Management of Big Data and Tools – DS8003 – Fall 2016

Final Exam Assignment

NAJLIS, BERNARDO - Student Number #500744793

Dataset: bbcsport.zip

- The dataset has articles from BBC
- The dataset one folder each for bbc articles for athletics, cricket, football, rugby, tennis
- **Reference:**
    - D. Greene and P. Cunningham. "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering", Proc. ICML 2006. [PDF] [BibTeX]
- **Note: For the exam you can choose any ONE of the folders.** Of course, you can use more folders if you want to.

Technologies: MapReduce, Spark, Hive, Pig, HBase, Mongo, HDFS. You can make use of python and python libraries as helpers.

Important Notes:

(a) Do not use any existing libraries to compute TFIDF. It has to be done from scratch.

(b) Make sure to test for number of reducers/executors > 1. If you have a specific reason to choose 1 reducer/executor then provide an explanation as to why that is the case.

Questions:

1. (Building Index) Compute TFIDF scores for all words in all documents and build an inverted index **(10)**

2. (Search) Given a new query (one or more words) and a value N, retrieve the top N matching documents with a score (use TFIDF scores to retrieve the matching documents) (10)

(a) If the system is able to take in a query with a value N and return top N matching documents with a score **(7)**

(b) Additionally, if the system can do it in real-time or near real-time **(3)**

3. Documentation **(5)**

- The document should contain both codes and explanations
- Summary section: 3-4 short bullet points of what your system does, why it is awesome, what tools are used etc
- Give your document a title
- Code portion should have comments
- You document portion should refer to line numbers of code when describing your method
- Need an overall architecture diagram: For example, You can come up with a diagram that shows your Data store, all steps in the middle (as boxes) and finally the data sink

- You should also write a section on why you made use of certain tools and contrast that with why you did not choose some others
- If you employ any tricks to make your search system real-time then highlight that
- Result Section: Choose 2 queries (with number of words > 1). For each query report results for following values of N
    - N=1
    - N=3
    - N=5
    - for the purposes of the document, copy and paste the actual contents of the resulting documents so that I can easily understand the results.
- Try to format your document nicely. Think! You could submit the document to some person who asks for an example of your previous work in Big Data Tools

Submission:

- Create a single pdf/doc file

- Submit into Assignment folder titled "Final Exam"

Resolution

Introduction

As part of the final exam assignment for my Masters in Data Science course “DS8003 – Management of Big Data Tools”, I created a Big Data TF-IDF index builder and query tool. The tool consists a script with functions to create a TF-IDF (term frequency-inverse document frequency) index and it is then used it to return matching queries for a list of terms provided and number of results expected.

Features Summary

- Developed with PySpark, SparkSQL and DataFrames API for maximum compatibility with Spark 2.0
- Documents to build the TF-IDF index can be on a local or HDFS path
- Index is stored in parquet format in HDFS
- Query terms and number of results are specified via command line arguments

The choice to use a combination of Spark, SparkSQL and the DataFrames API over all other possible tools that we covered in class and I could have used (including MapReduce, Hive, Pig, HBase, MongoDB) is purely personal: I believe that Spark is the best Big Data platform in terms of performance and long term, and through the course is the one that proved the easiest to work with and better performance dealing with large data sets.

The Data frame schema is as follows, please see the sample data set below:

doc_name	word	tf_idf
hdfs://sandbox.hortonworks…	spark	0.6999
hdfs://sandbox.hortonworks…	is	0.6999
hdfs://sandbox.hortonworks…	awesome	0.6999

Using the parquet format for the tf-idf index store is another design choice based on my personal experience with columnar format stores on other platforms. Considering the schema is a denormalized list of documents-words, the expected cardinality of these two columns makes them very good candidates for the high compression rates that columnar stores have.

The code is a python script to be used with spark-submit as a submit job, but it can easily be adapted to other uses.

PySpark Script

Here is the script, as you can see it is not very extensive, and comments provide an easy read to the code, which is fairly simple.

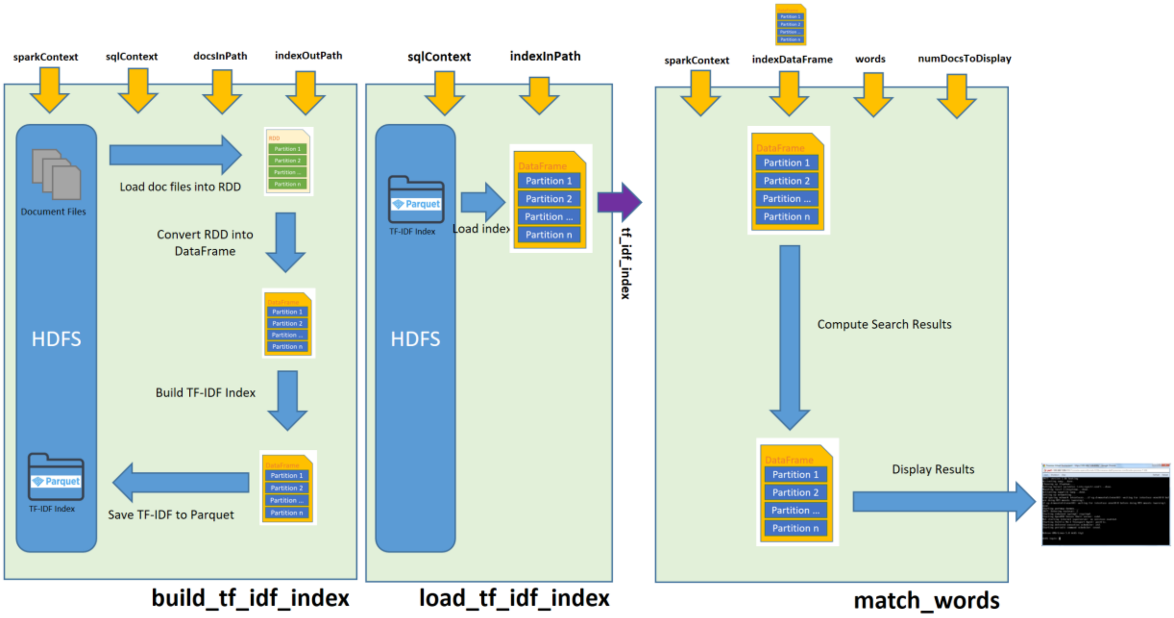
```
import sys, getopt
from pyspark import SparkConf, SparkContext
from pyspark.sql import SQLContext, DataFrameWriter
from pyspark.sql.functions import *

# ##########################################################################
# TF-IDF Index builder funtion
###########################################################################
# Input Parameters:
#                   sc: Spark Context (created in main function)
#                   sqlContext: SparkSql Context (created in main function)
#                   docs_path: string with path containing documents to index
# Returns:
#                   tf_idf: DataFrame containing TF-IDF index
#                             - Schema for DataFrame:
#                                       - doc_name: string containing original document name path
#                                       - word: word
#                                       - tf_idf: TF-IDF score for the word in doc_name
##########################################################################


def build_tf_idf_index(sc, sqlContext, docs_path, index_path):
          print("INFO Reading files from docs in " + docs_path + " ...")
          # Load files into RDD
          textFiles = sc.wholeTextFiles(docs_path)
          # Count number of documents in total to use in IDF calculations
          num_docs = textFiles.count() # ACTION

          print("INFO Building index for " + str(num_docs) + " files found...")
          ###### Build TF-IDF index ######

          # Split document into words (still in RDD)
          tmpFilesRDD = textFiles.map(lambda docs:(docs[0], docs[1].split(" "))) #transformation
          # Convert RDD into Data Frame, to use SparkSQL
          textFilesDF = tmpFilesRDD.toDF(["doc_name", "doc_content"]) # ACTION
          # Explode the Data Frame to get one row per document*word combination. This DF is the basis for TF and IDF
          words_by_docDF = textFilesDF.select(textFilesDF.doc_name, explode(textFilesDF.doc_content).alias("word"))
#transformation

          ### Build Term Frequency table ###

          # First count each word in all documents
          pre_tf = words_by_docDF.groupBy(words_by_docDF.doc_name, words_by_docDF.word).count() #transformation
          # Do sum() aggregation of counts and set as column 'tf'
          tf = pre_tf.groupBy(words_by_docDF.doc_name, words_by_docDF.word, ).agg(sum("count").alias("tf")) #transformation

          ### Build Inverse Document Frequency table ###

          # First do a count distinc group by words, to get the number of docs using each word
          pre_idf = words_by_docDF.distinct().groupby(words_by_docDF.word).count() #transformation
          # Now calculate IDF as log(num_docs / docs_using_word)
          idf = pre_idf.select(pre_idf.word, col("count"), log10(num_docs/(col("count"))).alias("idf")) #transformation

          # Join TF with IDF dataframes (TF left outer join IDF on word) so we have both TF and IDF per word*document
          pre_tf_idf = tf.join(idf, tf.word == idf.word, 'outer') #transformation

          # Build TF-IDF by multiplying TF * IDF
          tf_idf = pre_tf_idf.select(col("doc_name"), tf["word"], (col("tf") * col("idf")).alias("tf_idf")) #transformation
          print("Saving index to " + index_path + "spark-tfidf.parquet" + " ...")
          # Save index to HDFS parquet format
          try:
                    tf_idf.write.save(index_path + "spark-tfidf.parquet")
          except:
                    usage("Error saving index.")

          return tf_idf

def match_words(sc, tfidf_index, all_words, ndocs):

          print("INFO Finding matching documents...")
          # words are submittes as comma separated list, so we split the string into a python list first
          all_words = words.split(',')
          # count how many words are in the query for the normalization factor
          query_words_qty = len(all_words)
          # then we convert the list into a spark data frame (passing through a parallelized rdd and dummy lambda function)
          words_df = sc.parallelize(all_words).map(lambda x:(x,)).toDF(["query_word"])
          # get a subset data frame with just the query words
          joined = tfidf_index.join(words_df, words_df.query_word == tfidf_index.word, "inner")
          # counts how many matched words and aggregates (sum) the score per word
          pre_scored = joined.groupBy("doc_name").agg({"*":"count", "tf_idf":"sum"}).withColumnRenamed("count(1)",
"matched_words_qty").withColumnRenamed("sum(tf_idf)", "pre_score")
          # calculate the tf_idf score for the document as pre_score (summ of per-word score) times number of matched words
divided by num of words in query
          pre_scored2 = pre_scored.select(pre_scored.doc_name, (pre_scored.pre_score * pre_scored.matched_words_qty) /
query_words_qty)
          # rename the matching score calculated column
          scored = pre_scored2.withColumnRenamed("((pre_score * matched_words_qty) / " + str(query_words_qty) + ")",
"matching_score")
```

```
        # sort the documents by descending matching score and return only the top ndocs requested
        doc_matches = scored.sort("matching_score", ascending=False)
        if(int(ndocs) > 0):
                doc_matches = doc_matches.limit(int(ndocs))

        return doc_matches

def load_tf_idf_index(sql, tf_idf_index_path):
        print("INFO Reading index from "+ tf_idf_index_path + "spark-tfidf.parquet")
        tf_idf_index = sql.read.parquet(tf_idf_index_path + "spark-tfidf.parquet")
        return tf_idf_index

def usage(err):
        # Prints error message and command line usage help
        print("ERROR " + err)
        print("Builds a TF IDF index from documents, saves it in parquet format, then searches for terms in the index and
return the top X document matches.")
        print
        print("Usage: -w[words to search] -n[results] -d[docs to index path] -i[tfidf index path]")
        print
        print("Options:")
        print(" -d DOCUMENTS_TO_INDEX_PATH      hdfs:// or local file path with documents to index.")
        print("                                 Optional if path in -i contains a tf-idf index.")
        print(" -i TF_IDF_INDEX_PATH            hdfs:// or local path to parquet tfidf index.")
        print("                                 Path location has to end in '/' and contain no files.")
        print("                                 If -d is specified will save resulting index here.")
        print("                                 If -d is not specified, will try to load index from here.")
        print(" -w WORDS                        Comma-separated list of words to search")
        print(" -n RESULTS                      Optional: number of document results to return from index.")
        print("                                 If not specified returns all results.")
        sys.exit()

# Main entry function
# Sets Spark and Sql Context to call Index builder and Document matching query
def main(docs_path, words, ndocs, tfidf_index):
        c = SparkConf().setAppName("TF-IDF Indexer") # Set Spark configuration
        sc = SparkContext(conf = c)   # Set spark context with config
        sc.setLogLevel("ERROR")                  # Reduce the error logging to minimum
        sql = SQLContext(sc)

        if docs_path:       # If a path to docs is specified...
                # ...build index from the docs...
                tf_idf_index = build_tf_idf_index(sc, sql, docs_path, tfidf_index)
        else:
                # ... or load a previously created and saved index ...
                tf_idf_index = load_tf_idf_index(sql, tfidf_index)

        if words:
                matches = match_words(sc,tf_idf_index, words, ndocs)
                if(int(ndocs) > 0):
                        matches.show(int(ndocs), False)
                else:
                        matches.show(matches.count(), False)
        else:
                print("INFO No search terms specified, exiting...")

        sc.stop()                                          # Stop Spark context after done with main function


# Catch all to parse command line options and redirect to main function
if __name__ == "__main__":
        try:
                # Gets command line options required for script usage
                opts, args = getopt.getopt(sys.argv[1:],"d:w:n:i:")
        except getopt.GetoptError as err:
                usage(err);


        if(len(opts) == 0 ):
                usage("Error: Command line parameters not specified.")

        # Gets command line parameters
        tfidf_index = ""
        words = ""
        docs_path = ""
        ndocs = "-1"
        for opt,arg in opts:

                if opt == "-i":
                        tfidf_index = arg              # Path to tf_idf index
                        if not tfidf_index:
                                usage("Error: option -i missing.")


                if opt == "-w":
                        words = arg                                        # List of words to use for query
                        if not words:
                                usage("Error: option -w missing.")
```

```
177
178                     if opt == "-d":
179                               docs_path = arg               # Path containing documents to be indexed
180
181
182                     if opt == "-n":
183                               ndocs = arg                                       # Number of documents to return in matching query
184
185
186
187           main(docs_path, words, ndocs, tfidf_index)
```

Program flow Description

The following diagram shows the three main functions used in the script with their input and output parameters.



build_tf_idf_index (line 22)

This is the main function that builds the tf-idf index, and takes four parameters:

- **sparkContext**: required to load the document files to be indexed into an RDD
- **sqlContext**: required to convert the RDD into DataFrame and apply transformations to it
- **docsInPath**: string containing the HDFS path where the documents to index are stored
- **indexOutPath**: string containing the HDFS path where the parquet tf-idf index will be stored

load_tf_idf_index (line 96)

This function loads a previously saved tf-idf index and outputs a data frame containing the index. It only has two input parameters:

- **sqlContext**: to load the parquet data frame from HDFS
- **indexInPath**: string containing the HDFS path where the tf-idf index was stored

and just one output parameter:

- **tf_idf_index**: data frame containing the index

match_words (line 68)

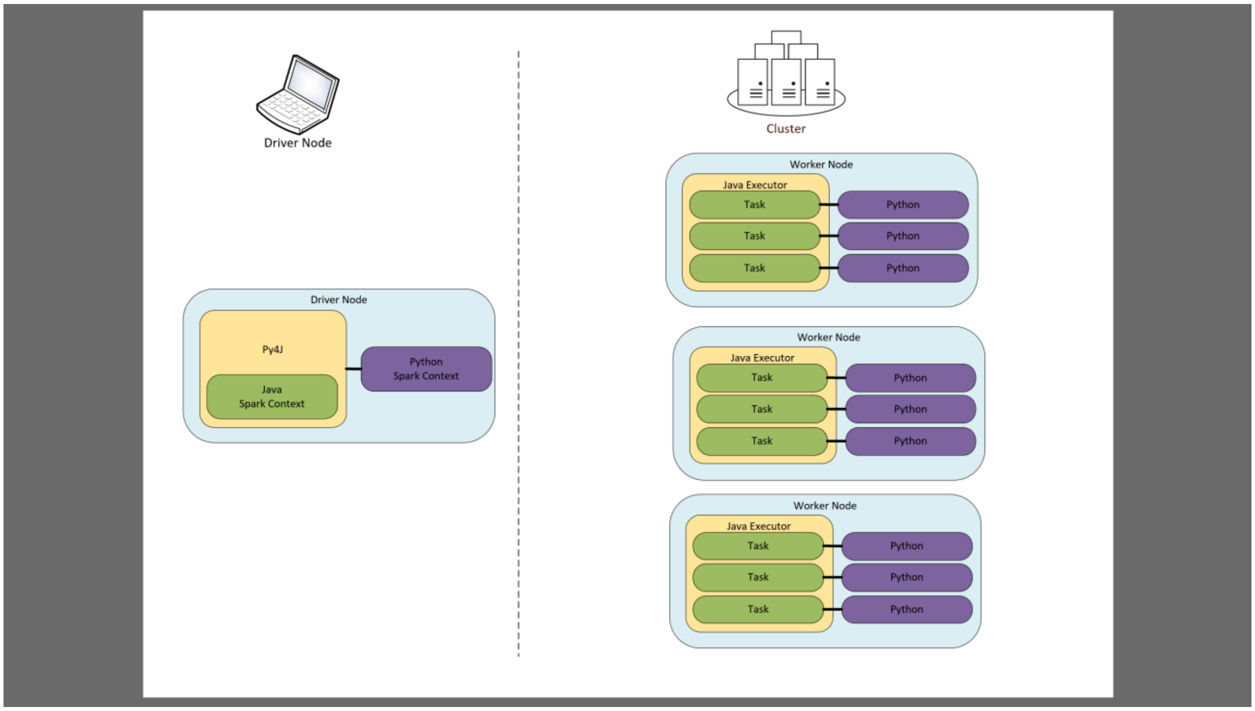
This function looks for word matches in an index. It takes four parameters:

- **sparkContext**: to apply transformations
- **indexDataFrame**: data frame containing the tf-idf-index
- **words**: list of words to search the index for
- **numDocsToDisplay:** number of results to show

and it prints a table with the results in the console.

