BBC Sport Document Search Engine

DS8003

Rafik Matta

Contents

[Summary 3](#_Toc531893147)

[Technology Choices + Performance Tuning 3](#_Toc531893148)

[Architecture 5](#_Toc531893149)

[Code Explanation 6](#_Toc531893150)

[Build Index 6](#_Toc531893151)

[Load Index + Query 6](#_Toc531893152)

[Results 7](#_Toc531893153)

[Appendix A - Code 9](#_Toc531893154)

[Initialize 9](#_Toc531893155)

[Build Index Function 9](#_Toc531893156)

[Load Index + Query Functions 10](#_Toc531893157)

[Call the Functions 11](#_Toc531893158)

[Appendix B – Screenshots of System Running in PySpark Console 13](#_Toc531893159)

[Load Data to HDFS and Start PySpark Console 13](#_Toc531893160)

[Load Libraries and set shuffle partitions 13](#_Toc531893161)

[Define Build Index Function 14](#_Toc531893162)

[Define Load Index + Query Function 14](#_Toc531893163)

[Run Build Rugby Index 15](#_Toc531893164)

[Run Load Index + Dummy Query to Action the Load 15](#_Toc531893165)

[Run the Queries 16](#_Toc531893166)

[Query1,N=1 16](#_Toc531893167)

[Query1,N=3 16](#_Toc531893168)

[Query1,N=5 16](#_Toc531893169)

[Query2,N=1 17](#_Toc531893170)

[Query2,N=3 17](#_Toc531893171)

[Query2,N=5 17](#_Toc531893172)

[Appendix C – Excerpts of the top Documents 18](#_Toc531893173)

[Query 1 18](#_Toc531893174)

[Rubgy/098.txt 18](#_Toc531893175)

[Rugby/127.txt 18](#_Toc531893176)

[Rugby/086.txt 19](#_Toc531893177)

[Query 2 19](#_Toc531893178)

[Rugby/001.txt 19](#_Toc531893179)

[Rugby/003.txt 20](#_Toc531893180)

[Rugby/134.txt 20](#_Toc531893181)

# Summary

This search system has been developed by building an index around the provided BBC Sport document data. The system was developed using Hortonworks HDP with HDFS as a persistence layer, Spark as the processing layer, and Livy + Jupyter as the interface layer. The system allows a user to choose which of the directories available in the dataset to index and then using Livy+Jupyter allows them to query and review results. The system performs quite well at searching once the data has been loaded. Queries on average take under a second.

# Technology Choices + Performance Tuning

Out of the possible technology choices available for this system, PySpark (2.3) on HDP 2.6.5 was the primary technology choice. Spark was chosen for its ease of us, and built in optimizations (especially for querying). MongoDB would have also served well for this data, given its unstructured nature as well as relatively small size. That being said, the main drawback to using Mongo was the overly complex aggregation pipeline mechanism required to actually do anything useful. Spark offers a pretty easy to use SQL like interface over Mongo’s JSON based structure, which due to the Catalyst engine is already optimized.

Very few performance tuning tricks needed to be done to achieve almost instant querying as DataFrames in Spark already do much of the query optimization. The slowest operation is loading the index. Therefore, the first query takes a performance penalty due to the lazy load nature of Spark, but every subsequent query thereafter is very rapid.

Here are the specifications of the machine/cluster used to perform the analysis:

* Windows 10 64-bit
* Hortonworks HDP 2.6.5 with Spark v 2.3
* 16 GB of DDR3 RAM
* 4-Core Intel Core i7 (2011 Model) – Hyperthreading Enabled

With this in mind, the following choices were made:

1. Run Spark in Local mode (as it’s running on a single laptop), with 8 cores assigned to driver (as it’s technically a 4-core multithreaded machine making it virtually 8 cores)
2. Set partition size to 8 (equal to the number of cores) to eliminate shuffles of data (most expensive Spark operation).
3. Make extensive use of DataFrames to leverage Catalyst optimizations.
4. In the Query logic, ensure the query vector was loaded using a Broadcast variable in Spark which ensures a balanced join.

