DS8003 Final: TFIDF SEARCH WITH SPARK

Anika Tabassum
ID: 500865054

Summary

- My system uses Apache Spark with HDFS for creating the TFIDF index and searching for queries. I am using the cricket corpus for this project.
- It first loads the documents as separate records and then tokenizes each record and calculates the count of each word per document (TF). Then it calculates the number of distinct documents for each term (DF).
  It calculates IDF: $$IDF(t,D) = \log \frac{|D|+1}{DF(t,D)+1},$$

  where $|D|$ = total number of documents

  Then it calculates the TFIDF index : $$TFIDF(t,d,D) = TF(t,d) \cdot IDF(t,D).$$
- Once the TFIDF index is built, my system is able to take any query and tokenize the query the same way it would tokenize any document in the corpus and then conduct the search.
- The system is awesome because it uses Apache Spark which is a blazing fast distributed in-memory data processing engine. It provides easy-to-read functional APIs and leverages lazy evaluation scheme. Also, I am using the DataFrame API which is more performance-optimized than the traditional RDD based API.

Why I made use of certain tools and why I did not choose others

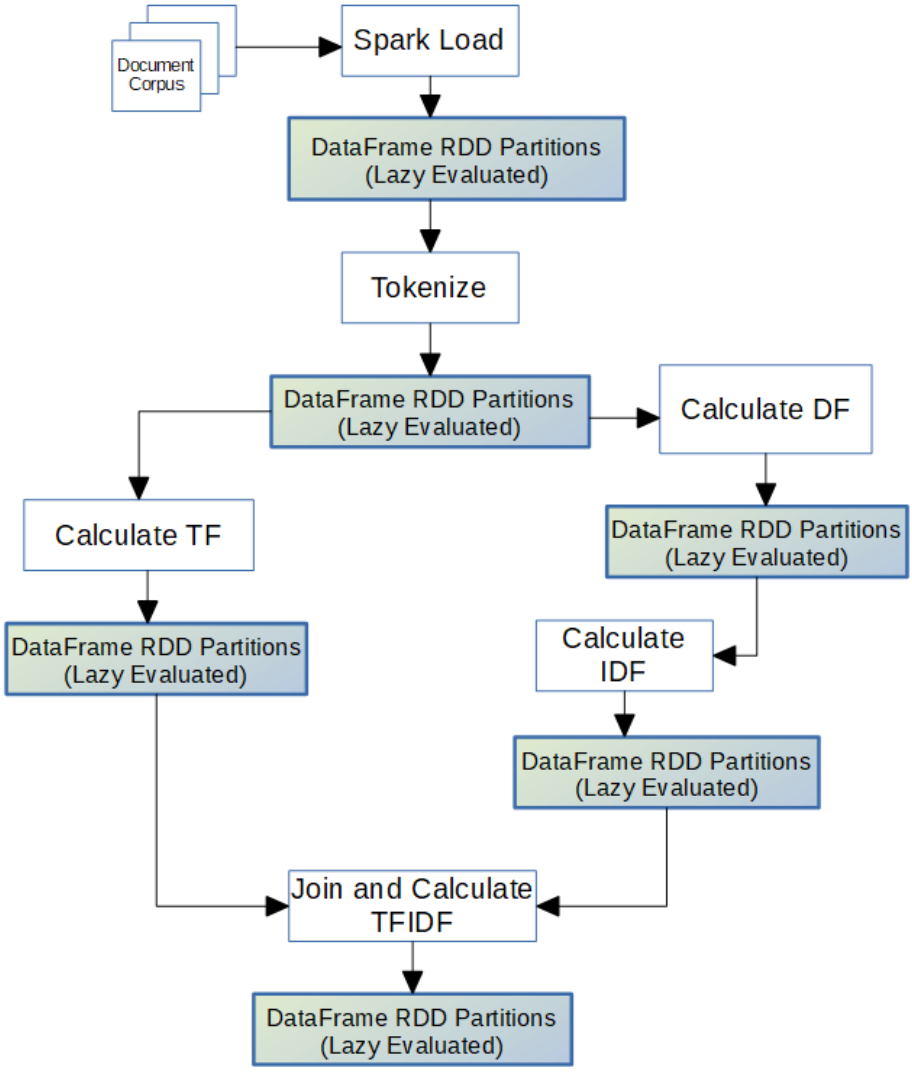
MapReduce, Hive, Pig, HBase, Mongo – all these other tools make use of disk access in some way or another.
Spark's processing is fully in-memory and much faster than traditional MapReduce. Since I am doing everything in-memory I am not using any disk-based storage and retrieval systems like Hive, Pig, etc. I am only using the HDFS disk space for the initial storage of the document corpus.