General Architecture

In terms of PySpark execution and cluster architecture, the process is kick-started from the driver node which commands the worker nodes. As the script was programmed using Python, PySpark adds a Python Spark Context to the regular Java Spark Context that regular Spark Java tasks use. This context adds a little bit of overhead (for object marshalling and serialization / deserialization). All transformations in RDDs are split into partitions and therefore, can be run in multiple worker nodes at the same time, effectively parallelizing the work of building the TF-IDF index.



How to use the tf-idf-indexer.py script

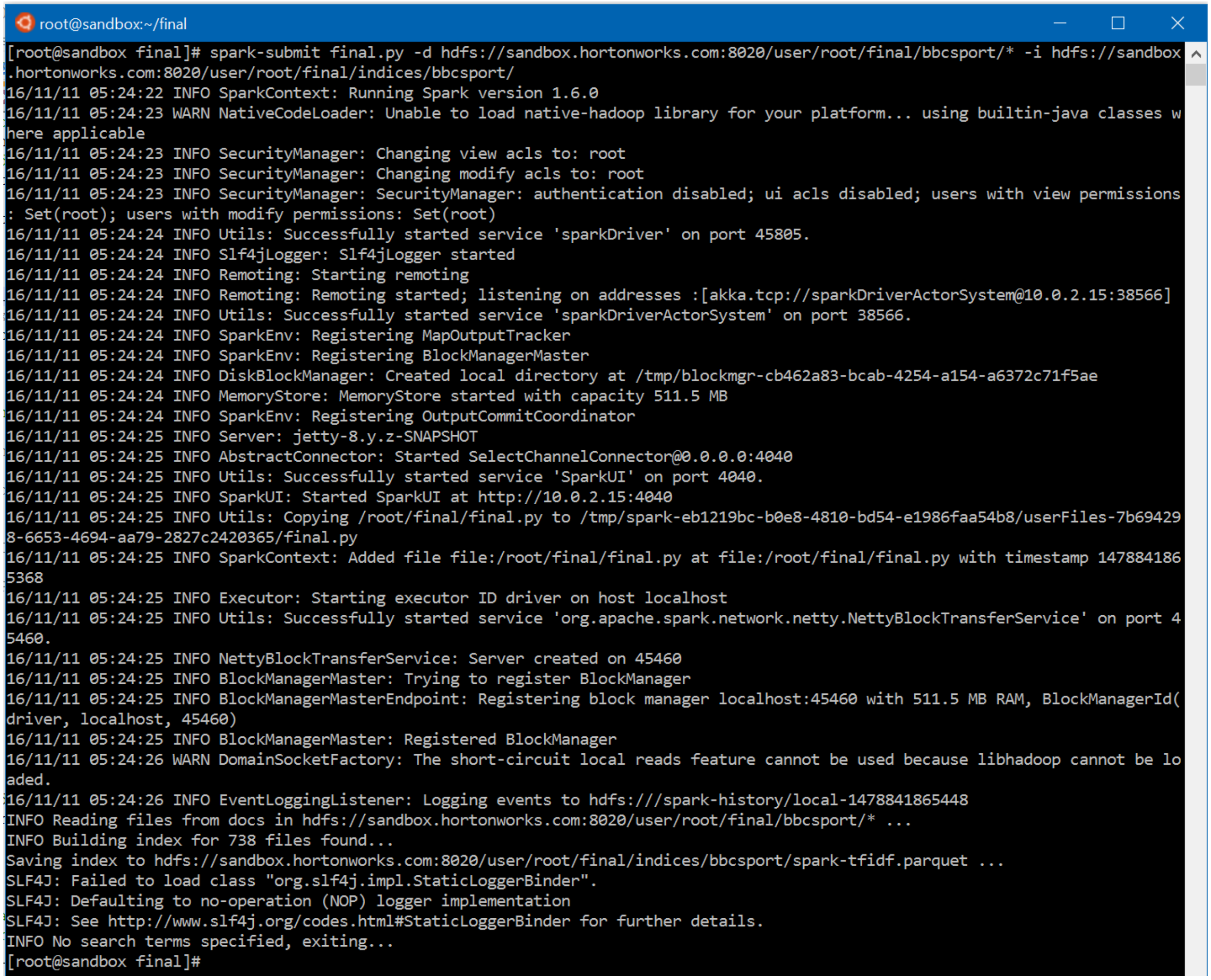
The index can be invoked using spark-submit with the following options:

```
Usage: -w[words to search] -n[results] -d[docs to index path] -i[tfidf index path]

Options:


-d DOCUMENTS_TO_INDEX_PATH    hdfs:// or local file path with documents to index.
                              Optional if path in -i contains a tf-idf index.

-i TF_IDF_INDEX_PATH          hdfs:// or local path to parquet tfidf index.
                              Path location has to end in '/' and contain no files.
                              If -d is specified will save resulting index here.
                              If -d is not specified, will try to load index from here.

-w WORDS                      Comma-separated list of words to search

-n RESULTS                    Optional: number of document results to return from index.
                              If not specified returns all results.
```

Some sample runs are included here. For testing purposes, I used the bbcsport data set[1] which consist of 737 document with 4613 terms from the BBC sport website corresponding to sport news articles in five topical areas from 2004-2005.

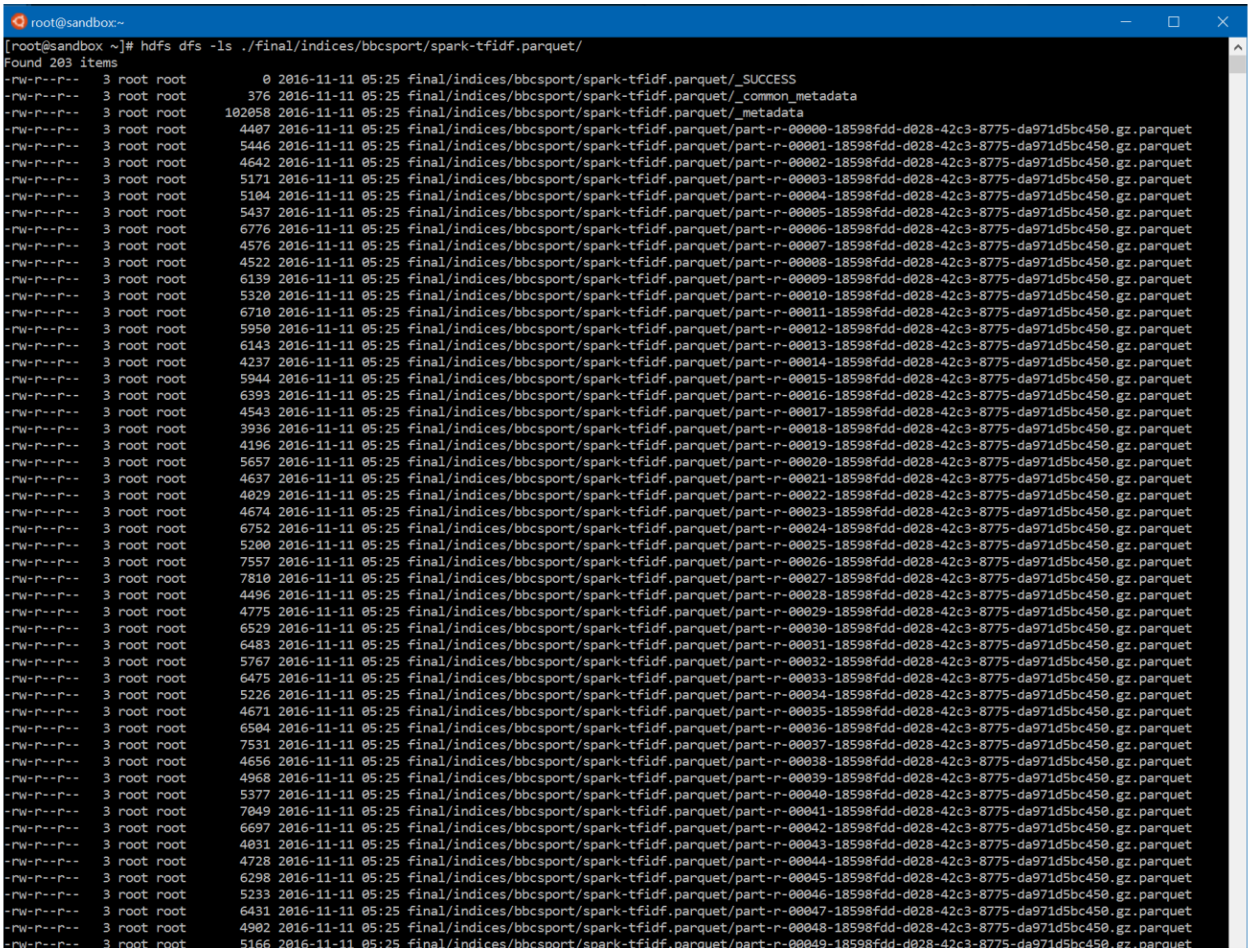
In this sample run the documents are stored in **hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/*** (note the final * to include all documents in the path) and the index will be stored in **hdfs://sandbox.hortonworks.com:8020/user/root/final/indices/bbcsport/ .**

[1] - D. Greene and P. Cunningham. "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering", Proc. ICML 2006. All rights, including copyright, in the content of the original articles are owned by the BBC. Contact Derek Greene <derek.greene@ucd.ie> for further information.
http://mlg.ucd.ie/datasets/bbc.html



```
[root@sandbox final]# spark-submit final.py -d hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/* -i hdfs://sandbox
.hortonworks.com:8020/user/root/final/indices/bbcsport/
16/11/11 05:24:22 INFO SparkContext: Running Spark version 1.6.0
16/11/11 05:24:23 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes w
here applicable
16/11/11 05:24:23 INFO SecurityManager: Changing view acls to: root
16/11/11 05:24:23 INFO SecurityManager: Changing modify acls to: root
16/11/11 05:24:23 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions
: Set(root); users with modify permissions: Set(root)
16/11/11 05:24:24 INFO Utils: Successfully started service 'sparkDriver' on port 45805.
16/11/11 05:24:24 INFO Slf4jLogger: Slf4jLogger started
16/11/11 05:24:24 INFO Remoting: Starting remoting
16/11/11 05:24:24 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkDriverActorSystem@10.0.2.15:38566]
16/11/11 05:24:24 INFO Utils: Successfully started service 'sparkDriverActorSystem' on port 38566.
16/11/11 05:24:24 INFO SparkEnv: Registering MapOutputTracker
16/11/11 05:24:24 INFO SparkEnv: Registering BlockManagerMaster
16/11/11 05:24:24 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-cb462a83-bcab-4254-a154-a6372c71f5ae
16/11/11 05:24:24 INFO MemoryStore: MemoryStore started with capacity 511.5 MB
16/11/11 05:24:24 INFO SparkEnv: Registering OutputCommitCoordinator
16/11/11 05:24:25 INFO Server: jetty-8.y.z-SNAPSHOT
16/11/11 05:24:25 INFO AbstractConnector: Started SelectChannelConnector@0.0.0.0:4040
16/11/11 05:24:25 INFO Utils: Successfully started service 'SparkUI' on port 4040.
16/11/11 05:24:25 INFO SparkUI: Started SparkUI at http://10.0.2.15:4040
16/11/11 05:24:25 INFO Utils: Copying /root/final/final.py to /tmp/spark-eb1219bc-b0e8-4810-bd54-e1986faa54b8/userFiles-7b69429
8-6653-4694-aa79-2827c2420365/final.py
16/11/11 05:24:25 INFO SparkContext: Added file file:/root/final/final.py at file:/root/final/final.py with timestamp 147884186
5368
16/11/11 05:24:25 INFO Executor: Starting executor ID driver on host localhost
16/11/11 05:24:25 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 4
5460.
16/11/11 05:24:25 INFO NettyBlockTransferService: Server created on 45460
16/11/11 05:24:25 INFO BlockManagerMaster: Trying to register BlockManager
16/11/11 05:24:25 INFO BlockManagerMasterEndpoint: Registering block manager localhost:45460 with 511.5 MB RAM, BlockManagerId(
driver, localhost, 45460)
16/11/11 05:24:25 INFO BlockManagerMaster: Registered BlockManager
16/11/11 05:24:26 WARN DomainSocketFactory: The short-circuit local reads feature cannot be used because libhadoop cannot be lo
aded.
16/11/11 05:24:26 INFO EventLoggingListener: Logging events to hdfs:///spark-history/local-1478841865448
INFO Reading files from docs in hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/* ...
INFO Building index for 738 files found...
Saving index to hdfs://sandbox.hortonworks.com:8020/user/root/final/indices/bbcsport/spark-tfidf.parquet ...
SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".
SLF4J: Defaulting to no-operation (NOP) logger implementation
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.
INFO No search terms specified, exiting...
[root@sandbox final]#
```

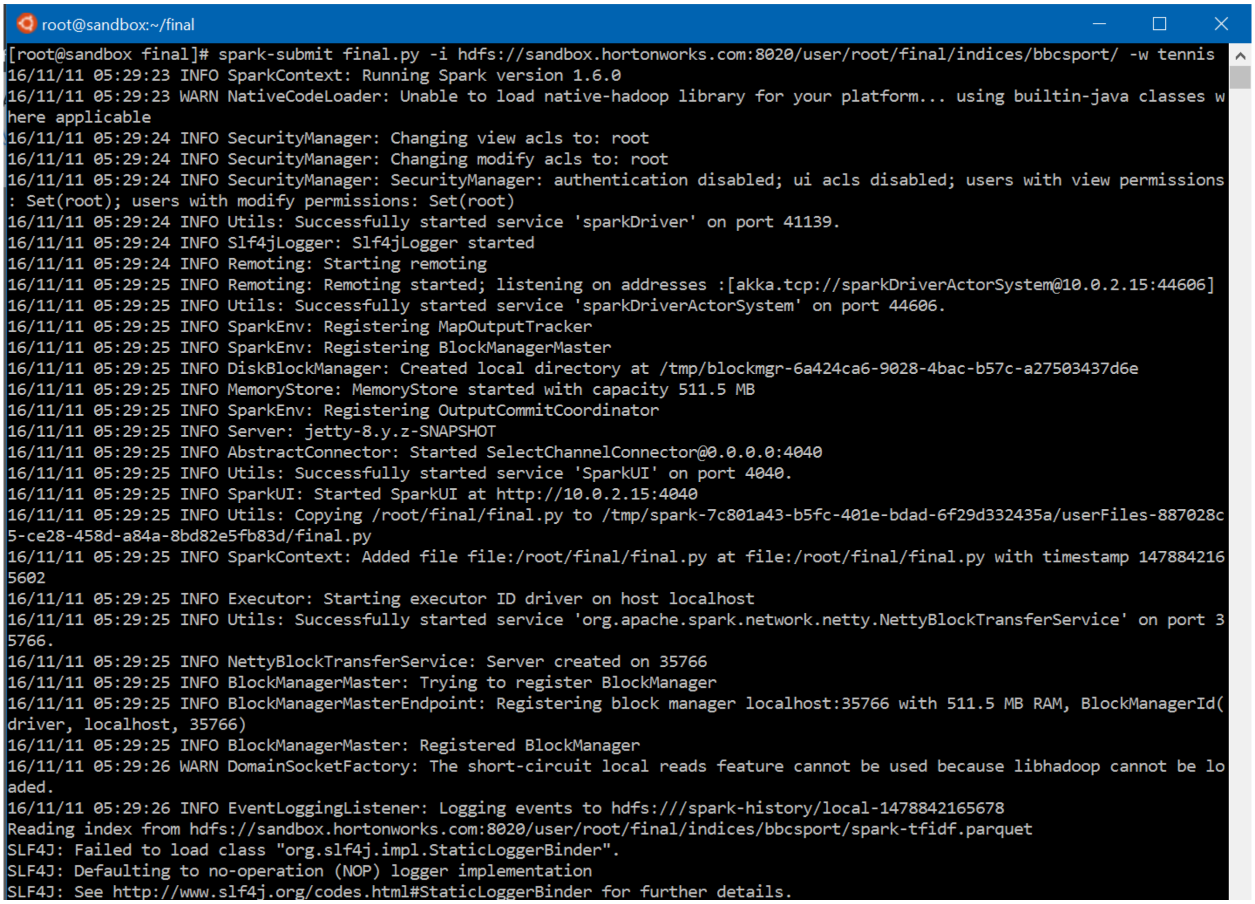
Some exploration in the output folder shows the structure of the parquet files, revealing the index is stored in about 200 files. This is the default value, which can be changed if the data frame is re partitioned before saving.



```
[root@sandbox ~]# hdfs dfs -ls ./final/indices/bbcsport/spark-tfidf.parquet/
Found 203 items
-rw-r--r--   3 root root          0 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/_SUCCESS
-rw-r--r--   3 root root        376 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/_common_metadata
-rw-r--r--   3 root root     102058 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/_metadata
-rw-r--r--   3 root root       4407 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00000-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       5446 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00001-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       4642 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00002-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       5171 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00003-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       5104 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00004-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       5437 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00005-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       6776 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00006-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       4576 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00007-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       4522 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00008-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       6139 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00009-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       5320 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00010-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       6710 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00011-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       5950 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00012-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       6143 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00013-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       4237 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00014-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       5944 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00015-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       6393 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00016-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       4543 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00017-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       3936 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00018-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       4196 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00019-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       5657 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00020-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       4637 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00021-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       4029 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00022-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       4674 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00023-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       6752 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00024-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       5200 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00025-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       7557 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00026-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       7810 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00027-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       4496 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00028-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       4775 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00029-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       6529 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00030-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       6483 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00031-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       5767 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00032-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       6475 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00033-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       5226 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00034-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       4671 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00035-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       6504 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00036-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       7531 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00037-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       4656 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00038-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       4968 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00039-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       5377 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00040-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       7049 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00041-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       6697 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00042-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       4031 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00043-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       4728 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00044-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       6298 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00045-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       5233 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00046-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       6431 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00047-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       4902 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00048-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
-rw-r--r--   3 root root       5166 2016-11-11 05:25 final/indices/bbcsport/spark-tfidf.parquet/part-r-00049-18598fdd-d028-42c3-8775-da971d5bc450.gz.parquet
```