Further to this, provided in Table 1 is a list of the technology choices available for the project and developing the system as well as their pro’s and con’s.

|  |  |  |  |
| --- | --- | --- | --- |
| **Technology** | **Used** | **Pros** | **Cons** |
| MapReduce | No | MapReduce is excellent for low level Map/Reduce operations on a Hadoop Cluster. It allows effectively full control over the process and data flow | MapReduce doesn’t offer out of the box optimization, which means unless one is well trained in using it, one would be better off using something like Spark. |
| Hive | No | Hive offers an excellent SQL like interface and with the recent updates made to it, it offers ACID transaction guarantees. With HiveServer2 Interactive, performance is significantly better as well (using Tez). | The major issue with Hive is that there’s no really solid interface to interact with it as opposed to something like Spark + Livy. It also still performs relatively worse to Spark. |
| Pig | No | Pig serves as a step up from MapReduce, and offers some minor out of the box optimizations. | Pig still does not do well with offering really well optimized operations as compared to Spark. Further, the interface to use it and the general syntax leaves much to be desired. |
| HBase | No | HBase is great for column-based operations. This actually would have made a lot of sense for the project. | Given the optimization capabilities of Spark and the generally marginal performance difference achievable, it was not worth going through the learning curve of HBase. |
| Mongo | No | As MongoDB is an object store, it would actually have been another excellent fit for this problem. As we had to build our own Index logic, Mongo would have done nothing here except serving as a persistence layer. Also Mongo doesn’t require a Hadoop Cluster. | The major con to using Mongo here would be that it’s aggregations aren’t particularly intuitive and having to use an aggregation pipeline wasn’t necessary here. |
| HDFS | Yes | Distributed File System plus natively supported by Spark | Requires a Hadoop Cluster to be setup (used the HDP Sandbox) |
| Spark | Yes | Powerful and easy to use distributed processing engine with built in optimizations and a data structure called DataFrame for doing SQL like querying operations | Requires a Hadoop Cluster to be setup (used the HDP Sandbox) |

**Table 1.** Technology Comparison Table

# Architecture

The below diagrams show the system design + dataflow.

Legend:

□ – Spark Step/Process

□ – HDFS Step/Process

□ – Livy Step/Process

**Build Index Architecture Query Architecture**

Documents

Index

Documents Selection

Output to User

Query + N

Read Index

Build Query Vector

Score Query vs. Doc

Produce output

Read Documents

Pre-process

Map Words to Docs

Calc TF/IDF

Join Data Together

Index

**Figure 1.** System Architecture

# Code Explanation

This explanation coincides with the System Architecture Diagram in Figure 1, as well as the code in Appendix A. All lines numbers can be found in the corresponding section in Appendix A.

## Build Index

1. Set the partitions to the appropriate amount for the cluster setup. *(Line 9)*
2. Load the text files belonging to a specific sport into Spark as an RDD using the wholeTextFiles function and count the total number of documents. *(Line 12-13)*
3. Pre-process the text. This includes removing punctuation, end-of-line characters, lower casing all the words to produce one large string per document. The end result is an RDD mapping where the key is the document name/reference and the values are the pre-processed document content. *(Line 16-20)*
4. Flatten the resulting string into words by separating based on whitespaces and convert to a DataFrame. From here, we stop using RDD’s and begin leveraging the power of DataFrames. *(Line 22- 23)*
5. Calculate the Term Frequency and the Inverse Document Frequency based on the following formulas *(Line 25-34)*:
6. Calculate TF-IDF by joining the resulting TF DataFrame and IDF Dataframe and using the follow formula *(Line 36):*
7. Write resulting TF-IDF index to HDFS using the ORC format. This format was chosen due to its excellent compression characteristics *(Line 38).*

## Load Index + Query

1. Load the Index that will be used for searching *(Line 41-46, 73)*
2. Select Query + result amount *(Line 48, 81-82).*
3. Take the query string and split it into words (using whitespace as a delimiter). *(Line 50)*
4. Calculate the TF and the IDF of the Query String and calculate the TF-IDF using the same logic and formulas given previously. *(Line 52-58)*
5. Compare the resulting Query vector against all of the documents and produce the score using the following formula *(Line 60):*
6. Output the result *(Line 61).*

# Results

The following analysis was run on the system:

