FlavorsOfSpark Writeup

Note that since I use Windows, I was unable to save the directories due to lack of Hadoop. That being said, this is the method I made to save to a directory (for Dataset). I have no idea if it works.

```
private static <T> void saveDataset(Dataset<T> dataset, String directoryName,
String format) {
   File directory = new File(directoryName);
   if (directory.exists()) {
       directory.delete();
   }
   dataset.write().format(format).save(directoryName);
}
```

You said on Piazza that you want all our code in the writeup (this is the longest writeup I have ever made), so this is my general code that calls methods for each step.

```
public static void main(String[] args) {
   // Note: I'm getting strange errors if I try to run this directly from
IntelliJ
   // If I run it directly from Maven through FakeTest.java, it works, but I
get strange warnings
   SparkConf sparkConf = new SparkConf().setAppName("RDD Flavor")
           .setMaster("local").set("spark.executor.memory", "2g");
   sc = new JavaSparkContext(sparkConf);
   JavaRDD<String> titles = step1();
   step2(titles);
   sc.close();

   // it looks like the dataframe stuff is from completely different API
   spark = SparkSession.builder()
           .appName("Dataset Flavor").master("local")
           .getOrCreate();

   Dataset<Row> bookRows = step3();
   Dataset<Book> bookSet = step4(bookRows);
   Dataset<Row> goodDateSet = step5(bookSet);

   step6(goodDateSet);
   step7(goodDateSet);
```

```
    step8(goodDateSet);

    spark.stop();

}
```

Step 1:

```
private static JavaRDD<String> step1() {
    JavaRDD<String> bookRows = sc.textFile("books.csv");
    bookRows = bookRows.map(bookRow -> {
        // there's probably a better way of doing things, but the easiest other
way I can find involves
        // making a dataframe and then converting back
        int firstComma = bookRow.indexOf(',');
        int titleStart = bookRow.indexOf(',', firstComma + 1) + 1;
        int titleEnd;
        if (bookRow.charAt(titleStart) == '"') {
            titleEnd = bookRow.indexOf('"', titleStart + 1) + 1;
        } else {
            titleEnd = bookRow.indexOf(',', titleStart);
        }
        return bookRow.substring(titleStart, titleEnd);
    });

    printRDD(bookRows);
    return bookRows;
}
```

title

"Fantastic Beasts and Where to Find Them: The Original Screenplay"

"Harry Potter and the Sorcerer's Stone: The Illustrated Edition (Harry Potter Book 1)"

"The Tales of Beedle the Bard Standard Edition (Harry Potter)"

"Harry Potter and the Chamber of Secrets: The Illustrated Edition (Harry Potter Book 2)"

"Informix 12.10 on Mac 10.12 with a dash of Java 8: The Tale of the Apple, the Coffee, and a Great Database"

"Development Tools in 2006: any Room for a 4GL-style Language?: An independent study by Jean Georges Perrin, IIUG Board Member"

"Adventures of Huckleberry Finn"

A Connecticut Yankee in King Arthur's Court

"Jacques le Fataliste"

"Diderot Encyclopedia"

"A Woman in Berlin"

"Spring Boot in Action"

"Spring in Action: Covers Spring 4"

"Soft Skills: The software developer's life manual"

"Of Mice and Men"

"Java 8 in Action: Lambdas, Streams, and functional-style programming"

Hamlet

Pensées

"Fables choisies, mises en vers par M. de La Fontaine"

Discourse on Method and Meditations on First Philosophy

Twelfth Night

Macbeth

Step 2:

```
private static void step2(JavaRDD<String> titles) {
    JavaRDD<String> harryPotterTitles = titles.filter(title ->
title.contains("Harry Potter"));
    System.out.println(); // to give space between the outputs
    printRDD(harryPotterTitles);

    // Note that the code below here isn't used
    String directoryName = "rddOutput";
    File directory = new File(directoryName);
    if (directory.exists()) {
        directory.delete();
    }
    // TODO: Figure out file saving
    //harryPotterTitles.saveAsTextFile(directoryName);
}
```

"Harry Potter and the Sorcerer's Stone: The Illustrated Edition (Harry Potter Book 1)"

"The Tales of Beedle the Bard Standard Edition (Harry Potter)"

"Harry Potter and the Chamber of Secrets: The Illustrated Edition (Harry Potter Book 2)"