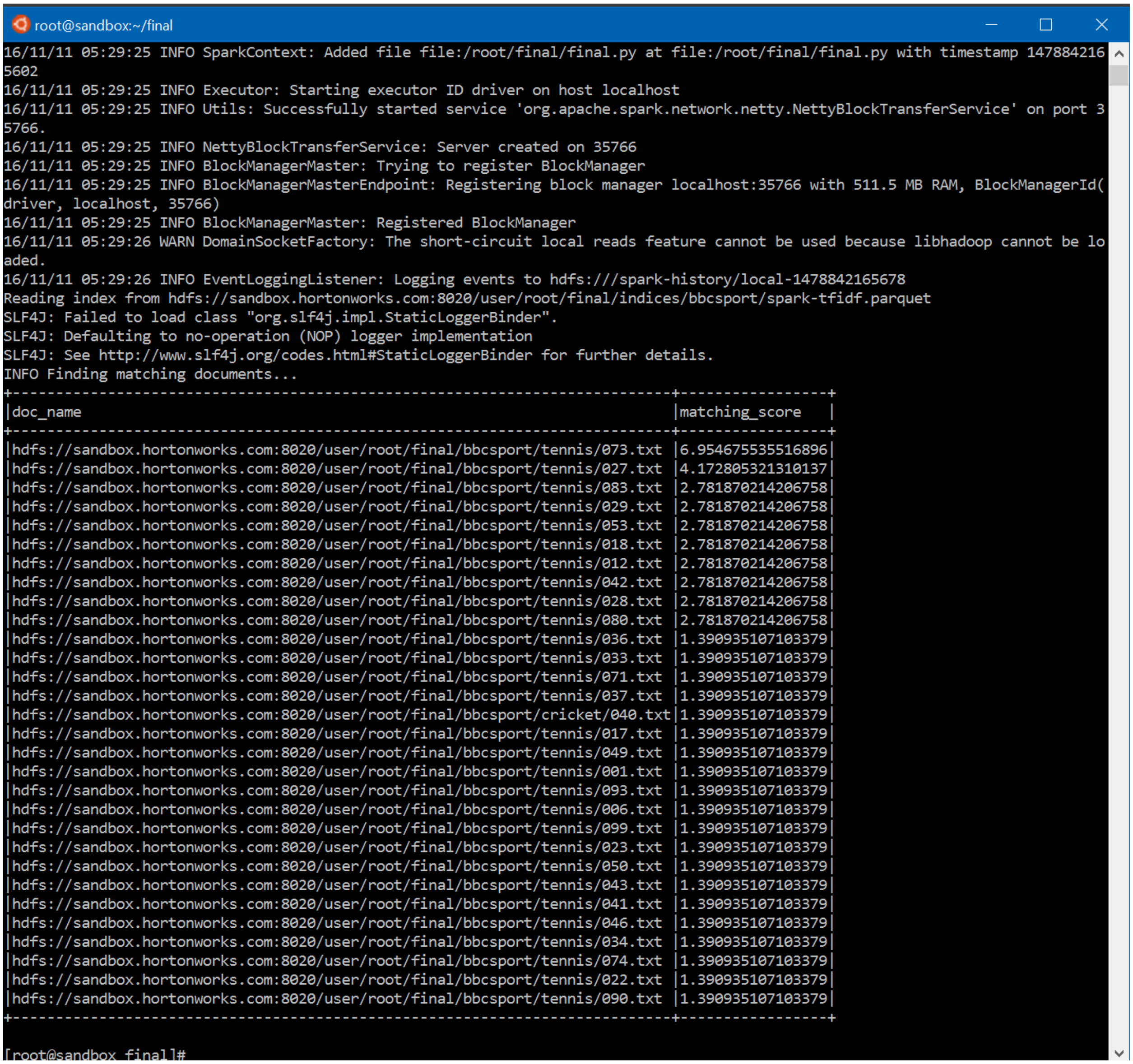
Another test run to obtain search results from the index is here, using the command:

```
spark-submit final.py -i hdfs://sandbox.hortonworks.com:8020/user/root/final/indices/bbcsport/ -w
tennis
```



```
[root@sandbox final]# spark-submit final.py -i hdfs://sandbox.hortonworks.com:8020/user/root/final/indices/bbcsport/ -w tennis
16/11/11 05:29:23 INFO SparkContext: Running Spark version 1.6.0
16/11/11 05:29:23 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes w
here applicable
16/11/11 05:29:24 INFO SecurityManager: Changing view acls to: root
16/11/11 05:29:24 INFO SecurityManager: Changing modify acls to: root
16/11/11 05:29:24 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions
: Set(root); users with modify permissions: Set(root)
16/11/11 05:29:24 INFO Utils: Successfully started service 'sparkDriver' on port 41139.
16/11/11 05:29:24 INFO Slf4jLogger: Slf4jLogger started
16/11/11 05:29:24 INFO Remoting: Starting remoting
16/11/11 05:29:25 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkDriverActorSystem@10.0.2.15:44606]
16/11/11 05:29:25 INFO Utils: Successfully started service 'sparkDriverActorSystem' on port 44606.
16/11/11 05:29:25 INFO SparkEnv: Registering MapOutputTracker
16/11/11 05:29:25 INFO SparkEnv: Registering BlockManagerMaster
16/11/11 05:29:25 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-6a424ca6-9028-4bac-b57c-a27503437d6e
16/11/11 05:29:25 INFO MemoryStore: MemoryStore started with capacity 511.5 MB
16/11/11 05:29:25 INFO SparkEnv: Registering OutputCommitCoordinator
16/11/11 05:29:25 INFO Server: jetty-8.y.z-SNAPSHOT
16/11/11 05:29:25 INFO AbstractConnector: Started SelectChannelConnector@0.0.0.0:4040
16/11/11 05:29:25 INFO Utils: Successfully started service 'SparkUI' on port 4040.
16/11/11 05:29:25 INFO SparkUI: Started SparkUI at http://10.0.2.15:4040
16/11/11 05:29:25 INFO Utils: Copying /root/final/final.py to /tmp/spark-7c801a43-b5fc-401e-bdad-6f29d332435a/userFiles-887028c
5-ce28-458d-a84a-8bd82e5fb83d/final.py
16/11/11 05:29:25 INFO SparkContext: Added file file:/root/final/final.py at file:/root/final/final.py with timestamp 147884216
5602
16/11/11 05:29:25 INFO Executor: Starting executor ID driver on host localhost
16/11/11 05:29:25 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 3
5766.
16/11/11 05:29:25 INFO NettyBlockTransferService: Server created on 35766
16/11/11 05:29:25 INFO BlockManagerMaster: Trying to register BlockManager
16/11/11 05:29:25 INFO BlockManagerMasterEndpoint: Registering block manager localhost:35766 with 511.5 MB RAM, BlockManagerId(
driver, localhost, 35766)
16/11/11 05:29:25 INFO BlockManagerMaster: Registered BlockManager
16/11/11 05:29:26 WARN DomainSocketFactory: The short-circuit local reads feature cannot be used because libhadoop cannot be lo
aded.
16/11/11 05:29:26 INFO EventLoggingListener: Logging events to hdfs:///spark-history/local-1478842165678
Reading index from hdfs://sandbox.hortonworks.com:8020/user/root/final/indices/bbcsport/spark-tfidf.parquet
SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".
SLF4J: Defaulting to no-operation (NOP) logger implementation
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.
```

(contd.)

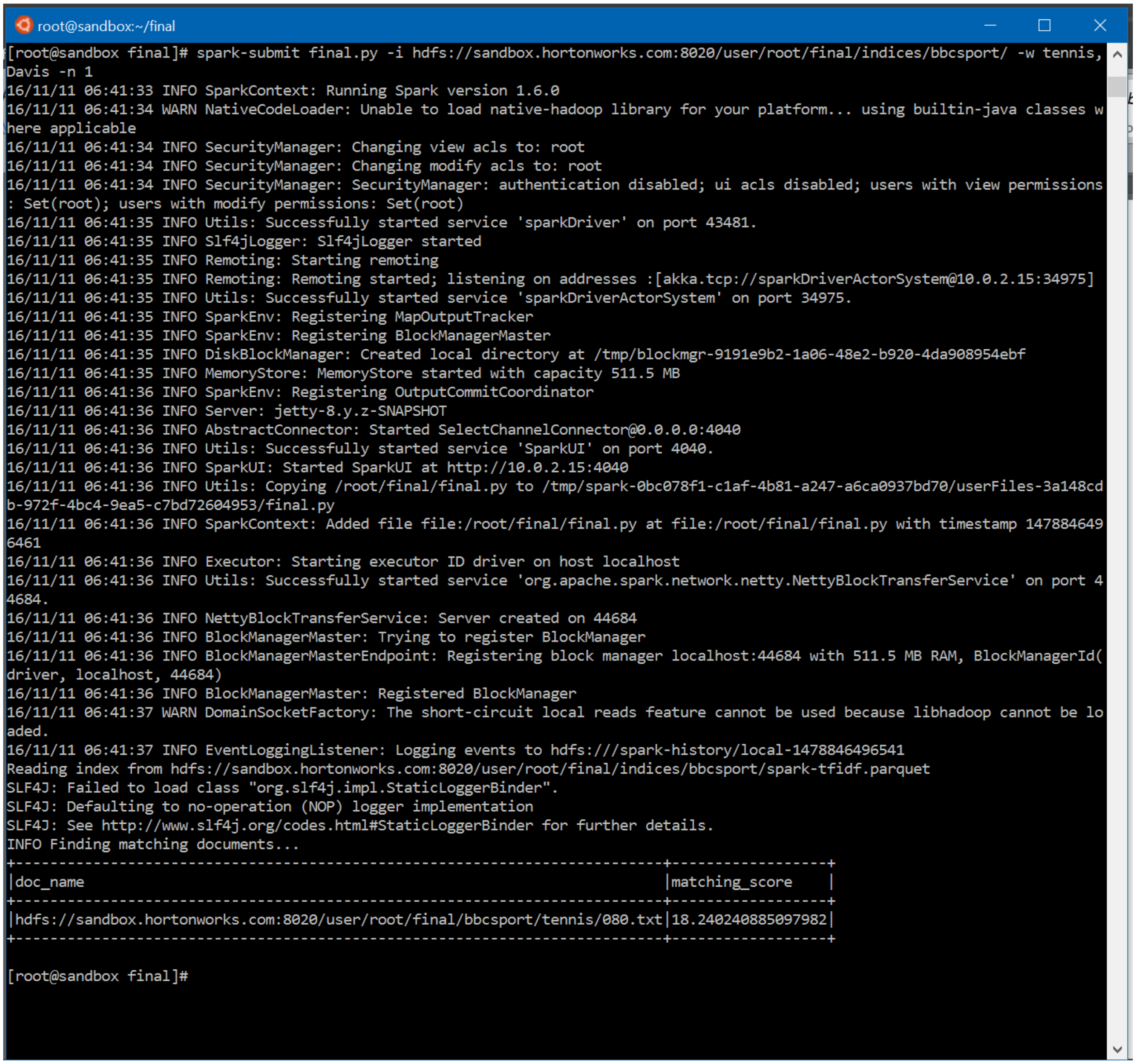


```
16/11/11 05:29:25 INFO SparkContext: Added file file:/root/final/final.py at file:/root/final/final.py with timestamp 147884216
5602
16/11/11 05:29:25 INFO Executor: Starting executor ID driver on host localhost
16/11/11 05:29:25 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 3
5766.
16/11/11 05:29:25 INFO NettyBlockTransferService: Server created on 35766
16/11/11 05:29:25 INFO BlockManagerMaster: Trying to register BlockManager
16/11/11 05:29:25 INFO BlockManagerMasterEndpoint: Registering block manager localhost:35766 with 511.5 MB RAM, BlockManagerId(
driver, localhost, 35766)
16/11/11 05:29:25 INFO BlockManagerMaster: Registered BlockManager
16/11/11 05:29:26 WARN DomainSocketFactory: The short-circuit local reads feature cannot be used because libhadoop cannot be lo
aded.
16/11/11 05:29:26 INFO EventLoggingListener: Logging events to hdfs:///spark-history/local-1478842165678
Reading index from hdfs://sandbox.hortonworks.com:8020/user/root/final/indices/bbcsport/spark-tfidf.parquet
SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".
SLF4J: Defaulting to no-operation (NOP) logger implementation
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.
INFO Finding matching documents...
+----------------------------------------------------------------------------+-----------------+
|doc_name                                                                    |matching_score   |
+----------------------------------------------------------------------------+-----------------+
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/073.txt |6.954675535516896|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/027.txt |4.172805321310137|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/083.txt |2.781870214206758|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/029.txt |2.781870214206758|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/053.txt |2.781870214206758|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/018.txt |2.781870214206758|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/012.txt |2.781870214206758|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/042.txt |2.781870214206758|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/028.txt |2.781870214206758|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/080.txt |2.781870214206758|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/036.txt |1.390935107103379|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/033.txt |1.390935107103379|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/071.txt |1.390935107103379|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/037.txt |1.390935107103379|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/cricket/040.txt|1.390935107103379|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/017.txt |1.390935107103379|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/049.txt |1.390935107103379|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/001.txt |1.390935107103379|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/093.txt |1.390935107103379|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/006.txt |1.390935107103379|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/099.txt |1.390935107103379|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/023.txt |1.390935107103379|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/050.txt |1.390935107103379|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/043.txt |1.390935107103379|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/041.txt |1.390935107103379|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/046.txt |1.390935107103379|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/034.txt |1.390935107103379|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/074.txt |1.390935107103379|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/022.txt |1.390935107103379|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/090.txt |1.390935107103379|
+----------------------------------------------------------------------------+-----------------+

[root@sandbox final]#
```