Query1, N = 1

+-------------+-------------------+

| file| score|

+-------------+-------------------+

|rugby/098.txt|0.12769736588535455|

+-------------+-------------------+

Query1, N = 3

+-------------+-------------------+

| file| score|

+-------------+-------------------+

|rugby/098.txt|0.12769736588535455|

|rugby/127.txt|0.06749421879927163|

|rugby/086.txt|0.05140068228743627|

+-------------+-------------------+

Query1, N = 5

+-------------+--------------------+

| file| score|

+-------------+--------------------+

|rugby/098.txt| 0.12769736588535455|

|rugby/127.txt| 0.06749421879927163|

|rugby/086.txt| 0.05140068228743627|

|rugby/092.txt|0.012255797343391425|

|rugby/060.txt|0.007656956682990293|

+-------------+--------------------+

Query2, N = 1

+-------------+--------------------+

| file| score|

+-------------+--------------------+

|rugby/001.txt|0.043268433667157026|

+-------------+--------------------+

Query2, N = 3

+-------------+--------------------+

| file| score|

+-------------+--------------------+

|rugby/001.txt|0.043268433667157026|

|rugby/003.txt| 0.02242535737241099|

|rugby/134.txt|0.020408392192625024|

+-------------+--------------------+

Query2, N = 5

+-------------+--------------------+

| file| score|

+-------------+--------------------+

|rugby/001.txt|0.043268433667157026|

|rugby/003.txt| 0.02242535737241099|

|rugby/134.txt|0.020408392192625024|

|rugby/141.txt| 0.01865745104894276|

|rugby/097.txt|0.016309428405249717|

+-------------+--------------------+

# Appendix A - Code

## Initialize



## Build Index Function



## Load Index + Query Functions



## Call the Functions



113 **print(**"Time taken: " **+** str**((**b**-**a**).**total\_seconds**())** **+** " seconds"**)** #prints time taken

114

115 a **=** datetime**.**datetime**.**now**()**

116 N **=** 3 # top N results

117 query **=** "Yachvili slotted over four penalties"

118 query\_call**(**query**,** N**,**index**)**

119 b **=** datetime**.**datetime**.**now**()**

120 **print(**"Time taken: " **+** str**((**b**-**a**).**total\_seconds**())** **+** " seconds"**)** #prints time taken

121

122 a **=** datetime**.**datetime**.**now**()**

123 N **=** 5 # top N results

124 query **=** "Yachvili slotted over four penalties"

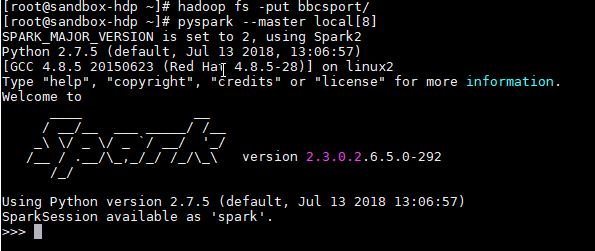
125 query\_call**(**query**,** N**,**index**)**

126 b **=** datetime**.**datetime**.**now**()**

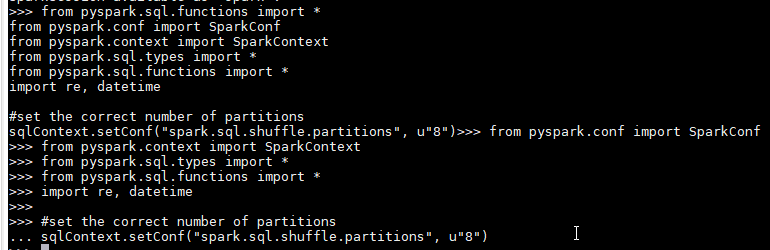
127 **print(**"Time taken: " **+** str**((**b**-**a**).**total\_seconds**())** **+** " seconds"**)** #prints time taken

