Máster en Ingeniería del Software: Cloud, Datos y Gestión TI.

Big Data

Práctica 1: MapReduce

Alumno: Agustín Núñez Arenas

Propuesta del alumno

La práctica consiste en implementar el cálculo del **TF-IDF** (Term Frequency - Inverse Document Frequency) en MapReduce dado un conjunto de documentos de texto.

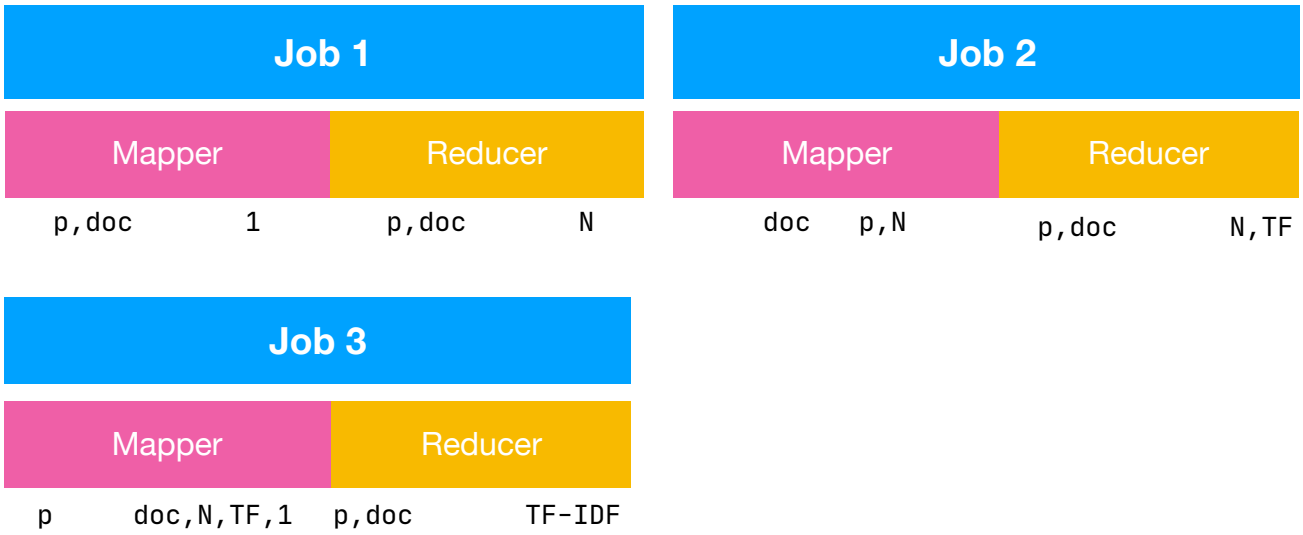
El **TF-IDF** de una palabra con respecto a un documento se utiliza en distintos campos como en: Aprendizaje automático, para identificar la relevancia de una palabra en un texto; en minería de datos, para la extracción de información e incluso se utiliza como medida de ranking en motores de búsqueda.

La fórmula del TF-IDF que se usará en la práctica es la siguiente:

$$TF \cdot IDF = f(p,d) \cdot \log \frac{|D|}{|\{d \in D : p \in d\}|}$$

Donde la primera parte del producto es la frecuencia de la palabra en un documento. Y la segunda parte es el número total de documentos partido del número de documentos donde aparece la palabra.

Esquema del proceso



- **Job 1:** Hacemos una cuenta de todas las palabras, sacando también el nombre del documento donde se sacó esa palabra.

- **Job 2:** Sacamos cuántas veces aparecen las palabras en el documento. Dado que teníamos el nombre del documento en cada fila, tan solo hay que contar de nuevo. Esta cuenta es el TF de nuestra fórmula.

- **Job 3:** Calculamos el TF-IDF al completo, haciendo el cálculo de la segunda parte de la fórmula: Dado el número de documentos totales, dividimos entre el número de documentos que contienen esa palabra. Luego calculamos el logaritmo en base 10.

Pasos seguidos

1. Preparación de entorno

Aquí tiene lugar la descarga de los documentos y los scripts que vamos a usar para la ejecución de la práctica.

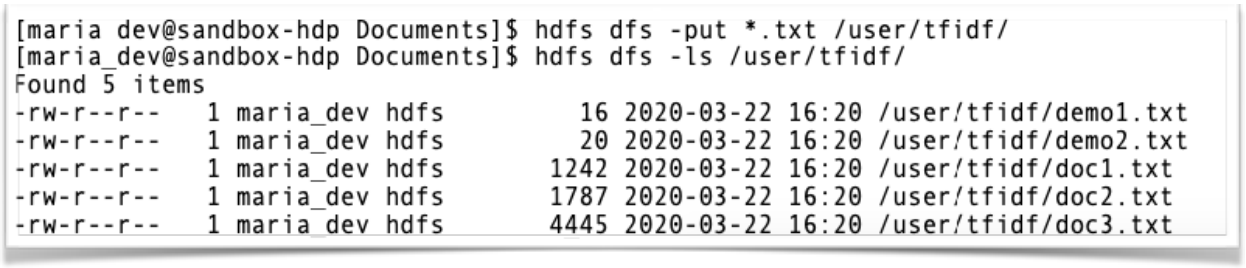
```
$ git clone https://github.com/Agusnez/tfidf-mapreduce-hadoop.git

$ cd tfidf-mapreduce-hadoop
```

2. Carga de documentos en el sistema HDFS

Con este comandos se cargarán en el sistema de ficheros HDFS los distintos documentos.

```
$ hdfs dfs -mkdir /user/tfidf
$ hdfs dfs -put Documents/*.txt /user/tfidf
```

Aquí se observa el resultado tras haberlos cargados correctamente.