Results

Query 1: tennis, Davis

Command: `spark-submit final.py -i hdfs://sandbox.hortonworks.com:8020/user/root/final/indices/bbcsport/ -w tennis,Davis -n 1`



```
[root@sandbox final]# spark-submit final.py -i hdfs://sandbox.hortonworks.com:8020/user/root/final/indices/bbcsport/ -w tennis,
Davis -n 1
16/11/11 06:41:33 INFO SparkContext: Running Spark version 1.6.0
16/11/11 06:41:34 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes w
here applicable
16/11/11 06:41:34 INFO SecurityManager: Changing view acls to: root
16/11/11 06:41:34 INFO SecurityManager: Changing modify acls to: root
16/11/11 06:41:34 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions
: Set(root); users with modify permissions: Set(root)
16/11/11 06:41:35 INFO Utils: Successfully started service 'sparkDriver' on port 43481.
16/11/11 06:41:35 INFO Slf4jLogger: Slf4jLogger started
16/11/11 06:41:35 INFO Remoting: Starting remoting
16/11/11 06:41:35 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkDriverActorSystem@10.0.2.15:34975]
16/11/11 06:41:35 INFO Utils: Successfully started service 'sparkDriverActorSystem' on port 34975.
16/11/11 06:41:35 INFO SparkEnv: Registering MapOutputTracker
16/11/11 06:41:35 INFO SparkEnv: Registering BlockManagerMaster
16/11/11 06:41:35 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-9191e9b2-1a06-48e2-b920-4da908954ebf
16/11/11 06:41:35 INFO MemoryStore: MemoryStore started with capacity 511.5 MB
16/11/11 06:41:36 INFO SparkEnv: Registering OutputCommitCoordinator
16/11/11 06:41:36 INFO Server: jetty-8.y.z-SNAPSHOT
16/11/11 06:41:36 INFO AbstractConnector: Started SelectChannelConnector@0.0.0.0:4040
16/11/11 06:41:36 INFO Utils: Successfully started service 'SparkUI' on port 4040.
16/11/11 06:41:36 INFO SparkUI: Started SparkUI at http://10.0.2.15:4040
16/11/11 06:41:36 INFO Utils: Copying /root/final/final.py to /tmp/spark-0bc078f1-c1af-4b81-a247-a6ca0937bd70/userFiles-3a148cd
b-972f-4bc4-9ea5-c7bd72604953/final.py
16/11/11 06:41:36 INFO SparkContext: Added file file:/root/final/final.py at file:/root/final/final.py with timestamp 147884649
6461
16/11/11 06:41:36 INFO Executor: Starting executor ID driver on host localhost
16/11/11 06:41:36 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 4
4684.
16/11/11 06:41:36 INFO NettyBlockTransferService: Server created on 44684
16/11/11 06:41:36 INFO BlockManagerMaster: Trying to register BlockManager
16/11/11 06:41:36 INFO BlockManagerMasterEndpoint: Registering block manager localhost:44684 with 511.5 MB RAM, BlockManagerId(
driver, localhost, 44684)
16/11/11 06:41:36 INFO BlockManagerMaster: Registered BlockManager
16/11/11 06:41:37 WARN DomainSocketFactory: The short-circuit local reads feature cannot be used because libhadoop cannot be lo
aded.
16/11/11 06:41:37 INFO EventLoggingListener: Logging events to hdfs:///spark-history/local-1478846496541
Reading index from hdfs://sandbox.hortonworks.com:8020/user/root/final/indices/bbcsport/spark-tfidf.parquet
SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".
SLF4J: Defaulting to no-operation (NOP) logger implementation
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.
INFO Finding matching documents...
+---------------------------------------------------------------------------+------------------+
|doc_name                                                                   |matching_score    |
+---------------------------------------------------------------------------+------------------+
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/080.txt|18.240240885097982|
+---------------------------------------------------------------------------+------------------+


[root@sandbox final]#
```

Results

Doc_name	Matching_score
/bbcsport/tennis/080.txt	18.24

/bbcsport/tennis/080.txt

Henman decides to quit Davis Cup

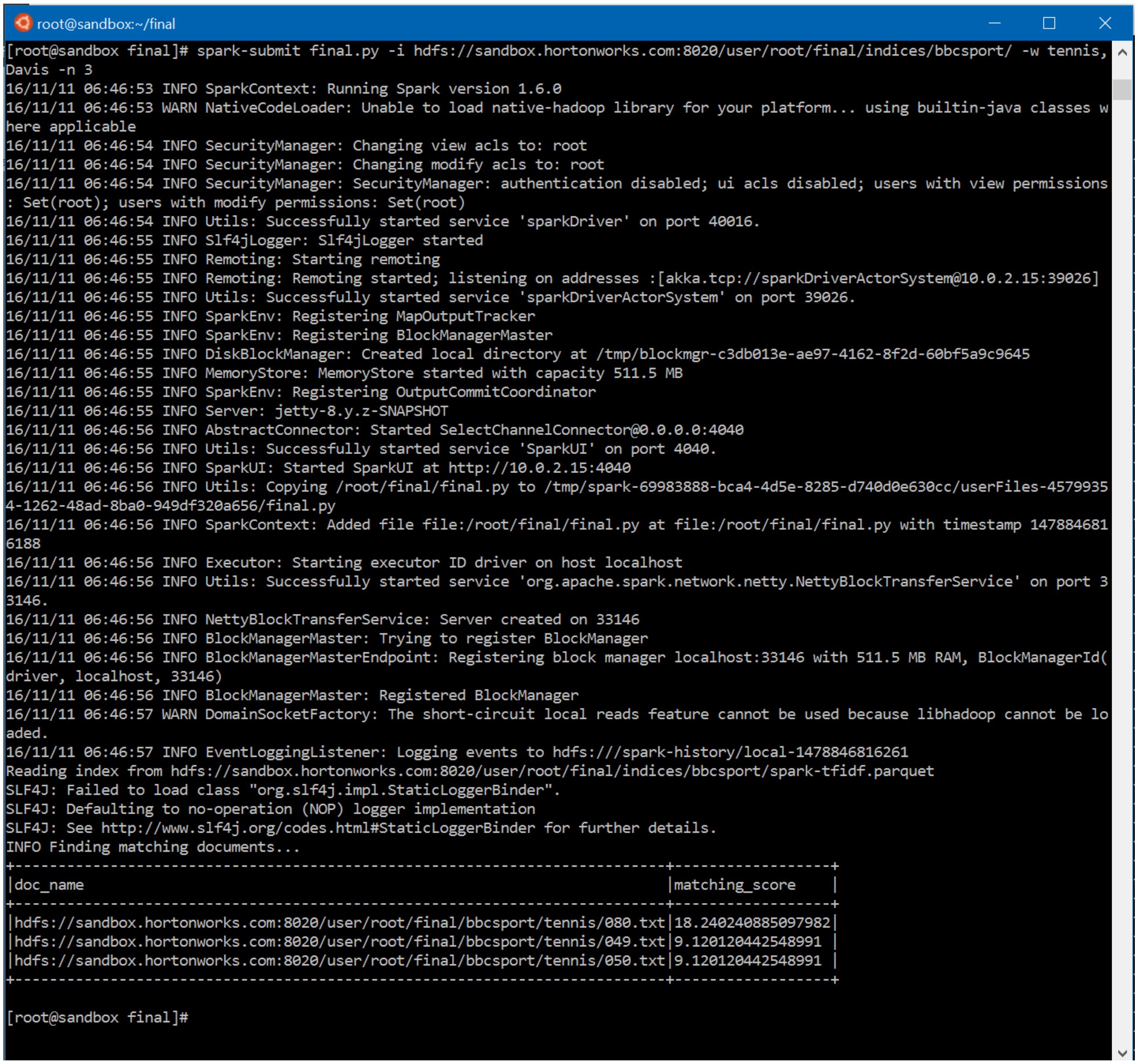
Tim Henman has retired from Great Britain's Davis Cup team.

The 30-year-old, who made his Davis Cup debut in 1994, is now set to fully focus on the ATP Tour and on winning his first Grand Slam event. "I've made no secret of the fact that representing Great Britain has always been a top priority for me throughout my career," Henman told his website. Captain Jeremy Bates has touted Alex Bogdanovic and Andrew Murray as possible replacements for the veteran. Henman added that he was available to help Britain in its bid for Davis Cup success, with the next tie against Israel in March . "Although I won't be playing, I would still like to make myself available to both Jeremy and the LTA in the future so that I can draw upon my experience in the hope of trying to help the British players develop their full potential," he added. "I've really enjoyed playing in front of the thousands of British fans both home and abroad and would like to thank every one of them for their unwavering support over the years." Henman leaves Davis Cup tennis with an impressive record, having won 36 of his 50 matches. Great Britain captain Jeremy Bates paid tribute to Henman's efforts over the years.

"Tim has quite simply had a phenomenal Davis Cup career and it has been an absolute privilege to have captained the team with him in it," said Bates. "Tim's magnificent record speaks for itself. While it's a great loss I completely understand and respect his decision to retire from Davis Cup and focus on the Grand Slams and Tour. " "Looking to the future this decision obviously marks a watershed in British Davis Cup tennis but it is also a huge opportunity for the next generation to make their mark. "We have a host of talented players coming through and despite losing someone of Tim's calibre, I remain very optimistic about the future." Henman made his Davis Cup debut in 1994 against Romania in Manchester. He and partner Bates won their doubles rubber on the middle Saturday of the tie. Britain eventually lost the contest 3-2. Henman and Britain had little luck in Davis Cup matches until 1999 when they qualified for the World Group. Britain drew the USA and lost the tie when Greg Rusedski fell to Jim Courier in the deciding rubber. They made the final stages again, in 2002, but this time lost out to the might of Sweden.

Query 2: tennis, Davis

Command: `spark-submit final.py -i hdfs://sandbox.hortonworks.com:8020/user/root/final/indices/bbcsport/ -w tennis,Davis -n 3`



```
[root@sandbox final]# spark-submit final.py -i hdfs://sandbox.hortonworks.com:8020/user/root/final/indices/bbcsport/ -w tennis,
Davis -n 3
16/11/11 06:46:53 INFO SparkContext: Running Spark version 1.6.0
16/11/11 06:46:53 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes w
here applicable
16/11/11 06:46:54 INFO SecurityManager: Changing view acls to: root
16/11/11 06:46:54 INFO SecurityManager: Changing modify acls to: root
16/11/11 06:46:54 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions
: Set(root); users with modify permissions: Set(root)
16/11/11 06:46:54 INFO Utils: Successfully started service 'sparkDriver' on port 40016.
16/11/11 06:46:55 INFO Slf4jLogger: Slf4jLogger started
16/11/11 06:46:55 INFO Remoting: Starting remoting
16/11/11 06:46:55 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkDriverActorSystem@10.0.2.15:39026]
16/11/11 06:46:55 INFO Utils: Successfully started service 'sparkDriverActorSystem' on port 39026.
16/11/11 06:46:55 INFO SparkEnv: Registering MapOutputTracker
16/11/11 06:46:55 INFO SparkEnv: Registering BlockManagerMaster
16/11/11 06:46:55 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-c3db013e-ae97-4162-8f2d-60bf5a9c9645
16/11/11 06:46:55 INFO MemoryStore: MemoryStore started with capacity 511.5 MB
16/11/11 06:46:55 INFO SparkEnv: Registering OutputCommitCoordinator
16/11/11 06:46:55 INFO Server: jetty-8.y.z-SNAPSHOT
16/11/11 06:46:56 INFO AbstractConnector: Started SelectChannelConnector@0.0.0.0:4040
16/11/11 06:46:56 INFO Utils: Successfully started service 'SparkUI' on port 4040.
16/11/11 06:46:56 INFO SparkUI: Started SparkUI at http://10.0.2.15:4040
16/11/11 06:46:56 INFO Utils: Copying /root/final/final.py to /tmp/spark-69983888-bca4-4d5e-8285-d740d0e630cc/userFiles-4579935
4-1262-48ad-8ba0-949df320a656/final.py
16/11/11 06:46:56 INFO SparkContext: Added file file:/root/final/final.py at file:/root/final/final.py with timestamp 147884681
6188
16/11/11 06:46:56 INFO Executor: Starting executor ID driver on host localhost
16/11/11 06:46:56 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 3
3146.
16/11/11 06:46:56 INFO NettyBlockTransferService: Server created on 33146
16/11/11 06:46:56 INFO BlockManagerMaster: Trying to register BlockManager
16/11/11 06:46:56 INFO BlockManagerMasterEndpoint: Registering block manager localhost:33146 with 511.5 MB RAM, BlockManagerId(
driver, localhost, 33146)
16/11/11 06:46:56 INFO BlockManagerMaster: Registered BlockManager
16/11/11 06:46:57 WARN DomainSocketFactory: The short-circuit local reads feature cannot be used because libhadoop cannot be lo
aded.
16/11/11 06:46:57 INFO EventLoggingListener: Logging events to hdfs:///spark-history/local-1478846816261
Reading index from hdfs://sandbox.hortonworks.com:8020/user/root/final/indices/bbcsport/spark-tfidf.parquet
SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".
SLF4J: Defaulting to no-operation (NOP) logger implementation
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.
INFO Finding matching documents...
+---------------------------------------------------------------------------+------------------+
|doc_name                                                                   |matching_score    |
+---------------------------------------------------------------------------+------------------+
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/080.txt|18.240240885097982|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/049.txt|9.120120442548991 |
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/050.txt|9.120120442548991 |
+---------------------------------------------------------------------------+------------------+

[root@sandbox final]#
```

Results

Doc_name	Matching_score
/bbcsport/tennis/080.txt	18.24
/bbcsport/tennis/049.txt	9.12
/bbcsport/tennis/0.50.txt	9.12

/bbcsport/tennis/080.txt

Henman decides to quit Davis Cup

Tim Henman has retired from Great Britain's Davis Cup team.

The 30-year-old, who made his Davis Cup debut in 1994, is now set to fully focus on the ATP Tour and on winning his first Grand Slam event. "I've made no secret of the fact that representing Great Britain has always been a top priority for me throughout my career," Henman told his website. Captain Jeremy Bates has touted Alex Bogdanovic and Andrew Murray as possible replacements for the veteran. Henman added that he was available to help Britain in its bid for Davis Cup success, with the next tie against Israel in March . "Although I won't be playing, I would still like to make myself available to both Jeremy and the LTA in the future so that I can draw upon my experience in the hope of trying to help the British players develop their full potential," he added. "I've really enjoyed playing in front of the thousands of British fans both home and abroad and would like to thank every one of them for their unwavering support over the years." Henman leaves Davis Cup tennis with an impressive record, having won 36 of his 50 matches. Great Britain captain Jeremy Bates paid tribute to Henman's efforts over the years.

"Tim has quite simply had a phenomenal Davis Cup career and it has been an absolute privilege to have captained the team with him in it," said Bates. "Tim's magnificent record speaks for itself. While it's a great loss I completely understand and respect his decision to retire from Davis Cup and focus on the Grand Slams and Tour. " "Looking to the future this decision obviously marks a watershed in British Davis Cup tennis but it is also a huge opportunity for the next generation to make their mark. "We have a host of talented players coming through and despite losing someone of Tim's calibre, I remain very optimistic about the future." Henman made his Davis Cup debut in 1994 against Romania in Manchester. He and partner Bates won their doubles rubber on the middle Saturday of the tie. Britain eventually lost the contest 3-2. Henman and Britain had little luck in Davis Cup matches until 1999 when they qualified for the World Group. Britain drew the USA and lost the tie when Greg Rusedski fell to Jim Courier in the deciding rubber. They made the final stages again, in 2002, but this time lost out to the might of Sweden.

/bbcsport/tennis/049.txt

Moya emotional after Davis Cup win

Carlos Moya described Spain's Davis Cup victory as the highlight of his career after he beat Andy Roddick to end the USA's challenge in Seville.

Moya made up for missing Spain's 2000 victory through injury by beating Roddick 6-2 7-6 (7-1) 7-6 (7-5) to give the hosts an unassailable 3-1 lead. "I have woken up so many nights dreaming of this day," said Moya. "All my energy has been focused on today. "What I have lived today I do not think I will live again." Spain's only other Davis Cup title came two years ago in Valencia, when they beat Australia. And Moya, nicknamed Charly, admitted: "The Davis Cup is my dream and I was a bit nervous at the outset. "Some people have said that I am obsessed but I think that it is better this way. It helps me reach my goals if I am obsessed. "It's really incredible - to get the winning point is really something." Spanish captain Jordi Arrese said: "Charly played a great game. It was his opportunity and he hasn't let us down. "He had lost three times to Roddick, and this was his day to beat him. "He had been waiting years to be in this position." Spain's victory was also remarkable for the performance of Rafael Nadal, who beat Roddick in the opening singles.

Aged 18 years and 185 days, the Mallorcan became the youngest player to win the Davis Cup. "What a great way to finish the year," said Nadal afterwards. US coach Patrick McEnroe wants Roddick and the rest of his team to play more tennis on clay and hone their skills on the surface. "I think it will help these guys even on slow hard courts to learn how to mix things up a little bit and to play a little bit smarter and tactically better." "Obviously it's unrealistic to say that we're going to just start playing constantly on clay, with the schedule. "But certainly I think we can put the work in at the appropriate time and play a couple more events and play against these guys who are the best on this stuff," said McEnroe. Roddick was left frustrated after losing both his singles on the slow clay of Seville's Olympic Stadium. "It's just tough because I felt like I was in it the whole time against one of the top three clay-courters in the world," said the American. "I had my chances and just didn't convert them. The bottom line is they were just better than us this weekend. "They came out, took care of business and they beat us. It's as simple as that."

