DS8003 Final Exam

Xin Zhao

500510757

Summary:

- This search query program will take a certain words separated by "," and retrieve the top N matching document with the word being search, how many word occurrence in the document, idf score, and tf-idf score
- This program is developed with PySpark, SparkSQL and Dataframes
- This program in not case sensitive, therefore user can enter the word in both uppercase and lowercase
- This program trims all unnecessary white space of user input

The reason why I chose the tools Pyspark, sparkSQL and dataframes is because in Pyspark I can use high level API's to perform tasks compare to complex mapper reducers. I also used dataframes instead of RDD is beucase dataframes have better space and speed efficiency than RDD format.

The tf-idf inverted index is built on the dataset is used from BBC sports website. The dataset consists of a total of 265 text files which contains football sports news[1].

Program Insruction:

In order to run the python file , the program takes 2 parameters from the user. The first parameter starts with "-w", the program will take in the words that need to be search in the documents, user "," to separate between words. The second parameter starts with "-n" which will take in how many result to output according to the tf-idf score.

Functions Description:

tf_idf_dataframe(sc,sqlContext):

This function will create a store the dataset in RDD format first, it counts the number of documents in the dataset, and clean the data by converting all letters to lower case and split each word by space. Then the dataset is converted to a dataframe. The program will calculate term frequency of each word, find distinct words in how many documents and calculate idf. In order to calculate tf_idf, tf dataframe is joined with idf dataframe, then tf_idf is calculated by multiplying tf with idf. In the end the tf_idf is sorted with descending order. The tf_idf index is saved as csv format

search_engine(sc,tf_idf_index,query,n):

This function takes in the tf idf index created in tf_idf_dataframe, the words need to be searched and the number of result will be output. First the program cleans the query by changing all words to lower case and split each word by space. Then find out the how many words in the query. The query is first taken as RDD format, then it is converted to dataframe format for further tasks. The dataframe is trimmed by cleaning out unnecessary white spaces. The program joins the tf_idf index created by tf_idf_dataframe function

[1] D. Greene and P. Cunningham. "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering", Proc. ICML 2006.

with the query word and find out the tf_idf of each word. Then each word's tf_idf is combined to see the relevance of all words in each document. Then the final tf_idf score is calculated by using the combined socre times number of matched words and divide by total number of words. The final result is sorted with descending order and output n number of result

Data preparation:

hadoop fs –ls /user/root/xin_zhao/final_exam/football

sample screen shoot of 265 text files in football folder

```
[root@sandbox-hdp final_exam]# hadoop fs -ls /user/root/xin_zhao/final_exam/football
Found 265 items
-rw-r--r--   1 root hdfs       3227 2018-11-23 01:25 /user/root/xin_zhao/final_exam/football/001.txt
-rw-r--r--   1 root hdfs       1455 2018-11-23 01:25 /user/root/xin_zhao/final_exam/football/002.txt
-rw-r--r--   1 root hdfs       2424 2018-11-23 01:25 /user/root/xin_zhao/final_exam/football/003.txt
-rw-r--r--   1 root hdfs        784 2018-11-23 01:25 /user/root/xin_zhao/final_exam/football/004.txt
-rw-r--r--   1 root hdfs       1374 2018-11-23 01:25 /user/root/xin_zhao/final_exam/football/005.txt
-rw-r--r--   1 root hdfs       2039 2018-11-23 01:25 /user/root/xin_zhao/final_exam/football/006.txt
-rw-r--r--   1 root hdfs       3639 2018-11-23 01:25 /user/root/xin_zhao/final_exam/football/007.txt
-rw-r--r--   1 root hdfs       2017 2018-11-23 01:25 /user/root/xin_zhao/final_exam/football/008.txt
-rw-r--r--   1 root hdfs        970 2018-11-23 01:25 /user/root/xin_zhao/final_exam/football/009.txt
-rw-r--r--   1 root hdfs       3688 2018-11-23 01:25 /user/root/xin_zhao/final_exam/football/010.txt
-rw-r--r--   1 root hdfs       3682 2018-11-23 01:25 /user/root/xin_zhao/final_exam/football/011.txt
-rw-r--r--   1 root hdfs       3413 2018-11-23 01:25 /user/root/xin_zhao/final_exam/football/012.txt
-rw-r--r--   1 root hdfs       2611 2018-11-23 01:25 /user/root/xin_zhao/final_exam/football/013.txt
-rw-r--r--   1 root hdfs       2966 2018-11-23 01:25 /user/root/xin_zhao/final_exam/football/014.txt
-rw-r--r--   1 root hdfs       2640 2018-11-23 01:25 /user/root/xin_zhao/final_exam/football/015.txt
-rw-r--r--   1 root hdfs       1218 2018-11-23 01:25 /user/root/xin_zhao/final_exam/football/016.txt
-rw-r--r--   1 root hdfs       2589 2018-11-23 01:25 /user/root/xin_zhao/final_exam/football/017.txt
-rw-r--r--   1 root hdfs       7324 2018-11-23 01:25 /user/root/xin_zhao/final_exam/football/018.txt
-rw-r--r--   1 root hdfs       1407 2018-11-23 01:25 /user/root/xin_zhao/final_exam/football/019.txt
-rw-r--r--   1 root hdfs       1676 2018-11-23 01:25 /user/root/xin_zhao/final_exam/football/020.txt
-rw-r--r--   1 root hdfs       3565 2018-11-23 01:25 /user/root/xin_zhao/final_exam/football/021.txt
-rw-r--r--   1 root hdfs        832 2018-11-23 01:25 /user/root/xin_zhao/final_exam/football/022.txt
-rw-r--r--   1 root hdfs       1201 2018-11-23 01:25 /user/root/xin_zhao/final_exam/football/023.txt
-rw-r--r--   1 root hdfs       6093 2018-11-23 01:25 /user/root/xin_zhao/final_exam/football/024.txt
-rw-r--r--   1 root hdfs       2749 2018-11-23 01:25 /user/root/xin_zhao/final_exam/football/025.txt
-rw-r--r--   1 root hdfs       1112 2018-11-23 01:25 /user/root/xin_zhao/final_exam/football/026.txt
-rw-r--r--   1 root hdfs       1359 2018-11-23 01:26 /user/root/xin_zhao/final_exam/football/248.txt
-rw-r--r--   1 root hdfs       2983 2018-11-23 01:26 /user/root/xin_zhao/final_exam/football/249.txt
-rw-r--r--   1 root hdfs       1308 2018-11-23 01:26 /user/root/xin_zhao/final_exam/football/250.txt
-rw-r--r--   1 root hdfs       2870 2018-11-23 01:26 /user/root/xin_zhao/final_exam/football/251.txt
-rw-r--r--   1 root hdfs       1500 2018-11-23 01:26 /user/root/xin_zhao/final_exam/football/252.txt
-rw-r--r--   1 root hdfs       3332 2018-11-23 01:26 /user/root/xin_zhao/final_exam/football/253.txt
-rw-r--r--   1 root hdfs       1062 2018-11-23 01:26 /user/root/xin_zhao/final_exam/football/254.txt
-rw-r--r--   1 root hdfs       1643 2018-11-23 01:26 /user/root/xin_zhao/final_exam/football/255.txt
-rw-r--r--   1 root hdfs       1602 2018-11-23 01:26 /user/root/xin_zhao/final_exam/football/256.txt
-rw-r--r--   1 root hdfs        877 2018-11-23 01:26 /user/root/xin_zhao/final_exam/football/257.txt
-rw-r--r--   1 root hdfs       2158 2018-11-23 01:26 /user/root/xin_zhao/final_exam/football/258.txt
-rw-r--r--   1 root hdfs       2779 2018-11-23 01:26 /user/root/xin_zhao/final_exam/football/259.txt
-rw-r--r--   1 root hdfs       3602 2018-11-23 01:26 /user/root/xin_zhao/final_exam/football/260.txt
-rw-r--r--   1 root hdfs       1232 2018-11-23 01:26 /user/root/xin_zhao/final_exam/football/261.txt
-rw-r--r--   1 root hdfs       1430 2018-11-23 01:26 /user/root/xin_zhao/final_exam/football/262.txt
-rw-r--r--   1 root hdfs       1178 2018-11-23 01:26 /user/root/xin_zhao/final_exam/football/263.txt
-rw-r--r--   1 root hdfs       1069 2018-11-23 01:26 /user/root/xin_zhao/final_exam/football/264.txt
-rw-r--r--   1 root hdfs        732 2018-11-23 01:26 /user/root/xin_zhao/final_exam/football/265.txt
```

