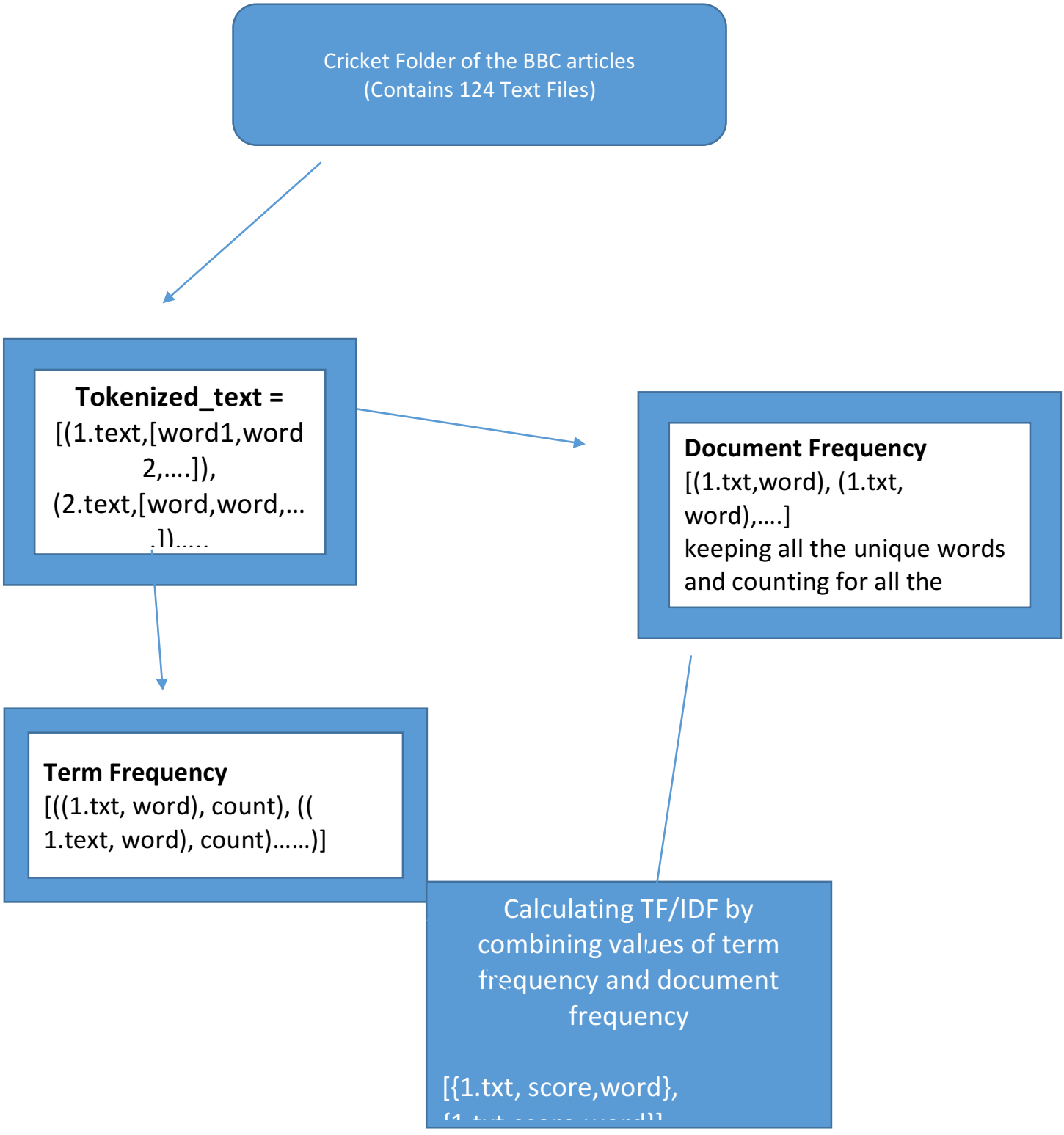
TF-IDF SEARCH USING SPARK

Mezbah Uddin

ID: 500793378

System does a TF-idf based search query on the cricket folder of bbc articles.
The tools that have been used is Spark.
Following is the system design .