How search system is made to work real-time

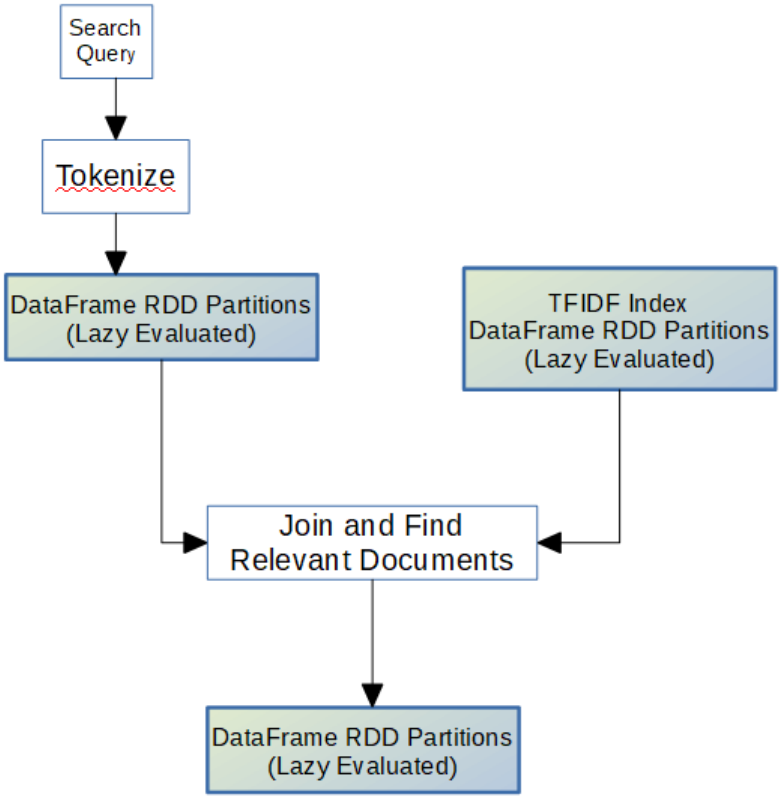
After building the TFIDF index, I used Spark's persist() method so that it caches the TFIDF index in memory after the next search. Thereafter, any further search will be near real-time.

Architecture Diagrams:

Part 1 (Computing the TFIDF score):



Part 2 (Search) :



Results Section:

Choose 2 queries (with number of words > 1). For each query report results for following values of N

- N=1

```
search('Bangladesh wins', 1)
search('Bangladesh vs India', 1)
```

```
18/12/06 18:55:18 INFO EventLoggingListener: Logging events
+--------------------------------+-----------------+
|doc                             |score            |
+--------------------------------+-----------------+
|/user/root/Final/cricket/115.txt|6.782796859741211|
+--------------------------------+-----------------+
only showing top 1 row

+--------------------------------+------------------+
|doc                             |score             |
+--------------------------------+------------------+
|/user/root/Final/cricket/057.txt|11.952113469441732|
+--------------------------------+------------------+
only showing top 1 row
```

- N=3

```
search('Bangladesh wins', 3)
search('Bangladesh vs India', 3)
```

```
+--------------------------------+-----------------+
|doc                             |score            |
+--------------------------------+-----------------+
|/user/root/Final/cricket/057.txt|6.782796859741211|
|/user/root/Final/cricket/115.txt|6.782796859741211|
|/user/root/Final/cricket/077.txt|5.813826084136963|
+--------------------------------+-----------------+
only showing top 3 rows
```

```
+--------------------------------+------------------+
|doc                             |score             |
+--------------------------------+------------------+
|/user/root/Final/cricket/057.txt|11.952113469441732|
|/user/root/Final/cricket/077.txt|11.387248357137045|
|/user/root/Final/cricket/058.txt|9.205960273742676 |
+--------------------------------+------------------+
only showing top 3 rows
```

- N=5

```
search('Bangladesh wins', 5)
search('Bangladesh vs India', 5)
```

```
+--------------------------------+------------------+
|doc                             |score             |
+--------------------------------+------------------+
|/user/root/Final/cricket/057.txt|6.782796859741211 |
|/user/root/Final/cricket/115.txt|6.782796859741211 |
|/user/root/Final/cricket/058.txt|5.813826084136963 |
|/user/root/Final/cricket/077.txt|5.813826084136963 |
|/user/root/Final/cricket/065.txt|2.9069130420684814|
+--------------------------------+------------------+
only showing top 5 rows
```

```
+--------------------------------+------------------+
|doc                             |score             |
+--------------------------------+------------------+
|/user/root/Final/cricket/057.txt|11.952113469441732|
|/user/root/Final/cricket/077.txt|11.387248357137045|
|/user/root/Final/cricket/058.txt|9.205960273742676 |
|/user/root/Final/cricket/060.txt|6.784268379211426 |
|/user/root/Final/cricket/065.txt|6.784268379211426 |
+--------------------------------+------------------+
only showing top 5 rows
```

Please find actual contents of the resulting documents appended at the end of the report. Relevant words have been highlighted in red with yellow background.

