Ejercicios MapReduce

Dalia Camacho

Estos ejercicios tuvieron como objetivo aprender a utilizar un código que mapeara una variable (*key*) a otra (value) y a partir de ese mapeo realizar funciones de reducción para mostrar resúmenes de datos de distintas maneras. Es importante notar que los datos están ordenados en términos de *key*, pero que no se mantiene el orden en *value*.

Ejercicio 1

¿Cuántos distritos electorales contiene el archivo votacion.csv?
En total hay 20 distritos electorales.
Para obtener este resultado *key=distrito*, *value* lo dejamos como "1", ya que no hacemos uso del valor asociado. El reducer es similar al que hicimos en clase, la diferencia se encuentra en el acumulador. En este caso el acumulador incrementa únicamente cuando el distrito actual es distinto al distrito anterior y no se reinicia.

Ejercicio 2

¿Cuántas encuestas se obtuvieron por distrito electoral? ¿En qué distrito secapturaron más encuestas? ¿En cuál menos? El distrito en que más encuestas se capturaron fue el distrito 8425 con 10192 encuestas. Mientras que el distrito con menos encuestas fue el 1232 con 1963 encuestas. En la Tabla 1 se muestran todos los resultados.

Encontrar el número de encuestas por distrito siguió la misma lógica que el ejemplo visto en clase. En este caso se agregaron comparaciones para encontrar el número minimo y máximo de encuestas. Cada que había un nuevo distrito se comparaba si el total de acumulados era mayor al número máximo de encuestas hasta ese momento o menor que el mínimo. Si alguna de estas ocurría se modificaba el máximo o el mínimo según fuera el caso.

Distrito	Encuestas
1048	4041
1232	1963
2078	2035
2279	3996
3135	3964
3972	3925
4137	3885
4682	4063
5208	1999
5432	9999
5572	2002
6002	3994
6592	2000
7373	9886
7674	4010
7932	4011
8425	10192
8999	4006
9184	9924
9600	10105

Table 1: Encuestas por distrito

Ejercicio 3

¿Cuántos votos obtuvo cada candidato en cada distrito electoral? ¿Cuántos obtuvo el candidato 1 en el distrito 1232? ¿Cuántos el candidato 5 en el distrito 9184?
El candidato 1 tuvo 188 votos en el distrito 1232 mientras que el candidato 5 tuvo 2519 votos en el distrito 9184. En las Tablas 2 y 3 se muestran los votos que tuvo cada candidato en cada distrito. Inicialmente había colocado candidato como *key* y cistrito como *value*. El problema es que a pesar de estar ordenados los candidatos, los distritos no estaban necesariamente en orden. Para resolver este problema cambié la llave por $candidato + "Distrito : " + distrito$ de esta forma se ordenaban tanto por candidato como por distrito y el código funcionaban sin problemas.

Para encontrar los votos del candidato uno en el distrito 1232 y los del candidato 5 en el 2519 se hacía una comparación cada que había una nueva llave y se evaluaba si ésta correspondía a uno de estos casos. Cuando así ocurrió se guardó el número de votos en variables auxiliares.

Candidato	Distrito	Votos
1	1048	402
1	1232	188
1	2078	214
1	2279	413
1	3135	363
1	3972	397
1	4137	379
1	4682	381
1	5208	197
1	5432	1016
1	5572	191
1	6002	386
1	6592	207
1	7373	976
1	7674	400
1	7932	453
1	8425	1052
1	8999	431
1	9184	989
1	9600	1021
2	1048	376
2	1232	165
2	2078	198
2	2279	384
2	3135	378
2	3972	371
2	4137	398
2	4682	387
2	5208	218
2	5432	994
2	5572	203
2	6002	426
2	6592	221
2	7373	961
2	7674	387
2	7932	414
2	8425	1009
2	8999	412
2	9184	958
2	9600	1024
3	1048	606
3	1232	324
3	2078	305
3	2279	628
3	3135	638
3	3972	584
3	4137	573
3	4682	626
3	5208	285
3	5432	1509
3	5572	313
3	6002	638
3	6592	274
3	7373	1561
3	7674	604
3	7932	527
3	8425	1532
3	8999	615
3	9184	1484
3	9600	1425

Table 2: Votos por candidato y distrito

Candidato	Distrito	Votos
4	1048	1671
4	1232	777
4	2078	790
4	2279	1588
4	3135	1567
4	3972	1604
4	4137	1545
4	4682	1656
4	5208	828
4	5432	4015
4	5572	814
4	6002	1614
4	6592	795
4	7373	3877
4	7674	1644
4	7932	1582
4	8425	4071
4	8999	1536
4	9184	3974
4	9600	4070
5	1048	986
5	1232	509
5	2078	528
5	2279	983
5	3135	1018
5	3972	969
5	4137	990
5	4682	1013
5	5208	471
5	5432	2465
5	5572	481
5	6002	930
5	6592	503
5	7373	2511
5	7674	975
5	7932	1035
5	8425	2528
5	8999	1012
5	9184	2519
5	9600	2565

Table 3: Votos por candidato y distrito

Ejercicio 4

Cómo se distribuyó el voto por género para cada candidato en cada distritoelectoral? ¿En qué distrito y para qué candidato se obtuvo la menor preferencia electoral de los varones?