In [ ]:
```
TF-IDF Search Using Spark
```

In [ ]:
```

```

In [ ]:
```
# Working with the cricket text files folder
```

In [3]:
```
#importing the data from HDFS into Spark
#mapping into a, b where a=text file name, b= content of the text file
from pyspark.sql import SQLContext, Row
cricket_text = sc.wholeTextFiles('/user/root/crc').map(lambda (a,b): Row(
le =a.replace('hdfs://sandbox.hortonworks.com:8020',''), text=b) )
```

In [5]:
```
number_of_docs = cricket_text.count()
number_of_docs

# output shows its dealing with 124 text files in the cricket folder
```

Out[5]:
```
124
```

In [28]:
```
import re
def tokenize(s):
  return re.split("\\W+", s.lower())



# definition that splits each word of the document and also keeping trac
k of the file name
```

In [13]:
```
# Calculating frequency of each word per document.
#Used flat map values function
#Pass each value in the key-value pair RDD through a flatMap function wi
thout changing the keys;
#this also retains the original RDD's partitioning.

term_frequency = tokenized_text.flatMapValues(lambda x:
x).countByValue()
term_frequency.items()[:5]
```

Out[13]:
```
[((u'/user/root/crc/067.txt', u'team'), 5),
 ((u'/user/root/crc/106.txt', u'taking'), 1),
 ((u'/user/root/crc/071.txt', u'ago'), 1),
 ((u'/user/root/crc/014.txt', u'now'), 1),
 ((u'/user/root/crc/103.txt', u'proved'), 1)]
```

In [51]:
```
document_frequency = tokenized_text.flatMapValues(lambda x: x).distinct()
filter(lambda x: x[1] != '').map(lambda (title,word): (word,title)).coun
tByKey()
document_frequency.items()[:5]

#Step 1: taking all the unique words in all the documents
# Step 2: filtering / discarding any null values
# Counting all the unique words in all the docs by creating dictionary
# count by key()
# the idea is is the any word's count is more than 1, it appeared in mor
e than 1 doc
```

Out[51]:
```
[(u'nudges', 1),
 (u'limited', 7),
 (u'devilliers', 1),
 (u'bidding', 1),
 (u'khalil', 9)]
```

In [52]:
```
document_frequency['nudges']
```

Out[52]:
```
1
```

In [ ]:
```
# explaining the tf_idf function

Step 1: taking each element of term_frequency which is in the format[((f
ile,word),TF)]
as a key value Pair

Step2: assigining (filename,word) into list doc and term
Step 3: collecting document frequency of each term of the docuemnt_frequ
ency
function already created
Step4: calculating tf-idf for each word in each document,along with term
 frequncy
Step 5: appending to result
```

In [57]:
```
# Calculating TF-IDF
import numpy as np
from __future__ import division
def tf_idf(N, tf, df):
    result = []
    for key, value in tf.items():
        doc = key[0]
        term = key[1]
        df = document_frequency[term]
        if (df>0):
          tf_idf = float(value)*np.log(number_of_docs/df)

        result.append({"doc":doc, "term":term, "score":tf_idf})
    return result
tf_idf_output = tf_idf(number_of_docs, term_frequency, document_frequenc
y)
tf_idf_output[:4]
```

Out[57]:
```
[{'doc': u'/user/root/crc/067.txt',
  'score': 3.2294714785469991,
  'term': u'team'},
 {'doc': u'/user/root/crc/106.txt',
  'score': 1.9299098077088723,
  'term': u'taking'},
 {'doc': u'/user/root/crc/071.txt',
  'score': 2.6230569882688175,
  'term': u'ago'},
 {'doc': u'/user/root/crc/014.txt',
  'score': 1.0590814499114745,
  'term': u'now'}]
```