/bbcsport/tennis/050.txt

Moya emotional at Davis Cup win

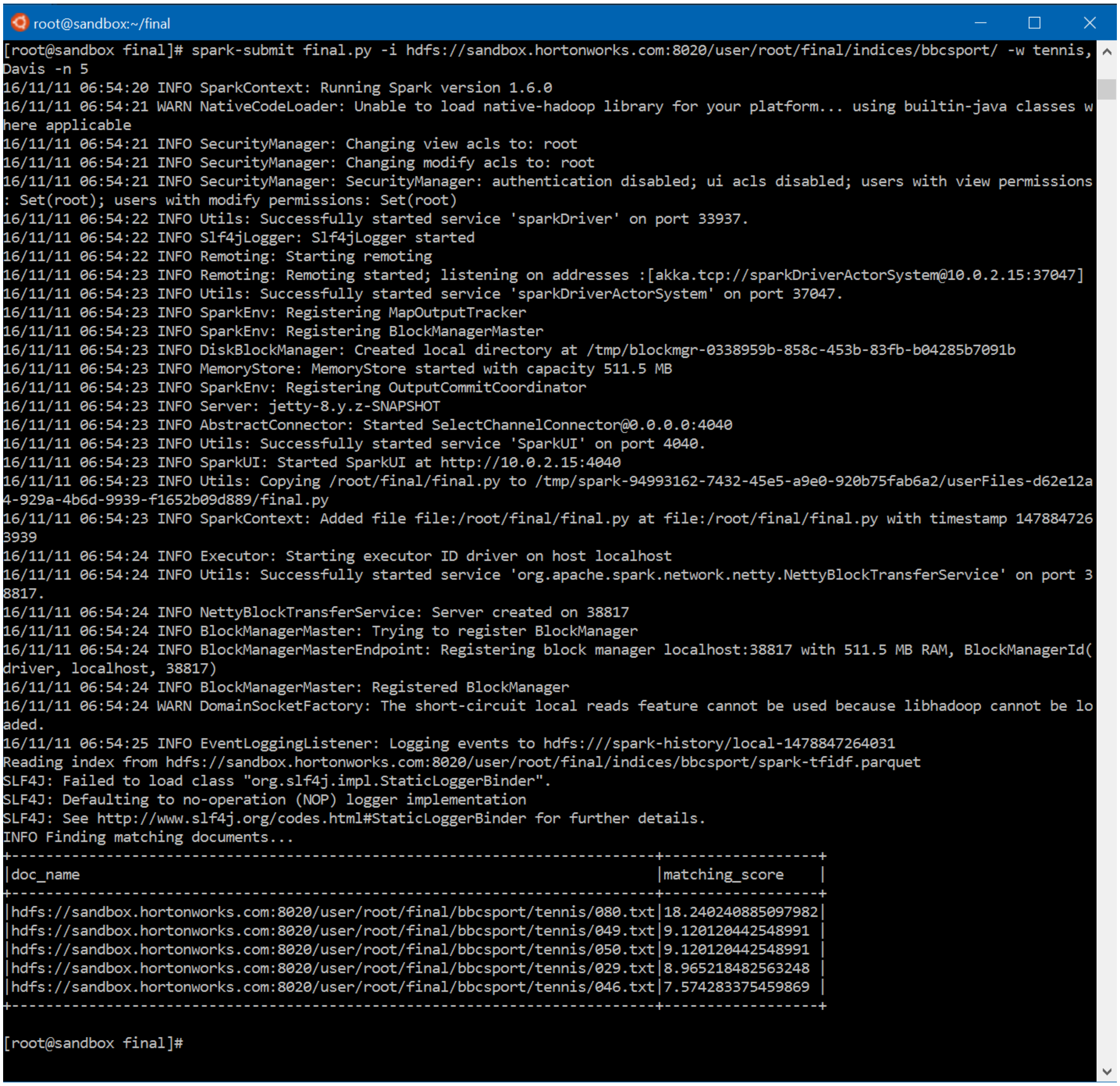
Carlos Moya described Spain's Davis Cup victory as the highlight of his career after he beat Andy Roddick to end the USA's challenge in Seville.

Moya made up for missing Spain's 2000 victory through injury by beating Roddick 6-2 7-6 (7-1) 7-6 (7-5) to give the hosts an unassailable 3-1 lead. "I have woken up so many nights dreaming of this day," said Moya. "All my energy has been focused on today. "What I have lived today I do not think I will live again." Spain's only other Davis Cup title came two years ago in Valencia, when they beat Australia. And Moya, nicknamed Charly, admitted: "The Davis Cup is my dream and I was a bit nervous at the outset. "Some people have said that I am obsessed but I think that it is better this way. It helps me reach my goals if I am obsessed. "It's really incredible - to get the winning point is really something." Spanish captain Jordi Arrese said: "Charly played a great game. It was his opportunity and he hasn't let us down. "He had lost three times to Roddick, and this was his day to beat him. "He had been waiting years to be in this position." Spain's victory was also remarkable for the performance of Rafael Nadal, who beat Roddick in the opening singles.

Aged 18 years and 185 days, the Mallorcan became the youngest player to win the Davis Cup. "What a great way to finish the year," said Nadal afterwards. US coach Patrick McEnroe wants Roddick and the rest of his team to play more tennis on clay and hone their skills on the surface. "I think it will help these guys even on slow hard courts to learn how to mix things up a little bit and to play a little bit smarter and tactically better." "Obviously it's unrealistic to say that we're going to just start playing constantly on clay, with the schedule. "But certainly I think we can put the work in at the appropriate time and play a couple more events and play against these guys who are the best on this stuff," said McEnroe. Roddick was left frustrated after losing both his singles on the slow clay of Seville's Olympic Stadium. "It's just tough because I felt like I was in it the whole time against one of the top three clay-courters in the world," said the American. "I had my chances and just didn't convert them. The bottom line is they were just better than us this weekend. "They came out, took care of business and they beat us. It's as simple as that.".

Query 3: tennis, Davis

Command: `spark-submit final.py -i hdfs://sandbox.hortonworks.com:8020/user/root/final/indices/bbcsport/ -w tennis,Davis -n 5`



```
[root@sandbox final]# spark-submit final.py -i hdfs://sandbox.hortonworks.com:8020/user/root/final/indices/bbcsport/ -w tennis,
Davis -n 5
16/11/11 06:54:20 INFO SparkContext: Running Spark version 1.6.0
16/11/11 06:54:21 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes w
here applicable
16/11/11 06:54:21 INFO SecurityManager: Changing view acls to: root
16/11/11 06:54:21 INFO SecurityManager: Changing modify acls to: root
16/11/11 06:54:21 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions
: Set(root); users with modify permissions: Set(root)
16/11/11 06:54:22 INFO Utils: Successfully started service 'sparkDriver' on port 33937.
16/11/11 06:54:22 INFO Slf4jLogger: Slf4jLogger started
16/11/11 06:54:22 INFO Remoting: Starting remoting
16/11/11 06:54:23 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkDriverActorSystem@10.0.2.15:37047]
16/11/11 06:54:23 INFO Utils: Successfully started service 'sparkDriverActorSystem' on port 37047.
16/11/11 06:54:23 INFO SparkEnv: Registering MapOutputTracker
16/11/11 06:54:23 INFO SparkEnv: Registering BlockManagerMaster
16/11/11 06:54:23 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-0338959b-858c-453b-83fb-b04285b7091b
16/11/11 06:54:23 INFO MemoryStore: MemoryStore started with capacity 511.5 MB
16/11/11 06:54:23 INFO SparkEnv: Registering OutputCommitCoordinator
16/11/11 06:54:23 INFO Server: jetty-8.y.z-SNAPSHOT
16/11/11 06:54:23 INFO AbstractConnector: Started SelectChannelConnector@0.0.0.0:4040
16/11/11 06:54:23 INFO Utils: Successfully started service 'SparkUI' on port 4040.
16/11/11 06:54:23 INFO SparkUI: Started SparkUI at http://10.0.2.15:4040
16/11/11 06:54:23 INFO Utils: Copying /root/final/final.py to /tmp/spark-94993162-7432-45e5-a9e0-920b75fab6a2/userFiles-d62e12a
4-929a-4b6d-9939-f1652b09d889/final.py
16/11/11 06:54:23 INFO SparkContext: Added file file:/root/final/final.py at file:/root/final/final.py with timestamp 147884726
3939
16/11/11 06:54:24 INFO Executor: Starting executor ID driver on host localhost
16/11/11 06:54:24 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 3
8817.
16/11/11 06:54:24 INFO NettyBlockTransferService: Server created on 38817
16/11/11 06:54:24 INFO BlockManagerMaster: Trying to register BlockManager
16/11/11 06:54:24 INFO BlockManagerMasterEndpoint: Registering block manager localhost:38817 with 511.5 MB RAM, BlockManagerId(
driver, localhost, 38817)
16/11/11 06:54:24 INFO BlockManagerMaster: Registered BlockManager
16/11/11 06:54:24 WARN DomainSocketFactory: The short-circuit local reads feature cannot be used because libhadoop cannot be lo
aded.
16/11/11 06:54:25 INFO EventLoggingListener: Logging events to hdfs:///spark-history/local-1478847264031
Reading index from hdfs://sandbox.hortonworks.com:8020/user/root/final/indices/bbcsport/spark-tfidf.parquet
SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".
SLF4J: Defaulting to no-operation (NOP) logger implementation
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.
INFO Finding matching documents...
+---------------------------------------------------------------------------+------------------+
|doc_name                                                                   |matching_score    |
+---------------------------------------------------------------------------+------------------+
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/080.txt|18.240240885097982|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/049.txt|9.12012044254899  |
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/050.txt|9.12012044254899  |
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/029.txt|8.96521848256324  |
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/tennis/046.txt|7.57428337545986  |
+---------------------------------------------------------------------------+------------------+

[root@sandbox final]#
```

Results

Doc_name	Matching_score
/bbcsport/tennis/080.txt	18.24
/bbcsport/tennis/049.txt	9.12
/bbcsport/tennis/050.txt	9.12
/bbcsport/tennis/029.txt	8.95
/bbcsport/tennis/046.txt	7.57

/bbcsport/tennis/080.txt

Henman decides to quit Davis Cup

Tim Henman has retired from Great Britain's Davis Cup team.

The 30-year-old, who made his Davis Cup debut in 1994, is now set to fully focus on the ATP Tour and on winning his first Grand Slam event. "I've made no secret of the fact that representing Great Britain has always been a top priority for me throughout my career," Henman told his website. Captain Jeremy Bates has touted Alex Bogdanovic and Andrew Murray as possible replacements for the veteran. Henman added that he was available to help Britain in its bid for Davis Cup success, with the next tie against Israel in March . "Although I won't be playing, I would still like to make myself available to both Jeremy and the LTA in the future so that I can draw upon my experience in the hope of trying to help the British players develop their full potential," he added. "I've really enjoyed playing in front of the thousands of British fans both home and abroad and would like to thank every one of them for their unwavering support over the years." Henman leaves Davis Cup tennis with an impressive record, having won 36 of his 50 matches. Great Britain captain Jeremy Bates paid tribute to Henman's efforts over the years.

"Tim has quite simply had a phenomenal Davis Cup career and it has been an absolute privilege to have captained the team with him in it," said Bates. "Tim's magnificent record speaks for itself. While it's a great loss I completely understand and respect his decision to retire from Davis Cup and focus on the Grand Slams and Tour. " "Looking to the future this decision obviously marks a watershed in British Davis Cup tennis but it is also a huge opportunity for the next generation to make their mark. "We have a host of talented players coming through and despite losing someone of Tim's calibre, I remain very optimistic about the future." Henman made his Davis Cup debut in 1994 against Romania in Manchester. He and partner Bates won their doubles rubber on the middle Saturday of the tie. Britain eventually lost the contest 3-2. Henman and Britain had little luck in Davis Cup matches until 1999 when they qualified for the World Group. Britain drew the USA and lost the tie when Greg Rusedski fell to Jim Courier in the deciding rubber. They made the final stages again, in 2002, but this time lost out to the might of Sweden.

/bbcsport/tennis/049.txt

Moya emotional after Davis Cup win

Carlos Moya described Spain's Davis Cup victory as the highlight of his career after he beat Andy Roddick to end the USA's challenge in Seville.

Moya made up for missing Spain's 2000 victory through injury by beating Roddick 6-2 7-6 (7-1) 7-6 (7-5) to give the hosts an unassailable 3-1 lead. "I have woken up so many nights dreaming of this day," said Moya. "All my energy has been focused on today. "What I have lived today I do not think I will live again." Spain's only other Davis Cup title came two years ago in Valencia, when they beat Australia. And Moya, nicknamed Charly, admitted: "The Davis Cup is my dream and I was a bit nervous at the outset. "Some people have said that I am obsessed but I think that it is better this way. It helps me reach my goals if I am obsessed. "It's really incredible - to get the winning point is really something." Spanish captain Jordi Arrese said: "Charly played a great game. It was his opportunity and he hasn't let us down. "He had lost three times to Roddick, and this was his day to beat him. "He had been waiting years to be in this position." Spain's victory was also remarkable for the performance of Rafael Nadal, who beat Roddick in the opening singles.

Aged 18 years and 185 days, the Mallorcan became the youngest player to win the Davis Cup. "What a great way to finish the year," said Nadal afterwards. US coach Patrick McEnroe wants Roddick and the rest of his team to play more tennis on clay and hone their skills on the surface. "I think it will help these guys even on slow hard courts to learn how to mix things up a little bit and to play a little bit smarter and tactically better." "Obviously it's unrealistic to say that we're going to just start playing constantly on clay, with the schedule. "But certainly I think we can put the work in at the appropriate time and play a couple more events and play against these guys who are the best on this stuff," said McEnroe. Roddick was left frustrated after losing both his singles on the slow clay of Seville's Olympic Stadium. "It's just tough because I felt like I was in it the whole time against one of the top three clay-courters in the world," said the American. "I had my chances and just didn't convert them. The bottom line is they were just better than us this weekend. "They came out, took care of business and they beat us. It's as simple as that."

/bbcsport/tennis/050.txt

What now for British tennis?

Tim Henman's decision to quit Davis Cup tennis has left the British team with a gargantuan void to fill.

The world number seven is tied for fourth among his countrymen for wins in the history of the tournament (he has 36 from his 50 rubbers). And Great Britain's last Davis Cup win without Henman came against Slovenia as far back as 1996. Worse could follow, according to former British team member Chris Bailey. Bailey told BBC Sport: "After Tim's announcement, I doubt Greg Rusedski will be that far behind him." But without their top two, where does that leave British ambitions in the sport's premier team event? Captain Jeremy Bates has singled out Alex Bogdanovic and Andrew Murray as potential replacements. The Yugoslavian-born Bogdanovic, though, is 184 places below Henman in the world rankings and has played just two cup ties - winning one and losing the other.

Murray, on the other hand, is 407th in the current ATP entry list and yet to make his cup debut. But Bailey does see some hope for the future. He said: "Now we've dropped down to the Euro-Africa zone, the time was right for him to step down and let the young guys come to the fore." Britain's next opponents, Israel, are hardly likely to be quaking in their boots ahead of the 4-6 March match against a likely trio of Bogdanovic, Murray and the 187th-ranked Arvind Parmar. Bailey said: "It will be tough for GB to move up, but there comes a time when our young players have to step up. This was always going to be inevitable with Tim and Greg's growing years. "I'm confident about the future. I wouldn't lay money on us getting back into the world group next year, but I'd imagine in five years time we'll be competing for the major honours." Of those lining up to replace Henman, the 17-year-old Murray, with four Futures titles under his belt last year, looks the best long-term bet. "Murray is the one that looks likeliest to take over Tim's mantle," said Bailey. "He has an enormous amount of self-confidence, judging by what he's said in the past." Bogdanovic, three years Murray's senior, has had a more troubled time under Britain's Davis Cup umbrella.

While Murray has been marked out as Britain's golden boy, Bogdanovic was warned by the Lawn Tennis Association for a lack of drive at the end of 2003. And Bailey said: "Despite that, Alex is clearly talented as well, while Arvind is another contender. "They're among the guys who have experienced the intensity of Davis Cup tennis - whether as players or on the sidelines. "The LTA has always done an exceptional job of ensuring that. "Now they'll finally get to play regularly in the cauldron of the cup. And I'm confident that will springboard Team GB to greater success."

/bbcsport/tennis/029.txt

Moya emotional at Davis Cup win

Carlos Moya described Spain's Davis Cup victory as the highlight of his career after he beat Andy Roddick to end the USA's challenge in Seville.

Moya made up for missing Spain's 2000 victory through injury by beating Roddick 6-2 7-6 (7-1) 7-6 (7-5) to give the hosts an unassailable 3-1 lead. "I have woken up so many nights dreaming of this day," said Moya. "All my energy has been focused on today. "What I have lived today I do not think I will live again." Spain's only other Davis Cup title came two years ago in Valencia, when they beat Australia. And Moya, nicknamed Charly, admitted: "The Davis Cup is my dream and I was a bit nervous at the outset. "Some people have said that I am obsessed but I think that it is better this way. It helps me reach my goals if I am obsessed. "It's really incredible - to get the winning point is really something." Spanish captain Jordi Arrese said: "Charly played a great game. It was his opportunity and he hasn't let us down. "He had lost three times to Roddick, and this was his day to beat him. "He had been waiting years to be in this position." Spain's victory was also remarkable for the performance of Rafael Nadal, who beat Roddick in the opening singles.

Aged 18 years and 185 days, the Mallorcan became the youngest player to win the Davis Cup. "What a great way to finish the year," said Nadal afterwards. US coach Patrick McEnroe wants Roddick and the rest of his team to play more tennis on clay and hone their skills on the surface. "I think it will help these guys even on slow hard courts to learn how to mix things up a little bit and to play a little bit smarter and tactically better." "Obviously it's unrealistic to say that we're going to just start playing constantly on clay, with the schedule. "But certainly I think we can put the work in at the appropriate time and play a couple more events and play against these guys who are the best on this stuff," said McEnroe. Roddick was left frustrated after losing both his singles on the slow clay of Seville's Olympic Stadium. "It's just tough because I felt like I was in it the whole time against one of the top three clay-courters in the world," said the American. "I had my chances and just didn't convert them. The bottom line is they were just better than us this weekend. "They came out, took care of business and they beat us. It's as simple as that.".

