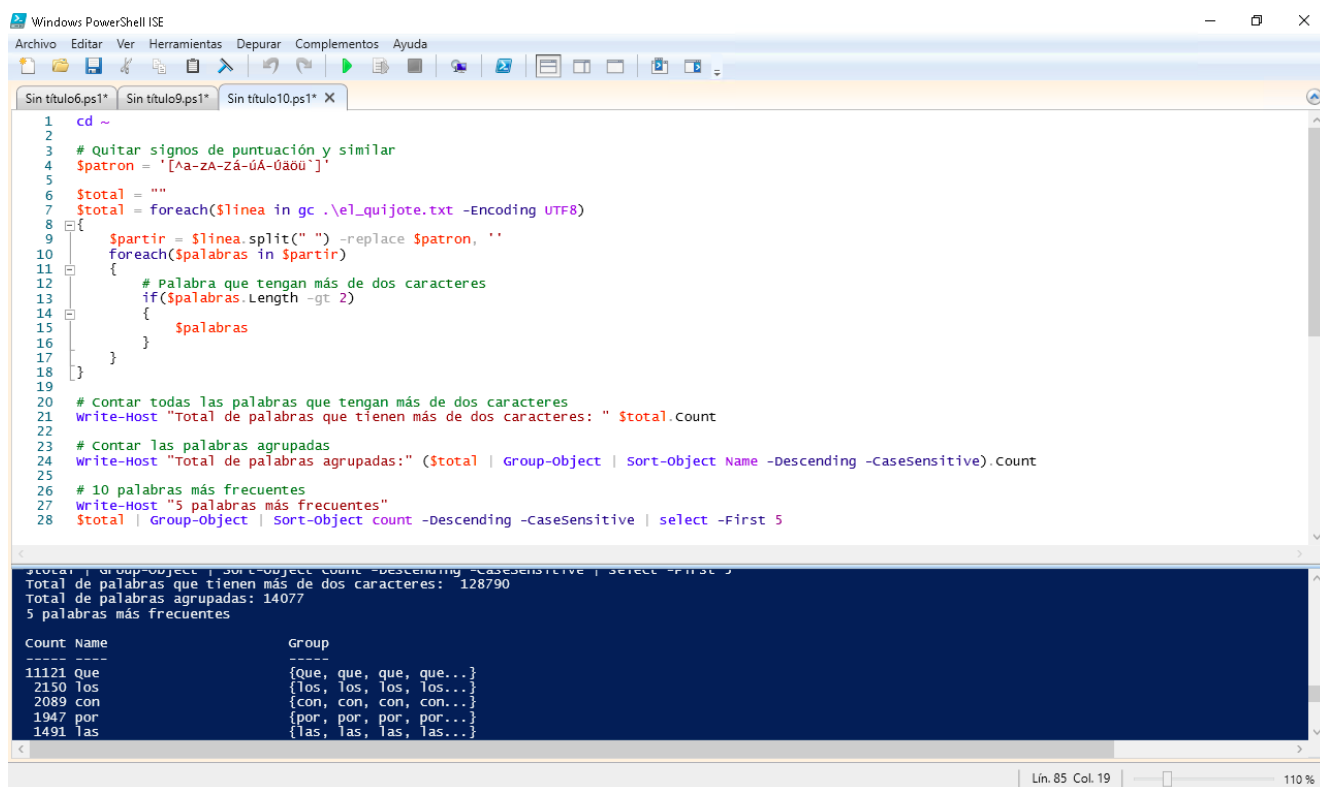
Total de palabras que tienen más de un carácter: 172843
Total de palabras agrupadas: 14129
5 palabras más frecuentes

Count Name                Group
-----
11121 que                  {que, que, que, que...}
9082  de                   {de, de, de, de...}
5015  la                   {la, la, la, la...}
4607  el                   {el, el, el, el...}
4050  en                   {en, en, en, en...}

```

Segundo script quitando las palabras de dos letras o menos

[crayon-6a9d836140f1e572186669/]