In [ ]:
```
Defining a search function
(1) Tokens= taking the query as string an splitting into words
(2) Word search of the  each word in the query in the rdd to create a jo
ined rdd
which gives word, no of times it appeared in that document and tf-idf sc
ore
(3) scount aggregates by key and returns sum of tfidf based on query for
 each document
(4) scores multiplies the sum multiplied with query doc existences in ea
ch document / len(query)
(5) Also does an inverted index
finally returns top score and document name
```

In [83]:
```
tfidf_RDD = sc.parallelize(tf_idf_output).map(lambda x: (x['term'],(x['doc'],x['score']) )) # the corpus with tfidf scores

def search(query, topN):
  tokens = sc.parallelize(tokenize(query)).map(lambda x: (x,1) ).collectAsMap()
  bcTokens = sc.broadcast(tokens)

  joined_tfidf = tfidf_RDD.map(lambda (k,v): (k,bcTokens.value.get(k,'-')) ).filter(lambda (a,b,c): b != '-' )

  scount = joined_tfidf.map(lambda a: a[2]).aggregateByKey((0,0),
  (lambda acc, value: (acc[0] +value,acc[1]+1)),
  (lambda acc1,acc2: (acc1[0]+acc2[0],acc1[1]+acc2[1])) )

  scores = scount.map(lambda (k,v): ( v[0]*v[1]/len(tokens), k) ).top(topN)

  return scores
```

In [84]:
```
# returns the result in less than 5 seconds
search('bangladesh win',5 )
```

Out[84]:
```
[(19.454654995321306, u'/user/root/crc/115.txt'),
 (13.308703690412814, u'/user/root/crc/077.txt'),
 (8.1330195264573124, u'/user/root/crc/039.txt'),
 (7.3474990257663526, u'/user/root/crc/065.txt'),
 (6.9547387754208723, u'/user/root/crc/057.txt')]
```

In [85]:
```
search('australia plays india',3)
```

Out[85]:
```
[(11.36766318009561, u'/user/root/crc/045.txt'),
 (9.1542441822941516, u'/user/root/crc/044.txt'),
 (7.080891939236551, u'/user/root/crc/026.txt')]
```

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

Search 1 Results(txt documents)

search('bangladesh win',5)

115.txt

Bangladesh delighted at Test win

Bangladeshi players and fans celebrated after the side's historic first Test victory, over Zimbabwe in Chittagong.

Thousands of fans, armed with drums and flags, ran into the streets in the capital Dhaka within minutes of the end of the game, halting traffic. "It's the best day in my life. I won't forget the day I was a member of Bangladesh's winning team. "I don't want to remember those hard days, I only want to think about the victory," said captain Habibul Bashar. Bangladesh President Iajuddin Ahmed, Prime Minister Khaleda Zia and opposition leader Sheikh Hasina have all congratulated the team. The win by 226 runs in Chittagong was Bangladesh's first Test win at their 35th attempt since being granted Test status in 2000. Bangladesh managed three draws in their previous 34 matches - two of them against Zimbabwe and one against West Indies. Bangladesh coach Dav Whatmore described the victory as "a wonderful feeling". "You can see the joy and the relief of lots of other people," Whatmore told BBC World Service. "We've taken our share of hammerings in the last year and a half and, putting this win in perspective, there's probably a few more down the track. "But I sense there's a bit more self-belief when they come to play tougher opposition."

Whatmore led Sri Lanka to victory in the 1996 World Cup and said this could not compare. But he continued: "It was important for the whole country that the sport of cricket stand up and show that we're progressing. "There's been a lot of frustration for a long time here in Bangladesh that the team is not pushing the opposition enough." The former Australia Test batsman played down the status of Zimbabwe, whose weakened side have just returned from a seven-month suspension of their Test status. "Yes, our opponents are ranked pretty much near us at the moment so right from the outset that would suggest we had a chance of winning," he added. "But to actually go out there and do it is another matter." Zimbabwe captain Tatenda Taibu, who had made 92 in the first innings, but was dismissed for a duck in the second, was disappointed with the performance. "There was some bad cricket on our side and good cricket by the Bangladeshis," he said. "Our top order batsmen didn't come to the party and we dropped about four to five catches." The second and final Test of the series against Zimbabwe begins on Friday in Dhaka.