In order to illustrate my work line by line in pyspark I created a test sample which contains the first 10 files in the football files

hadoop fs –ls /user/root/xin_zhao/final_exam/test/test

```
[root@sandbox-hdp final_exam]# hadoop fs -ls /user/root/xin_zhao/final_exam/test/test
Found 10 items
-rw-r--r--   1 root hdfs       3227 2018-11-23 04:13 /user/root/xin_zhao/final_exam/test/test/001.txt
-rw-r--r--   1 root hdfs       1455 2018-11-23 04:13 /user/root/xin_zhao/final_exam/test/test/002.txt
-rw-r--r--   1 root hdfs       2424 2018-11-23 04:13 /user/root/xin_zhao/final_exam/test/test/003.txt
-rw-r--r--   1 root hdfs        784 2018-11-23 04:13 /user/root/xin_zhao/final_exam/test/test/004.txt
-rw-r--r--   1 root hdfs       1374 2018-11-23 04:13 /user/root/xin_zhao/final_exam/test/test/005.txt
-rw-r--r--   1 root hdfs       2039 2018-11-23 04:13 /user/root/xin_zhao/final_exam/test/test/006.txt
-rw-r--r--   1 root hdfs       3639 2018-11-23 04:13 /user/root/xin_zhao/final_exam/test/test/007.txt
-rw-r--r--   1 root hdfs       2017 2018-11-23 04:13 /user/root/xin_zhao/final_exam/test/test/008.txt
-rw-r--r--   1 root hdfs        970 2018-11-23 04:13 /user/root/xin_zhao/final_exam/test/test/009.txt
-rw-r--r--   1 root hdfs       3688 2018-11-23 04:13 /user/root/xin_zhao/final_exam/test/test/010.txt
```

Python Code

```
from pyspark import SparkConf, SparkContext
from pyspark.sql import SQLContext
from pyspark.sql.functions import *
from pyspark.sql import Row

import sys
import argparse

# custom argument input
parser = argparse.ArgumentParser(description = "DS8003 Final Exam" )
# starts with "-w", the program takes in user's search word
parser.add_argument('-w', '--word',type = str,  help = "Words need to searched, use , to separate")
# starts with "-n", the program takes in top n tf-idf score
parser.add_argument('-n', '--number',type = int , help = "top n document")
# custom argument input
args = parser.parse_args()


def tf_idf_dataframe(sc,sqlContext):
    # sores all documents from the document path
    data = sc.wholeTextFiles("/user/root/xin_zhao/final_exam/football")
    # count the number of document
    num_docs = data.count()
    # change all text to lower case and spilt each word by space
    temp = data.map(lambda docs:(docs[0], docs[1].lower().split(" ")))
    # create dataframe for each document with two columns file and text
    df = sqlContext.createDataFrame(temp,['file','text'])
    # each word in each document (tf)
    tf = df.withColumn('text', explode(col('text'))).groupBy('file', 'text').count()
    # find each distinct word in all documents
    wc = df.withColumn("text",explode(col('text')))
    # find each distinct word in how many documents
    distinct_word = wc.distinct().groupBy('text').count()
```

```
    # calculate idf by using the formula
    idf = distinct_word.withColumn('idf', log10(num_docs/ (col('count')))).drop('count')
    # join tf dataframe with idf dataframe to calculate tf_idf
    new_dataframe=tf.join(idf, tf.text==idf.text,'outer').drop(idf.text)
    # calculate tf_idf by multiplying tf with idf
    tf_idf = new_dataframe.withColumn('tf_idf', (col('idf')* (col('count'))))
    # order tf_idf scores with descending order
    sorted_tf_idf=tf_idf.orderBy('tf_idf',ascending=False)
    # save tf_idf as csv format
    sorted_tf_idf.write.csv("/user/root/xin_zhao/final_exam/tf_idf")
    return(sorted_tf_idf)
```

```
def search_engine(sc,tf_idf_index,query,n):
    # change query words to lower case and split each word with space
    words = query.lower().split(",")
    # get the length of the word
    query_length = len(words)
    # parallelize words into RDD format
    query_rdd = sc.parallelize(words).map(lambda x: Row(x))
    #covert query to dataframe format
    query_dataframe = sqlContext.createDataFrame(query_rdd,['text'])
    # trim query without unnecessary white space
    trim_dataframe = query_dataframe.withColumn('text',trim(col('text')))
    # search query within the documents (tf)
    search_dataframe= tf_idf_index.join(trim_dataframe, tf_idf_index.text==trim_dataframe.text,'inner').drop(trim_dataframe.text)
    # combine tf_idf scores of each query word
    sum_tf_idf = search_dataframe.groupBy("file").agg({"*":"count", "tf_idf":"sum"})
    # change column names
    sum_tf_idf = sum_tf_idf.withColumnRenamed("count(1)", "num_match_word").withColumnRenamed("sum(tf_idf)", "sum_score")
    #  calculate final tf_idf score for the document by using sum_score times number of matched words and divide by total number of query words
    final_result = sum_tf_idf.select(sum_tf_idf.file, (sum_tf_idf.sum_score * sum_tf_idf.num_match_word) / query_length)
    # change column names
    final_result = final_result.withColumnRenamed("((sum_score * num_match_word) / " + str(query_length) + ")", "tf_idf")
    # sort tf_idf scores with descending order, and output n result
    final_result = final_result.orderBy("tf_idf", ascending=False)
    final_result.show(n,False)


def main(sc):
    # store user input as query word
    search_word = args.word
    # store user input as top n document
    topn = args.number
    # build tf_idf index
    tf_idf = tf_idf_dataframe(sc,sqlContext)
    # search query and return tf_idf
    search_engine(sc,tf_idf,search_word,topn)

if __name__  == "__main__":
    conf = SparkConf().setAppName("MyApp")
    sc = SparkContext(conf = conf)
    sqlContext = SQLContext(sc)
    main(sc)
    sc.stop()
```