3. Ejecución de los comandos

Gracias al script bash que se ofrece en el repositorio, podemos ejecutar los 3 pasos del proceso de forma automática con el siguiente comando. No olvidemos darle permisos de ejecución al fichero.

```
$ chmod 777 tfidf-start.sh
$ ./tfidf-start.sh
```

Esto al final es lo mismo que ejecutar 3 veces lo comandos de MapReduce vistos en clase. A continuación se muestran la salida del paso 1 hecha a mano:

```
[maria_dev@sandbox-hdp tfidf-mapreduce-hadoop]$ hadoop jar
py -file Job1/reducer.py -mapper Job1/mapper.py -reducer Jo
20/03/22 18:09:29 WARN streaming.StreamJob: -file option is
packageJobJar: [Job1/mapper.py, Job1/reducer.py] [/usr/hdp/
20/03/22 18:09:30 INFO client.RMProxy: Connecting to Resour
20/03/22 18:09:30 INFO client.AHSProxy: Connecting to Appli
20/03/22 18:09:30 INFO client.RMProxy: Connecting to Resour
20/03/22 18:09:30 INFO client.AHSProxy: Connecting to Appli
20/03/22 18:09:31 INFO mapred.FileInputFormat: Total input
20/03/22 18:09:31 INFO mapreduce.JobSubmitter: number of sp
20/03/22 18:09:31 INFO mapreduce.JobSubmitter: Submitting t
20/03/22 18:09:31 INFO impl.YarnClientImpl: Submitted appli
20/03/22 18:09:31 INFO mapreduce.Job: The url to track the
20/03/22 18:09:31 INFO mapreduce.Job: Running job: job_1584
20/03/22 18:09:37 INFO mapreduce.Job: Job job_1584879372462
20/03/22 18:09:37 INFO mapreduce.Job:  map 0% reduce 0%
20/03/22 18:09:44 INFO mapreduce.Job:  map 75% reduce 0%
20/03/22 18:09:45 INFO mapreduce.Job:  map 100% reduce 0%
20/03/22 18:09:50 INFO mapreduce.Job:  map 100% reduce 100%
20/03/22 18:09:50 INFO mapreduce.Job: Job job_1584879372462
20/03/22 18:09:50 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=80439
                FILE: Number of bytes written=943626
```

4. Resultados

Los resultados obtenidos como salida del último paso son los siguientes:

```
550   tablet,hdfs://sandbox-hdp.hortonworks.com:8020/user/tfidf/doc3.txt  0.000697545694035
551   technology,hdfs://sandbox-hdp.hortonworks.com:8020/user/tfidf/doc3.txt  0.0020926370821
552   term,hdfs://sandbox-hdp.hortonworks.com:8020/user/tfidf/doc1.txt    0.00486858423183
553   termweighting,hdfs://sandbox-hdp.hortonworks.com:8020/user/tfidf/doc1.txt   0.00243429211592
554   text,hdfs://sandbox-hdp.hortonworks.com:8020/user/tfidf/doc1.txt    0.00486858423183
555   textbased,hdfs://sandbox-hdp.hortonworks.com:8020/user/tfidf/doc1.txt   0.00243429211592
556   tfidf,hdfs://sandbox-hdp.hortonworks.com:8020/user/tfidf/doc1.txt   0.00243429211592
557   tf-idf,hdfs://sandbox-hdp.hortonworks.com:8020/user/tfidf/doc1.txt  0.0146057526955
558   tf-idf2,hdfs://sandbox-hdp.hortonworks.com:8020/user/tfidf/doc1.txt 0.00243429211592
559   than,hdfs://sandbox-hdp.hortonworks.com:8020/user/tfidf/doc3.txt    0.000697545694035
560   that,hdfs://sandbox-hdp.hortonworks.com:8020/user/tfidf/doc1.txt    0.0
561   that,hdfs://sandbox-hdp.hortonworks.com:8020/user/tfidf/doc2.txt    0.0
562   that,hdfs://sandbox-hdp.hortonworks.com:8020/user/tfidf/doc3.txt    0.0
```

Como se puede observar, «tf-idf» es bastante relevante en el documento 1 ya que es un artículo de este tema sacado de Wikipedia. Palabras del tipo «than», «that» que se repiten en todos los documentos obtienen una puntuación mucho más baja, llegando incluso a 0.