Step 3:

```
private static Dataset<Row> step3() {
   Dataset<Row> bookRows = spark.read().format("csv")
           .option("header", "true").load("books.csv");

   System.out.println();
   System.out.println(bookRows.showString(5, 0, true));

   return bookRows;
}
```

-RECORD
0----------
id | 1
authorId | 1
title | Fantastic Beasts and Where to Find Them: The Original Screenplay
releaseDate | 11/18/16
link | http://amzn.to/2kup94P
-RECORD
1----------
id | 2
authorId | 1
title | Harry Potter and the Sorcerer's Stone: The Illustrated Edition (Harry Potter Book 1)
releaseDate | 10/6/15
link | http://amzn.to/2l2lSwP
-RECORD
2----------
id | 3
authorId | 1
title | The Tales of Beedle the Bard Standard Edition (Harry Potter)
releaseDate | 12/4/08
link | http://amzn.to/2kYezqr

-RECORD
3--
 id | 4
 authorId | 1
 title | Harry Potter and the Chamber of Secrets: The Illustrated Edition (Harry Potter Book 2)
 releaseDate | 10/4/16
 link | http://amzn.to/2kYhL5n
-RECORD
4--
 id | 5
 authorId | 2
 title | Informix 12.10 on Mac 10.12 with a dash of Java 8: The Tale of the Apple, the Coffee, and a Great Database
 releaseDate | 4/23/17
 link | http://amzn.to/2i3mthT
only showing top 5 rows

Step 4:

```
private static Dataset<Book> step4(Dataset<Row> bookRows) {
   bookRows.printSchema();
   Dataset<Book> bookSet = bookRows.map((MapFunction<Row, Book>) row->
       new Book(new String[]{row.getAs("id"), row.getAs("authorId"),
row.getAs("title"), row.getAs("releaseDate"), row.getAs("link")}),
           bookEncoder);

   System.out.println(bookSet.showString(5, 0, true));
   bookSet.printSchema();

   // TODO: Save this also
   //saveDataset(bookSet, "dataSet", "json");

   return bookSet;
}
```

-RECORD 0----------
 authorId | 1
 id | 1
 link | http://amzn.to/2kup94P
 releaseDate | {18, 5, 0, 0, 10, 0, 1479445200000, 300, 116}
 title | Fantastic Beasts and Where to Find Them: The Original Screenplay
-RECORD 1----------
 authorId | 1
 id | 2
 link | http://amzn.to/2l2lSwP
 releaseDate | {6, 2, 0, 0, 9, 0, 1444104000000, 240, 115}
 title | Harry Potter and the Sorcerer's Stone: The Illustrated Edition (Harry Potter Book 1)
-RECORD 2----------
 authorId | 1
 id | 3
 link | http://amzn.to/2kYezqr
 releaseDate | {4, 4, 0, 0, 11, 0, 1228366800000, 300, 108}
 title | The Tales of Beedle the Bard Standard Edition (Harry Potter)
-RECORD 3----------
 authorId | 1
 id | 4
 link | http://amzn.to/2kYhL5n
 releaseDate | {4, 2, 0, 0, 9, 0, 1475553600000, 240, 116}
 title | Harry Potter and the Chamber of Secrets: The Illustrated Edition (Harry Potter Book 2)
-RECORD 4----------
 authorId | 2

```
 id          | 5
 link        | http://amzn.to/2i3mthT
 releaseDate | {23, 0, 0, 0, 3, 0, 1492920000000, 240, 117}
 title       | Informix 12.10 on Mac 10.12 with a dash of Java 8: The Tale of the Apple, the Coffee,
and a Great Database
only showing top 5 rows

root
 |-- authorId: integer (nullable = false)
 |-- id: integer (nullable = false)
 |-- link: string (nullable = true)
 |-- releaseDate: struct (nullable = true)
 |    |-- date: integer (nullable = false)
 |    |-- day: integer (nullable = false)
 |    |-- hours: integer (nullable = false)
 |    |-- minutes: integer (nullable = false)
 |    |-- month: integer (nullable = false)
 |    |-- seconds: integer (nullable = false)
 |    |-- time: long (nullable = false)
 |    |-- timezoneOffset: integer (nullable = false)
 |    |-- year: integer (nullable = false)
 |-- title: string (nullable = true)
```