*Original python file is submitted with this document

Code Description

tf_idf_dataframe(sc,sqlContext):

sores all documents from the document path

data = sc.wholeTextFiles("/user/root/xin_zhao/final_exam/test/test")

count the number of document

num_docs = data.count()

print(num_docs)

```
>>> data = sc.wholeTextFiles("/user/root/xin_zhao/final_exam/test/test")
>>> num_docs = data.count()
>>> print(num_docs)
10
```

change all text to lower case and spilt each word by space

temp = data.map(lambda docs:(docs[0], docs[1].lower().split(" ")))

create dataframe for each document with two columns file and text

df = sqlContext.createDataFrame(temp,['file','text'])

df.show()

```
>>> from pyspark import SparkConf, SparkContext
>>> from pyspark.sql import SQLContext
>>> from pyspark.sql.functions import *
>>> from pyspark.sql import Row
>>> sqlContext = SQLContext(sc)
>>> temp = data.map(lambda docs:(docs[0], docs[1].lower().split(" ")))
>>> df = sqlContext.createDataFrame(temp,['file','text'])
>>> df.show()
+--------------------+--------------------+
|                file|                text|
+--------------------+--------------------+
|hdfs://sandbox-hd...|[man, utd, stroll...|
|hdfs://sandbox-hd...|[van, nistelrooy,...|
|hdfs://sandbox-hd...|[moyes, u-turn, o...|
|hdfs://sandbox-hd...|[ronaldo, conside...|
|hdfs://sandbox-hd...|[smith, keen, on,...|
|hdfs://sandbox-hd...|[mido, makes, thi...|
|hdfs://sandbox-hd...|[man, city, 0-2, ...|
|hdfs://sandbox-hd...|[gerrard, plays, ...|
|hdfs://sandbox-hd...|[duff, ruled, out...|
|hdfs://sandbox-hd...|[chelsea, clinch,...|
+--------------------+--------------------+
```

each word in each document (tf)

tf = df.withColumn('text', explode(col('text'))).groupBy('file', 'text').count()

tf.show()

```
>>> tf = df.withColumn('text', explode(col('text'))).groupBy('file', 'text').count()
>>> tf.show()
+--------------------+----------+-----+
|                file|      text|count|
+--------------------+----------+-----+
|hdfs://sandbox-hd...|free-kick.|    1|
|hdfs://sandbox-hd...|     doing|    1|
|hdfs://sandbox-hd...|      made|    1|
|hdfs://sandbox-hd...|        my|    1|
|hdfs://sandbox-hd...|  problems|    1|
|hdfs://sandbox-hd...|      axed|    1|
|hdfs://sandbox-hd...|        he|    6|
|hdfs://sandbox-hd...|       ben|    1|
|hdfs://sandbox-hd...|     about|    2|
|hdfs://sandbox-hd...|       for|    3|
|hdfs://sandbox-hd...|        an|    1|
|hdfs://sandbox-hd...|  previous|    1|
|hdfs://sandbox-hd...|     saved|    1|
|hdfs://sandbox-hd...|   hapless|    1|
|hdfs://sandbox-hd...|    chance|    1|
|hdfs://sandbox-hd...|   rethink|    1|
|hdfs://sandbox-hd...|      this|    1|
|hdfs://sandbox-hd...|     groin|    1|
|hdfs://sandbox-hd...| tottenham|    1|
|hdfs://sandbox-hd...|   managed|    1|
+--------------------+----------+-----+
only showing top 20 rows
```

find each distinct word in all documents

wc = df.withColumn("text",explode(col('text')))

wc.show()

```
>>> wc = df.withColumn("text",explode(col('text')))
>>> wc.show()
+--------------------+----------+
|                file|      text|
+--------------------+----------+
|hdfs://sandbox-hd...|       man|
|hdfs://sandbox-hd...|       utd|
|hdfs://sandbox-hd...|    stroll|
|hdfs://sandbox-hd...|        to|
|hdfs://sandbox-hd...|       cup|
|hdfs://sandbox-hd...|win

wayne|
|hdfs://sandbox-hd...|    rooney|
|hdfs://sandbox-hd...|      made|
|hdfs://sandbox-hd...|         a|
|hdfs://sandbox-hd...|   winning|
|hdfs://sandbox-hd...|    return|
|hdfs://sandbox-hd...|        to|
|hdfs://sandbox-hd...|   everton|
|hdfs://sandbox-hd...|        as|
|hdfs://sandbox-hd...|manchester|
|hdfs://sandbox-hd...|    united|
|hdfs://sandbox-hd...|   cruised|
|hdfs://sandbox-hd...|      into|
|hdfs://sandbox-hd...|       the|
|hdfs://sandbox-hd...|        fa|
+--------------------+----------+
only showing top 20 rows
```

find each distinct word in how many documents

distinct_word = wc.distinct().groupBy('text').count()

distinct_word.show()

```
>>> distinct_word = wc.distinct().groupBy('text').count()
>>> distinct_word.show()
+------------+-----+
|        text|count|
+------------+-----+
|       nunez|    1|
|       still|    5|
|        some|    3|
|     persist|    1|
|   connected|    1|
|         ...|    1|
|portsmouth.
|    1|
|  television|    1|
|    received|    1|
|       equal|    1|
|  criticised|    1|
|     jarosik|    1|
|    fletcher|    1|
|   tomorrow.|    1|
|   naysmith.|    1|
|    portugal|    1|
| apologised.|    1|
| striker.

-|    1|
|   character|    1|
|     player.|    1|
+------------+-----+
only showing top 20 rows
```

calculate idf by using the formula

idf = distinct_word.withColumn('idf', log10(num_docs/ (col('count')))).drop('count')

idf.show()

```
>>> idf = distinct_word.withColumn('idf', log10(num_docs/ (col('count')))).drop('count')
>>> idf.show()
+------------+-----------------+
|        text|              idf|
+------------+-----------------+
|       nunez|              1.0|
|       still|0.3010299956639812|
|        some|0.5228787452803376|
|     persist|              1.0|
|   connected|              1.0|
|         ...|              1.0|
|portsmouth.
|             1.0|
|  television|              1.0|
|    received|              1.0|
|       equal|              1.0|
|  criticised|              1.0|
|     jarosik|              1.0|
|    fletcher|              1.0|
|   tomorrow.|              1.0|
|   naysmith.|              1.0|
|    portugal|              1.0|
| apologised.|              1.0|
| striker.

-|             1.0|
|   character|              1.0|
|     player.|              1.0|
+------------+-----------------+
only showing top 20 rows
```

join tf dataframe with idf dataframe to calculate tf_idf

new_dataframe=tf.join(idf, tf.text==idf.text,'outer').drop(idf.text)