El distrito y candidato para los que menos votos de hombres hubo fue en el distrito 1232 para el candidato 2 con únicamente 86 votos. Los resultados obtenidos para cada candidato y distrito por género se encuentran en las Tablas 4 hast 7.

Este fue el único caso en que realmente utilicé el *value* y no sólo *key*. La llave la tomé como $candidato + ' Distrito : ' + distrito$ y $value = genero$. En este caso definí dos acumuladores: uno para hombres y otro para mujeres. Estos se reiniciaban cada que se tenía una llave distinta. El acumulador de cada genero aumentaba en una unidad cuando el votante fuera del género correspondiente.

Para guardar el candidato y distrito con mínima participación de los hombres se definió una variable auxiliar que se iba comparando con los acumulados de hombres. Si el acumulado era menor que el mínimo, entonces éste se actualizaba.

En este caso la llave también pudo haber incluído el género para no tener distintos acumuladores y evitar el uso de condicionales.

Ejercicio 5

Este ejercicio es muy similar al que vimos originalmente en clase. Sin embargo hay que entregar los valores ordenados numéricamente. En caso contrario los valores serán ordenados alfabéticamente y las últimas horas en aparecer serán las 8 y las 9 de la mañana. Para hacer esta modificación no se ajustaron los códigos de mapper o reducer, lo que cambió fue la instrucción en la terminal. A esta se le agregó:

```
-D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator
-D mapred.text.key.comparator.options=-n
```

como se indica en:

`https://stackoverflow.com/questions/13331722/how-to-sort-numerically-in-hadoops-shuffle-sort-phase`

Con lo que la indicación para hadoop queda como en la Figura 1.

```
ubuntu@ip-172-31-42-210:~/Prs/code$ hadoop jar $STRJAR -D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator -D
mapred.text.key.comparator.options=-n -files EjMapper_ej5.py,EjREducer_ej5.py -input votacion.csv -output Ej5_1 -mapper EjMapper_ej5.py -reducer EjREd
ucer_ej5.py
```

Figure 1: Comando de hadoop para indicar que el ordenamiento de la salida es numérico

Los resultados obtenidos se muestran en la Tabla 8 se muestran los resultados.

Candidato	Distrito	Género	Votos
1	1048	Hombres	205
1	1048	Mujeres	197
1	1232	Hombres	102
1	1232	Mujeres	86
1	2078	Hombres	104
1	2078	Mujeres	110
1	2279	Hombres	185
1	2279	Mujeres	228
1	3135	Hombres	180
1	3135	Mujeres	183
1	3972	Hombres	190
1	3972	Mujeres	207
1	4137	Hombres	177
1	4137	Mujeres	202
1	4682	Hombres	205
1	4682	Mujeres	176
1	5208	Hombres	93
1	5208	Mujeres	104
1	5432	Hombres	519
1	5432	Mujeres	497
1	5572	Hombres	90
1	5572	Mujeres	101
1	6002	Hombres	206
1	6002	Mujeres	180
1	6592	Hombres	99
1	6592	Mujeres	108
1	7373	Hombres	491
1	7373	Mujeres	485
1	7674	Hombres	193
1	7674	Mujeres	207
1	7932	Hombres	224
1	7932	Mujeres	229
1	8425	Hombres	532
1	8425	Mujeres	520
1	8999	Hombres	217
1	8999	Mujeres	214
1	9184	Hombres	523
1	9184	Mujeres	466
1	9600	Hombres	525
1	9600	Mujeres	496
2	1048	Hombres	185
2	1048	Mujeres	191
2	1232	Hombres	86
2	1232	Mujeres	79
2	2078	Hombres	99
2	2078	Mujeres	99
2	2279	Hombres	196
2	2279	Mujeres	188
2	3135	Hombres	183
2	3135	Mujeres	195
2	3972	Hombres	188
2	3972	Mujeres	183
2	4137	Hombres	194
2	4137	Mujeres	204