```
1 cd ~
2
3 # quitar signos de puntuación y similar
4 $patron = '[^a-zA-Zá-úÁ-0à0ü]'
```

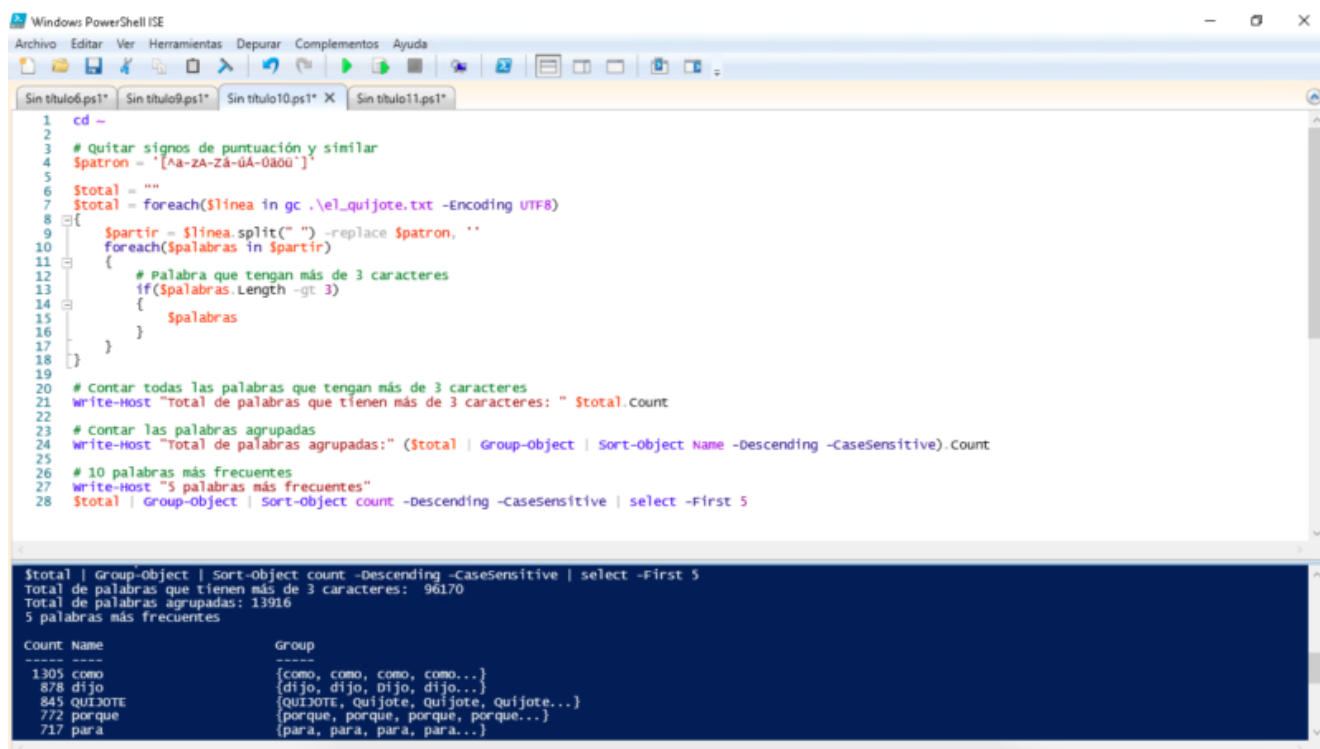
```
6 $total = ""
7 $total = foreach($linea in gc .\el_quijote.txt -Encoding UTF8)
8 {
9     $partir = $linea.split(" ") -replace $patron, ''
10     foreach($palabras in $partir)
11     {
12         # Palabra que tengan más de dos caracteres
13         if($palabras.Length -gt 2)
14         {
15             $palabras
16         }
17     }
18 }
19
20 # Contar todas las palabras que tengan más de dos caracteres
21 Write-Host "Total de palabras que tienen más de dos caracteres: " $total.Count
22
23 # Contar las palabras agrupadas
24 Write-Host "Total de palabras agrupadas:" ($total | Group-Object | Sort-Object Name -Descending -CaseSensitive).Count
25
26 # 10 palabras más frecuentes
27 Write-Host "5 palabras más frecuentes"
28 $total | Group-Object | Sort-Object count -Descending -CaseSensitive | select -First 5
```

```
$total | Group-Object | Sort-Object count -Descending -CaseSensitive | select -First 5
Total de palabras que tienen más de dos caracteres: 128790
Total de palabras agrupadas: 14077
5 palabras más frecuentes
```

Count	Name	Group
11121	que	{que, que, que, que...}
2150	los	{los, los, los, los...}
2089	con	{con, con, con, con...}
1947	por	{por, por, por, por...}
1491	las	{las, las, las, las...}

Tercer script quitando las palabras de tres letras o menos

[crayon-6a9d836140f20187742528/]



```
1 cd ~
2
3 # quitar signos de puntuación y similar
4 $patron = '[^a-zA-Zá-úÁ-0à0ü]'
```

```
6 $total = ""
7 $total = foreach($linea in gc .\el_quijote.txt -Encoding UTF8)
8 {
9     $partir = $linea.split(" ") -replace $patron, ''
10     foreach($palabras in $partir)
11     {
12         # Palabra que tengan más de 3 caracteres
13         if($palabras.Length -gt 3)
14         {
15             $palabras
16         }
17     }
18 }
19
20 # Contar todas las palabras que tengan más de 3 caracteres
21 Write-Host "Total de palabras que tienen más de 3 caracteres: " $total.Count
22
23 # Contar las palabras agrupadas
24 Write-Host "Total de palabras agrupadas:" ($total | Group-Object | Sort-Object Name -Descending -CaseSensitive).Count
25
26 # 10 palabras más frecuentes
27 Write-Host "5 palabras más frecuentes"
28 $total | Group-Object | Sort-Object count -Descending -CaseSensitive | select -First 5
```

```
$total | Group-Object | Sort-Object count -Descending -CaseSensitive | select -First 5
Total de palabras que tienen más de 3 caracteres: 96170
Total de palabras agrupadas: 13916
5 palabras más frecuentes
```

Count	Name	Group
1305	como	{como, como, como, como...}
878	dijo	{dijo, dijo, dijo, dijo...}
845	QUIJOTE	{QUIJOTE, Quijote, Quijote, Quijote...}
772	porque	{porque, porque, porque, porque...}
717	para	{para, para, para, para...}