/bbcsport/tennis/046.txt

Moya clinches Cup for Spain

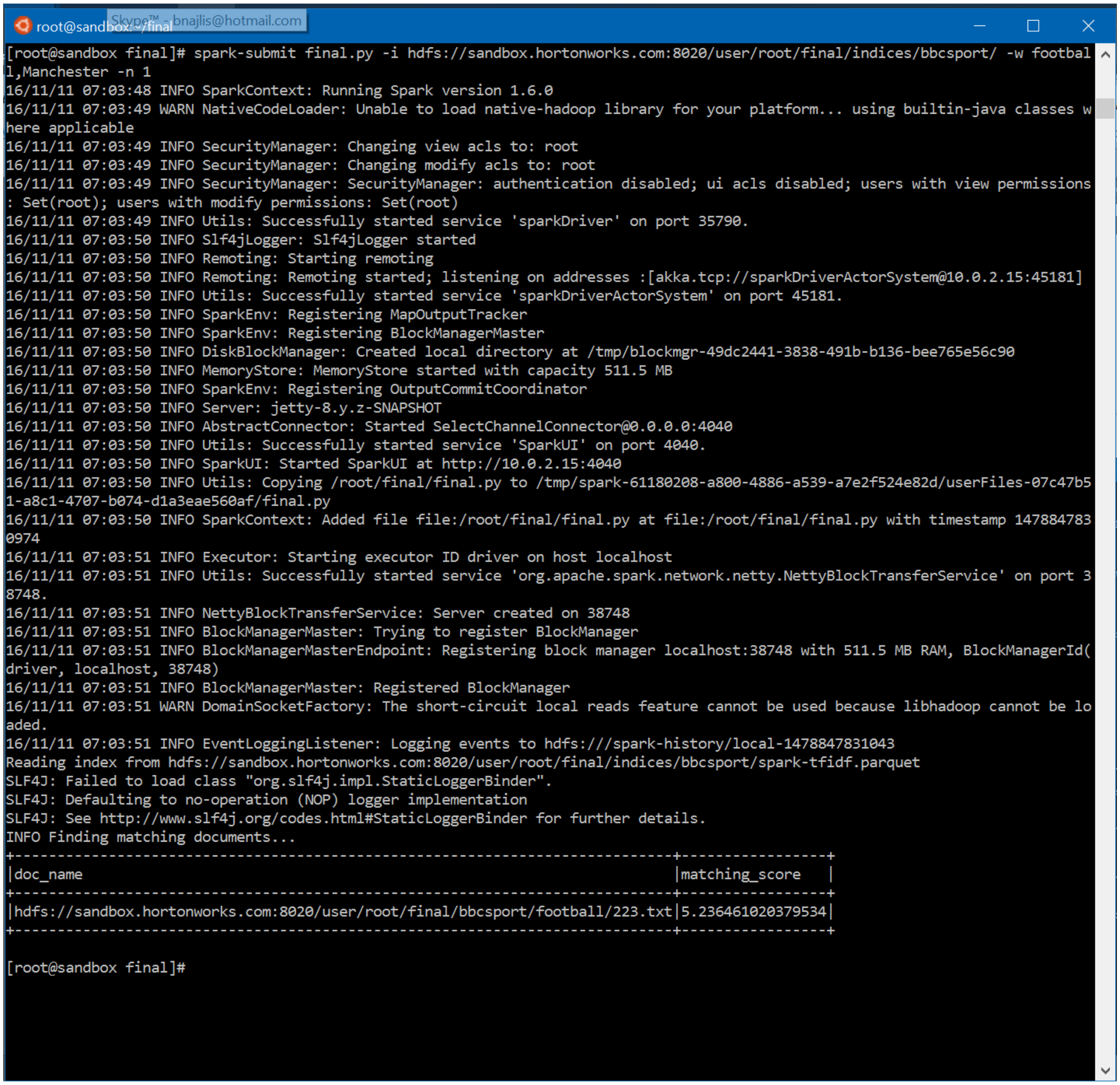
Spain won the Davis Cup for the second time in their history when Carlos Moya beat the USA's Andy Roddick in the fourth rubber in Seville.

Moya won 6-2 7-6 (7-1) 7-6 (7-5) to give the hosts an unassailable 3-1 lead with only one singles rubber remaining. Roddick battled hard and had chances in the second set, but Moya's clay-court expertise proved the difference. Mardy Fish beat Tommy Robredo 7-6 (8-6) 6-2 in the final dead rubber to cut Spain's winning margin to 3-2. Spain's only other Davis Cup title came in 2000, when they beat Australia in Barcelona. This time they chose to play the final in Seville and the city's Olympic Stadium was revamped to allow for a record crowd for a competitve tennis event of 27,000 spectators. And the home fans gave vociferous support to their players, with 18-year-old Nadal and Moya winning both Friday's singles rubbers. American twins Mike and Bob Bryan gave the visitors hope with victory over Juan Carlos Ferrero and Tommy Robredo in Saturday's doubles. But it remained an uphill task for a US team far happier on hard courts than clay, and 1998 French Open champion Moya had too much for world number two Roddick. "This has been incredible - the moment I've been waiting for for years," said Moya, who missed out in 2000 through injury. "I've prepared myself for this day. I knew that playing on clay I would have my chances to win.

"The Davis Cup is my dream. I can't ask for more. There is nothing bigger than what I've lived today." Moya stormed into a 4-0 lead on his way to taking the first set against Roddick and recovered immediately from dropping serve in the second. The match came down to two tie-breaks and Moya dominated both, clinching victory on his third match point. Wild celebrations followed, with an emotional Moya congratulated by his team-mates, including Nadal, who at 18 years and 187 days becomes the youngest ever victor in Davis Cup history. "I think we put up a better fight here than in Paris two years ago," said Roddick, who was on the US team which lost to France in the 2002 semi-finals at Roland Garros. "They were just better than us this weekend. I have nothing to be ashamed of, I gave it my all. "I am not going to walk out of here with my head down, that's for sure. "There's no miracle answer. We have to improve."

Query 4: football,Manchester

Command: `spark-submit final.py -i hdfs://sandbox.hortonworks.com:8020/user/root/final/indices/bbcsport/ -w football,Manchester -n 1`



```
[root@sandbox final]# spark-submit final.py -i hdfs://sandbox.hortonworks.com:8020/user/root/final/indices/bbcsport/ -w footbal
l,Manchester -n 1
16/11/11 07:03:48 INFO SparkContext: Running Spark version 1.6.0
16/11/11 07:03:49 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes w
here applicable
16/11/11 07:03:49 INFO SecurityManager: Changing view acls to: root
16/11/11 07:03:49 INFO SecurityManager: Changing modify acls to: root
16/11/11 07:03:49 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions
: Set(root); users with modify permissions: Set(root)
16/11/11 07:03:49 INFO Utils: Successfully started service 'sparkDriver' on port 35790.
16/11/11 07:03:50 INFO Slf4jLogger: Slf4jLogger started
16/11/11 07:03:50 INFO Remoting: Starting remoting
16/11/11 07:03:50 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkDriverActorSystem@10.0.2.15:45181]
16/11/11 07:03:50 INFO Utils: Successfully started service 'sparkDriverActorSystem' on port 45181.
16/11/11 07:03:50 INFO SparkEnv: Registering MapOutputTracker
16/11/11 07:03:50 INFO SparkEnv: Registering BlockManagerMaster
16/11/11 07:03:50 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-49dc2441-3838-491b-b136-bee765e56c90
16/11/11 07:03:50 INFO MemoryStore: MemoryStore started with capacity 511.5 MB
16/11/11 07:03:50 INFO SparkEnv: Registering OutputCommitCoordinator
16/11/11 07:03:50 INFO Server: jetty-8.y.z-SNAPSHOT
16/11/11 07:03:50 INFO AbstractConnector: Started SelectChannelConnector@0.0.0.0:4040
16/11/11 07:03:50 INFO Utils: Successfully started service 'SparkUI' on port 4040.
16/11/11 07:03:50 INFO SparkUI: Started SparkUI at http://10.0.2.15:4040
16/11/11 07:03:50 INFO Utils: Copying /root/final/final.py to /tmp/spark-61180208-a800-4886-a539-a7e2f524e82d/userFiles-07c47b5
1-a8c1-4707-b074-d1a3eae560af/final.py
16/11/11 07:03:50 INFO SparkContext: Added file file:/root/final/final.py at file:/root/final/final.py with timestamp 147884783
0974
16/11/11 07:03:51 INFO Executor: Starting executor ID driver on host localhost
16/11/11 07:03:51 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 3
8748.
16/11/11 07:03:51 INFO NettyBlockTransferService: Server created on 38748
16/11/11 07:03:51 INFO BlockManagerMaster: Trying to register BlockManager
16/11/11 07:03:51 INFO BlockManagerMasterEndpoint: Registering block manager localhost:38748 with 511.5 MB RAM, BlockManagerId(
driver, localhost, 38748)
16/11/11 07:03:51 INFO BlockManagerMaster: Registered BlockManager
16/11/11 07:03:51 WARN DomainSocketFactory: The short-circuit local reads feature cannot be used because libhadoop cannot be lo
aded.
16/11/11 07:03:51 INFO EventLoggingListener: Logging events to hdfs:///spark-history/local-1478847831043
Reading index from hdfs://sandbox.hortonworks.com:8020/user/root/final/indices/bbcsport/spark-tfidf.parquet
SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".
SLF4J: Defaulting to no-operation (NOP) logger implementation
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.
INFO Finding matching documents...
+-----------------------------------------------------------------------------+-----------------+
|doc_name                                                                     |matching_score   |
+-----------------------------------------------------------------------------+-----------------+
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/football/223.txt|5.236461020379534|
+-----------------------------------------------------------------------------+-----------------+

[root@sandbox final]#
```

Results

Doc_name	Matching_score
/bbcsport/football/223.txt	5.23

/bbcsport/football/224.txt

The wonder of Ronaldinho

Ronaldinho has the most famous smile in football - but it is the grins his incredible talent has put on the faces of fans that ensured he picked up the World Player of the Year award on Monday night.

The Brazilian landed the prize ahead of strikers Thierry Henry of Arsenal and AC Milan's Andriy Shevchenko after a year in which his dancing feet dazzled defenders and delighted fans. Henry and Shevchenko led their clubs to the league titles in England and Italy while Ronaldinho ended last season trophy-less. But while Ronaldinho's achievements may not have won him any medals, they have won him the hearts of a football world. He has turned despair into delight at Barcelona since deciding to move to the Nou Camp instead of Manchester United and, in the process, brought a joy back to the sport. Barcelona were an ailing club thirsty for former glories to be restored as they hung heavily in the shadows of the "galacticos" of arch-rivals Real Madrid.

But the arrival of Ronaldinho in July 2003 has spearheaded a rapid rise which now sees the Catalan club playing the kind of flamboyant football encapsulated by the dashing skills of their playmaker and inspiration. He has fans on the edge of their seats as they wonder what marvels the boy from Brazil will produce to amaze them. Ronaldinho's magic rarely disappoints and, while he possesses all the feints, step-overs, shoulder-drops and vision one could wish for, he also has the crucial ability of complementing his skills with an end product. His victory in Fifa's annual vote may be hard for Henry and Shevchenko, who have amazed with their exploits, but Ronaldinho just has that little extra va-va voom. Ronaldinho can add the award to the World Cup winners medal he earned with Brazil at the 2002 World Cup and the 1999 Copa America title.

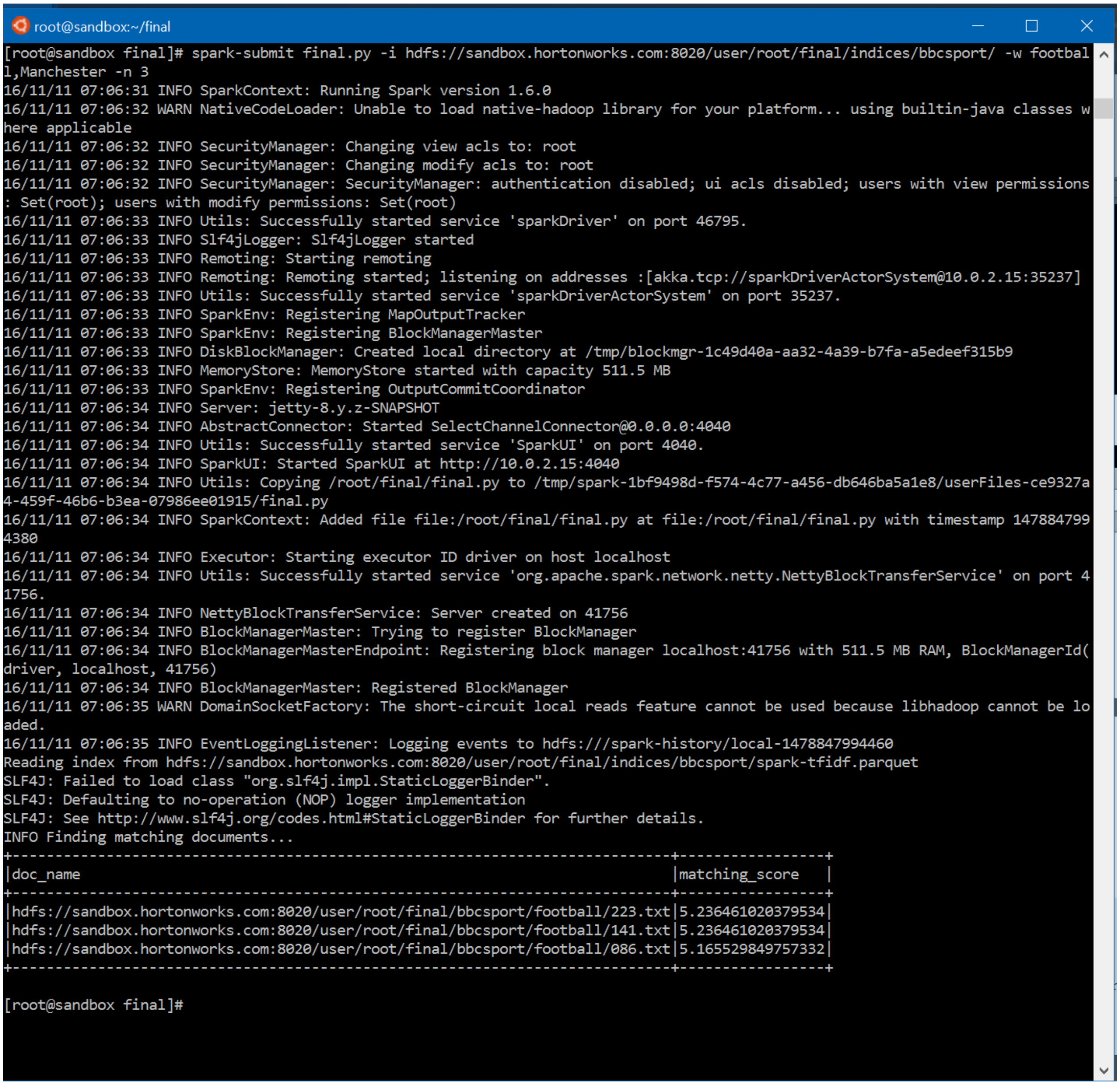
Ronaldinho - full name Ronaldo de Assis Moreira - has come a long way from his humble origins in Porto Allegre, where he was spotted by his hometown club Gremio at the age of 18. His flamboyance and vision was a key factor as Brazil won the World under-17 title in 1997, and in 1999 he captured the attention of then-Brazil coach Wanderley Luxemburgo with 15 goals in 14 matches for Gremio. Ronaldinho made his international debut that year and announced himself on the world stage with a sensational solo effort against Venezuela in the Copa America, which Brazil went on to win. A protracted transfer saga saw him move to Paris St Germain where his relationship with the French club was far from harmonious, and coach Luis Fernandez was not amused when he turned up late following a trip home at Christmas. There was also discontent about the samba star's penchant for dancing at Parisian night spots as much as waltzing past opponents.

But Barcelona were confirmed admirers and a year later his club career - which had been somewhat grounded - really took off after a move to Spain. Sir Alex Ferguson had been confident of tempting him to Manchester United, but the lure of the Nou Camp was too strong. Things did not go that well at the start of his life in Spain, but following the winter break, Ronaldinho and Barcelona hit top gear, finishing the campaign strongly and earning a Champions League place.

Transfer speculation has followed him throughout his young career and almost inevitably he was linked with Chelsea this summer. But even the Blues' billionaire owner Roman Abramovich could not prise him away from Barca, who rewarded him with an improved contract with a buy-out clause of £100m.