new_dataframe.show()

```
>>> new_dataframe=tf.join(idf, tf.text==idf.text,'outer').drop(idf.text)
>>> new_dataframe.show()
+--------------------+------------+-----+-----------------+
|                file|        text|count|              idf|
+--------------------+------------+-----+-----------------+
|hdfs://sandbox-hd...|         ...|    1|              1.0|
|hdfs://sandbox-hd...|   connected|    1|              1.0|
|hdfs://sandbox-hd...|       nunez|    2|              1.0|
|hdfs://sandbox-hd...|     persist|    1|              1.0|
|hdfs://sandbox-hd...|        some|    2|0.5228787452803376|
|hdfs://sandbox-hd...|        some|    1|0.5228787452803376|
|hdfs://sandbox-hd...|        some|    1|0.5228787452803376|
|hdfs://sandbox-hd...|       still|    1|0.3010299956639812|
|hdfs://sandbox-hd...|       still|    2|0.3010299956639812|
|hdfs://sandbox-hd...|       still|    2|0.3010299956639812|
|hdfs://sandbox-hd...|       still|    1|0.3010299956639812|
|hdfs://sandbox-hd...|       still|    3|0.3010299956639812|
|hdfs://sandbox-hd...|portsmouth.
|    1|              1.0|
|hdfs://sandbox-hd...|    received|    1|              1.0|
|hdfs://sandbox-hd...|  television|    1|              1.0|
|hdfs://sandbox-hd...|  criticised|    1|              1.0|
|hdfs://sandbox-hd...|       equal|    1|              1.0|
|hdfs://sandbox-hd...|    fletcher|    1|              1.0|
|hdfs://sandbox-hd...|     jarosik|    1|              1.0|
|hdfs://sandbox-hd...|   naysmith.|    1|              1.0|
+--------------------+------------+-----+-----------------+
only showing top 20 rows
```

calculate tf_idf by multiplying tf with idf

tf_idf = new_dataframe.withColumn('tf_idf', (col('idf')* (col('count'))))

tf_idf.show()

```
>>> tf_idf = new_dataframe.withColumn('tf_idf', (col('idf')* (col('count'))))
>>> tf_idf.show()
+--------------------+------------+-----+------------------+------------------+
|                file|        text|count|               idf|            tf_idf|
+--------------------+------------+-----+------------------+------------------+
|hdfs://sandbox-hd...|         ...|    1|               1.0|               1.0|
|hdfs://sandbox-hd...|   connected|    1|               1.0|               1.0|
|hdfs://sandbox-hd...|       nunez|    2|               1.0|               2.0|
|hdfs://sandbox-hd...|     persist|    1|               1.0|               1.0|
|hdfs://sandbox-hd...|        some|    2|0.5228787452803376|1.0457574905606752|
|hdfs://sandbox-hd...|        some|    1|0.5228787452803376|0.5228787452803376|
|hdfs://sandbox-hd...|        some|    1|0.5228787452803376|0.5228787452803376|
|hdfs://sandbox-hd...|       still|    1|0.3010299956639812|0.3010299956639812|
|hdfs://sandbox-hd...|       still|    2|0.3010299956639812|0.6020599913279624|
|hdfs://sandbox-hd...|       still|    2|0.3010299956639812|0.6020599913279624|
|hdfs://sandbox-hd...|       still|    1|0.3010299956639812|0.3010299956639812|
|hdfs://sandbox-hd...|       still|    3|0.3010299956639812|0.9030899869919435|
|hdfs://sandbox-hd...|portsmouth.
|    1|              1.0|              1.0|
|hdfs://sandbox-hd...|    received|    1|               1.0|               1.0|
|hdfs://sandbox-hd...|  television|    1|               1.0|               1.0|
|hdfs://sandbox-hd...|  criticised|    1|               1.0|               1.0|
|hdfs://sandbox-hd...|       equal|    1|               1.0|               1.0|
|hdfs://sandbox-hd...|    fletcher|    1|               1.0|               1.0|
|hdfs://sandbox-hd...|     jarosik|    1|               1.0|               1.0|
|hdfs://sandbox-hd...|   naysmith.|    1|               1.0|               1.0|
+--------------------+------------+-----+------------------+------------------+
only showing top 20 rows
```

order tf_idf scores with descending order

sorted_tf_idf=tf_idf.orderBy('tf_idf',ascending=False)

sorted_tf_idf.show()

```
>>> sorted_tf_idf=tf_idf.orderBy('tf_idf',ascending=False)
>>> sorted_tf_idf.show()
+--------------------+---------------+-----+------------------+------------------+
|                file|           text|count|               idf|            tf_idf|
+--------------------+---------------+-----+------------------+------------------+
|hdfs://sandbox-hd...|           mido|    8|               1.0|               8.0|
|hdfs://sandbox-hd...|         martyn|    7|               1.0|               7.0|
|hdfs://sandbox-hd...|       national|    5|               1.0|               5.0|
|hdfs://sandbox-hd...|          moyes|    5|               1.0|               5.0|
|hdfs://sandbox-hd...|            van|    5|               1.0|               5.0|
|hdfs://sandbox-hd...|        fortune|    4|               1.0|               4.0|
|hdfs://sandbox-hd...|         united|   10|3.9794000086720376| 3.979400086720376|
|hdfs://sandbox-hd...|      liverpool|    7|0.5228787452803376|3.6601512169623636|
|hdfs://sandbox-hd...|      champions|    6|0.5228787452803376|3.1372724716820257|
|hdfs://sandbox-hd...|wright-phillips|    3|               1.0|               3.0|
|hdfs://sandbox-hd...|        doubled|    3|               1.0|               3.0|
|hdfs://sandbox-hd...|        shalaby|    3|               1.0|               3.0|
|hdfs://sandbox-hd...|          dunne|    3|               1.0|               3.0|
|hdfs://sandbox-hd...|      initially|    3|               1.0|               3.0|
|hdfs://sandbox-hd...|     nistelrooy|    3|               1.0|               3.0|
|hdfs://sandbox-hd...|      everton's|    3|               1.0|               3.0|
|hdfs://sandbox-hd...|          press|    3|               1.0|               3.0|
|hdfs://sandbox-hd...|       ferguson|    3|               1.0|               3.0|
|hdfs://sandbox-hd...|        despite|    3|               1.0|               3.0|
|hdfs://sandbox-hd...|          again|    3|               1.0|               3.0|
+--------------------+---------------+-----+------------------+------------------+
only showing top 20 rows
```

search_engine(sc,tf_idf_index,query,n):

For illustration purpose in pyspark I put “United , Manchester” as the user query input, and 5 as the top 5 result to be output

query = “United , Manchester”

n=5

change query words to lower case and split each word with space

words = query.lower().split(",")

print(words)

```
>>> query = "United  ,Manchester"
>>> n = 5
>>> words = query.lower().split(",")
>>> print(words)
['united  ', 'manchester']
```

parallelize words into RDD format

query_rdd = sc.parallelize(words).map(lambda x: Row(x))

#covert query to dataframe format

query_dataframe = sqlContext.createDataFrame(query_rdd,['text'])

query_dataframe.show()

```
>>> query_rdd = sc.parallelize(words).map(lambda x: Row(x))
>>> query_dataframe = sqlContext.createDataFrame(query_rdd,['text'])
>>> query_dataframe.show()
+-----------+
|      text|
+-----------+
|  united  |
|manchester|
+-----------+
```