# Appendix B – Screenshots of System Running in PySpark Console

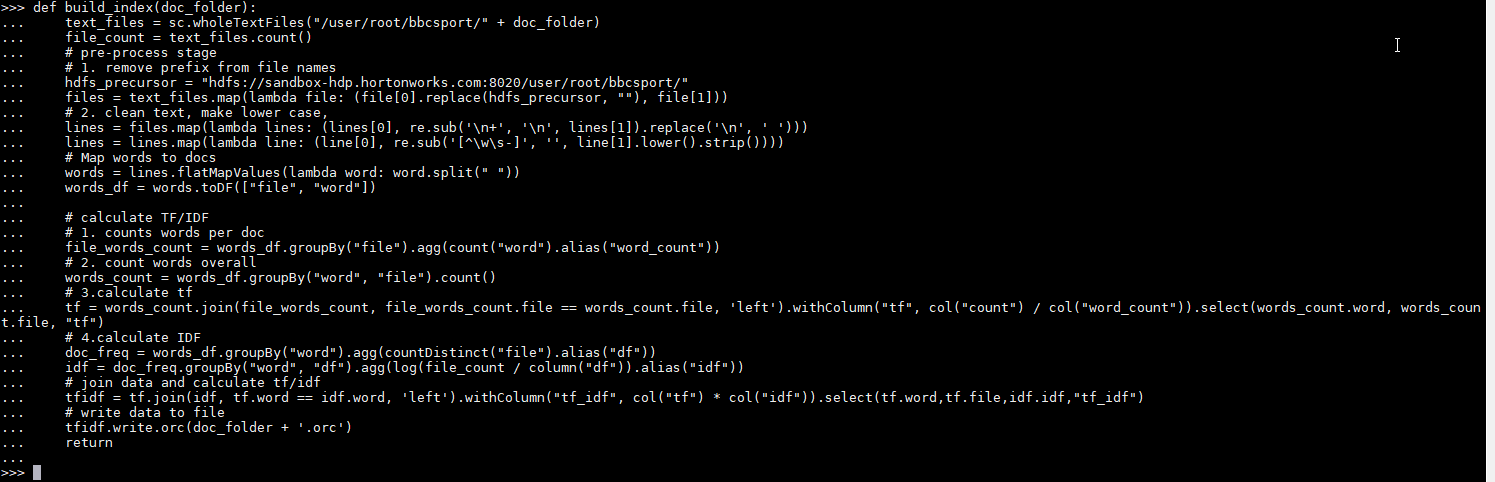
## Load Data to HDFS and Start PySpark Console



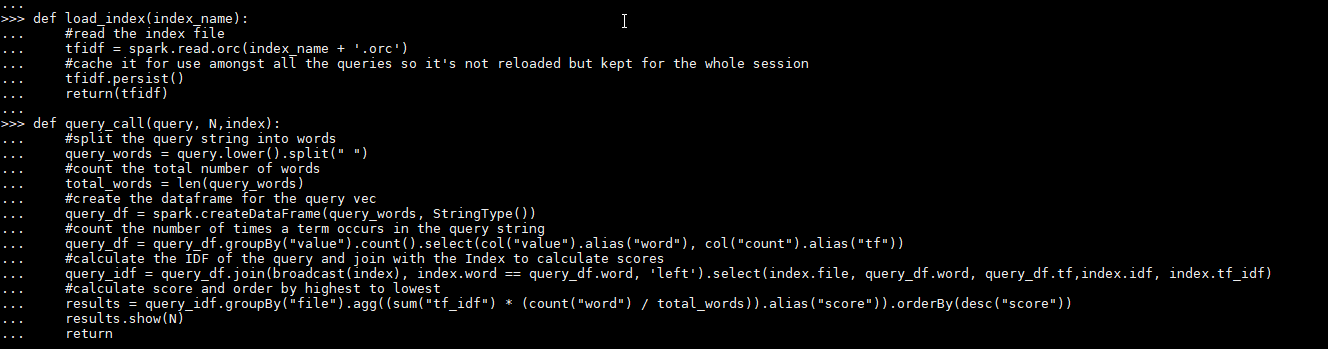
## Load Libraries and set shuffle partitions