Query 5: football,Manchester

Command: `spark-submit final.py -i hdfs://sandbox.hortonworks.com:8020/user/root/final/indices/bbcsport/ -w football,Manchester -n 3`



```
[root@sandbox final]# spark-submit final.py -i hdfs://sandbox.hortonworks.com:8020/user/root/final/indices/bbcsport/ -w footbal
l,Manchester -n 3
16/11/11 07:06:31 INFO SparkContext: Running Spark version 1.6.0
16/11/11 07:06:32 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes w
here applicable
16/11/11 07:06:32 INFO SecurityManager: Changing view acls to: root
16/11/11 07:06:32 INFO SecurityManager: Changing modify acls to: root
16/11/11 07:06:32 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions
: Set(root); users with modify permissions: Set(root)
16/11/11 07:06:33 INFO Utils: Successfully started service 'sparkDriver' on port 46795.
16/11/11 07:06:33 INFO Slf4jLogger: Slf4jLogger started
16/11/11 07:06:33 INFO Remoting: Starting remoting
16/11/11 07:06:33 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkDriverActorSystem@10.0.2.15:35237]
16/11/11 07:06:33 INFO Utils: Successfully started service 'sparkDriverActorSystem' on port 35237.
16/11/11 07:06:33 INFO SparkEnv: Registering MapOutputTracker
16/11/11 07:06:33 INFO SparkEnv: Registering BlockManagerMaster
16/11/11 07:06:33 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-1c49d40a-aa32-4a39-b7fa-a5edeef315b9
16/11/11 07:06:33 INFO MemoryStore: MemoryStore started with capacity 511.5 MB
16/11/11 07:06:33 INFO SparkEnv: Registering OutputCommitCoordinator
16/11/11 07:06:34 INFO Server: jetty-8.y.z-SNAPSHOT
16/11/11 07:06:34 INFO AbstractConnector: Started SelectChannelConnector@0.0.0.0:4040
16/11/11 07:06:34 INFO Utils: Successfully started service 'SparkUI' on port 4040.
16/11/11 07:06:34 INFO SparkUI: Started SparkUI at http://10.0.2.15:4040
16/11/11 07:06:34 INFO Utils: Copying /root/final/final.py to /tmp/spark-1bf9498d-f574-4c77-a456-db646ba5a1e8/userFiles-ce9327a
4-459f-46b6-b3ea-07986ee01915/final.py
16/11/11 07:06:34 INFO SparkContext: Added file file:/root/final/final.py at file:/root/final/final.py with timestamp 147884799
4380
16/11/11 07:06:34 INFO Executor: Starting executor ID driver on host localhost
16/11/11 07:06:34 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 4
1756.
16/11/11 07:06:34 INFO NettyBlockTransferService: Server created on 41756
16/11/11 07:06:34 INFO BlockManagerMaster: Trying to register BlockManager
16/11/11 07:06:34 INFO BlockManagerMasterEndpoint: Registering block manager localhost:41756 with 511.5 MB RAM, BlockManagerId(
driver, localhost, 41756)
16/11/11 07:06:34 INFO BlockManagerMaster: Registered BlockManager
16/11/11 07:06:35 WARN DomainSocketFactory: The short-circuit local reads feature cannot be used because libhadoop cannot be lo
aded.
16/11/11 07:06:35 INFO EventLoggingListener: Logging events to hdfs:///spark-history/local-1478847994460
Reading index from hdfs://sandbox.hortonworks.com:8020/user/root/final/indices/bbcsport/spark-tfidf.parquet
SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".
SLF4J: Defaulting to no-operation (NOP) logger implementation
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.
INFO Finding matching documents...
+-----------------------------------------------------------------------------+-----------------+
|doc_name                                                                     |matching_score   |
+-----------------------------------------------------------------------------+-----------------+
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/football/223.txt|5.236461020379534|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/football/141.txt|5.236461020379534|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/football/086.txt|5.165529849757332|
+-----------------------------------------------------------------------------+-----------------+

[root@sandbox final]#
```

Results

Doc_name	Matching_score
/bbcsport/football/223.txt	5.23
/bbcsport/football/141.txt	5.23
/bbcsport/football/086.txt	5.16

/bbcsport/football/224.txt

The wonder of Ronaldinho

Ronaldinho has the most famous smile in football - but it is the grins his incredible talent has put on the faces of fans that ensured he picked up the World Player of the Year award on Monday night.

The Brazilian landed the prize ahead of strikers Thierry Henry of Arsenal and AC Milan's Andriy Shevchenko after a year in which his dancing feet dazzled defenders and delighted fans. Henry and Shevchenko led their clubs to the league titles in England and Italy while Ronaldinho ended last season trophy-less. But while Ronaldinho's achievements may not have won him any medals, they have won him the hearts of a football world. He has turned despair into delight at Barcelona since deciding to move to the Nou Camp instead of Manchester United and, in the process, brought a joy back to the sport. Barcelona were an ailing club thirsty for former glories to be restored as they hung heavily in the shadows of the "galacticos" of arch-rivals Real Madrid.

But the arrival of Ronaldinho in July 2003 has spearheaded a rapid rise which now sees the Catalan club playing the kind of flamboyant football encapsulated by the dashing skills of their playmaker and inspiration. He has fans on the edge of their seats as they wonder what marvels the boy from Brazil will produce to amaze them. Ronaldinho's magic rarely disappoints and, while he possesses all the feints, step-overs, shoulder-drops and vision one could wish for, he also has the crucial ability of complementing his skills with an end product. His victory in Fifa's annual vote may be hard for Henry and Shevchenko, who have amazed with their exploits, but Ronaldinho just has that little extra va-va voom. Ronaldinho can add the award to the World Cup winners medal he earned with Brazil at the 2002 World Cup and the 1999 Copa America title.

Ronaldinho - full name Ronaldo de Assis Moreira - has come a long way from his humble origins in Porto Allegre, where he was spotted by his hometown club Gremio at the age of 18. His flamboyance and vision was a key factor as Brazil won the World under-17 title in 1997, and in 1999 he captured the attention of then-Brazil coach Wanderley Luxemburgo with 15 goals in 14 matches for Gremio. Ronaldinho made his international debut that year and announced himself on the world stage with a sensational solo effort against Venezuela in the Copa America, which Brazil went on to win. A protracted transfer saga saw him move to Paris St Germain where his relationship with the French club was far from harmonious, and coach Luis Fernandez was not amused when he turned up late following a trip home at Christmas. There was also discontent about the samba star's penchant for dancing at Parisian night spots as much as waltzing past opponents.

But Barcelona were confirmed admirers and a year later his club career - which had been somewhat grounded - really took off after a move to Spain. Sir Alex Ferguson had been confident of tempting him to Manchester United, but the lure of the Nou Camp was too strong. Things did not go that well at the start of his life in Spain, but following the winter break, Ronaldinho and Barcelona hit top gear, finishing the campaign strongly and earning a Champions League place.

Transfer speculation has followed him throughout his young career and almost inevitably he was linked with Chelsea this summer. But even the Blues' billionaire owner Roman Abramovich could not prise him away from Barca, who rewarded him with an improved contract with a buy-out clause of £100m.

/bbcsport/football/141.txt

Wenger steps up row

Arsene Wenger has stepped up his feud with Sir Alex Ferguson by claiming the Manchester United manager is guilty of bringing football into disrepute.

The pair's long-running row was put back in the headlines on Saturday when Ferguson said his Arsenal counterpart was "a disgrace". Wenger initially refused to bite back, saying only: "I will never answer any questions any more about this man." But now he claims Ferguson should be punished by the Football Association. The latest twist in the Ferguson-Wenger saga came on Saturday when the United boss, in an interview with The Independent newspaper, discussed the events after the game between the two sides in October. United won 2-0 that day, at Old Trafford, but the game was followed by a now notorious food fight which saw Ferguson's clothes covered in soup and pizza. The sides meet again at Highbury on 1 February. "In the tunnel Wenger was criticising my players, calling them cheats, so I told him to leave them alone and behave himself," Ferguson said on Saturday. "He ran at me with hands raised saying 'what do you want to do about it?' "To not apologise for the behaviour of the players to another manager is unthinkable. It's a disgrace, but I don't expect Wenger to ever apologise, he's that type of person."

Those allegations were put to Wenger after Saturday's game at Bolton, which Arsenal lost to slip 10 points behind Chelsea in the title race. At first he said only: "I've always been consistent with that story and told you nothing happened. "If he has to talk, he talks. If he wants to make a newspaper article, he makes a newspaper article. "He doesn't interest me and doesn't matter to me at all. I will never answer to any provocation from him any more. "He does what he likes in England anyway. He can go abroad one day and see how it is." But later on Saturday, according to The Independent, Wenger spoke to a smaller group of reporters and expanded on his reaction. "I have no diplomatic relations with him," the Arsenal boss is quoted as saying. "What I don't understand is that he does what he wants and you (the press) are all at his feet.

"The situation (concerning the food fight) has been judged and there is a game going on in a month. "The managers have a responsibility to protect the game before the game. But in England you are only punished for what you say after the game. "Now the whole story starts again. I don't go into that game. We play football. I am a football manager and I love football above all ... no matter what people say." Reminded that Ferguson called him "a disgrace", Wenger added: "I don't respond to anything. In England you have a good phrase. It is 'bringing the game into disrepute'. "But that is not only after a game, it is as well before a game."

Ferguson had also claimed that United chief executive David Gill and Arsenal vice-chairman David Dein had agreed at boardroom level not to discuss the incident in public. But Ferguson added: "In the ensuing weeks all you got was a diatribe from Arsenal about being kicked off the pitch and all that nonsense. Gill phoned Dein three times to complain but nothing was done. "The return is on 1 February and they will come out with another diatribe. "David Gill and I feel we should set the record straight because Arsenal have not written to us to apologise and we would not let that happen here." Meanwhile, the League Managers Association have offered to act as peacemakers in the hope of resolving the on-going row. During that stormy game in October, United striker Ruud van Nistelrooy caught Arsenal's Ashley Cole with one particularly strong tackle. Wenger later accused Van Nistelrooy of "cheating" and was fined £15,000 and "severely reprimanded" by the Football Association. Ferguson admitted on Saturday that Van Nistelrooy's tackle, which earned the Dutchman a ban, "could have given (Cole) a serious injury", but he believes Arsenal were the main aggressors.

"Wenger is always complaining the match was not played in the right spirit," he added. "They are the worst losers of all time, they don't know how to lose. Maybe it is just Manchester United, they don't lose many games to other teams. "We tend to forget the worst disciplinary record of all time was Arsenal's up until last season. In fairness it has improved and now they are seen as paragons of virtue. "But to Wenger it never happens, it is all some dream or nightmare.".

/bbcsport/football/086.txt

Brentford v Southampton

Griffin Park

Tuesday, 1 March

1945 GMT

Barry Knight (Kent)

home to Manchester United in the quarter-finals

Midfielder Andrew Frampton and striker Deon Burton are both slight doubts with hamstring injuries, but should be fit Saints are missing their entire first-choice midfield of Jamie Redknapp, Graeme Le Saux, Nigel Quashie and David Prutton. Anders Svensson and Matt Oakley are likely replacements with Kevin Phillips also scheduled to start, with Henri Camara rested.

- Brentford boss Martin Allen: "After conceding eight goals in our last three matches, I have to admit I'm not very confident. "There's no doubt we're the underdogs and after defending so poorly recently it's not looking good. "Southampton have just drawn with the Premiership champions and that makes our task harder than it was already."
- Southampton boss Harry Redknapp: "We know they can give us problems. "Brentford have done well but we are the Premiership side and should have the better players. "Staying in the Premiership is our priority. We want to win, of course we do. We'll battle but if it comes to a football match I think we'll win." KEY MATCH STATS
- BRENTFORD are the lowest ranked club left in the FA Cup. They're on their best run in the competition since reaching the quarter-finals for the fourth time in their history 16 years ago. Now they have the carrot of the plum draw in the last eight dangling before them. Victory over Premiership strugglers Southampton, would bring the mighty Manchester United to Griffin Park and a gigantic pay day for the sole League One survivors.
- Martin Allen's brave side came back from two goals down at St Mary's to earn a deserved replay. Southampton striker Henri Camara scored twice from close range to put the Saints in command, but Isaiah Rankin hit back just before half time, and Sam Sodje headed past a creaky defence on 58 minutes.
- The Londoners have conceded six goals in their two subsequent League outings - three each in losing away to Hartlepool and drawing at home to Sheffield Wednesday. But they haven't lost in six League and Cup games on home turf - winning three and drawing three since the reverse to Torquay on Boxing Day.
- SOUTHAMPTON go into this tie on the back of an eventful Premiership match with Arsenal on Saturday. An angry David Prutton pushed referee Alan Wiley after being shown the red card, but his side still came back to draw 1-1. It was Saints' fourth stalemate in succession in all competitions, but didn't lift them out of the relegation zone. The retention of their ever present Premier League status must be the number one priority, irrespective of the rewards that success against Brentford would bring.
- Victory here would set up a repeat of the 1976 final, when Saints astounded the football world by defeating Tommy Docherty's Manchester United courtesy of Bobby Stokes' famous winner. They also knocked out the Red Devils in 1991 on penalties in the fourth round. But to write another chapter in their FA Cup history, the Solent side must avoid succumbing to lower division opposition for the first time since Rotherham, from the second level, beat them 2-1 in a third round tie at Millmoor on 16 January 2002.

Southampton were last humbled by a club from the third tier six years ago. Fulham were then in the Second Division, when they won a third round replay at Craven Cottage 1-0 on 13 January 1999.
 - To get to within two matches of a second visit to the Millennium Stadium in three years for the final, Harry Redknapp must guide his side past a club 36 places inferior on the League ladder, and a manager 19 years his junior, who played under him at West Ham. HEAD TO HEAD

10th League One

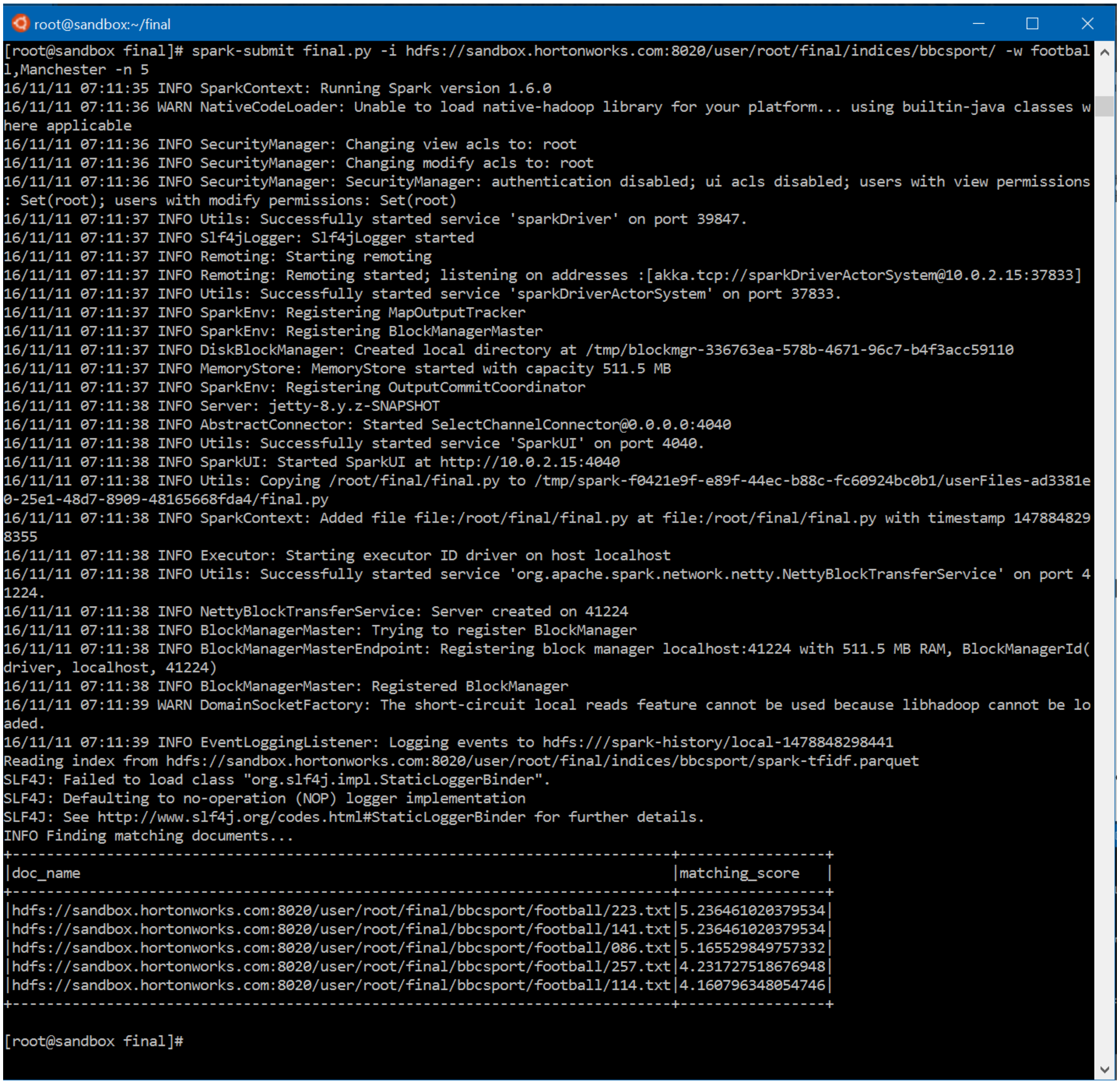
QUARTER-FINALS (four times)

