



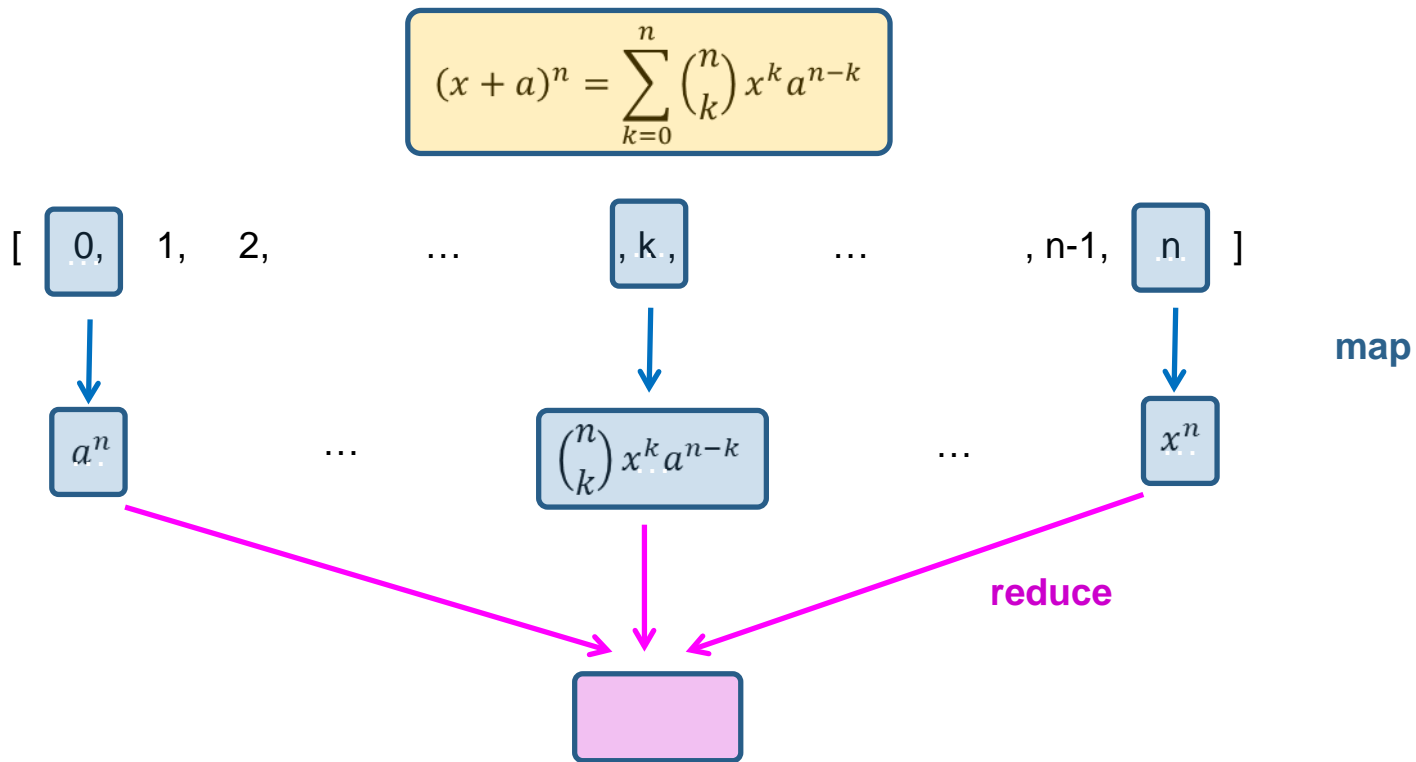
Programación. Python

MapReduce



- Procesamiento de big data en el sistema de archivos distribuido *hadoop* mediante un esquema funcional a base de las operaciones "map" y "reduce".

MapReduce, ejemplo 1



$\text{índices} = \text{range}(n+1) = [0, \dots, n]$

$\text{términos} = \text{map}(f, \text{índices})$, donde $f(k) = \binom{n}{k} x^k a^{n-k}$

$\text{suma} = \text{reduce}(\text{suma}, \text{términos})$, donde $\text{suma}(a, b) = a + b$

MapReduce, ejemplo 2

"My dear Mr. Bennet," said his lady to him one day, "have you heard that Netherfield Park is let at last?"