Code

Steps to setup:

hdfs dfs -mkdir /user/root/Final

hdfs dfs -put /root/Final/cricket /user/root/Final

```
[root@sandbox-hdp ~]# hdfs dfs -mkdir /user/root/Final
[root@sandbox-hdp ~]#
```

```
[root@sandbox-hdp ~]# hdfs dfs -put /root/Final/cricket /user/root/Final
[root@sandbox-hdp ~]#
```

I have run the following commands in Pyspark console. However, I have also provided the python file just in case you want to run it with spark-submit: spark-submit --num-executors=4 /root/Final/final.py

TFIDF and search

```
import re
import math
import pyspark.sql.functions as F
from pyspark.sql.types import ArrayType, StringType, FloatType
from pyspark.sql import SparkSession

# Q1 - Computing TFIDF scores
# --------------------------
# Build Spark session
spark = SparkSession.builder.getOrCreate()
# Omit all logs except errors
spark.sparkContext.setLogLevel('ERROR')
# Read each file in cricket folder as a separate record
rdd = spark.sparkContext.wholeTextFiles('/user/root/Final/cricket/')
# Suppress hortonworks path prefix from filename and create
# data frame with 2 columns ('doc' and 'text')
data = rdd.map(lambda x: (x[0].replace('hdfs://sandbox-hdp.hortonworks.com:8020', ''),
x[1])).toDF(['doc', 'text'])
# Get total document count
total_docs = data.count()

# utility method for tokenizing a piece of text
def tokenize(text):
    return re.findall('\\w+', text.lower())
# Register the tokenize method as a udf
```

```
tokenize_udf = F.udf(tokenize, ArrayType(StringType()))
# tokenize all the text
data = data.select(['doc', tokenize_udf('text').alias('text')])
# make 1 separate row for each token
data_tokens = data.withColumn("token", F.explode('text'))

# calculate term frequency
tf = data_tokens.groupBy('doc', 'token').agg(F.count('text').alias('tf'))
# calculate document frequency
df = data_tokens.groupBy('token').agg(F.countDistinct('doc').alias('df'))

# utility method for calculating inverse document frequency
def inverse_doc_frequency(doc_frequency):
    return math.log((total_docs + 1) * 1.0 / (doc_frequency + 1))

# register inverse document frequency as a udf
inverse_doc_frequency_udf = F.udf(inverse_doc_frequency, FloatType())
# calculate the inverse document frequency
idf = df.withColumn('idf', inverse_doc_frequency_udf('df'))
# calculate tfidf
tfidf = tf.join(idf, 'token').withColumn('tfidf', F.col('tf') * F.col('idf'))
# show 10 rows from tfidf index
# tfidf.show(10, False)
```

```
>>> import re
>>> import math
>>> import pyspark.sql.functions as F
>>> from pyspark.sql.types import ArrayType, StringType, FloatType
>>> from pyspark.sql import SparkSession
>>> rdd = spark.sparkContext.wholeTextFiles('/user/root/Final/cricket/')

>>> data = rdd.map(lambda x: (x[0].replace('hdfs://sandbox-hdp.hortonworks.com:8020', ''), x[1])).toDF(['doc', 'text'])
>>>
```

```
>>> total_docs = data.count()
>>> def tokenize(text):
...     return re.findall('\\w+', text.lower())
...
>>> tokenize_udf = F.udf(tokenize, ArrayType(StringType()))
>>> data = data.select(['doc', tokenize_udf('text').alias('text')])
>>> data_tokens = data.withColumn("token", F.explode('text'))
>>> tf = data_tokens.groupBy('doc', 'token').agg(F.count('text').alias('tf'))
>>> df = data_tokens.groupBy('token').agg(F.countDistinct('doc').alias('df'))
>>> def inverse_doc_frequency(doc_frequency):
...     return math.log((total_docs + 1) * 1.0 / (doc_frequency + 1))
...
>>> inverse_doc_frequency_udf = F.udf(inverse_doc_frequency, FloatType())
>>> idf = df.withColumn('idf', inverse_doc_frequency_udf('df'))
>>> tfidf = tf.join(idf, 'token').withColumn('tfidf', F.col('tf') * F.col('idf'))
>>> tfidf.show(10)
+-------+--------------------+---+---+---------+---------+
|  token|                 doc| tf| df|      idf|    tfidf|
+-------+--------------------+---+---+---------+---------+
|  1970s|/user/root/Final/...|  1|  1|4.1351666|4.1351666|
|    296|/user/root/Final/...|  2|  1|4.1351666| 8.270333|
|bastman|/user/root/Final/...|  1|  1|4.1351666|4.1351666|
| doubts|/user/root/Final/...|  1|  2|3.7297015|3.7297015|
| doubts|/user/root/Final/...|  1|  2|3.7297015|3.7297015|
|elevate|/user/root/Final/...|  1|  1|4.1351666|4.1351666|
|    few|/user/root/Final/...|  1| 16|1.9951004|1.9951004|
|    few|/user/root/Final/...|  1| 16|1.9951004|1.9951004|
|    few|/user/root/Final/...|  1| 16|1.9951004|1.9951004|
|    few|/user/root/Final/...|  1| 16|1.9951004|1.9951004|
+-------+--------------------+---+---+---------+---------+
only showing top 10 rows

>>>
```

