

Ficheros de entrada para las prácticas

En este y en los siguientes ejercicios utilizaremos dos ficheros de entrada que han sido descargados del National Bureau of Economic Research (NBER) de EEUU (<http://www.nber.org/patents/>).

En concreto, usaremos los ficheros cite75_99.txt y apat63_99.txt. Podéis descargar estos dos ficheros, junto con el fichero country_codes.txt, desde [aquí](#).

En este otro [enlace](#) podéis descargar una versión reducida de los datos para hacer pruebas rápidas.

Una descripción detallada de estos ficheros puede encontrarse en:

Hall, B. H., A. B. Jaffe, and M. Trajtenberg (2001). "The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools." NBER Working Paper 8498.

Fichero cite75_99.txt

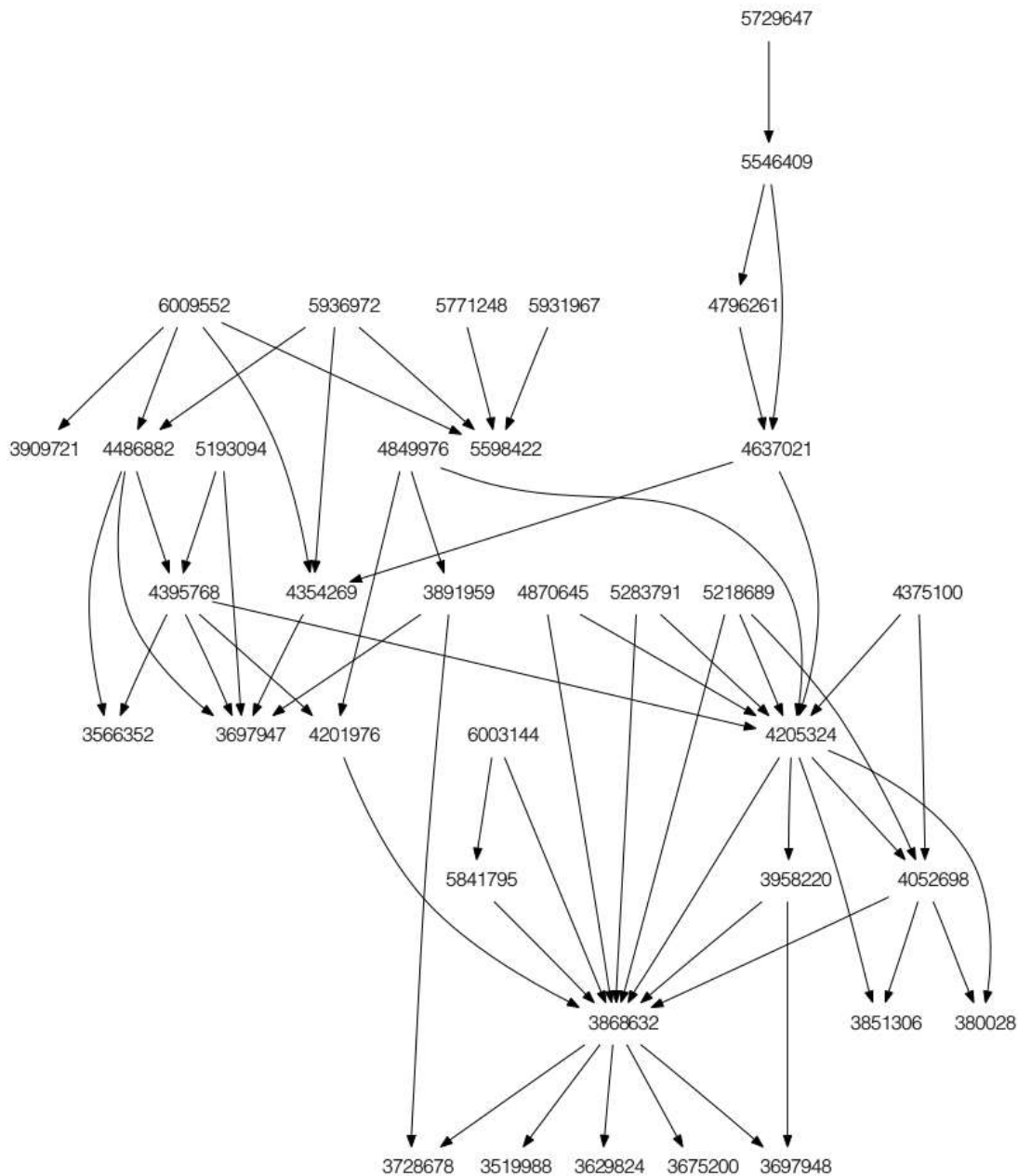
Este fichero contiene citas de patentes emitidas entre 1975 y 1990 en los EEUU. Es un fichero CSV (*comma-separated values*) con más de 16,5 millones de filas, y las primeras líneas son como sigue:

```
"CITING", "CITED"
3858241, 956203
3858241, 1324234
3858241, 3398406
3858241, 3557384
3858241, 3634889
3858242, 1515701
3858242, 3319261
3858242, 3668705
.....
```