...

Now, Kitty, you may cough as much as you choose," said Mr. Bennet.

...

Elizabeth Bennet had been obliged, by the scarcity of gentlemen, to sit down for two dances; and during part of that time, Mr. Darcy had been standing near enough for her to hear a conversation between him and Mr. Bingley, who came from the dance for a few minutes, to press his friend to join it.

[["Mr", "Bennet"], 1
["Mr", "Netherfield"], 1
["Mr", "Park"], 1
["Bennet", "Netherfield"], 1
["Bennet", "Park"], 1
["Netherfield", "Park"], 1]

[["Kitty", "Mr", "Bennet"]]

[["Elizabeth", "Bennet", "Mr", "Darcy", "Mr", "Bingley"]]

map

...

...

reduce

[["Mr", "Bennet"], 1173
["Mr", "Netherfield"], 2001
["Mr", "Park"], 1200
["Bennet", "Netherfield"], 541
["Bennet", "Park"], 258
["Netherfield", "Park"], 785
]

Map y reduce

INPUT

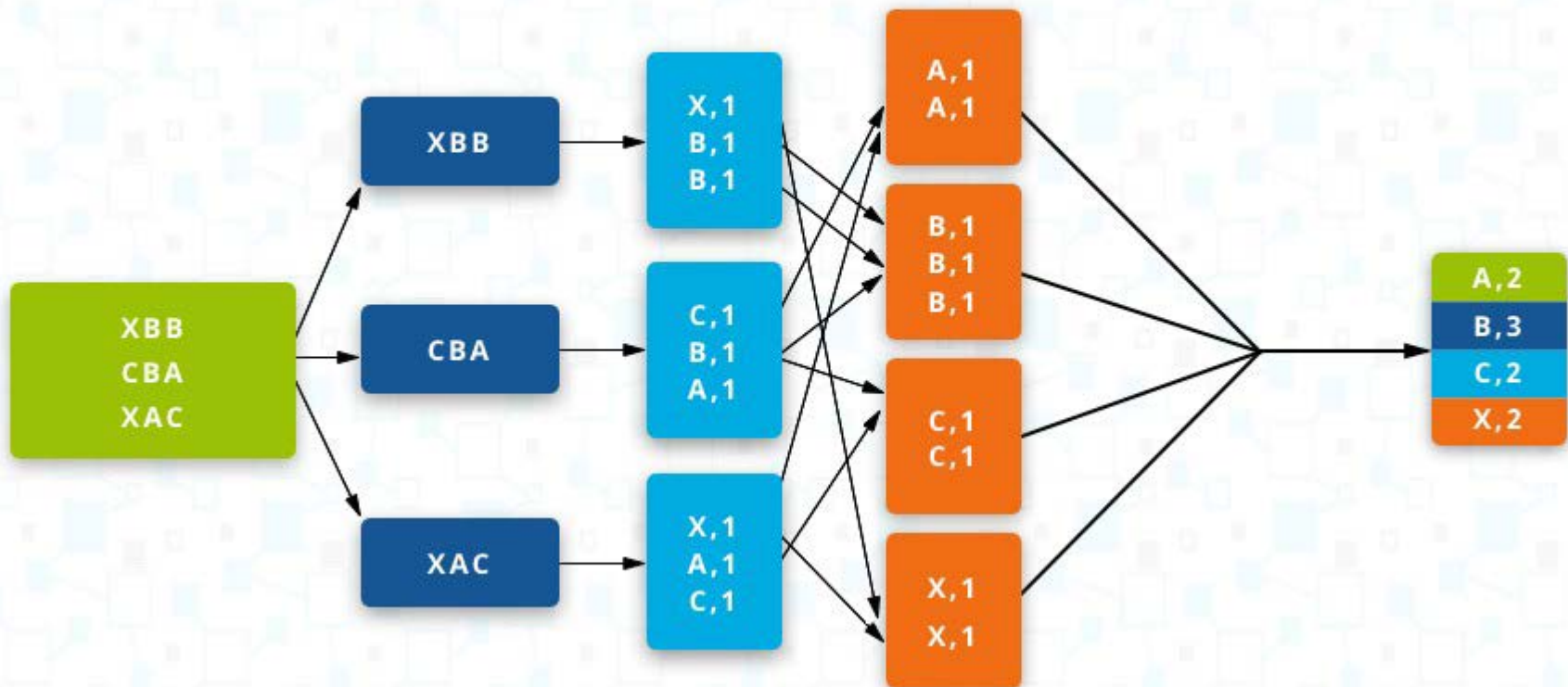
SPLIT

MAP

COMBINE









PARTITION

REDUCE



Cuenta-palabras secuencial

```
a 1
aros 3
cociente 2
cuadrados 2
daros 1
de 2
definible 2
diametral 1
inmedible 1
los 3
mi 1
nombre 1
que 1
redondos 1
seré 2
siempre 2
soy 3
también 1
tengo 1
todos 2
y 4
```

	1-wordcount.py	28/12/2018 11:30	Archivo PY	3 KB
	2-wordcount.py	19/04/2016 18:27	Archivo PY	1 KB
<input checked="" type="checkbox"/> 	Command Prompt	05/03/2016 13:51	Acceso directo	2 KB
	map-reduce.pptx	28/12/2018 11:36	Presentación de ...	2.271 KB
	palabras.txt	28/12/2018 11:42	Documento de tex...	1 KB
	pride_and_prejudice.txt	24/03/2016 11:29	Documento de tex...	701 KB
	promptMR	11/05/2016 13:32	Acceso directo	2 KB
	quijote.txt	13/04/2016 13:24	Documento de tex...	2.105 KB

Objetivo