077.txt

Kaif shines in India win

First one-day international, Chittagong: India 245-8 (50 overs) v Bangladesh 234-8 (50 overs) by 11 runs

Mohammad Kaif (80) and Rahul Dravid (53) shared a stand of 128 as the tourists posted a total of 245-8. Skipper Habibul Bashar kept Bangladesh in the hunt with 65, but his departure left them with too much to do. Khaled Mashud hit an unbeaten 50 but Sridharan Sriram took 3-43 as the home side were restricted to 234-8. After winning the Test series 2-0, India took the opportunity to rest batsman Virender Sehwag and fast bowler Zaheer Khan and give debuts to wicket-keeper Mahendra Dhoni and seamer Joginder Sharma. But skipper Sourav Ganguly lost the toss and opposite number Bashar's decision to put them in paid off initially as India were reduced to 45-3. Ganguly was bowled for nought by the second ball of the match from Tapash Baisya and 17-year-old Nazmul Hossain then claimed the prized scalp of Sachin Tendulkar, who was caught behind for 19.

Mushfiqur Rahman trapped Yuvraj Singh lbw for 21, but Kaif and Dravid steadied the innings and Bangladesh had to wait 28 overs for their next success. Both batsmen reached their half centuries off 74 balls, but Dravid edged a catch to the keeper off Khaled Mahmud soon after and Sriram was stumped for three off spinner Mohammad Rafique. Dhoni's first innings for India lasted one delivery as he was run out for nought and when Kaif gave a return catch to Nazmul in the 47th over, the total had only just passed 200. But Ajit Agarkar made 25 and Irfan Pathan hit two sixes in his 21 not out off 11 balls, runs which ultimately made all the difference.

Bangladesh were soon in trouble in reply as Rafique (eight), Nafis Iqbal (nine) and Mohammad Ashraful (two) all failed - the latter becoming Sharma's first international victim when he was caught by Ganguly. Bashar and Aftab Ahmed put on 64 in 14 overs before both fell victim to Sriram's left-arm spin, along with Rajin Saleh (14), as the home side slumped from 108-3 to 156-6. Mushfique was lbw to Agarkar for two but Mashud and Mahmud did their best to revive their side, adding 40 for the eighth wicket in six overs. The target was out of reach, however, and Mahmud perished for 21 to a catch by Man of the Match Kaif as Bangladesh's hopes were finally extinguished. Mashud had the consolation of reaching his fifth one-day half century before Baisya drove the final ball of the game to extra cover for four, but it was too late for Bangladesh.

Nafis Iqbal, Habibul Bashar (Capt), Mohammad Ashraful Aftab Ahmed, Khaled Mashud (Wkt), Mushfiqur Rahman, Khaled Mahmud Manjural Islam Rana, Mohammad Rafique, Tapash Baisya Nazmul Hossain.

S Sriram, S R Tendulkar, S C Ganguly (Capt), R Dravid Yuvraj Singh, M Kaif, M S Dhoni (Wkt), I K Pathan, Harbhajan Singh J Sharma, A B Agarkar.

Aleem Dar and Mahbubur Rahman

039.txt

Bashar delighted after series win

Bangladesh skipper Habibul Bashar was thrilled after the win in the decider against Zimbabwe gave his country their first-ever one-day series triumph.

"Yes, our morale was down after losing the first two matches," he said, after the eight-wicket victory in Dhaka. "But we knew we could come back and win the series. "We worked hard and tried to rectify our mistakes and I am happy to have won the match chasing a target, which was not easy." Taibu's counterpart, Tatenda Taibu, was left to reflect on what might have been after losing the early initiative in the five-match series.