#search query within the documents (tf)

search_dataframe= sorted_tf_idf.join(trim_dataframe, sorted_tf_idf.text==trim_dataframe.text,'inner').drop(trim_dataframe.text)

search_dataframe.show()

```
>>> search_dataframe= sorted_tf_idf.join(trim_dataframe, sorted_tf_idf.text==trim_dataframe.text,'inner').drop(trim_datafr
ame.text)
>>> search_dataframe.show()
+--------------------+----------+-----+-----------------+-----------------+
|              file|      text|count|              idf|           tf_idf|
+--------------------+----------+-----+-----------------+-----------------+
|hdfs://sandbox-hd...|    united|   10|0.3979400086720376| 3.979400086720376|
|hdfs://sandbox-hd...|    united|    7|0.3979400086720376|2.7855800607042633|
|hdfs://sandbox-hd...|    united|    4|0.3979400086720376|1.5917600346881504|
|hdfs://sandbox-hd...|    united|    3|0.3979400086720376| 1.193820026016113|
|hdfs://sandbox-hd...|manchester|    3|0.5228787452803376|1.5686362358410129|
|hdfs://sandbox-hd...|manchester|    3|0.5228787452803376|1.5686362358410129|
|hdfs://sandbox-hd...|manchester|    2|0.5228787452803376|1.0457574905606752|
+--------------------+----------+-----+-----------------+-----------------+
```

combine tf_idf scores of each query word

sum_tf_idf = search_dataframe.groupBy("file").agg({"*":"count", "tf_idf":"sum"})

sum_tf_idf.show()

```
>>> sum_tf_idf = search_dataframe.groupBy("file").agg({"*":"count", "tf_idf":"sum"})
>>> sum_tf_idf.show()
+--------------------+------------------+--------+
|                file|       sum(tf_idf)|count(1)|
+--------------------+------------------+--------+
|hdfs://sandbox-hd...| 4.354216296545276|       2|
|hdfs://sandbox-hd...| 5.548036322561389|       2|
|hdfs://sandbox-hd...| 1.193820026016113|       1|
|hdfs://sandbox-hd...|2.6375175252488257|       2|
+--------------------+------------------+--------+
```

change column names

sum_tf_idf = sum_tf_idf.withColumnRenamed("count(1)", "num_match_word").withColumnRenamed("sum(tf_idf)", "sum_score")

```
>>> sum_tf_idf = sum_tf_idf.withColumnRenamed("count(1)", "num_match_word").withColumnRenamed("sum(tf_idf)", "sum_score")
>>> sum_tf_idf.show()
+--------------------+------------------+-------------+
|                file|         sum_score|num_match_word|
+--------------------+------------------+-------------+
|hdfs://sandbox-hd...| 4.354216296545276|            2|
|hdfs://sandbox-hd...| 5.548036322561389|            2|
|hdfs://sandbox-hd...| 1.193820026016113|            1|
|hdfs://sandbox-hd...|2.6375175252488257|            2|
+--------------------+------------------+-------------+
```

get the length of the word

query_length = len(words)

print(query_length)

```
>>> query_length = len(words)
>>> print(query_length)
2
```

calculate final tf_idf score for the document by using sum_score times number of matched words and divide by total number of words

final_result = sum_tf_idf.select(sum_tf_idf.file, (sum_tf_idf.sum_score * sum_tf_idf.num_match_word) / query_length)

final_result.show()

```
>>> final_result = sum_tf_idf.select(sum_tf_idf.file, (sum_tf_idf.sum_score * sum_tf_idf.num_match_word) / query_length)
>>> final_result.show()
+--------------------+----------------------------------+
|                file|((sum_score * num_match_word) / 2)|
+--------------------+----------------------------------+
|hdfs://sandbox-hd...|                 4.354216296545276|
|hdfs://sandbox-hd...|                 5.548036322561389|
|hdfs://sandbox-hd...|                0.5969100130080565|
|hdfs://sandbox-hd...|                2.6375175252488257|
+--------------------+----------------------------------+
```

change column names

final_result = final_result.withColumnRenamed("((sum_score * num_match_word) / " + str(query_length) + ")", "tf_idf")

final_result.show()

```
>>> final_result = final_result.withColumnRenamed("((sum_score * num_match_word) / " + str(query_length) + ")", "tf_idf")
>>> final_result.show()
+--------------------+------------------+
|                file|            tf_idf|
+--------------------+------------------+
|hdfs://sandbox-hd...| 4.354216296545276|
|hdfs://sandbox-hd...| 5.548036322561389|
|hdfs://sandbox-hd...|0.5969100130080565|
|hdfs://sandbox-hd...|2.6375175252488257|
+--------------------+------------------+
```

sort tf_idf scores with descending order, and output n result

final_result = final_result.orderBy("tf_idf", ascending=False)

final_result.show(n,False)

```
>>> final_result = final_result.orderBy("tf_idf", ascending=False)
>>> final_result.show(n,False)
+-----------------------------------------------------------------------------------------+------------------+
|file                                                                                     |tf_idf            |
+-----------------------------------------------------------------------------------------+------------------+
|hdfs://sandbox-hdp.hortonworks.com:8020/user/root/xin_zhao/final_exam/test/test/007.txt|5.548036322561389 |
|hdfs://sandbox-hdp.hortonworks.com:8020/user/root/xin_zhao/final_exam/test/test/001.txt|4.354216296545276 |
|hdfs://sandbox-hdp.hortonworks.com:8020/user/root/xin_zhao/final_exam/test/test/002.txt|2.6375175252488257|
|hdfs://sandbox-hdp.hortonworks.com:8020/user/root/xin_zhao/final_exam/test/test/004.txt|0.5969100130080565|
+-----------------------------------------------------------------------------------------+------------------+
```

Result Section:

Search query: win, Manchester

N: 1

*The top 5 text with highlighted query are listed together after all queries done.

spark-submit --master yarn-client --executor-memory 512m --num-executors 5 --executor-cores 1 --driver-memory 512m finalexam_final.py -w win,Manchester -n 1

```
[root@sandbox-hdp final_exam]# spark-submit --master yarn-client --executor-memory 512m --num-executors 5 --executor-cores
 1 --driver-memory 512m finalexam_final.py -w win,Manchester -n 1
```

```
18/11/30 20:56:59 INFO CodeGenerator: Code generated in 53.44787 ms
+-----------------------------------------------------------------------------------------+------------------+
|file                                                                                     |tf_idf            |
+-----------------------------------------------------------------------------------------+------------------+
|hdfs://sandbox-hdp.hortonworks.com:8020/user/root/xin_zhao/final_exam/football/207.txt|4.697769812453673|
+-----------------------------------------------------------------------------------------+------------------+
only showing top 1 row
```

Search query: win, Manchester

N: 3

spark-submit --master yarn-client --executor-memory 512m --num-executors 5 --executor-cores 1 --driver-memory 512m finalexam_final.py -w win,Manchester -n 3

```
[root@sandbox-hdp final_exam]# spark-submit --master yarn-client --executor-memory 512m --num-executors 5 --executor-cores
1 --driver-memory 512m finalexam_final.py -w win,Manchester -n 3
```

```
18/11/30 21:12:21 INFO CodeGenerator: Code generated in 27.18619 ms
+-----------------------------------------------------------------------------------------+------------------+
|file                                                                                     |tf_idf            |
+-----------------------------------------------------------------------------------------+------------------+
|hdfs://sandbox-hdp.hortonworks.com:8020/user/root/xin_zhao/final_exam/football/207.txt|4.697769812453673 |
|hdfs://sandbox-hdp.hortonworks.com:8020/user/root/xin_zhao/final_exam/football/250.txt|3.7881938513565006|
|hdfs://sandbox-hdp.hortonworks.com:8020/user/root/xin_zhao/final_exam/football/209.txt|2.634878668835703 |
+-----------------------------------------------------------------------------------------+------------------+
only showing top 3 rows
```