palabras.txt: Bloc de notas

Archivo Edición Formato Ver Ayuda

Soy y seré a todos definible
mi nombre tengo que daros
cociente diametral siempre inmedible
soy de los redondos aros

y seré también todos los aros cuadrados
y soy definible y cociente siempre de lo

Cuenta-palabras secuencial

Command Prompt

```
C:\Users\CPAREJA\Documents\docencia\Taller de algoritmos\05 - Estudio de casos\map-reduce>python 1-wordcount.py --count palabras.txt
```

1-wordcount.py 28/12/2018 11:30 Archivo PY 3 KB
2-wordcount.py 19/04/2016 18:27 Archivo PY 1 KB
☒ Command Prompt 05/03/2016 13:51 Acceso directo 2 KB
map-reduce.pptx 28/12/2018 11:36 Presentación de ... 2.271 KB
palabras.txt 28/12/2018 11:42 Documento de tex... 1 KB
pride_and_prejudice.txt 24/03/2016 11:29 Documento de tex... 701 KB
promptMR 11/05/2016 13:32 Acceso directo 2 KB
quijote.txt 13/04/2016 13:24 Documento de tex... 2.105 KB

C:\Users\CPAREJA\Documents\docencia\Taller de algoritmos\05 - Estudio de casos\map-reduce>

palabras.txt: Bloc de notas

Archivo Edición Formato Ver Ayuda

Soy y seré a todos definible
mi nombre tengo que daros
cociente diametral siempre inmedible
soy de los redondos aros

y seré también todos los aros cuadrados
y soy definible y cociente siempre de lo

Cuenta-palabras secuencial

Command Prompt

```
C:\Users\CPAREJA\Documents\docencia\Taller de algoritmos\05 - Estudio de casos\map-reduce>python 1-wordcount.py --count palabras.txt
```

```
a 1
aros 3
cociente 2
cuadrados 2
daros 1
de 2
definible 2
diametral 1
inedible 1
los 3
mi 1
nombre 1
que 1
redondos 1
seré 2
siempre 2
soy 3
también 1
tengo 1
todos 2
y 4
```