Step 5: Note that my date isn't in the format you want it in, though I'm convinced that your format isn't possible.

```
private static Dataset<Row> step5(Dataset<Book> bookSet) {
   Dataset<Row> newSet = bookSet.withColumn("niceDate",

functions.concat(functions.col("releaseDate.year").plus(functions.lit(1900)),
                                    functions.lit("-"),

functions.col("releaseDate.month").plus(functions.lit(1)),
```

```
                                    functions.lit('-'),
functions.col("releaseDate.date"))
                            .cast("date"))
           .drop("releaseDate");
   // since it's stored as number of years since 1900, I have to add 1900 to
get the full date
   // while month is stored starting from 0, so I have to add 1 to it
   // it sounds like this is what you want for date, though it isn't clear
   System.out.println(newSet.showString(25, 0, true));
   newSet.printSchema();

   // TODO: Save this also
   //saveDataset(newSet, "niceDate", "csv");

   return newSet;
}
```

-RECORD
0----------

---

authorId | 1

id | 1

link | http://amzn.to/2kup94P

title | Fantastic Beasts and Where to Find Them: The Original Screenplay

niceDate | 2016-11-18

-RECORD
1----------

---

authorId | 1

id | 2

link | http://amzn.to/2l2lSwP

title | Harry Potter and the Sorcerer's Stone: The Illustrated Edition (Harry Potter Book 1)

niceDate | 2015-10-06

-RECORD
2----------
 authorId | 1
 id | 3
 link | http://amzn.to/2kYezqr
 title | The Tales of Beedle the Bard Standard Edition (Harry Potter)
 niceDate | 2008-12-04
-RECORD
3----------
 authorId | 1
 id | 4
 link | http://amzn.to/2kYhL5n
 title | Harry Potter and the Chamber of Secrets: The Illustrated Edition (Harry Potter Book 2)
 niceDate | 2016-10-04
-RECORD
4----------
 authorId | 2
 id | 5
 link | http://amzn.to/2i3mthT
 title | Informix 12.10 on Mac 10.12 with a dash of Java 8: The Tale of the Apple, the Coffee, and a Great Database
 niceDate | 2017-04-23
-RECORD
5----------
 authorId | 2
 id | 6
 link | http://amzn.to/2vBxOe1

title | Development Tools in 2006: any Room for a 4GL-style Language?: An independent study by Jean Georges Perrin, IIUG Board Member
niceDate | 2016-12-28
-RECORD 6----------
authorId | 3
id | 7
link | http://amzn.to/2wOeOav
title | Adventures of Huckleberry Finn
niceDate | 1994-05-26
-RECORD 7----------
authorId | 3
id | 8
link | http://amzn.to/2x1NuoD
title | A Connecticut Yankee in King Arthur's Court
niceDate | 2017-06-17
-RECORD 8----------
authorId | 4
id | 10
link | http://amzn.to/2uZj2KA
title | Jacques le Fataliste
niceDate | 2000-03-01
-RECORD 9----------
authorId | 4

```
 id       | 11
 link     | http://amzn.to/2i2zo3I
 title    | "Diderot Encyclopedia": The Complete Illustrations 1762-1777
 niceDate | NULL
-RECORD
10----------------------------------------------------------------------------------------------------------------
--------------
 authorId | 5
 id       | 12
 link     | http://amzn.to/2i472WZ
 title    | A Woman in Berlin
 niceDate | NULL
-RECORD
11----------------------------------------------------------------------------------------------------------------
--------------
 authorId | 6
 id       | 13
 link     | http://amzn.to/2hCPktW
 title    | Spring Boot in Action
 niceDate | 2016-01-03
-RECORD
12----------------------------------------------------------------------------------------------------------------
--------------
 authorId | 6
 id       | 14
 link     | http://amzn.to/2yJLyCk
 title    | Spring in Action: Covers Spring 4
 niceDate | 2014-11-28
-RECORD
13----------------------------------------------------------------------------------------------------------------
--------------
```

```
 authorId | 7
 id       | 15
 link     | http://amzn.to/2zNnSyn
 title    | Soft Skills: The software developer's life manual
 niceDate | 2014-12-29
-RECORD
14-------------------------------------------------------------------------------------------------------------------
--------------
 authorId | 8
 id       | 16
 link     | http://amzn.to/2zJjXoc
 title    | Of Mice and Men
 niceDate | NULL
-RECORD
15-------------------------------------------------------------------------------------------------------------------
--------------
 authorId | 9
 id       | 17
 link     | http://amzn.to/2isdqoL
 title    | Java 8 in Action: Lambdas, Streams, and functional-style programming
 niceDate | 2014-08-28
-RECORD
16-------------------------------------------------------------------------------------------------------------------
--------------
 authorId | 12
 id       | 18
 link     | http://amzn.to/2yRbewY
 title    | Hamlet
 niceDate | 2012-06-08
```