La primera línea contiene una cabecera con la descripción de las columnas. Cada una de las otras líneas indica una cita que la patente con el número de la primera columna ha hecho a la patente con el número en la segunda.

Por ejemplo, la segunda fila indica que la patente nº 3858241 ("citing" o *citante*) hace una cita a la patente nº 956203 ("cited" o citada). El fichero está ordenado por las patentes citantes. Así podemos ver que la patente nº 3858241 cita a otras 5 patentes.

Este fichero permite extraer conclusiones sobre las patentes que a primera vista están ocultas. Por ejemplo, en el siguiente gráfico¹ se muestra una vista parcial del grafo de citaciones entre patentes:



En este grafo, los vértices son los números de patente y las aristas dirigidas indican la cita (la arista apunta de la patente *citante* a la citada).

Se puede observar que algunas patentes son citadas múltiples veces mientras que otras no tienen ninguna cita².

Podemos observar hechos curiosos, como que aunque las patentes 5936972 y 6009552 citan a un conjunto similar de patentes (4354269, 4486882, 5598422), no se citan una a la otra.

Fichero apat63_99.txt

Este fichero contiene una descripción de las patentes. Es, de nuevo, un fichero CSV e incluye, entre otros campos, el número de patente, el año de solicitud ("APYEAR"), el año de concesión ("GYEAR"), el país del primer inventor ("COUNTRY"), el número de reivindicaciones ("CLAIMS") y otros metadatos acerca de las patentes. Tiene más de 2,9 millones de filas, y las primeras de estas son:

```

"PATENT","GYEAR","GDATE","APYEAR","COUNTRY","POSTATE","ASSIGNEE","ASSCODE","CLAIMS","NCLASS","CAT","SI
3070801,1963,1096,, "BE", "", ,1,,269,6,69,,1,,0,,,,,
3070802,1963,1096,, "US", "TX", ,1,,2,6,63,,0,,,,,
3070803,1963,1096,, "US", "IL", ,1,,2,6,63,,9,,0.3704,,,,,
3070804,1963,1096,, "US", "OH", ,1,,2,6,63,,3,,0.6667,,,,,
3070805,1963,1096,, "US", "CA", ,1,,2,6,63,,1,,0,,,,,
3070806,1963,1096,, "US", "PA", ,1,,2,6,63,,0,,,,,
3070807,1963,1096,, "US", "OH", ,1,,623,3,39,,3,,0.4444,,,,,

```

```
3070808,1963,1096,, "US", "IA", ,1,, 623,3,39,,4,,0.375,,,,,,,,,
.....
```

Al igual que en muchos datasets reales, algunos de los valores de este fichero están vacíos.

Fichero country_codes.txt

Es un pequeño fichero conteniendo el nombre completo de los países correspondientes a los códigos COUNTRY del fichero apat63_99.txt. Este fichero NO lo debéis copiar en HPFS.

Uso del cluster del CESGA

La máquina de acceso al cluster BigData del CESGA es login.hdp.cesga.es. Se accede mediante ssh (no es necesario tener activa la VPN del CESGA):

```
ssh cursoxxx@hadoop3.cesga.es
```

Desde esa máquina tenemos acceso a los comandos hdfs, para subir los datos al HDFS, y yarn, para ejecutar las aplicaciones, de forma idéntica a como lo hicimos en nuestro cluster pequeño.