1-wordcount.py

28/12/2018 11:30

Archivo PY

3 KB

2-wordcount.py

19/04/2016 18:27

Archivo PY

1 KB

☒ Command Prompt

05/03/2016 13:51

Acceso directo

2 KB

map-reduce.pptx

palabras.txt

pride_and_prejudice

promptMR

quijote.txt

```
4
5 import sys
6
7 def freq(x,xx):
8     # num. de veces que x is in xx
9     n = 0
10    for a in xx:
11        if a == x:
12            n = n+1
13    return n
14
15 def list_words(filename):
16     words = []
17     f = open(filename,'r')
18     for line in f:
19         for w in line.split():
20             words.append(w)
21     words = list(map(lambda x : x.lower(),words))
22     words.sort()
23     wordsFreq = []
24     while words != []:
25         w = words[0]
26         n = freq(w,words)
27         wordsFreq.append((w,n))
28         words = list(filter(lambda e: e != w , words))
29     return wordsFreq
30
31 def print_words(filename):
32     for (x,y) in list_words(filename):
33         print(x,y)
```

```
2.2 # This basic command line argument parsing code
2.3 # calls the print_words() and print_top() functions
2.4
2.5 def main():
2.6     if len(sys.argv) != 3:
2.7         print("uso: ./wordcount.py {--count | --top | --numlines | --numPalabras | --palabrasConSuLong}")
2.8         sys.exit(1)
2.9
2.10    option = sys.argv[1]
2.11    filename = sys.argv[2]
2.12    if option == '--count':
2.13        print_words(filename)
2.14    elif option == '--topcount':
2.15        print_top(filename)
2.16    elif option == '--numlines':
2.17        print(num_lines(filename))
2.18    elif option == '--numPalabras':
2.19        print(num_palabras(filename))
2.20    elif option == '--palabrasConSuLong':
2.21        print(palabrasConSuLong(filename))
2.22    else:
2.23        print('unknown option: ' + option)
2.24        sys.exit(1)
2.25
2.26 if __name__ == '__main__':
2.27     main()
```

palabras.txt: Bloc de notas

Archivo Edición Formato Ver Ayuda

Soy y seré a todos definible
mi nombre tengo que daros
cociente diametral siempre inedible
soy de los redondos aros

y seré también todos los aros cuadr
y soy definible y cociente siempre

Cuenta-palabras con map-reduce

```
"Soy" 1
"a" 1
"aros" 3
"cociente" 2
"cuadrados" 2
"daros" 1
"de" 2
"definible" 2
"diametral" 1
"inedible" 1
"los" 3
"mi" 1
"nombre" 1
"que" 1
"redondos" 1
"ser\u00e9" 2
"siempre" 2
"soy" 2
"tambi\u00e9n" 1
"tengo" 1
"todos" 2
"y" 4
```

```
1 from mrjob.job import MRJob
2
3 class MRCharCount(MRJob):
4
5     def mapper(self, _, line):
6         for w in line.split():
7             yield w, 1
8
9     def reducer(self, key, values):
10        yield key, sum(values)
11
12 if __name__ == '__main__':
13     MRCharCount.run()
```

palabras.txt: Bloc de notas

Archivo Edición Formato Ver Ayuda

Soy y seré a todos definible
mi nombre tengo que daros
cociente diametral siempre inedible
soy de los redondos aros

y seré también todos los aros cuadrados
y soy definible y cociente siempre de lo

Cuenta-palabras con map-reduce

Administrador: promptMR

Microsoft Windows [Versión 10.0.18362.267]
(c) 2019 Microsoft Corporation. Todos los derechos reservados.

C:\WINDOWS\system32>cd "C:\Users\CPAREJA\Documents\docencia\Taller de algoritmos\jupyter\E7 - map-reduce"

C:\Users\CPAREJA\Documents\docencia\Taller de algoritmos\jupyter\E7 - map-reduce>python 2-wordcount.py palabras.txt

No configs found; falling back on auto-configuration
No configs specified for ini
Creating temp directory C:\US
Running step 1 of 1...
job output is in C:\Users\CPA
Streaming final output from C

Soy" 1
"a" 1
"aros" 3
"cociente" 2
"cuadrados" 2
"daros" 1
"de" 2
"definible" 2
"diametral" 1
"inedible" 1
"los" 3
"mi" 1
"nombre" 1
"que" 1
"redondos" 1
"ser\u00e9" 2
"siempre" 2
"soy" 2
"tambi\u00e9n" 1
"tengo" 1
"todos" 2
"y" 4

1 from mrjob.job import MRJob
2
3 class MRCharCount(MRJob):
4
5 def mapper(self, _, line):
6 for w in line.split():
7 yield w, 1
8
9 def reducer(self, key, values):
10 yield key, sum(values)
11
12 if __name__ == '__main__':
13 MRCharCount.run()

Propiedades: promptMR

1-wordcount.py 28/12/2018 11:30
2-wordcount.py 19/04/2016 18:27
Command Prompt 05/03/2016 13:51
map-reduce.pptx 28/12/2018 11:36
palabras.txt 28/12/2018 11:42
pride_and_prejudice.txt
promptMR
quijote.txt 13/04/2016 13:24

Ejecutar como administrador

Esta opción le permite ejecutar este acceso directo como administrador, al tiempo que el equipo se protege contra cualquier actividad no autorizada.