Search query: win, Manchester

N: 5

spark-submit --master yarn-client --executor-memory 512m --num-executors 5 --executor-cores 1 --driver-memory 512m finalexam_final.py -w win,Manchester -n 5

```
[root@sandbox-hdp final_exam]# spark-submit --master yarn-client --executor-memory 512m --num-executors 5 --executor-cores
 1 --driver-memory 512m finalexam_final.py -w win,Manchester -n 5
```

```
18/11/30 21:04:50 INFO CodeGenerator: Code generated in 36.223043 ms
+--------------------------------------------------------------------------------------+------------------+
|file                                                                                  |tf_idf            |
+--------------------------------------------------------------------------------------+------------------+
|hdfs://sandbox-hdp.hortonworks.com:8020/user/root/xin_zhao/final_exam/football/207.txt|4.697769812453673 |
|hdfs://sandbox-hdp.hortonworks.com:8020/user/root/xin_zhao/final_exam/football/250.txt|3.7881938513565006|
|hdfs://sandbox-hdp.hortonworks.com:8020/user/root/xin_zhao/final_exam/football/209.txt|2.634878668835703 |
|hdfs://sandbox-hdp.hortonworks.com:8020/user/root/xin_zhao/final_exam/football/202.txt|2.5223491902092228|
|hdfs://sandbox-hdp.hortonworks.com:8020/user/root/xin_zhao/final_exam/football/007.txt|2.1754206222444505|
+--------------------------------------------------------------------------------------+------------------+
only showing top 5 rows
```

207.txt

Arsenal players. Arsenal's Lauren was banned for four games, Martin Keown three matches and Ray Parlour one after the incident. Manchester United pair Ryan Giggs and Cristiano Ronaldo were also fined for their part in the fracas. Arsenal's title triumph made up for more disappointment on the European stage, where they lost to Chelsea in the Champions League quarter-finals. Arsene Wenger's side looked on course to finally end their Champions League drought, particularly after winning 5-1 away to Inter Milan.

But in a twist on their domestic domination, Chelsea won 2-1 at Highbury to clinch a 3-2 aggregate victory. Manchester United fared even worse, going out to Porto when Francisco Costina scored a last-minute Old Trafford equaliser. United finished third in the league, but had an FA Cup win for consolation. Porto's success introduced manager Jose Mourinho to a wider audience, and particularly Chelsea. Chelsea manager Claudio Ranieri lived under a cloud of speculation all season, particulary claims he would be replaced by Sven-Goran Eriksson. It intensified when Eriksson was caught holding talks with Chelsea chief executive Peter Kenyon. An embarrassed England coach was then awarded a new four-year contract by the Football Association. Ranieri sealed his fate with a series of bizarre substitutions as Chelsea lost 3-1 in the first leg of the Champions League semi-final in Monaco.

And when he inevitably lost his job, it was Mourinho, who went on to win the Champions League with Porto, who stepped in to take over. Mourinho's reign began well, with Chelsea topping the Premiership at Christmas and joining Arsenal, Manchester United and Liverpool in the next phase of the Champions League. Another manager to lose his job was Liverpool's Gerard Houllier, whose reign ended after six years. Houllier paid the price for finishing fourth and without a trophy last term. He was replaced by Valencia's Rafael Benitez, whose impressive credentials included winning Spain's La Liga and the Uefa Cup last season. Valencia beat Marseille 2-0 to cement Benitez's growing reputation. Manchester United enjoyed their

figure in the drama. Teenager Rooney returned to Everton after Euro 2004 with superstar status assured after stunning performances. And when he refused to sign a new five-year contract, Newcastle United opened the bidding at £20m. A transfer request followed and Manchester United completed a £27m deal just hours before the transfer window closed at the end of August. Rooney confirmed his worth with a hat-trick on his debut in the Champions League against Fenerbahce.

Chelsea, inevitably, were among the big-spenders again, splashing out £24m on Marseille's Didier Drogba and £20m on Porto defender Ricardo Carvalho. Last and by no means least, the other major piece of domestic silverware went to Middlesbrough, who ended 128 barren years by winning the Carling Cup. They beat Bolton 2-1 in the final with goals from Joseph-Desire Job and a penalty from Boudewijn Zenden. The pressures of the top-flight were cruelly illustrated at Southampton. Paul Sturrock was sacked only two games into the new season, one of which was a win against Blackburn, and only 13 games in total at St Mary's. Another manager to pay the price early-season was West Brom's Gary Megson, who was sacked after revealing he would not renew his contract. He guided West Brom back into the Premiership, but his relationship with chairman Jeremy Peace was fragile. Bryan Robson succeeded him as he returned to his old club. In Scotland, Henrik Larsson bid an emotional and successful farewell to Celtic after seven glorious seasons. Celtic won the league and also the Scottish Cup, beating Dunfermline 3-1 in the final, with the Swede scoring twice to take his season's tally to 41. And it took his overall Celtic goals record to 242 goals in 315 appearances before joining Barcelona. Underdogs Livingston claimed the CIS Insurance Cup with a 2-0 win against Hibs, a victory for the romantics.

250.txt

Mourinho sends out warning shot

Chelsea boss Jose Mourinho believes his team's Carling Cup win over Manchester United has shown they have the strength to win the Premiership.

"It was important for us, not because we got into a final, but also the way we played," he said. "The mentality and the strength we showed here was a message we sent. "It is still difficult, we still have to win 11 matches to be champions but we have left a message here that we are really strong." Chelsea gave their manager a win on his 42nd birthday but Mourinho was prepared for a possible loss at Old Trafford. But the Blues are still on course for a four-pronged trophy assault as they are leading the Premiership title race and are in the FA Cup and Champions League.

"We can win four, we can lose four, but it would be normal to win something. To win the four is very, very difficult but it is still possible. "There is a long way to do it but if you could give me the Premiership I would be very happy. "This is just the final though, we have not won the competition and we have to now face another great team in Liverpool." "I was ready to lose the game and leave Old Trafford with a smile just to pass a message of confidence. "But my team would never lose their confidence or mentality just because of a defeat here."

209.txt

Owen displayed his world-class ability once again with another goal after coming on as a substitute in Real's win at Osasuna on Sunday. And therein lies his problem. Michael may have made a rod for his own back with his fantastic performances coming on as substitute. His many coaches at Real Madrid may have decided this is his best role. If that is the case, that is no good to him and he will be coming home sooner rather than later. Michael must hope his performances earn him a regular starting opportunity - and you can rest assured he will take that chance when it comes. I said when he was on the bench earlier this season that Michael's pride would ensure he stuck it out. He would not want to be branded a failure. But he is still on the bench and we are into February - and he could leave now and no-one could call him a flop after the terrific performances he has turned in. If and when he decides to leave, I don't think he will go elsewhere in Europe. I think he will come back to the Premiership. And there would be no shortage of takers. In an ideal world I would obviously love to see him return to Liverpool but you can rest assured that Arsenal and Chelsea would consider what Michael Owen could offer them as well. Owen is a great goalscorer but he is also a very good footballer as well. But one thing that is vital to Michael's game is sharpness - and he will know himself that you miss a vital ingredient when you are not starting games. Also Michael is a fine professional who did not sign for Real Madrid to sit on the sidelines - no matter how many Galacticos are around him. I would expect some serious interest should he fail to win a regular place.

- Chelsea have given the lie to claims that they are enjoying good fortune in their quest for the title - but they will do well just to have a look in their rear-view mirror at Manchester United.