"It was so near and yet so far for us," said the Zimbabwe captain. "I think we did not play spin well enough and lost wickets at crucial stages. They played well, but I think our boys did a good job. We worked hard and fought it out." Bangladesh have won only nine of their 106 matches since making their one-day debut in 1986 and clinched their maiden Test series against Zimbabwe earlier in January.

065.txt

India wrap up victory in Dhaka

First Test, Dhaka: Bangladesh 184 & 202 v India 526

India win by an innings and 140 runs

Left-arm paceman Irfan Pathan removed Tapash Baisya for 29 to finish with figures of 6-51, and 11-96 overall. Zaheer Khan claimed the final wicket when he had the diligent Manjural Islam Rana caught behind for 69. The home side, 170-8 overnight, subsided for 202 to slump to defeat by an innings and 140 runs. Bangladesh were left with a daunting task after Sachin Tendulkar's record unbeaten 248 helped India to a total of 526, a lead of 342.

Only Nafis Iqbal (54) and Islam Rana offered any real resistance as the hosts were routed in double-quick time. In their 33 Tests since 2000, Bangladesh have now accumulated 30 defeats, with only three draws to their credit. The second and final Test of the series starts in Chittagong on Friday.

Habibul Bashar (capt), Nafis Iqbal, Javed Omar, Mohammad Ashraful, Rajin Saleh, Khaled Mashud (wkt), Mushfiqur Rahman, Mohammad Rafique, Tapash Baisya, Mashrafe Mortaza, Manjurul Islam Rana.

S Ganguly (capt), V Sehwag, G Gambhir, S Tendulkar, R Dravid, M Kaif, D Karthik (wkt), I Pathan, A Kumble, Harbhajan Singh, Z Khan.

057.txt

Ganguly plays down fears

India captain Sourav Ganguly has attempted to play down safety fears over their tour to Bangladesh.

The Indian squad arrived in Dhaka on Wednesday for a 19-day tour featuring two Tests and three one-day matches. The first Test has already been put back a day to Friday after the Indian embassy received threats purporting to come from Islamic militants. "Security is an important factor but we as a team are concentrating on cricket and nothing else," Ganguly insisted. A hand-written fax allegedly sent by the Harkat-ul-Zihad group threatened to kill Indian cricketers, but has been dismissed as a hoax by the Bangladesh authorities. They are suspected of carrying out the assasination of poet Shamsur Rahman six years ago. The group's hostility towards India stems from riots in the western state of Gujarat in 2002, which left 2,000 people dead, many of them muslims. The Board of Control for Cricket in India is leaving nothing to chance and are sending security experts to assess the situation in Chittagong, where the second Test is due to start on 16 December.

Despite Bangladesh's mediocre record of 29 defeats in 32 matches at Test level since 2000, Ganguly said his team would take nothing for granted. "I don't think Bangladesh are pushovers. I always respect the opposition and Bangladesh are no exception. "I don't think any side has has gone and played in Bangladesh with a sense of complacency." India were Bangladesh's first Test opponents four years ago, winning by nine wickets in Dhaka despite the home side making 400 in their first innings.

Search 2 Results(txt documents)
Australia Plays India

45.txt
Australia dominate India

Third Test, Nagpur, day two (stumps):
Australia 398 v India 146-5

The home side closed on 146-5, 252 runs in arrears and still 53 short of avoiding the follow-on. Glenn McGrath, playing in his 100th Test, took 2-18 in 20 overs as Mohammad Kaif (47 not out) propped up an innings that crawled at two runs per over. Australia were earlier dismissed for 398, with Michael Clarke falling to Zaheer Khan (4-95) nine short of a ton. India did not take long in wrapping up Australia's innings at the start of the day, claiming the three remaining overnight wickets for 36 runs. Jason Gillespie, lbw to Zaheer, and Kasprowicz, bowled by Agit Agarkar, both fell cheaply, leaving McGrath as Clarke's last hope of a second Test century. The veteran enjoyed his stay at the crease with two hooked boundaries, and in the end it was he

who remained unbeaten when Clarke edged behind to end an entertaining knock that included 13 boundaries.