Opciones avanzadas...

Acceso directo

palabras.txt: Bloc de notas

Archivo Edición Formato Ver Ayuda

Soy y seré a todos definible
mi nombre tengo que daros
cociente diametral siempre inmedible
soy de los redondos aros

y seré también todos los aros cuadrados
y soy definible y cociente siempre de lo

Ejercicio propuesto

```
"""
Enunciado:

Tenemos un archivo de texto plano (ej.: "pride_and_prejudice.txt")
No tiene formato de líneas; esto es, cada párrafo está en una única línea.

Se plantea diseñar un programa que contabiliza cada par de palabras
con mayúscula que aparecen en un mismo párrafo,
con la esperanza de que esta contabilidad dé la relación entre los personajes
de una obra literaria...

Hazlo usando la técnica de map-reduce,
para que nos sirva para obras de gran tamaño.
```

El programa se usará así:

```
"""
>>> cuenta_pares.py < pride_and_prejudice.txt > resultado.txt
"""
```

Una solución con map-reduce:

```
from mrjob.job import MRJob
```

```
class MRCharCount(MRJob):
```

```
    def mapper(self, _, line):
        sin_valor = ["a", "after", "all", "and", "as",
                     "away", "but", "for", "if", "in",
                     # etcétera
                     "the", "this", "these", "of", "on"]
```

Se pide completar esto

```
    def reducer(self, key, values):
        yield key, sum(values)

if __name__ == '__main__':
    MRCharCount.run()
```

Administrador: promptMR

```
C:\Users\CPAREJA\Documents\investigación\Investigación - activa\2016 - Análisis de datos\quijote> cuenta_pares.py < pride_and_prejudice.txt > resultado.txt
```

```
Creating temp directory C:\Users\CPAREJA\AppData\Local\Temp\cuenta_pares.CPAREJA.20190122.120244.638949
Running step 1 of 1...
reading from STDIN
Streaming final output from C:\Users\CPAREJA\AppData\Local\Temp\cuenta_pares.CPAREJA.20190122.120244.638949\output...
Removing temp directory C:\Users\CPAREJA\AppData\Local\Temp\cuenta_pares.CPAREJA.20190122.120244.638949...
```