Table 4: Votos por distrito, candidato y género

Candidato	Distrito	Género	Votos
2	4682	Hombres	192
2	4682	Mujeres	195
2	5208	Hombres	113
2	5208	Mujeres	105
2	5432	Hombres	491
2	5432	Mujeres	503
2	5572	Hombres	97
2	5572	Mujeres	106
2	6002	Hombres	221
2	6002	Mujeres	205
2	6592	Hombres	105
2	6592	Mujeres	116
2	7373	Hombres	480
2	7373	Mujeres	481
2	7674	Hombres	186
2	7674	Mujeres	201
2	7932	Hombres	202
2	7932	Mujeres	212
2	8425	Hombres	509
2	8425	Mujeres	500
2	8999	Hombres	225
2	8999	Mujeres	187
2	9184	Hombres	500
2	9184	Mujeres	458
2	9600	Hombres	525
2	9600	Mujeres	499
3	1048	Hombres	301
3	1048	Mujeres	305
3	1232	Hombres	166
3	1232	Mujeres	158
3	2078	Hombres	163
3	2078	Mujeres	142
3	2279	Hombres	318
3	2279	Mujeres	310
3	3135	Hombres	312
3	3135	Mujeres	326
3	3972	Hombres	287
3	3972	Mujeres	297
3	4137	Hombres	288
3	4137	Mujeres	285
3	4682	Hombres	308
3	4682	Mujeres	318
3	5208	Hombres	143
3	5208	Mujeres	142
3	5432	Hombres	781
3	5432	Mujeres	728
3	5572	Hombres	146
3	5572	Mujeres	167
3	6002	Hombres	340
3	6002	Mujeres	298
3	6592	Hombres	144
3	6592	Mujeres	130
3	7373	Hombres	766
3	7373	Mujeres	795
3	7674	Hombres	306
3	7674	Mujeres	298
3	7932	Hombres	264

Table 5: Votos por distrito, candidato y género

Candidato	Distrito	Género	Votos
3	7932	Mujeres	263
3	8425	Hombres	774
3	8425	Mujeres	758
3	8999	Hombres	315
3	8999	Mujeres	300
3	9184	Hombres	740
3	9184	Mujeres	744
3	9600	Hombres	710
3	9600	Mujeres	715
4	1048	Hombres	824
4	1048	Mujeres	847
4	1232	Hombres	412
4	1232	Mujeres	365
4	2078	Hombres	408
4	2078	Mujeres	382
4	2279	Hombres	787
4	2279	Mujeres	801
4	3135	Hombres	779
4	3135	Mujeres	788
4	3972	Hombres	772
4	3972	Mujeres	832
4	4137	Hombres	782
4	4137	Mujeres	763
4	4682	Hombres	864
4	4682	Mujeres	792
4	5208	Hombres	422
4	5208	Mujeres	406
4	5432	Hombres	1997
4	5432	Mujeres	2018
4	5572	Hombres	419
4	5572	Mujeres	395
4	6002	Hombres	781
4	6002	Mujeres	833
4	6592	Hombres	402
4	6592	Mujeres	393
4	7373	Hombres	1979
4	7373	Mujeres	1898
4	7674	Hombres	814
4	7674	Mujeres	830
4	7932	Hombres	781
4	7932	Mujeres	801
4	8425	Hombres	1982
4	8425	Mujeres	2089
4	8999	Hombres	768
4	8999	Mujeres	768
4	9184	Hombres	1984
4	9184	Mujeres	1990
4	9600	Hombres	2036
4	9600	Mujeres	2034

Table 6: Votos por distrito, candidato y género

Candidato	Distrito	Género	Votos
5	1048	Hombres	496
5	1048	Mujeres	490
5	1232	Hombres	247
5	1232	Mujeres	262
5	2078	Hombres	267
5	2078	Mujeres	261
5	2279	Hombres	486
5	2279	Mujeres	497
5	3135	Hombres	534
5	3135	Mujeres	484
5	3972	Hombres	502
5	3972	Mujeres	467
5	4137	Hombres	499
5	4137	Mujeres	491
5	4682	Hombres	500
5	4682	Mujeres	513
5	5208	Hombres	234
5	5208	Mujeres	237
5	5432	Hombres	1194
5	5432	Mujeres	1271
5	5572	Hombres	249
5	5572	Mujeres	232
5	6002	Hombres	470
5	6002	Mujeres	460
5	6592	Hombres	257
5	6592	Mujeres	246
5	7373	Hombres	1252
5	7373	Mujeres	1259
5	7674	Hombres	458
5	7674	Mujeres	517
5	7932	Hombres	509
5	7932	Mujeres	526
5	8425	Hombres	1252
5	8425	Mujeres	1276
5	8999	Hombres	495
5	8999	Mujeres	517
5	9184	Hombres	1292
5	9184	Mujeres	1227
5	9600	Hombres	1249
5	9600	Mujeres	1316

Table 7: Votos por distrito, candidato y género

Hora	Votos
8	3899
9	5101
10	8045
11	15119
12	14042
13	9833
14	5988
15	9995
16	13999
17	13979

Table 8: Caption