-RECORD 17----------
 authorId | 13
 id | 19
 link | http://amzn.to/2jweHOG
 title | Pensées
 niceDate | 2013-06-09
-RECORD 18----------
 authorId | 14
 id | 20
 link | http://amzn.to/2yRH10W
 title | Fables choisies, mises en vers par M. de La Fontaine
 niceDate | 1999-09-01
-RECORD 19----------
 authorId | 15
 id | 21
 link | http://amzn.to/2hwB8zc
 title | Discourse on Method and Meditations on First Philosophy
 niceDate | 1999-06-15
-RECORD 20----------
 authorId | 12
 id | 22
 link | http://amzn.to/2zPYnwo
 title | Twelfth Night

```
 niceDate | 2003-07-01
-RECORD
21-------------------------------------------------------------------------------------------------------------------
--------------
 authorId | 12
 id       | 23
 link     | http://amzn.to/2zPYnwo
 title    | Macbeth
 niceDate | 2003-07-01


root
 |-- authorId: integer (nullable = false)
 |-- id: integer (nullable = false)
 |-- link: string (nullable = true)
 |-- title: string (nullable = true)
 |-- niceDate: date (nullable = true)
```

Step 6:

```
private static void step6(Dataset<Row> bookSet) {
   Dataset<Row> sortedSet = bookSet.sort(functions.desc("niceDate"));
   System.out.println(sortedSet.showString(25, 0, true));
}
```

```
-RECORD
0-------------------------------------------------------------------------------------------------------------------
--------------
 authorId | 3
 id       | 8
 link     | http://amzn.to/2x1NuoD
 title    | A Connecticut Yankee in King Arthur's Court
 niceDate | 2017-06-17
```

-RECORD 1----------
authorId | 2
id | 5
link | http://amzn.to/2i3mthT
title | Informix 12.10 on Mac 10.12 with a dash of Java 8: The Tale of the Apple, the Coffee, and a Great Database
niceDate | 2017-04-23
-RECORD 2----------
authorId | 2
id | 6
link | http://amzn.to/2vBxOe1
title | Development Tools in 2006: any Room for a 4GL-style Language?: An independent study by Jean Georges Perrin, IIUG Board Member
niceDate | 2016-12-28
-RECORD 3----------
authorId | 1
id | 1
link | http://amzn.to/2kup94P
title | Fantastic Beasts and Where to Find Them: The Original Screenplay
niceDate | 2016-11-18
-RECORD 4----------
authorId | 1
id | 4

```
 link     | http://amzn.to/2kYhL5n
 title    | Harry Potter and the Chamber of Secrets: The Illustrated Edition (Harry Potter Book 2)
 niceDate | 2016-10-04
-RECORD
5-------------------------------------------------------------------------------------------------------------
--------------
 authorId | 6
 id       | 13
 link     | http://amzn.to/2hCPktW
 title    | Spring Boot in Action
 niceDate | 2016-01-03
-RECORD
6-------------------------------------------------------------------------------------------------------------
--------------
 authorId | 1
 id       | 2
 link     | http://amzn.to/2l2lSwP
 title    | Harry Potter and the Sorcerer's Stone: The Illustrated Edition (Harry Potter Book 1)
 niceDate | 2015-10-06
-RECORD
7-------------------------------------------------------------------------------------------------------------
--------------
 authorId | 7
 id       | 15
 link     | http://amzn.to/2zNnSyn
 title    | Soft Skills: The software developer's life manual
 niceDate | 2014-12-29
-RECORD
8-------------------------------------------------------------------------------------------------------------
--------------
 authorId | 6
```

```
 id       | 14
 link     | http://amzn.to/2yJLyCk
 title    | Spring in Action: Covers Spring 4
 niceDate | 2014-11-28
-RECORD
9----------------------------------------------------------------------------------------------------------------
--------------
 authorId | 9
 id       | 17
 link     | http://amzn.to/2isdqoL
 title    | Java 8 in Action: Lambdas, Streams, and functional-style programming
 niceDate | 2014-08-28
-RECORD
10---------------------------------------------------------------------------------------------------------------
--------------
 authorId | 13
 id       | 19
 link     | http://amzn.to/2jweHOG
 title    | Pensées
 niceDate | 2013-06-09
-RECORD
11---------------------------------------------------------------------------------------------------------------
--------------
 authorId | 12
 id       | 18
 link     | http://amzn.to/2yRbewY
 title    | Hamlet
 niceDate | 2012-06-08
-RECORD
12---------------------------------------------------------------------------------------------------------------
--------------
```

```
 authorId | 1
 id       | 3
 link     | http://amzn.to/2kYezqr
 title    | The Tales of Beedle the Bard Standard Edition (Harry Potter)
 niceDate | 2008-12-04
-RECORD
13-------------------------------------------------------------------------------------------------------------------
--------------
 authorId | 12
 id       | 22
 link     | http://amzn.to/2zPYnwo
 title    | Twelfth Night
 niceDate | 2003-07-01
-RECORD
14-------------------------------------------------------------------------------------------------------------------
--------------
 authorId | 12
 id       | 23
 link     | http://amzn.to/2zPYnwo
 title    | Macbeth
 niceDate | 2003-07-01
-RECORD
15-------------------------------------------------------------------------------------------------------------------
--------------
 authorId | 4
 id       | 10
 link     | http://amzn.to/2uZj2KA
 title    | Jacques le Fataliste
 niceDate | 2000-03-01
```