India were given a typically blustery start by Virender Sehwag, who took four boundaries off Gillespie's first over. It forced skipper Adam Gilchrist to protect the boundary with a sweeper after just five overs, but having forced Australia on the defensive Sehwag soon gave his wicket away. Flashing hard at McGrath, the opener became the Aussie paceman's 447th victim when edging into the gloves of the athletic Gilchrist, diving high to his right. Gillespie, belted out of the attack by Sehwag, returned from the other end and struck immediately to have Aakash Chopra caught by Warne at slip for eight.

It left India 35-2 at lunch, and a very quiet period followed the interval as Rahul Dravid and the returning Sachin Tendulkar tried to block India to parity. The pair put on 15 runs in one ball shy of 13 dawdling overs before Tendulkar was trapped in front of stumps by Gillespie (2-47) for eight. The grassy Nagpur pitch meant Shane Warne's introduction was delayed until the 36th over, but two balls were all it took for the Test cricket's leading wicket-taker to add to his tally. In having VVS Laxman caught by Clarke in gully, Warne ousted him for the third time this series and reduced India to a perilous 75-4 in the process. Dravid's stay at the crease was a lengthy one but largely unproductive, and after consuming 140 balls for 21 runs he edged McGrath to slip. It was a predictable end after the veteran paceman had worked India's stand-in skipper over with an array of balls which leapt off the seam on a pitch more akin to Australian grounds.

At 103-5 India looked vulnerable, but Kaif and Parthiv Patel - fresh from their century stand in the second Test - showed heart in the fading light. It was not plain sailing for Kaif - a spurned run-out chance and a catch off a no-ball offered two reprieves - but he grew into his task and even managed a six off Warne over long-off. Although India's sixth wicket offered the home side hope, Australia were in a strong position to push for the victory that would award them a first series victory in India for 35 years.

A Chopra, V Sehwag, R Dravid (Capt), S R Tendulkar V V S Laxman, M Kaif, P A Patel (Wkt), A B Agarkar, A Kumble M Kartik, Z Khan.

M L Hayden, J L Langer, S M Katich, D R Martyn D S Lehmann, M J Clarke, A C Gilchrist (Capt, Wkt), S K Warne M S Kasprowicz, J N Gillespie, G D McGrath.

Aleem Dar, D R Shepherd.

44.txt

Australia build imposing lead

Third Test, Nagpur, day three (stumps):
Australia 202-3 & 398 v India 185

India were bowled out for just 185 in the morning session, with paceman Gillespie returning 5-56 - his eighth five-wicket haul in Test cricket. Katich then made 99 as the Aussies established an intimidating 415-run lead in reaching 202-3 at stumps. It places Australia on the brink of a first series win in India for 35 years. Murali Kartik showed some spirit with two wickets in the evening, but with Damien Martyn (41) and Michael Clarke (10) at the crease and batting still to come Australia look set to push on come Friday. India's tail fared no better than the top order on Nagpur's lively pitch, losing five wickets for 39 in the morning after resuming play on 146-5.

Eight balls into the day, Shane Warne claimed his 539th Test victim when fading one away from Parthiv Patel and taking the edge en route to slip. Skipper Adam Gilchrist delayed taking the new ball for a few overs, but was rewarded almost instantly when he did. Firstly, Gillespie gave Clarke the first of two catches at third slip to get rid of Agit Agarkar. And soon India's last line of defence was dismantled when Glenn McGrath and first slip Warne combined to oust Kaif for 55 - a second successive Test match fifty. India were nine down when Kartik fell to Gillespie - one of eight to perish to catches off the edge of the bat - and all that remained was for the paceman to bag his first five-wicket haul against India with the dismissal of last man Zaheer. With plenty of time left in the Test and an opportunity to rest his attack presenting itself, Gilchrist chose not to enforce the follow-on despite Australia's 213-run first-innings lead.