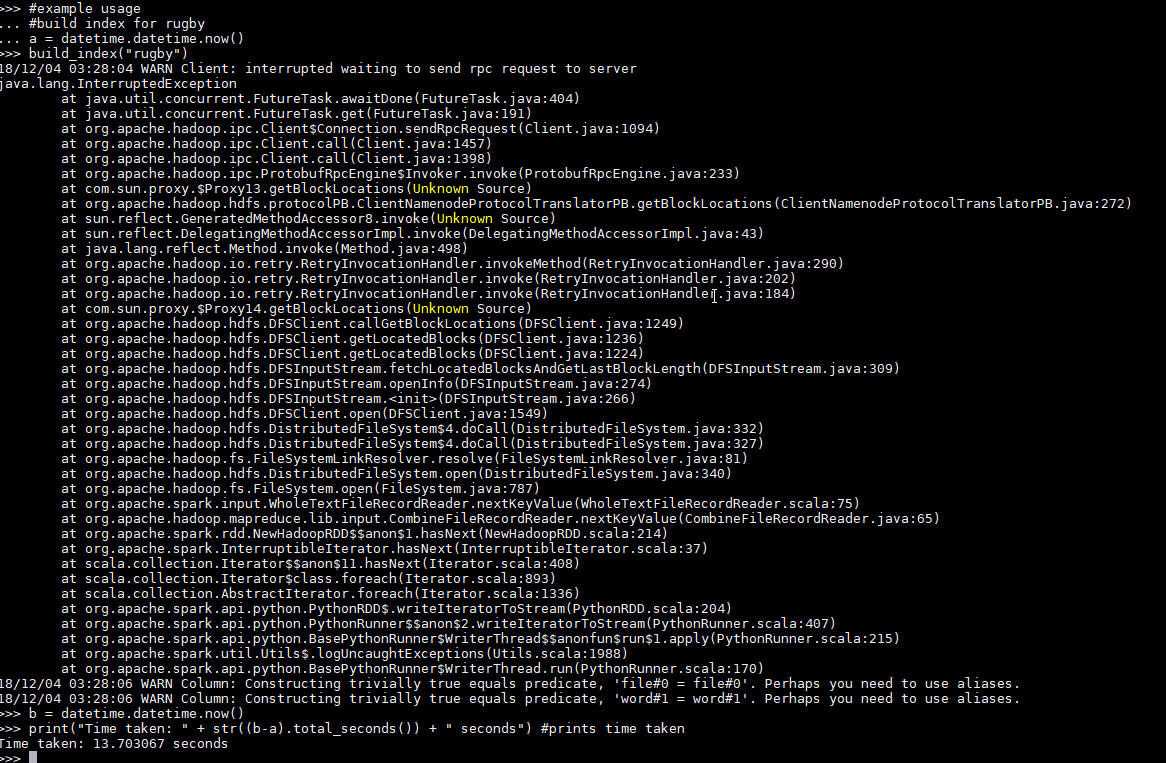
## Define Build Index Function



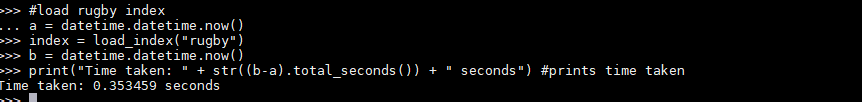
## Define Load Index + Query Function

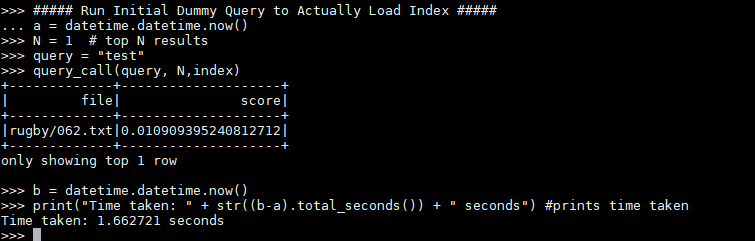


## Run Build Rugby Index



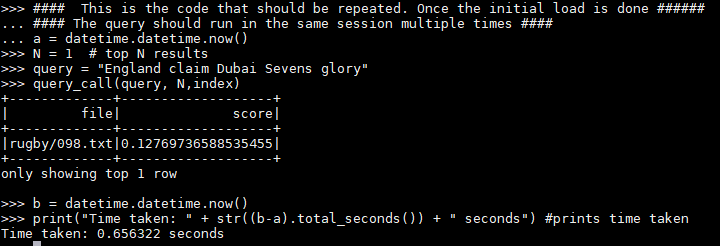
## Run Load Index + Dummy Query to Action the Load