-RECORD 16----------
authorId | 14
id | 20
link | http://amzn.to/2yRH10W
title | Fables choisies, mises en vers par M. de La Fontaine
niceDate | 1999-09-01
-RECORD 17----------
authorId | 15
id | 21
link | http://amzn.to/2hwB8zc
title | Discourse on Method and Meditations on First Philosophy
niceDate | 1999-06-15
-RECORD 18----------
authorId | 3
id | 7
link | http://amzn.to/2wOeOav
title | Adventures of Huckleberry Finn
niceDate | 1994-05-26
-RECORD 19----------
authorId | 4
id | 11
link | http://amzn.to/2i2zo3I
title | "Diderot Encyclopedia": The Complete Illustrations 1762-1777

```
 niceDate | NULL
-RECORD
20----------------------------------------------------------------------------------------------------------------
--------------
 authorId | 5
 id       | 12
 link     | http://amzn.to/2i472WZ
 title    | A Woman in Berlin
 niceDate | NULL
-RECORD
21----------------------------------------------------------------------------------------------------------------
--------------
 authorId | 8
 id       | 16
 link     | http://amzn.to/2zJjXoc
 title    | Of Mice and Men
 niceDate | NULL

-RECORD 0-------------------------------------------------------------
 authorId | 4
 id       | 11
 link     | http://amzn.to/2i2zo3I
 title    | "Diderot Encyclopedia": The Complete Illustrations 1762-1777
 niceDate | NULL
-RECORD 1-------------------------------------------------------------
 authorId | 5
 id       | 12
 link     | http://amzn.to/2i472WZ
 title    | A Woman in Berlin
 niceDate | NULL
-RECORD 2-------------------------------------------------------------
```

authorId | 8
id | 16
link | http://amzn.to/2zJjXoc
title | Of Mice and Men
niceDate | NULL

Step 7:

```
private static void step7(Dataset<Row> bookSet) {
   Dataset<Row> nullSet = bookSet.filter(functions.col("niceDate").isNull());
   System.out.println(nullSet.showString(25, 0, true));
}
```

-RECORD 0----------
authorId | 4
id | 11
link | http://amzn.to/2i2zo3I
title | "Diderot Encyclopedia": The Complete Illustrations 1762-1777
niceDate | NULL
-RECORD 1----------
authorId | 5
id | 12
link | http://amzn.to/2i472WZ
title | A Woman in Berlin
niceDate | NULL
-RECORD 2----------
authorId | 8
id | 16
link | http://amzn.to/2zJjXoc
title | Of Mice and Men
niceDate | NULL

Step 8:

```
private static void step8(Dataset<Row> bookSet) {
```

```
    Dataset<Row> mostBooks = bookSet.groupBy("authorId")
            .count()
            .withColumnRenamed("authorId", "id")
            .sort(functions.desc("count"))
            .limit(1);
    System.out.println(mostBooks.showString(1, 0, true));
}
```

```
-RECORD 0----
 id    | 1
 count | 4
```