The early loss of Matthew Hayden aside - bowled during an excellent Zaheer Khan spell - it proved a sensible decision. It was largely one-way traffic once Australia had negotiated a tough, run-free 40-minute period after lunch, with Justin Langer surviving a couple of close lbw appeals to put on 80 for the second wicket with Katich. Langer's scratchy stay came to an end just after tea when trying to hit Kartik out of the park but succeeding only in holing out at deep mid-wicket. Another meaningful partnership was in the offing, however, and Katich and the in-form Martyn upped the scoring rate in a deflating period for India. Katich was supreme against the spinners, playing patiently square of the wicket and using his feet to move into the 90s with sucessive on-driven boundaries off the hugely disappointing Anil Kumble (0-62). But fate - and Kartik - conspired against Katich, who stepped back in his crease and was plumb lbw to register the 74th score of 99 in Test cricket history. Martyn and Clarke were untroubled in the closing overs against a tiring attack, the latter announcing his arrival at the crease with two glorious boundaries in one Kumble over.

26.txt

Pakistan arrive for tour of India

Pakistan arrived in Delhi on Monday amid tight security for their first full tour of India for six years.

The 16-man squad, along with coach Bob Woolmer and support staff, were met at the heavily guarded Indira Gandhi airport after a 45-minute flight. Armed guards kept vigil as the tourists were taken to their hotel along a six-mile route lined with security men. Hindu fundamentalists have threatened to protest over perceived Pakistani backing to militancy in Kashmir. "I am delighted to be back in India," captain Inzamam-ul-Haq said. "I have always enjoyed playing here because people are crazy about cricket. "I am not worried about security. My only concern is how my team plays on the tour." Series between India and Pakistan are always eagerly awaited, largely because politics often restrict the amount of cricket the two teams can play against each other.

This particular engagement, which will include three Tests and six one-dayers, was twice threatened. The schedule has already been put back four days over a dispute which centred on Pakistan's refusal to play the second Test in Ahmedabad. The first Test in Mohali will now get under way on 8 March. Then, following a protracted row over television rights, the Madras High Court issued interim orders to ensure the matches would be broadcast live by state=owned channel Doordarshan. Inzamam beseeched his side to make up for the 2004 defeat to India on home soil, where they lost 2-1 in the Tests and 3-2 in the one-day internationals. He said: "We want to make amends for last year. "The Australian tour was a big learning experience and I think we are in a position to reap benefits in India." After the 2004 series, Pakistan sacked their coach Javed Miandad and replaced him with Woolmer. Inzamam believes Woolmer has improved the side even if a string of positive results have not yet come.

He said: "A lot has changed since that series. Woolmer has helped change the attitude of the boys, which is more positive and professional and I think they can take the pressure of playing in India." Pakistan will be without strike bowler Shoaib Akhtar (hamstring), leaving recalled off-spinner Arshad Khan and leg-spinner Danish Kaneria with much work to do. Inzamam said: "Shoaib's absence will be felt. But our strength is playing as a unit. "[Khan and Kaneria] have a big role to play. They are the ones on whom we will be depending a lot in the Test matches. "We are expecting slow, turning tracks and we have been practicing for such conditions." Pakistan's first match is a three-day warm-up against the Indian Board President's XI in Dharamsala from March 3.

Sourav Ganguly (captain), Virender Sehwag, Gautam Gambhir, Rahul Dravid, Sachin Tendulkar, VVS Laxman, Yuvraj Singh, Dinesh Karthik (wicketkeeper), Irfan Pathan, Anil Kumble, Harbhajan Singh, Zaheer Khan, Ashish Nehra, Lakshmipathy Balaji.

Salman Butt, Yasir Hameed, Taufeeq Umar, Younis Khan, Inzamam-ul-Haq (captain), Yousuf Youhana, Asim Kamal, Kamran Akmal (wkt), Abdul Razzaq, Shoaib Malik, Shahid Afridi, Arshad Khan, Danish Kaneria, Muhammad Sami, Rana Naved, Muhammad Khalil.