Q2 Retrieve the top N matching documents with a score

```
# ------------------------------------------------------
# 2a search
# utility method for searching query
def search(query, N):
    # tokenize query into terms
    terms = tokenize(query)
    # create a dataframe with each term as a separate row
    query_tokens = spark.createDataFrame(terms, StringType()).withColumnRenamed('value', 'token')
    # get aggregated score and count for each document for all the matched tokens
    result = query_tokens.join(tfidf, 'token').groupBy('doc').agg(F.sum('tfidf').alias('score_sum'),
F.count('tfidf').alias('matched_terms'))
    # calculate document score
    result = result.withColumn('score', F.col('score_sum') * F.col('matched_terms') / len(terms))
    # show top N documents
    result.select('doc', 'score').sort(F.col('score').desc()).show(N, False)


# for searching
# search('Bangladesh', 10)
```

```
>>> def search(query, N):
...     terms = tokenize(query)
...     query_tokens = spark.createDataFrame(terms, StringType()).withColumnRenamed('value', 'token')
...     result = query_tokens.join(tfidf, 'token').groupBy('doc').agg(F.sum('tfidf').alias('score_sum'), F.count('tfidf').alias('matched_terms'))
...     result = result.withColumn('score', F.col('score_sum') * F.col('matched_terms') / len(terms))
...     result.select('doc', 'score').sort(F.col('score').desc()).show(N, False)
...
>>> search('Bangladesh')
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: search() takes exactly 2 arguments (1 given)
>>> search('Bangladesh',10)
+-------------------------------+-----------------+
|doc                            |score            |
+-------------------------------+-----------------+
|/user/root/Final/cricket/115.txt|13.565593719482422|
|/user/root/Final/cricket/057.txt|13.565593719482422|
|/user/root/Final/cricket/058.txt|11.627652168273926|
|/user/root/Final/cricket/077.txt|11.627652168273926|
|/user/root/Final/cricket/060.txt|5.813826084136963 |
|/user/root/Final/cricket/065.txt|5.813826084136963 |
|/user/root/Final/cricket/061.txt|5.813826084136963 |
|/user/root/Final/cricket/039.txt|3.8758840560913086|
|/user/root/Final/cricket/090.txt|3.8758840560913086|
|/user/root/Final/cricket/111.txt|3.8758840560913086|
+-------------------------------+-----------------+
only showing top 10 rows

>>>
```

# 2b - real-time search
# for near real time
# persist the tfidf index into memory after the next search
tfidf.persist()

# and then search
search('Bangladesh wins', 1)
search('Bangladesh vs India', 1)

search('Bangladesh wins', 3)
search('Bangladesh vs India', 3)