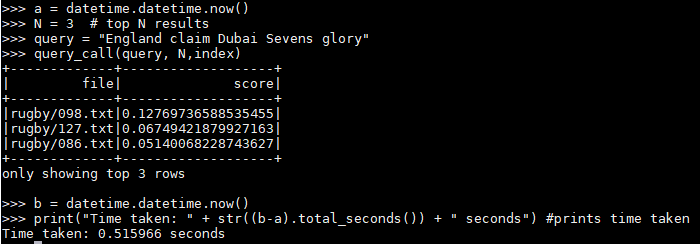


## Run the Queries

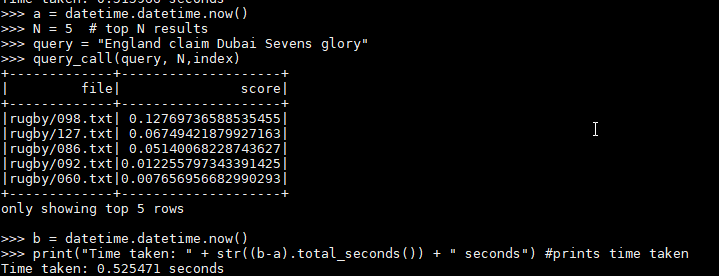
### Query1,N=1



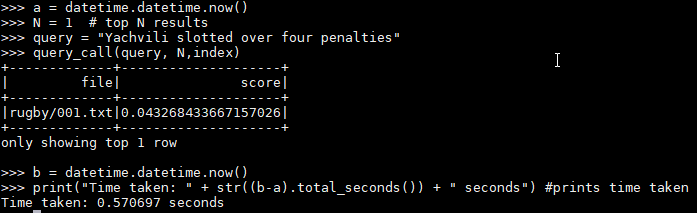
### Query1,N=3



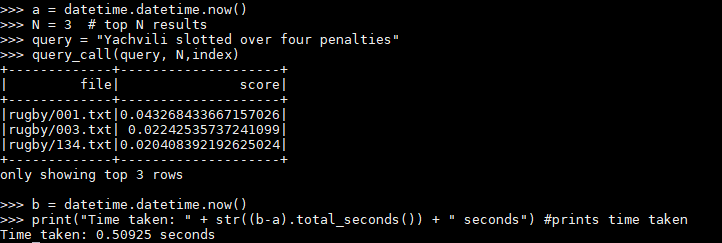
### Query1,N=5



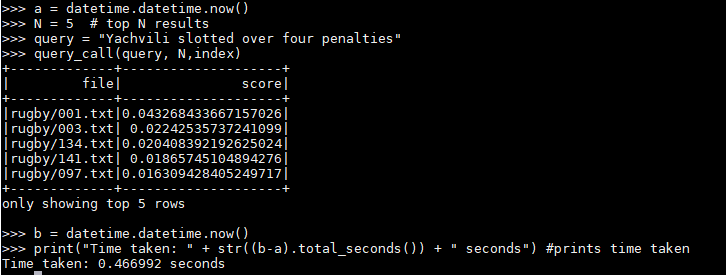
### Query2,N=1



### Query2,N=3



### Query2,N=5



# Appendix C – Excerpts of the top Documents

## Query 1

### Rubgy/098.txt

England claim Dubai Sevens glory

England beat Fiji 26-21 in a dramatic final in Dubai to win the first IRB Sevens event of the season.

Having beaten Australia and South Africa to reach the final, England fell behind to an early try against Fiji. They then took charge with scores from Pat Sanderson, Kai Horstman, Mathew Tait and Rob Thirlby, but Fiji rallied to force a tense finale. Scotland were beaten 33-15 by Samoa in the plate semi-final and Ireland lost 17-5 to Tunisia in the shield final. Mike Friday's England side matched their opponents for pace, power and skill in the final and led 19-7 at half-time. But Neumi Nanuku and Marika Vakacegu touched down for Fiji, only for a needless trip by Tuidriva Bainivalu on Geoff Appleford to allow England to run down the clock. "To be honest, England have wanted to win in Dubai for a very long time now, and the people here have wanted us to win for just as long," said Friday.

"We didn't want to put pressure on ourselves but we are thankful we have achieved that and brought through some young talent at the same time that can hopefully play for the England '15s' in a few years." Portugal confirmed their impressive progress in Sevens rugby by recording a sudden-death win over France in the bowl final. Samoa won the plate title by edging out Argentina 21-19.