Usaremos [Maven](#) para compilar los códigos. Para poder usarlo, en el terminal en el CESGA, el siguiente comando

```
module load maven
```

Para acceder al interfaz web de YARN o a HUE es necesario tener activar la VPN del CESGA. En <https://portalusuarios.cesga.es/> tenéis las instrucciones para activarla. Una vez activa, podéis acceder a través de <https://bigdata.cesga.es/>

Copia de los ficheros a HDFS

Una vez subidos los ficheros a vuestro directorio del CESGA (usando, por ejemplo, `scp cite75_99.txt cursoxxx@hadoop3.cesga.es:~`) a la hora de subirlos al HDFS (con `hdfs dfs -put`), y dado que son ficheros relativamente pequeños, suele interesar especificar un tamaño más pequeño de bloques. Esto va a aumentar el paralelismo, al tener más bloques por fichero y, por tanto, lanzar más maps. Para hacer esto:

```
hdfs dfs -mkdir patentes
hdfs dfs -D dfs.block.size=32M -put cite75_99.txt apat63_99.txt patentes
```

Recuerda que el fichero country_codes.txt no debéis subirlo a HDFS.

Uso de la cola urgent

Los trabajos propuestos llevan muy poco tiempo, por lo que podemos usar la cola para trabajos cortos (urgent) del CESGA. Para esto, en el momento de lanzar la aplicación yarn, tenéis que definir la variable `mapred.job.queue.name` al valor `urgent`. Por ejemplo:

```
yarn jar target/citingpatents-0.0.1-SNAPSHOT.jar -Dmapred.job.queue.name=urgent path_a_cite75_99.txt_en_HDFS dir_salida_en_HDFS
```

Actividad: Programación en Hadoop

- (35%) Plantilla **01-citingpatents**: programa MapReduce escrito en Java que, para cada patente de cite75_99.txt, obtenga la lista de las que la citan
 - Formato salida: *patente patente1,patente2...* (la separación entre la clave y los valores debe ser un tabulado)
 - El mapper debe obtener cada línea del fichero de entrada, separar los campos y darle la vuelta (para obtener como clave intermedia la patente citada y como valor intermedio la patente que la cita):
 - 3858245,3755824 → 3755824 3858245
 - El reducer, para cada patente recibe como valor una lista de las que la citan, y tiene que convertir esa lista en un string:
 - 3755824 {3858245 3858247...} → 3755824 3858245,3858247...
 - La salida debe de estar **numéricamente** ordenada tanto para las patentes citadas como para las citantes.
 - El carácter de separación entre clave y valor en la salida debe de ser el por defecto (tabulado).
 - La lista de números de patente en el campo valor de la salida debe de estar separada por comas, y no debe de haber una coma al final.
 - La salida debe de guardarse en formato comprimido gzip, para lo que debéis utilizar los métodos estáticos `setCompressOutput` y `setOutputCompressorClass` de la clase `FileOutputFormat`.
 - IMPORTANTE:
 - Los ficheros de entrada no deben modificarse de ningún modo.
 - La cabecera del fichero no debe aparecer en la salida
 - Utilizad como formato de entrada **KeyValueTextInputFormat**, indicando que el formato separador de campos es una coma.

- El carácter de separación entre clave y valor en la salida debe de ser el por defecto (tabulado)
- Para compilar la práctica y generar el fichero .jar usad maven: mvn package (el fichero .jar se crea en el directorio target)

◦ OPCIONAL:

- Crea un **Combiner** (modifica los códigos si es necesario). La salida con el Combiner debe de ser igual que antes. Recuerda que el Combiner debe ser opcional (el código debe funcionar igual se se usa el Combiner o si no se usa).

2. (30%) Plantilla **02-citationnumberbypatent_chained**: programa MapReduce que usa ChainMapper y ChainReducer para concatenar trabajos MapReduce

- Obtener el número de citas de una patente, combinando el programa anterior **01-citingpatents** con un mapper adicional (CCMapper) que, a partir de la salida del reducer del 01-citingpatents, para cada patente, cuente el número de patentes que la citan.
- Al igual que en la práctica anterior, usar como formato de entrada **KeyValueTextInputFormat**, indicando que el formato separador de campos es una coma.
- La salida debe guardarse **como un fichero binario de tipo Sequence** (formato clave/valor). Podéis ver el contenido de los ficheros de salida usando `hdfs dfs -text`
- **Para compilar**, copiad el fichero citingpatents-0.0.1-SNAPSHOT.jar generado en la práctica 1 al directorio src/resources de esta práctica, y usad maven para generar el nuevo .jar
- **Para ejecutar**, haced lo siguiente:

```
export HADOOP_CLASSPATH="./src/resources/citingpatents-0.0.1-SNAPSHOT.jar"
yarn jar target/citationnumberbypatent_chained-0.0.1-SNAPSHOT.jar -libjars $HADOOP_CLASSPATH
path_a_cite75_99.txt_en_HDFS dir_salida_en_HDFS
```