search('Bangladesh wins', 5)
search('Bangladesh vs India', 5)

```
>>> tfidf.persist()
DataFrame[token: string, doc: string, tf: bigint, df: bigint, idf: float, tfidf: float]
>>> search('Bangladesh', 10)
+--------------------------------+------------------+
|doc                             |score             |
+--------------------------------+------------------+
|/user/root/Final/cricket/057.txt|13.565593719482422|
|/user/root/Final/cricket/115.txt|13.565593719482422|
|/user/root/Final/cricket/077.txt|11.627652168273926|
|/user/root/Final/cricket/058.txt|11.627652168273926|
|/user/root/Final/cricket/065.txt|5.813826084136963 |
|/user/root/Final/cricket/060.txt|5.813826084136963 |
|/user/root/Final/cricket/061.txt|5.813826084136963 |
|/user/root/Final/cricket/039.txt|3.8758840560913086|
|/user/root/Final/cricket/111.txt|3.8758840560913086|
|/user/root/Final/cricket/090.txt|3.8758840560913086|
+--------------------------------+------------------+
only showing top 10 rows

>>> search('Bangladesh', 10)
+--------------------------------+------------------+
|doc                             |score             |
+--------------------------------+------------------+
|/user/root/Final/cricket/115.txt|13.565593719482422|
|/user/root/Final/cricket/057.txt|13.565593719482422|
|/user/root/Final/cricket/058.txt|11.627652168273926|
|/user/root/Final/cricket/077.txt|11.627652168273926|
|/user/root/Final/cricket/065.txt|5.813826084136963 |
|/user/root/Final/cricket/060.txt|5.813826084136963 |
|/user/root/Final/cricket/061.txt|5.813826084136963 |
|/user/root/Final/cricket/039.txt|3.8758840560913086|
|/user/root/Final/cricket/090.txt|3.8758840560913086|
|/user/root/Final/cricket/111.txt|3.8758840560913086|
+--------------------------------+------------------+
only showing top 10 rows
```

Files in the Results section:

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.

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

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.

058.txt

India clear for Chittagong games

Indian cricket officials have given clearance for the national team to play in Chittagong during their tour to Bangladesh, which began on Wednesday.

The team received alleged death threats from a radical Muslim group. But the foreign ministry were satisfied with security arrangements in Dhaka, where the first Test begins on Friday. And the team have now given the green light for the games scheduled for Chittagong, which will stage the second Test and first one-day international. "We have now received assessment of the Indian security team that went to Chittagong. "They made a full assessment of the security arrangements, in consultation with the Bangladesh authorities. "On the basis of security team's report, the government is advising BCCI to go ahead with the matches in Chittagong," said foreign ministry spokesman Navtej Sarna.

The inspection covered the venue, hotel and other areas where the Indian cricket team will be present. Specific recommendations for tightening security arrangements and these have been accepted by the Bangladesh authorities. A letter, apparently from a previously unknown Islamic group Harkat-ul-Jihad, arrived at the Indian embassy last Thursday. It threatened to kill the players in revenge for the riots in Gujarat in India three years ago which left many Muslims dead. However, the Bangladesh high

commission said the threats were a hoax. Explosions and other attacks are not uncommon in Bangladesh - more than 20 people were killed when grenades were thrown at an opposition rally in August. There have also been blasts at cinemas. Few arrests have been made, but Islamic extremists are widely blamed. Indian cricketers have visited Bangladesh several times, most recently in 2003 for a triangular tournament that also involved South Africa.

060.txt

Kumble breaks Kapil's record

First Test, Dhaka, day one (stumps): Bangladesh 184 all out v India

Kumble overtook the mark set by Kapil Dev when he had Mohammad Rafique lbw. And he followed up with a wicket next ball before Bangladesh were bowled out for 184 in 58 overs in Dhaka. After the first session was lost to rain, Irfan Pathan took five wickets to reduce the hosts to 106-7 before Mohammad Ashraful dug in. Ashraful ended unbeaten on 60, having hit six fours and faced 135 balls. Kumble had a chance of a hat-trick after removing Tapash Baisya via a catch at first slip but Mashrafe Mortaza safely defended the fifth ball of his 12th over. But a run out ended the innings not long afterwards.

India did not get chance to begin their reply as openers Virender Sehwag and Gautam Gambhir were immediately offered the light on stepping to the wicket. India won the toss and Pathan soon got stuck into the top order, dismissing Javed Omar lbw in his second over with one that nipped back. Nafis Iqbal and Rajin Saleh were also ajudged lbw by umpire Jeremy Lloyds off consecutive balls in Pathan's fifth over. Captain Habibul Bashar then pulled Zaheer Khan straight to square leg and when the same bowler had Khaled Mashud caught behind, Bangladesh were 50-5 after just 16 overs. Ashraful, largely in partnership with Rafique (47), did his best to build a recovery but India will expect to amass a huge lead on Saturday. Kumble is now fifth in the all-time list. Aged 34, he may still be able to reach the 500-mark, passed by only three men. Fellow leg-spinner Shane Warne tops the list with 552 wickets.

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

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.