18th PREM

WINNERS (once)

Query 6: football,Manchester

Command: `spark-submit final.py -i hdfs://sandbox.hortonworks.com:8020/user/root/final/indices/bbcsport/ -w football,Manchester -n 5`



```
[root@sandbox final]# spark-submit final.py -i hdfs://sandbox.hortonworks.com:8020/user/root/final/indices/bbcsport/ -w footbal
l,Manchester -n 5
16/11/11 07:11:35 INFO SparkContext: Running Spark version 1.6.0
16/11/11 07:11:36 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes w
here applicable
16/11/11 07:11:36 INFO SecurityManager: Changing view acls to: root
16/11/11 07:11:36 INFO SecurityManager: Changing modify acls to: root
16/11/11 07:11:36 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions
: Set(root); users with modify permissions: Set(root)
16/11/11 07:11:37 INFO Utils: Successfully started service 'sparkDriver' on port 39847.
16/11/11 07:11:37 INFO Slf4jLogger: Slf4jLogger started
16/11/11 07:11:37 INFO Remoting: Starting remoting
16/11/11 07:11:37 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkDriverActorSystem@10.0.2.15:37833]
16/11/11 07:11:37 INFO Utils: Successfully started service 'sparkDriverActorSystem' on port 37833.
16/11/11 07:11:37 INFO SparkEnv: Registering MapOutputTracker
16/11/11 07:11:37 INFO SparkEnv: Registering BlockManagerMaster
16/11/11 07:11:37 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-336763ea-578b-4671-96c7-b4f3acc59110
16/11/11 07:11:37 INFO MemoryStore: MemoryStore started with capacity 511.5 MB
16/11/11 07:11:37 INFO SparkEnv: Registering OutputCommitCoordinator
16/11/11 07:11:38 INFO Server: jetty-8.y.z-SNAPSHOT
16/11/11 07:11:38 INFO AbstractConnector: Started SelectChannelConnector@0.0.0.0:4040
16/11/11 07:11:38 INFO Utils: Successfully started service 'SparkUI' on port 4040.
16/11/11 07:11:38 INFO SparkUI: Started SparkUI at http://10.0.2.15:4040
16/11/11 07:11:38 INFO Utils: Copying /root/final/final.py to /tmp/spark-f0421e9f-e89f-44ec-b88c-fc60924bc0b1/userFiles-ad3381e
0-25e1-48d7-8909-48165668fda4/final.py
16/11/11 07:11:38 INFO SparkContext: Added file file:/root/final/final.py at file:/root/final/final.py with timestamp 147884829
8355
16/11/11 07:11:38 INFO Executor: Starting executor ID driver on host localhost
16/11/11 07:11:38 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 4
1224.
16/11/11 07:11:38 INFO NettyBlockTransferService: Server created on 41224
16/11/11 07:11:38 INFO BlockManagerMaster: Trying to register BlockManager
16/11/11 07:11:38 INFO BlockManagerMasterEndpoint: Registering block manager localhost:41224 with 511.5 MB RAM, BlockManagerId(
driver, localhost, 41224)
16/11/11 07:11:38 INFO BlockManagerMaster: Registered BlockManager
16/11/11 07:11:39 WARN DomainSocketFactory: The short-circuit local reads feature cannot be used because libhadoop cannot be lo
aded.
16/11/11 07:11:39 INFO EventLoggingListener: Logging events to hdfs:///spark-history/local-1478848298441
Reading index from hdfs://sandbox.hortonworks.com:8020/user/root/final/indices/bbcsport/spark-tfidf.parquet
SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".
SLF4J: Defaulting to no-operation (NOP) logger implementation
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.
INFO Finding matching documents...
+-----------------------------------------------------------------------------+-----------------+
|doc_name                                                                     |matching_score   |
+-----------------------------------------------------------------------------+-----------------+
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/football/223.txt|5.236461020379534|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/football/141.txt|5.236461020379534|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/football/086.txt|5.165529849757332|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/football/257.txt|4.231727518676948|
|hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/football/114.txt|4.160796348054746|
+-----------------------------------------------------------------------------+-----------------+

[root@sandbox final]#
```

Results

Doc_name	Matching_score
/bbcsport/football/223.txt	5.23
/bbcsport/football/141.txt	5.23
/bbcsport/football/086.txt	5.16
/bbcsport/football/257.txt	4.23
/bbcsport/football/114.txt	4.16

/bbcsport/football/224.txt

The wonder of Ronaldinho

Ronaldinho has the most famous smile in football - but it is the grins his incredible talent has put on the faces of fans that ensured he picked up the World Player of the Year award on Monday night.

The Brazilian landed the prize ahead of strikers Thierry Henry of Arsenal and AC Milan's Andriy Shevchenko after a year in which his dancing feet dazzled defenders and delighted fans. Henry and Shevchenko led their clubs to the league titles in England and Italy while Ronaldinho ended last season trophy-less. But while Ronaldinho's achievements may not have won him any medals, they have won him the hearts of a football world. He has turned despair into delight at Barcelona since deciding to move to the Nou Camp instead of Manchester United and, in the process, brought a joy back to the sport. Barcelona were an ailing club thirsty for former glories to be restored as they hung heavily in the shadows of the "galacticos" of arch-rivals Real Madrid.

But the arrival of Ronaldinho in July 2003 has spearheaded a rapid rise which now sees the Catalan club playing the kind of flamboyant football encapsulated by the dashing skills of their playmaker and inspiration. He has fans on the edge of their seats as they wonder what marvels the boy from Brazil will produce to amaze them. Ronaldinho's magic rarely disappoints and, while he possesses all the feints, step-overs, shoulder-drops and vision one could wish for, he also has the crucial ability of complementing his skills with an end product. His victory in Fifa's annual vote may be hard for Henry and Shevchenko, who have amazed with their exploits, but Ronaldinho just has that little extra va-va voom. Ronaldinho can add the award to the World Cup winners medal he earned with Brazil at the 2002 World Cup and the 1999 Copa America title.

Ronaldinho - full name Ronaldo de Assis Moreira - has come a long way from his humble origins in Porto Allegre, where he was spotted by his hometown club Gremio at the age of 18. His flamboyance and vision was a key factor as Brazil won the World under-17 title in 1997, and in 1999 he captured the attention of then-Brazil coach Wanderley Luxemburgo with 15 goals in 14 matches for Gremio. Ronaldinho made his international debut that year and announced himself on the world stage with a sensational solo effort against Venezuela in the Copa America, which Brazil went on to win. A protracted transfer saga saw him move to Paris St Germain where his relationship with the French club was far from harmonious, and coach Luis Fernandez was not amused when he turned up late following a trip home at Christmas. There was also discontent about the samba star's penchant for dancing at Parisian night spots as much as waltzing past opponents.

But Barcelona were confirmed admirers and a year later his club career - which had been somewhat grounded - really took off after a move to Spain. Sir Alex Ferguson had been confident of tempting him to Manchester United, but the lure of the Nou Camp was too strong. Things did not go that well at the start of his life in Spain, but following the winter break, Ronaldinho and Barcelona hit top gear, finishing the campaign strongly and earning a Champions League place.

Transfer speculation has followed him throughout his young career and almost inevitably he was linked with Chelsea this summer. But even the Blues' billionaire owner Roman Abramovich could not prise him away from Barca, who rewarded him with an improved contract with a buy-out clause of £100m.

/bbcsport/football/141.txt

Wenger steps up row

Arsene Wenger has stepped up his feud with Sir Alex Ferguson by claiming the Manchester United manager is guilty of bringing football into disrepute.

The pair's long-running row was put back in the headlines on Saturday when Ferguson said his Arsenal counterpart was "a disgrace". Wenger initially refused to bite back, saying only: "I will never answer any questions any more about this man." But now he claims Ferguson should be punished by the Football Association. The latest twist in the Ferguson-Wenger saga came on Saturday when the United boss, in an interview with The Independent newspaper, discussed the events after the game between the two sides in October. United won 2-0 that day, at Old Trafford, but the game was followed by a now notorious food fight which saw Ferguson's clothes covered in soup and pizza. The sides meet again at Highbury on 1 February. "In the tunnel Wenger was criticising my players, calling them cheats, so I told him to leave them alone and behave himself," Ferguson said on Saturday. "He ran at me with hands raised saying 'what do you want to do about it?' "To not apologise for the behaviour of the players to another manager is unthinkable. It's a disgrace, but I don't expect Wenger to ever apologise, he's that type of person."

Those allegations were put to Wenger after Saturday's game at Bolton, which Arsenal lost to slip 10 points behind Chelsea in the title race. At first he said only: "I've always been consistent with that story and told you nothing happened. "If he has to talk, he talks. If he wants to make a newspaper article, he makes a newspaper article. "He doesn't interest me and doesn't matter to me at all. I will never answer to any provocation from him any more. "He does what he likes in England anyway. He can go abroad one day and see how it is." But later on Saturday, according to The Independent, Wenger spoke to a smaller group of reporters and expanded on his reaction. "I have no diplomatic relations with him," the Arsenal boss is quoted as saying. "What I don't understand is that he does what he wants and you (the press) are all at his feet.

"The situation (concerning the food fight) has been judged and there is a game going on in a month. "The managers have a responsibility to protect the game before the game. But in England you are only punished for what you say after the game. "Now the whole story starts again. I don't go into that game. We play football. I am a football manager and I love football above all ... no matter what people say." Reminded that Ferguson called him "a disgrace", Wenger added: "I don't respond to anything. In England you have a good phrase. It is 'bringing the game into disrepute'. "But that is not only after a game, it is as well before a game."

Ferguson had also claimed that United chief executive David Gill and Arsenal vice-chairman David Dein had agreed at boardroom level not to discuss the incident in public. But Ferguson added: "In the ensuing weeks all you got was a diatribe from Arsenal about being kicked off the pitch and all that nonsense. Gill phoned Dein three times to complain but nothing was done. "The return is on 1 February and they will come out with another diatribe. "David Gill and I feel we should set the record straight because Arsenal have not written to us to apologise and we would not let that happen here." Meanwhile, the League Managers Association have offered to act as peacemakers in the hope of resolving the on-going row. During that stormy game in October, United striker Ruud van Nistelrooy caught Arsenal's Ashley Cole with one particularly strong tackle. Wenger later accused Van Nistelrooy of "cheating" and was fined £15,000 and "severely reprimanded" by the Football Association. Ferguson admitted on Saturday that Van Nistelrooy's tackle, which earned the Dutchman a ban, "could have given (Cole) a serious injury", but he believes Arsenal were the main aggressors.

"Wenger is always complaining the match was not played in the right spirit," he added. "They are the worst losers of all time, they don't know how to lose. Maybe it is just Manchester United, they don't lose many games to other teams. "We tend to forget the worst disciplinary record of all time was Arsenal's up until last season. In fairness it has improved and now they are seen as paragons of virtue. "But to Wenger it never happens, it is all some dream or nightmare.".

/bbcsport/football/086.txt

Brentford v Southampton

Griffin Park

Tuesday, 1 March

1945 GMT

Barry Knight (Kent)

home to Manchester United in the quarter-finals

Midfielder Andrew Frampton and striker Deon Burton are both slight doubts with hamstring injuries, but should be fit Saints are missing their entire first-choice midfield of Jamie Redknapp, Graeme Le Saux, Nigel Quashie and David Prutton. Anders Svensson and Matt Oakley are likely replacements with Kevin Phillips also scheduled to start, with Henri Camara rested.

- Brentford boss Martin Allen: "After conceding eight goals in our last three matches, I have to admit I'm not very confident. "There's no doubt we're the underdogs and after defending so poorly recently it's not looking good. "Southampton have just drawn with the Premiership champions and that makes our task harder than it was already."
- Southampton boss Harry Redknapp: "We know they can give us problems. "Brentford have done well but we are the Premiership side and should have the better players. "Staying in the Premiership is our priority. We want to win, of course we do. We'll battle but if it comes to a football match I think we'll win." KEY MATCH STATS
- BRENTFORD are the lowest ranked club left in the FA Cup. They're on their best run in the competition since reaching the quarter-finals for the fourth time in their history 16 years ago. Now they have the carrot of the plum draw in the last eight dangling before them. Victory over Premiership strugglers Southampton, would bring the mighty Manchester United to Griffin Park and a gigantic pay day for the sole League One survivors.
- Martin Allen's brave side came back from two goals down at St Mary's to earn a deserved replay. Southampton striker Henri Camara scored twice from close range to put the Saints in command, but Isaiah Rankin hit back just before half time, and Sam Sodje headed past a creaky defence on 58 minutes.
- The Londoners have conceded six goals in their two subsequent League outings - three each in losing away to Hartlepool and drawing at home to Sheffield Wednesday. But they haven't lost in six League and Cup games on home turf - winning three and drawing three since the reverse to Torquay on Boxing Day.
- SOUTHAMPTON go into this tie on the back of an eventful Premiership match with Arsenal on Saturday. An angry David Prutton pushed referee Alan Wiley after being shown the red card, but his side still came back to draw 1-1. It was Saints' fourth stalemate in succession in all competitions, but didn't lift them out of the relegation zone. The retention of their ever present Premier League status must be the number one priority, irrespective of the rewards that success against Brentford would bring.
- Victory here would set up a repeat of the 1976 final, when Saints astounded the football world by defeating Tommy Docherty's Manchester United courtesy of Bobby Stokes' famous winner. They also knocked out the Red Devils in 1991 on penalties in the fourth round. But to write another chapter in their FA Cup history, the Solent side must avoid succumbing to lower division opposition for the first time since Rotherham, from the second level, beat them 2-1 in a third round tie at Millmoor on 16 January 2002.

Southampton were last humbled by a club from the third tier six years ago. Fulham were then in the Second Division, when they won a third round replay at Craven Cottage 1-0 on 13 January 1999.
 - To get to within two matches of a second visit to the Millennium Stadium in three years for the final, Harry Redknapp must guide his side past a club 36 places inferior on the League ladder, and a manager 19 years his junior, who played under him at West Ham. HEAD TO HEAD

10th League One

QUARTER-FINALS (four times)

18th PREM

WINNERS (once)

/bbcsport/football/257.txt

Man Utd women's team to be axed

Manchester United will scrap their women's team once the current season ends, just three months before the North West hosts the Women's Euro 2005.

From next season, the club's commitment to women's football will only stretch as far as coaching up to the age of 16. "Our aim is best served concentrating on youngsters," said club director of communications Philip Townsend. "Our resources are better deployed at the level of school-age children rather than adults." Football Association vice-chairman Ray Kiddell, who heads the FA women's football committee, greeted the news with dismay. "It is very disappointing," he said. "The progress of women's football can be really helped by professional clubs taking women's teams under their umbrella. "It is a blow to the game that a great club like Manchester United will no longer be doing this.")

/bbcsport/football/114.txt

Souness backs Smith for Scotland

Graeme Souness believes Walter Smith would be the perfect choice to succeed Berti Vogts as Scotland manager.

Souness's former assistant at Rangers is hot favourite to take over from Vogts, who resigned on Monday. "Walter is most definitely the ideal candidate for that job. He'd be perfect for it," Souness told BBC Sport. The Scottish Football Association has appointed Tommy Burns as provisional caretaker-boss for the friendly against Sweden on 17 November. "He fits the bill because of his knowledge and understanding of the Scotland team and football. He is experienced and has been successful." Souness added: "Walter is a real football person, as I know from working with him at Ibrox. "On top of all that he is a proper human being who would command the instant respect of the players and everyone involved in Scottish football." Souness joined Sir Alex Ferguson in backing Smith's claims. The Scottish Football Association is about to embark on the search for Vogts successor after appointing Tommy Burns in a caretaker capacity. Ferguson said: "He (Smith) would be the outstanding candidate as far as I'm concerned. "You need somebody who knows what they're doing and Walter would bring a wealth of experience to the job." The Man Utd boss continued: "I don't know what credentials are needed to do the job but it's a job that needs a lot of experience. "He was my assistant with Scotland and here at Manchester United and he has also managed Glasgow Rangers. "He would need to change the whole shape of Scottish football and radical changes are needed." Smith was assistant to Ferguson at the World Cup in Mexico in 1986. The former Everton and Rangers boss has been out of the game since a spell as Manchester United assistant last term. BBC Sport understands that Smith would be willing to discuss taking over if he was approached by the Scottish FA.

If he is tempted to take over, it seems almost certain Smith's long-time right-hand man Archie Knox would also play a part in the national team set-up. Smith already has the backing of many pundits and fans, including former Scotland manager Craig Brown. Brown said: "Walter is an outstanding candidate without doubt. "He would be admirable choice. I spoke to him on Sunday and I got the impression he would take it. He was asking me about it and I was positive." Other candidates for the job include former Scotland midfielders Gordon Strachan and Gary McAllister and Vogts' assistant Tommy Burns.