Pràctica Scala

Enric Rodriguez, Marc Cané

29 de novembre de 2018

Continguts

Chapter 1

Resultats

Resultats segons els llindars de simil·litud considerats

Per comparar els resultats amb diferents llindars hem agafat un nombre de fitxers que contingués suficients documents perquè hi haguessin parelles similars pero alhora no massa gran perquè el càlcul no tardés molt. Hem decidit agafar 300 documents.

La seguent taula mostra el nombre de parelles de documents que obtenim que compleixin cada restricció amb diferents llindars.

Llindar	Parells de pàgines similars no interrelacionades	Parells de pàgines no similars interrelacionades
0.1	127	30849
0.2	24	31015
0.3	7	31047
0.4	4	31060
0.5	2	31068

A continuació es mostra la sortida de les execucions amb els diferents llindars:

Llindar = 0.1

10 primeres parelles de pàgines similars que no es referencien una a l'altra:
William Sholto Douglas, Otto Hoffmann von Waldau ->0.1288872009766826
Castell japonès, Ninja ->0.24025948403874114
William Sholto Douglas, Herbert Otto Gille ->0.10360432690191622
Nantes, William Sholto Douglas ->0.10875873369929959
William Sholto Douglas, Orde Virtuti Militari ->0.16418468360425356
Copa Volpi per la millor interpretació masculina, Christopher Lee ->0.18328534652636125
Ofensiva de Prússia Oriental, Front Oriental de la Segona Guerra Mundial ->0.3728678286341218
William Sholto Douglas, Charles Elwood Yeager ->0.1392076132407418
Andrew Browne Cunningham, Aleksandr Ivànovitx Pokrixkin ->0.11123539067004148
Messerschmitt Me 262, Charles Elwood Yeager ->0.11234735736587806

10 primeres parelles de pàgines que es referencien però no són similars:
Ruth Benedict, Simfonia núm. 8 (Mahler) ->0.009823870046741895
Smith, Història de l'Orient Mitjà ->0.005021916293628367
Sempre en Galiza, Enginyeria inversa ->0.0015703616271788405
Història de Mali, Palau Reial de Caserta ->0.005265578590939846
Neonazisme, Corbeta ->0.0029017094460869566
Memòries d'una geisha (pel·lícula), Reagrupament del Poble Francès ->0.005632770711926015
Rebecca Clarke, Superheroi ->0.005630297770149071
Federació Luterana Mundial, Gran Purga ->0.0017906154961118467
Tres estudis per a figures al peu d'una crucifixió, Història de Kosovo ->0.004374072163271377
Sivaix, Monarquia ->0.00543964030225733

Llindar = 0.2

10 primeres parelles de pàgines similars que no es referencien una a l'altra:
Castell japonès, Ninja ->0.24025948403874114
Ofensiva de Prússia Oriental, Front Oriental de la Segona Guerra Mundial ->0.3728678286341218
William Sholto Douglas, Galeazzo Ciano ->0.23427707057165204
William Sholto Douglas, William Duthie Morgan ->0.22460002270186138
Romania durant la Segona Guerra Mundial, Carles II de Romania ->0.4688662370901408
Andrew Browne Cunningham, HMS Illustrious (87) ->0.21220274194748687
William Sholto Douglas, Andrew McPherson ->0.24777401361516543
Aleksei Antónov, Vasili Marguèlov ->0.2175658681827359
Medalla dels Treballadors Distingits, Orde Virtuti Militari ->0.22175752010250174
Castell japonès, Castell de Malbork ->0.20047780310800487

10 primeres parelles de pàgines que es referencien però no són similars:
Ruth Benedict, Simfonia núm. 8 (Mahler) ->0.009823870046741895
Smith, Història de l'Orient Mitjà ->0.005021916293628367
Sempre en Galiza, Enginyeria inversa ->0.0015703616271788405
Història de Mali, Palau Reial de Caserta ->0.005265578590939846
Neonazisme, Corbeta ->0.0029017094460869566
Memòries d'una geisha (pel·lícula), Reagrupament del Poble Francès ->0.005632770711926015
Rebecca Clarke, Superheroi ->0.005630297770149071
Federació Luterana Mundial, Gran Purga ->0.0017906154961118467
Tres estudis per a figures al peu d'una crucifixió, Història de Kosovo ->0.004374072163271377
Sivaix, Monarquia ->0.00543964030225733

Llindar = 0.3

10 primeres parelles de pàgines similars que no es referencien una a l'altra:
Ofensiva de Prússia Oriental, Front Oriental de la Segona Guerra Mundial ->0.3728678286341218
Romania durant la Segona Guerra Mundial, Carles II de Romania ->0.4688662370901408
Operació Bagration, Ofensiva de Prússia Oriental ->0.4248229547539587
Batalla del Cap Matapan, HMS Illustrious (87) ->0.31551579001078023