"M," 1["Fitzwilliam.", "He"] 1["Fitzwill", "m"], 1["ive", "m"], 1["ive", "m"], 1["For,", "I"] 1["Fordyce's", "Lydia"] 1["Fordyce's", "Gute"] 6["Foundation", "Literary"] 8["Foundation", "Michael"] 1["Foundation", "Project"] 5["Foundation", "Section"] 1["Foundation", "States"] 1["Foundation", "To"] From", 3["From", "Pemberley"] 1["FULL", "GUTENBERG"] 1["Full", "Gutenberg\")."] 1["FULL", "LICENSE"] 2["FULL", "PROJECT"] 1["Full", "Project"] 1["FULL", "START:"] 1["Ge rdiner.", "You"] 1["Gardiner;", "Mrs."] 1["Gardiner?", "He"] 1["Gardiner?", "Mr."] 1["Gardiniers", "Jane;"] 1["Gardiniers", "Longbourn;"] 1["Gardiniers", "IVE"] 1["INCIDENTAL"] 1["GIVE", "NOTICE"] 1["GIVE", "POSSIBILITY"] 1["GIVE", "SUCH"] 1["Give", "What"] 1["GIVE", "YOU"] 1["God!", "I"] 1["God!", "Thank"] 1["Gc el"] 1["Gutenberg-tm", "Mission"] 1["Gutenberg-tm", "Nearly"] 1["Gutenberg-tm", "Professor"] 1["Gutenberg-tm", "Project"] 11["GUTENBERG-tm", "PROJECT"] 1["Gutenberg-tm", "Project"] ["Gutenbe", "To"] 1["Gutenberg", "Web"] 1["Gutenberg:", "Project"] 1["Gutenberg:", "Unless"] 1["Gutenberg?"),", "Project"] 1["Had", "He"] 1["Had", "I"] 7["Had", "Jane"] 1["He", "London;"] 1["He", "London?")"] 1["He", "Lydia"] 1["He", "Lydia."] 1["He", "Make"] 1["He", "Mary"] 1["He", "Miss"] 3["He", "Mr."] 16["He", "Mrs."] 3["He", "Meth Sir"] 1["Her", "They"] 1["Her", "William"] 1["Her", "With"] 1["Here", "Hill!"] 1["Here", "Hill;"] 1["Here", "I"] 1["Here", "Jane."] 1["Here", "Lady"] 1["He ", "Jane"] 1["His", "King"] 1["His", "Lucas"] 1["His", "Lydia."] 1["His", "Miss"] 4["His", "Mr."] 5["His", "Mrs."] 1["His", "Oh!"] 1["His", "Oh;"] 1["His", "Par nsford;"] 1["hen"] 1["Hunsford.", "William"] 1["Hunsford.", "Pray"] 1["Hunsford.", "She"] 1["Hunsford.", "Sir"] 1["Hunsford.", "Such"] 1["Hunsford.", "William"] 1["Hunsford; "] 1["I", "London"] 1["I", "I", "London."] 3["I", "London?"] 1["I", "Long"] 2["I", "Long;"] 1["I", "Long."] 1["I", "Longbourn"] 2["I", "Nor"] 1["I", "Nothing"] 1["I", "Noven 1["I", "Newcastle."] 1["I", "I", "Next"] 2["I", "Nicholls"] 1["I", "No;"] 4["I", "Nobody"] 2["I", "Nor"] 2["I", "Not"] 7["I", "Nothing"] 1["I", "Noven 1["I", "There"] 7["I", "I"] 13["I", "Thoug"] 2["I", "Thoughtless"] 1["I", "Till"] 5["I", "Times"] 1["I", "To"] 5["I", "Towards"] 2["I", "Tuesday;"] 1["I", "Tuesday."] "SUCH"] 1["INCIDENTAL", "YOU"] 1["INCLUDING", "KIND;"] 1["INCLUDING", "LIMITED"] 1["INCLUDING", "NOT"] 1["INCLUDING", "TO"] 1["INCLUDING", "WARRANTIES"] 1["Indeed", "You"] 1["Ir , "Lizzy!"] 1["Jane", "Lizzy;"] 1["Jane", "Lizzy?"] 1["Jane", "London;"] 1["Jane", "London."] 1["Jane", "Longbourn"] 1["Jane", "Lucas;"] 1["Jane", "Lydia's"] 1["Jane", "Ly Miss"] 1["Jane.", "No"] 1["Jane.", "Now"] 1["Jane.", "She"] 2["Jane.", "Though"] 1["Jane;", "Kitty"] 1["Jane;", "Lucases;"] 1["Jane;", "Miss"] 2["Jane?", "Shall"] King.", "Wickham's"] 1["King?", "Miss"] 1["Kitty's", "Lydia;"] 1["Kitty", "Kitty;"] 1["Kitty", "Kitty?\\""] 1["Kitty", "Lucases;"] 1["Kitty", "Lydia"] 4["Kitty", "Lydia;"] 1["Ki ady", "Lucas"] 13["Lady", "Lucas;"] 2["Lady", "Lucas."] 1["Lady", "Lucas.\\""] 1["Lady", "Maria;"] 1["Lady", "Metcalfe's"] 1["Lady", "Miss"] 3["Lady", "Monday"] 1["Lady", "Mr

- Evitar las palabras completamente en mayúsculas
- Eliminar los signos de puntuación junto a las palabras
- Seleccionar únicamente con los pares que aparecen más de ... veces

- Evitar las palabras completamente en mayúsculas
- Eliminar los signos de puntuación junto a las palabras
- Seleccionar únicamente con los pares que aparecen más de ... veces

Bibliografía

- MapReduce, entrada en la Wikipedia:

<https://es.wikipedia.org/wiki/MapReduce>

- MapReduce tutorial, en la web oficial de Hadoop:

https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html