3. (35%) Programa **simplereducesidejoin**: Unir datos de dos entradas usando un Reduce Side Join

- (a) Entrada 1: Ficheros binarios de salida del programa **02-citationnumberbypatent_chained**
- (b) Entrada 2: Fichero de texto **apat63_99.txt**

Salida: Fichero de texto en el que en cada línea aparezca

- patente país, año, n_citas
- Entre la patente y el país debe haber **un tabulado**, y entre el país, año y número de citas **una coma** (sin espacios en blanco).
- La salida debe quedar **en un único fichero de texto sin comprimir y ordenada por patente (numéricamente)**.

Mappers

- Deberemos utilizar un mapper diferente para cada entrada
- (a) Mapper-a: Obtiene el número de citas por patente y etiqueta cada salida con el string "ncitas"
3755824 9 → 3755824 ncitas,9
- (b) Mapper-b: Para cada patente, obtiene el país,año y etiqueta cada salida con el string "pais"
3755824,1973,4995,1971,"US","NY",.... → 3755824 pais, US,1973
- "ncitas" y "pais" son simples etiquetas que nos permitirán diferenciar en el reducer si el valor es un número de citas o un país. La salida de los mappers debe ser, por lo tanto, un valor Writable compuesto.

Reducer

- Hace un join de los dos mappers utilizando como clave el número de patente

```
3755824 ncitas, 9
→ 3755824 US,1973,9
3755824 pais, US,1973
```

IMPORTANTE: Queremos hacer un outer join, por lo que en el caso de que no dispongamos de información del país para una patente debería aparecer la expresión No disponible, y un 0 en el caso de que no dispongamos de información de citas.

- Ejemplo de salida:

..... 3070798 No disponible,5 3070799 No disponible,1 3070801 BE,1963,1 3070802 US,1963,0 3070803 US,1963,9 3070804 US,1963,3
.....

OPCIONAL:

Reemplazar en la salida el código del país por su nombre, obtenido del fichero country_codes.txt.

- El fichero country_codes.txt **debe residir en el disco local (no en HDFS)** y se debe utilizar la localización de dependencias para copiarlo a los nodos del cluster.
- Para hacer que el fichero country_codes.txt sea accesible mediante localización de dependencias, indicadlo con la opción **-files** al lanzar para lanzar la tarea:

```
yarn jar target/simplereducesidejoin*.jar -Dmapred.job.queue.name=urgent -files path_a_country_codes.txt_en_disco_local ....
```

Entrega



- Para cada código, enviar un fichero comprimido incluyendo todo el proyecto desarrollado, con todas las fuentes y preparado para compilar usando Maven.
- Incluir también un README con instrucciones de compilación y ejecución de cada uno de los programass.

1. Gráfico extraído de: Chuck Lam, *Hadoop in action*, Manning Publications Co., 2011.

2. Al igual que con otros tipos de datos, debemos de tener cuidado a la hora de sacar conclusiones cuando trabajamos con un conjunto limitado de información. Que una patente no cite a ninguna otra podría deberse a que es una patente antigua, para la que no tenemos información. Y las patentes más actuales tendrán pocas citas, ya que solo pueden ser citadas por otras patentes más nuevas. Este tipo de problemas es común en el análisis de datos.

-  [01-citingpatents.zip](#) 15 de outubro de 2023, 20:34
-  [02-citationnumberbypatent_chained.zip](#) 15 de outubro de 2023, 14:07
-  [03-simplereducesidejoin.zip](#) 15 de outubro de 2023, 20:53

Estado da entrega

Estado da entrega	Entregado para cualificaci3n
Estado das cualificaci3n	Cualificado
Tempo restante	A tarefa foi enviada 11 horas 26 mins en prazo
Última modificaci3n	domingo, 19 de novembro de 2023, 12:32
Entregas de ficheiros	<div> AbrahamTrashorrasRivas.zip 19 de novembro de 2023, 12:32</div>
Comentarios a entrega	<div> Comentarios (0)</div>

Comentarios

Cualificaci3n	6,50 / 10,00
---------------	--------------

Cualificado o

mércoles, 27 de diciembre de 2023, 18:48

Cualificado por



Anselmo Tomás Fernández Pena

Comentarios



P1: No funciona con el combiner (el reducer no sirve de combiner):
java.io.IOException: wrong value class: class org.apache.hadoop.io.Text...