Història de l'Argentina, Argentina ->0.5759558134798274
William Sholto Douglas, James O'Meara ->0.30860727097192947
Fußball-Club Bayern München, Cruzeiro Esporte Clube ->0.5139419149640321

10 primeres parelles de pàgines que es referencien però no són similars:
Ruth Benedict, Simfonia núm. 8 (Mahler) ->0.009823870046741895
Smith, Història de l'Orient Mitjà ->0.005021916293628367
Sempre en Galiza, Enginyeria inversa ->0.0015703616271788405
Història de Mali, Palau Reial de Caserta ->0.005265578590939846
Neonazisme, Corbeta ->0.0029017094460869566
Memòries d'una geisha (pel·lícula), Reagrupament del Poble Francès ->0.005632770711926015
Rebecca Clarke, Superheroi ->0.005630297770149071
Federació Luterana Mundial, Gran Purga ->0.0017906154961118467
Tres estudis per a figures al peu d'una crucifixió, Història de Kosovo ->0.004374072163271377
Sivaix, Monarquia ->0.00543964030225733

Llindar = 0.4

10 primeres parelles de pàgines similars que no es referencien una a l'altra:
Romania durant la Segona Guerra Mundial, Carles II de Romania ->0.4688662370901408
Operació Bagration, Ofensiva de Prússia Oriental ->0.4248229547539587
Història de l'Argentina, Argentina ->0.5759558134798274
Fußball-Club Bayern München, Cruzeiro Esporte Clube ->0.5139419149640321

10 primeres parelles de pàgines que es referencien però no són similars:
Ruth Benedict, Simfonia núm. 8 (Mahler) ->0.009823870046741895
Smith, Història de l'Orient Mitjà ->0.005021916293628367
Sempre en Galiza, Enginyeria inversa ->0.0015703616271788405
Història de Mali, Palau Reial de Caserta ->0.005265578590939846
Neonazisme, Corbeta ->0.0029017094460869566
Memòries d'una geisha (pel·lícula), Reagrupament del Poble Francès ->0.005632770711926015
Rebecca Clarke, Superheroi ->0.005630297770149071
Federació Luterana Mundial, Gran Purga ->0.0017906154961118467
Tres estudis per a figures al peu d'una crucifixió, Història de Kosovo ->0.004374072163271377
Sivaix, Monarquia ->0.00543964030225733

Llindar = 0.5

10 primeres parelles de pàgines similars que no es referencien una a l'altra:
Història de l'Argentina, Argentina ->0.5759558134798274
Fußball-Club Bayern München, Cruzeiro Esporte Clube ->0.5139419149640321

10 primeres parelles de pàgines que es referencien però no són similars:
Ruth Benedict, Simfonia núm. 8 (Mahler) ->0.009823870046741895
Smith, Història de l'Orient Mitjà ->0.005021916293628367
Sempre en Galiza, Enginyeria inversa ->0.0015703616271788405

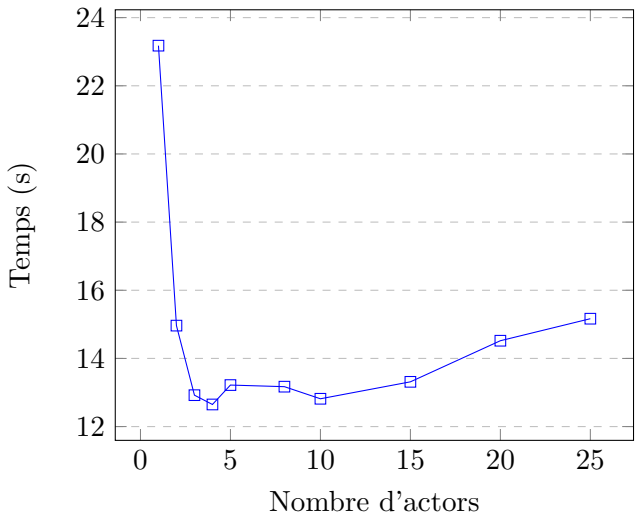
Història de Mali, Palau Reial de Caserta ->0.005265578590939846
Neonazisme, Corbeta ->0.0029017094460869566
Memòries d'una geisha (pel·lícula), Reagrupament del Poble Francès ->0.005632770711926015
Rebecca Clarke, Superheroi ->0.005630297770149071
Federació Luterana Mundial, Gran Purga ->0.0017906154961118467
Tres estudis per a figures al peu d'una crucifixió, Història de Kosovo ->0.004374072163271377
Sivaix, Monarquia ->0.00543964030225733

Rendiment segons el nombre d'actors en el MapReduce

Per fer aquestes proves de rendiment hem usat el mètode SecondHalf.MapReduceTfIdf.computeSimilarities(...) amb els 100 primers documents i hem usat el mateix nombre d'actors tant per el *mapper* com per el *reducer*.