### Rugby/127.txt

England given tough Sevens draw

England will have to negotiate their way through a tough draw if they are to win the Rugby World Cup Sevens in Hong Kong next month.

The second seeds have been drawn against Samoa, France, Italy, Georgia and Chinese Taipei. The top two sides in each pool qualify but England could face 2001 winners New Zealand in the quarter-finals if they stumble against Samoa. Scotland and Ireland are in Pool A together with the All Blacks. England won the first event of the International Rugby Board World Sevens series in Dubai but have slipped to fourth in the table after failing to build on that victory.

However, they beat Samoa in the recent Los Angeles Sevens before losing to Argentina in the semi-finals. "England have the ability and determination to win this World Cup and create sporting history by being the only nation to hold both the 15s and Sevens World Cups at the same time," said England sevens coach Mike Friday. "England have a fantastic record in Hong Kong and have won there the last three years, but the World Cup is on a different level. "Every pool contains teams who have caused upsets before and we will have to work hard to ensure we progress from our group. "We have not performed consistently to our true potential so far in the IRB Sevens which has been disappointing - but we can only look forward." England won the first Rugby World Cup Sevens in 1993 with a side that included the likes of Lawrence Dallaglio and Matt Dawson. In 1997 and 2001, England lost in the quarter-finals.

(seeds in brackets)

New Zealand (1), Scotland (8), Tonga, Ireland, Korea, USA.

England (2), Samoa (7), France, Italy, Georgia, Chinese Taipei.

Fiji (3), Australia (6), Canada, Portugal, Japan, Hong Kong.

Argentina (4), South Africa (5), Kenya, Tunisia, Russia, Uruguay.

### Rugby/086.txt

South Africa sweep top awards

South Africa's Schalk Burger was named player of the year as the Tri-Nations champions swept the top honours at the International Rugby Board's awards.

The flanker topped a list which included Ireland star Gordon D'Arcy and Australian sensation Matt Giteau. Jake White claimed the coaching award while his side held off Grand Slam winners France to take the team award. England player Simon Amor beat team-mate Ben Gollings and Argentine Lucio Lopez Fleming to win the sevens award. Burger's award came just a week after he won the equivalent prize from his fellow international players and White, who also coached Burger at under-21 level, paid tribute to him. "Schalk's emergence as a major force has meant a lot to South African rugby, but has also influenced world rugby," said White. "He's become to South African rugby what Jonty Rhodes was to South African cricket. It's amazing what he has achieved in such a short time so far in his international career." Amor, who will captain England in this season's opening IRB Sevens tournament, the Dubai Sevens, which start on Thursday, was delighted with his award. "There are so many great sevens players on the circuit at the moment that this is a genuine honour," said the Gloucester fly-half.

## Query 2

### Rugby/001.txt

Hodgson shoulders England blame

Fly-half Charlie Hodgson admitted his wayward kicking played a big part in England's 18-17 defeat to France.

Hodgson failed to convert three penalties and also missed a relatively easy drop goal attempt which would have given England a late win. "I'm very disappointed with the result and with my myself," Hodgson said. "It is very hard to take but it's something I will have to get through and come back stronger. My training's been good but it just didn't happen." Hodgson revealed that Olly Barkley had taken three penalties because they were "out of my range" but the centre could not convert his opportunities either, particularly the drop goal late on. "It wasn't a good strike," he added. "I felt as soon as it hit my boot it had missed. It's very disappointing, but I must recover." Andy Robinson said he would "keep working on the kicking" with his squad. However, the England coach added that he would take some positives from the defeat.

"We went out to play and played some very good rugby and what have France done?" he said. "They won the game from kicking penalties from our 10m line. "It's very frustrating. The lads showed a lot of ambition in the first half, they went out to sustain it in the second but couldn't build on it. "We took the ball into contact, and you know when you do that it is a lottery whether the referee is going to give the penalty to your side or the other side. "We have lost a game we should have won. There is a fine line between winning and losing, and for the second week we've been on the wrong side of that line and it hurts."