Sir Alex Ferguson played the mind games by suggesting Chelsea would struggle in the north - but they've answered that one. They beat Blackburn, who didn't half put a foot in against them, and then came through a real tough one at Everton, albeit aided by James Beattie's sending off. Chelsea have done brilliantly and you do not keep clean sheets like they do by luck alone. But all Manchester United can do is keep up the pressure - and boy are they doing that. No team has ever been better at chasing down the leaders than Manchester United. Ferguson's team will stay totally committed to the cause. It is still very much Chelsea's to lose and they have shown

202.txt

From Liverpool's point of view, the defeat was a bitter disappointment, but when the disappointment has subsided, they can take heart from a week of encouragement both at home and abroad. Liverpool had an excellent win against Bayer Leverkusen in the Champions League, when they got it down, played and scored goals. And in Sunday's Carling Cup final, they showed real defensive resilience when they were pinned back for long periods. I think Rafael Benitez is on the right lines and speaks with a lot of confidence about his team and what he wants from them. But there is no doubt Liverpool's next two games will shape their season, at Newcastle away in the league and then Bayer away in the Champions League second leg. What they cannot afford to do is produce any performances like they produced at Burnley, Southampton or Birmingham. If they slip up at Newcastle then Everton beat Blackburn 24 hours later, that will be an 11-point gap and that's an awful long way back for them in the race for the Champions League place. There is added spice because Everton are fourth. They had an impressive win at Aston Villa, and you cannot take away from them what they have done.

They've had an uncertain spell recently, but they've picked up points here and there and that is a great tribute to manager David Moyes and his players. And in Tim Cahill, they've paid £2m for a player from outside the Premiership who has proved himself in the top-flight. Liverpool will still be a massive magnet for top players, but they may also need to seek out the type of signings that Moyes has pulled off with Cahill. He has been excellent since arriving from Millwall and has been a very sound purchase by Moyes. While the battle for fourth hots up, Manchester United

turned the screw a little tighter on leaders Chelsea by beating Portsmouth and reducing the gap to six points - albeit with a game in hand for Jose Mourinho's side. The Carling Cup win against Liverpool was massive for Chelsea, because it stopped all the inevitable questions that would have been posed if they had lost three games in a week. I don't think they answered all the questions, because for all their long periods of possession they were struggling to score until Gerrard's unfortunate intervention. Obviously a lot of focus has been centred on Mourinho for events on and off the pitch, but I think he will be more than happy with that because it means the heat is taken off his players. If people are asking questions about the manager, they are leaving the players in peace, so Mourinho will settle for that. And while United are showing once again there is no-one better when it comes to the chase, I don't think there is any shift in the balance of power in the Premiership. It is all Chelsea's to lose, with a six-point lead and a game in hand. Throw in that their next four league games are against the bottom four sides in the table, and you can see they are in a strong position. They must keep their eye on the ball because Manchester United are masters of this situation - but the balance of power still lies with Chelsea.

007.txt

Manchester United reduced Chelsea's Premiership lead to nine points after a scrappy victory over Manchester City.

Wayne Rooney met Gary Neville's cross to the near post with a low shot, which went in via a deflection off Richard Dunne, to put United ahead. Seven minutes later, the unfortunate Dunne hooked a volley over David James' head and into his own net. Steve McManaman wasted City's best chance when he shot wide from three yards in the first half. In the opening 45 minutes United had looked unlikely to earn the win they needed to maintain any chance of catching Chelsea in the title race. Their approach play was more laboured than patient and they managed to fashion just one chance - a Paul Scholes header over the bar. And City seemed to be content to sit back and try and hit their rivals on the break as the game settled into a tepid pattern. Only Shaun Wright-Phillips appeared capable of interrupting the monotony, looking lively down the right and causing Gabriel Heinze problems.

- Manchester City boss Kevin Keegan: "We had a great chance to take the lead and the first goal was always going to be crucial. "We started off with a good tempo but then we allowed them to dictate the pace a bit too much. "But we still had four good chances, two after we'd gone 2-0 down, the one McManaman missed was very similar to the one Wayne Rooney scored from."

- Manchester United boss Sir Alex Ferguson: "It wasn't our best performance of the last three months but I think we're deserved winners. "At times, especially in the first half, we didn't play with enough speed. But with (Cristiano) Ronaldo and (Ryan) Giggs on, the speed improved.

Search query: SATURDAY,FOOTBALL

N: 1

spark-submit --master yarn-client --executor-memory 512m --num-executors 5 --executor-cores 1 --driver-memory 512m finalexam_final.py -w SATURDAY,FOOTBALL -n 1

```
[root@sandbox-hdp final_exam]# spark-submit --master yarn-client --executor-memory 512m --num-executors 5 --executor-cores
1 --driver-memory 512m finalexam_final.py -w SATURDAY,FOOTBALL -n 1
```

```
18/11/30 21:36:37 INFO CodeGenerator: Code generated in 38.939835 ms
+-----------------------------------------------------------------------------------------+------------------+
|file                                                                                     |tf_idf            |
+-----------------------------------------------------------------------------------------+------------------+
|hdfs://sandbox-hdp.hortonworks.com:8020/user/root/xin_zhao/final_exam/football/141.txt|6.4871098710073465|
+-----------------------------------------------------------------------------------------+------------------+
only showing top 1 row
```

Search query: SATURDAY,FOOTBALL

N: 3

spark-submit --master yarn-client --executor-memory 512m --num-executors 5 --executor-cores 1 --driver-memory 512m finalexam_final.py -w SATURDAY,FOOTBALL -n 3

```
[root@sandbox-hdp final_exam]# spark-submit --master yarn-client --executor-memory 512m --num-executors 5 --executor-cores
 1 --driver-memory 512m finalexam_final.py -w SATURDAY,FOOTBALL -n 3
```

```
18/11/30 21:29:34 INFO CodeGenerator: Code generated in 84.900743 ms
+-----------------------------------------------------------------------------------------+------------------+
|file                                                                                     |tf_idf            |
+-----------------------------------------------------------------------------------------+------------------+
|hdfs://sandbox-hdp.hortonworks.com:8020/user/root/xin_zhao/final_exam/football/141.txt|6.4871098710073465|
|hdfs://sandbox-hdp.hortonworks.com:8020/user/root/xin_zhao/final_exam/football/198.txt|3.6165382190097946|
|hdfs://sandbox-hdp.hortonworks.com:8020/user/root/xin_zhao/final_exam/football/128.txt|1.8082691095048973|
+-----------------------------------------------------------------------------------------+------------------+
only showing top 3 rows
```

Search query: SATURDAY,FOOTBALL

N: 5

spark-submit --master yarn-client --executor-memory 512m --num-executors 5 --executor-cores 1 --driver-memory 512m finalexam_final.py -w SATURDAY,FOOTBALL -n 5

```
[root@sandbox-hdp final_exam]# spark-submit --master yarn-client --executor-memory 512m --num-executors 5 --executor-cores
 1 --driver-memory 512m finalexam_final.py -w SATURDAY,FOOTBALL -n 5
```

```
18/11/30 21:21:31 INFO CodeGenerator: Code generated in 68.005663 ms
+-----------------------------------------------------------------------------------------+------------------+
|file                                                                                     |tf_idf            |
+-----------------------------------------------------------------------------------------+------------------+
|hdfs://sandbox-hdp.hortonworks.com:8020/user/root/xin_zhao/final_exam/football/141.txt|6.4871098710073465|
|hdfs://sandbox-hdp.hortonworks.com:8020/user/root/xin_zhao/final_exam/football/198.txt|3.6165382190097946|
|hdfs://sandbox-hdp.hortonworks.com:8020/user/root/xin_zhao/final_exam/football/128.txt|1.8082691095048973|
|hdfs://sandbox-hdp.hortonworks.com:8020/user/root/xin_zhao/final_exam/football/238.txt|1.3278781781158187|
|hdfs://sandbox-hdp.hortonworks.com:8020/user/root/xin_zhao/final_exam/football/224.txt|1.3278781781158187|
+-----------------------------------------------------------------------------------------+------------------+
only showing top 5 rows
```

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

"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."