Nombre d'actors	Temps
1	23.177229516
2	14.962622731
3	12.921794558
4	12.648637579
5	13.218281647
8	13.171002857
10	12.816006747
15	13.313055689
20	14.519156598
25	15.165376866

Ho podem veure visualment en el seguent gràfic:



Les proves de rendiment s'han fet en un intel i5 M430 (2010), 2.26 Ghz, 2 cores, 4 threads, 6 GB de RAM.

Chapter 2

Exemple de titol

Quan parlem de matrius disperses ens referim a matrius de grans dimensions en la qual la majoria d'elements son zero. Direm que una matriu és dispersa, quan hi hagi benefici en aplicar els mètodes propis d'aquestes.

Per identificar si una matriu és dispersa, podem usar el següent:

> Una matriu $n \times n$ serà dispersa si el número de coeficients no nuls es $n^{\gamma+1}$, on $\gamma < 1$.

En funció del problema, decidim el valor del paràmetre γ. Aquí hi ha els valors típics de γ:

- $\gamma = 0.2$ per problemes d'anàlisi de sistemes elèctrics de generació i de transport d'energia.
- $\gamma = 0.5$ per matrius en bandes associades a problemes d'anàlisi d'estructures.

Chapter 3

Titol 2

Per Coordenades

És la primera aproximació que podríem pensar i és bastant intuïtiva. Per cada element no nul guardem una tupla amb el valor i les seves coordenades: (a_{ij}, i, j).

A la realitat però, aquest mètode d'emmagatzemar les dades és poc eficient quan hem de fer operacions amb les matrius.

Per files

També conegut com a *Compressed Sparse Rows (CSR)*, *Compressed Row Storage (CRS)*, o format *Yale*. És el mètode més estès.

Consisteix en guardar els elements ordenats per files, guardar la columna on es troben, i la posició del primer element de cada fila en el vector de valors. Així ens quedaran tres vectors:

- **valors:** de mida n_z, conté tots els valors diferents.
- **columnes:** també de mida n_z, conté la columna on es troba cada un dels elements anteriors.
- **iniFiles:** de mida $m + 1$, conté la posició on comença cada fila en els vectors valors i columnes, sent m el nombre de files de la matriu.

Exemple

Si es canvien files per columnes, dona la implementació per columnes, o també anomenada *Compressed Sparse Columns (CSC)*.

Implementació del mètode CSR

Hem implementat un script Matlab amb una classe `CSRSparseMatrix` que guardi les dades necessàries. Aquestes les tenim en "l'atribut" `Matrix` dins del bloc `properties` (línia 10 del codi següent). Aquestes dades consisteixen en el següent:

- `Matrix.nColumns`: número de columnes de la matriu, necessari per recrear les files posteriorment.
- `Matrix.values`: vector valors comentat anteriorment, amb els valors no nuls de la matriu.
- `Matrix.columns`: vector de columnes, amb la columna corresponent a cada valor amb el mateix índex.
- `Matrix.beginningRow`: vector amb els índex comença cada fila en el vector de valors i de columnes.

Codi

Exemple codi Scala:

```
object MapReduceEnric{
 def mapping1(file_name: String, file: (String, List[String])): List[(String, (String,
     Int))] = {
   val wordList = FirstHalf.readFile(file._1).split("
       +").toList.filterNot(file._2.contains(_))

   val x = (for(word <- wordList) yield (file_name, (word, 1)))//.groupBy(_._1)
   x
 }

 //key-> Filename, values-> list of (Word, count)
 def reducing1(key: String, values: List[(String, Int)]): List[(String, Int)] = {
   val res = for( (word, count_list) <- values.groupBy(_._1).toList )
     //For every pair of word and list of counts, add up its counts
     yield (word, count_list.map( {case (_, count) => count } ).reduceLeft( _ + _))

   res.sortWith(FirstHalf.moreFrequent)
 }

 def main1() = {
   val stopwords = FirstHalf.readFile("test/english-stop.txt").split(" +").toList
   val files = Main.openPgTxtFiles("test", "pg", ".txt")

   val input = ( for( file <- files) yield (file.getName, (file.getAbsolutePath,
       stopwords)) ).toList

   val system = ActorSystem("TextAnalizer2")

   val master = system.actorOf(Props(new MapReduceActor[String, (String, List[String]),
       String, (String, Int)](input, MapReduceEnric.mapping1, MapReduceEnric.reducing1, 2,
       2)))
   implicit val timeout = Timeout(10 days)
   val futureResponse = master ? "start"
   val result = Await.result(futureResponse, timeout.duration)
   system.shutdown
   print(result)

 }
```