England went in at half-time with a 17-6 lead but they failed to score in the second half and Dimitri Yachvili slotted over four penalties as France overhauled the deficit. England skipper Jason Robinson admitted his side failed to cope with France's improved second-half display.

"We controlled the game in the first half but we knew that they would come out and try everything after half-time," he said. "We made a lot of mistakes in the second half and they punished us. They took their chances when they came. "It's very disappointing. Last week we lost by two points, now one point."

### Rugby/003.txt

Yachvili savours France comeback

France scrum-half Dimitri Yachvili praised his team after they fought back to beat England 18-17 in the Six Nations clash at Twickenham.

Yachvili kicked all of France's points as they staged a second-half revival. "We didn't play last week against Scotland and we didn't play in the first half against England," he said. "But we're very proud to beat England at Twickenham. We were just defending in the first half and we said we had to put them under pressure. We did well." Yachvili admitted erratic kicking from England's Charlie Hodgson and Olly Barkley, who missed six penalties and a drop goal chance between them, had been decisive. "I know what it's like with kicking. When you miss some it's very hard mentally, but it went well for us," he said. France captain Fabien Pelous insisted his side never doubted they could secure their first win against England at Twickenham since 1997. France were 17-6 down at half-time, but Pelous said: "No-one was down at half-time, we were still confident. "We said we only had 11 points against us, which was not much. "The plan was to keep hold of possession and pressure England to losing their composure." France coach Bernard Laporte accepted his side had not played well. "We know we have to play better to defend the title," he said. "I'm not happy we didn't score a try but we're happy because we won."

### Rugby/134.txt

Scotland 18-10 Italy

Six Chris Paterson penalties gave Scotland victory in a dour but clinical encounter against Italy at Murrayfield.

Coach Matt Williams' side were outmuscled and outplayed in a tense first half but led 6-3 at the break. Paterson slotted four more second-half penalties and Scotland were denied a try when wing Sean Lamont's touchdown was ruled out for a forward pass. A late Andrea Masi try was small consolation for Italy, chasing their first away win in the Six Nations. Scotland came out on top of the early exchanges and took a quick 3-0 lead through the boot of full-back Paterson. But the more powerful Azzurri pack eventually rumbled into life.

A series of drives into Scottish territory set up a penalty attempt, missed by full-back Roland De Marigny, and a wayward drop-goal effort from Luciano Orquera. Scotland defended the initial thrusts but on 20 minutes Italy, after coming up yards short of the line, equalised through a De Marigny penalty. Italy were offered another penalty when Scottish flanker Simon Taylor was offside but the left-footed De Marigny pushed his kick wide. Scotland finally made the most of a rare foray into the Italian half and snatched three points from Paterson when an Italian forward handled the ball in a ruck.

As the half wore on, both sides squandered promising spells of momentum with sloppy penalties, and the period fizzled out with Scotland numerically, if not psychologically, on top. Italy's De Marigny narrowly missed a chance to level the scores again shortly after the break but his long-range kick shaved the right upright. And Scotland capitalised with a third Paterson penalty on 50 minutes. Williams' side seemed to have found a spark from somewhere and, after a couple of probing attacks, Paterson was able to slot another three points to widen the gap.

With the pendulum of possession swinging towards Scotland, Lamont thought he had wriggled over in the left corner after 65 minutes but play was recalled for a marginal forward pass from Paterson. Another Paterson penalty on 70 minutes kept the pressure on the wilting visitors. But John Kirwan's men had the last laugh when Gordon Ross' attempted clearance was charged down and Masi pounced for the try, converted by De Marigny.

: C Paterson; S Webster, A Craig, H Southwell, S Lamont; D Parks, C Cusiter; T Smith, G Bulloch (capt), G Kerr; S Grimes, S Murray; S Taylor, J Petrie, A Hogg.

R Russell, B Douglas, N Hines, J Dunbar, M Blair, G Ross, B Hinshelwood.

R de Marigny; Mirco Bergamasco, C Stoica, A Masi, L Nitoglia; L Orquera, A Troncon; A Lo Cicero, F Ongaro, M Castrogiovanni; S Dellape, M Bortolami (capt); A Persico, D Dal Maso, S Parisse.

G Intoppa, S Perugini, CA del Fava, S Orlando, P Griffen, R Pedrazzi, KP Robertson.