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.

We lost 1-0 but our centre-half Joe Shaw was, for me, the best player in the world that day. On Saturday my son, who at three-and-a-half-years is a similar age to what I was, is going to be on the pitch before the game, while my daughter is the mascot. I think it will have the same effect on him - I don't think you forget things like that. I think the next time I went to Highbury was when I took a team there. When I was at Notts County we went there in the top flight and lost 2-0, but I took Huddersfield down there for the second leg of the League Cup and got a draw. I think they expected to get a cricket score because they beat us 5-0 in the first leg. Ian Wright was in the team and George Graham made one or two comments that made us pretty determined and had they not equalised late on we would have won that leg. They are a club I admire a lot. After we played Arsenal in the semi-final in 2003 the number of letters and messages I got from Arsenal fans regarding the behaviour of our fans and our team was quite outstanding and it just shows they are a great set of supporters. Everybody will be making an issue of that game at Old Trafford. I felt personally hard done by. You don't mind getting beat but in this particular case I think everybody saw their goal - scored by Freddie Ljungberg - was a travesty and should not have been allowed. I got fined for comments I made about referee Graham Poll, though I would not take any of those back. But that's how it has gone for us. The higher you go means you have to have Premier League referees and we have never really done very well with them. I would have liked a Football League official but Neale Barry, who is an experienced referee, has been awarded the game and you have to get on with it. That said, I think we will need more than help from the referee on Saturday after watching them demolish Crystal Palace on Monday.

People ask me if we have done anything different in our preparations this week? We have...nothing! The players have been too exhausted after playing for almost an hour with 10 men on Sunday that they have not been able to do anything. They have done stretches and warm downs but very little football because of tiredness. On Thursday we will work on something but when it comes to dealing with Bergkamp and Henry I will probably say to my defenders just close your eyes and get your fingers crossed. When they are on their game there is not much you can do about them.

128.txt

Middlesbrough boss Steve McClaren has praised the way his side have got to grips with European football after the 2-0 Uefa Cup win against Lazio.

Boro, who are playing in Europe for the first time in their 128-year history, are top of Group E with maximum points. "I think we have taken to Europe really well," said McClaren. "We got about Lazio, didn't let them settle or play. And in possession, we controlled it and looked threatening every time we went forward." Before the match, McClaren had said that a win over the Italian giants would put Boro firmly on the European footballing map. And after they did just that he said: "It was a perfect European night. For the team to give the fans a performance like that was the icing on the cake. "There have been many good performances but this was something special.

"You can see that the experience we have in the squad is showing. To win in Europe you need to defend well, and we have done that because we have conceded only one goal in four games. "We can also score goals, and again that is something you can see from the performances we have had, so we have good balance. McClaren's only criticism of his side was that their dominance should have been resulted in more goals. "It should have been more convincing," said McClaren. "But I had watched Lazio in recent weeks and I saw them score a late equaliser against Inter Milan on Saturday so I knew we needed a second goal.

238.txt

As usual, the issue is probably blown out of proportion, but I don't think anyone in football will deny there is a problem with the rules as they apply to recruiting players. I read somewhere at the weekend that they did a straw poll and questioned every player at a particular club as to how he got there. Just about every one said the first approach was through their agent, or a third party or somebody involved with the club.

On that basis, under the rules as they stand, they all got there illegally. That's the name of the game these days, I'm afraid. Not that I have ever tapped a player up - I wouldn't dream of it! I know there is a school of thought that says the rules that apply in football just wouldn't be tolerated in the outside business world. In business, if you want to change jobs, you can simply go and have a chat with another prospective employer. But in football you're not allowed to do that. Football does have strange anomalies. For example, the game has a disciplinary procedure where there is no evidence but you can still find yourself in trouble. It's the sort of thing that wouldn't happen in a court of law.

Compared to the outside world, football does have some very restrictive practices, and a lot of them have to be looked at, but if you want to be part of it, you have to adhere to the rules. You try and do things the right way, but it's like buying a house. If you do things properly and play by

224.txt

Club football in South America continues to suffer from the continent's economic crisis.

The best players are lured across the Atlantic - 900 players left Brazil alone in 2004. And crowds are low, with money tight and the fear of violence proving a powerful incentive to stay at home. The average attendance in this year's Brazilian Championship was a record low of 8,139. Furthermore, in 2004 no new talent exploded on the continent with quite the same force as Diego, Robinho and Carlos Tevez in the two previous years. There is no doubt that the highlight of South America's footballing year has to come from the international game. There were some dramatic tournaments in 2004. The Copa America, held in Peru last July, could hardly have come to a more exciting conclusion. With the last kick of the final Brazil scored one of the most decisive equalising goals in the history of football.

Argentina thought they had the game won and were in no state to take penalties when Adriano's 93rd-minute blast forced a shoot-out. Earlier in the year there was intense disappointment for Brazil when they amazingly failed to qualify for the Olympic football tournament. The gold medal is the only title they lack but they will have to wait four more years after missing out on the two qualification slots up for grabs in South America's Under-23 Championships, held in Chile in January. Argentina and Paraguay made it through and went on to secure gold and silver in Athens. These tournaments were terrific but the real highlights of the year in South American football could only come from the World Cup qualification campaign. The reason? It is the only time that all the great players come back home and play

was about to get under way. The thought struck me that the accumulation of talent on the Belo Horizonte pitch was at least as outstanding as that about to go into action in Portugal. For 90 minutes Brazil was the capital of the football world. Ronaldo obviously felt the sense of occasion.

He will probably never play a bigger match in Brazil and it was taking place in the stadium where he had made his name over a decade earlier. He was unstoppable, suffering and scoring three penalties in Brazil's 3-1 win. If Ronaldo's performance was the individual highlight of the year, my collective prize goes to Argentina in a match they played a few months later. In October I was down in Buenos Aires to see their first match under new coach Jose Pekerman. Nerves were settled with an early goal and old rivals Uruguay were blasted away. Soon after half-time Argentina were four up and although Uruguay hit back with two consolation goals, it was the performance of the hosts that lingers in the memory. Argentina's goals were magnificent team efforts involving all of the fundamentals of well-played football. The ball moved quickly around the field, with plenty of running off the ball to provide options and lots of crisp, accurate

passing, mixing long and short to find gaps in the defence. The crowd responded in fine style, creating the kind of atmosphere familiar to all those old enough to remember the 1978 World Cup. These, then, were my highlights of 2004. With seven vital rounds of World Cup qualifiers scheduled for next year, South American football should come up with much more to celebrate in 2005.

sample screen shoot of tf_idf index output as csv format

Hadoop fs –ls /user/root/xin_zhao/final/tf_idf