**Project 2**

**Chandra Sekhar Padhy**  **CSP180002**

**Introduction and Problem Description** :

We were given a dataset which was hosted on AWS. After exploring the dataset, many columns were explored like customer\_id, review\_id, product\_id, helpful\_votes, total\_votes, star\_rating, review\_body, etc.

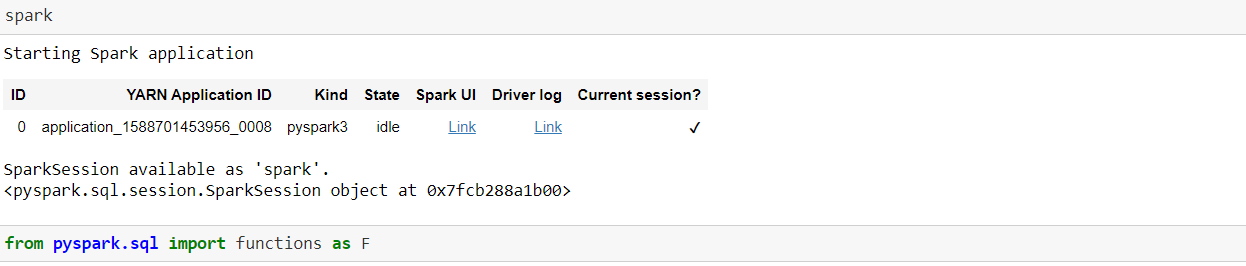
I tried to develop relationships between some of these coulmns mentioned above and thus draw effective conclusions on how various products performed, wheter the one which performed better were the digital or the printed articles to name a few.

I decided to filter the data before analysing to remove multiple reviews given by the same user and thus draw effective and accurate assumptions.

The most effective weeks for the performance of digital and printed books with the best ratings were analysed along the way. LDA was performed on the data set, topic modeling was done to analyse average star ratings below 3 as well as above 3 star ratings, so as covering most of the topics. Stop words which were a hinderance were dealt with for accurate results.

**Project Solution :**

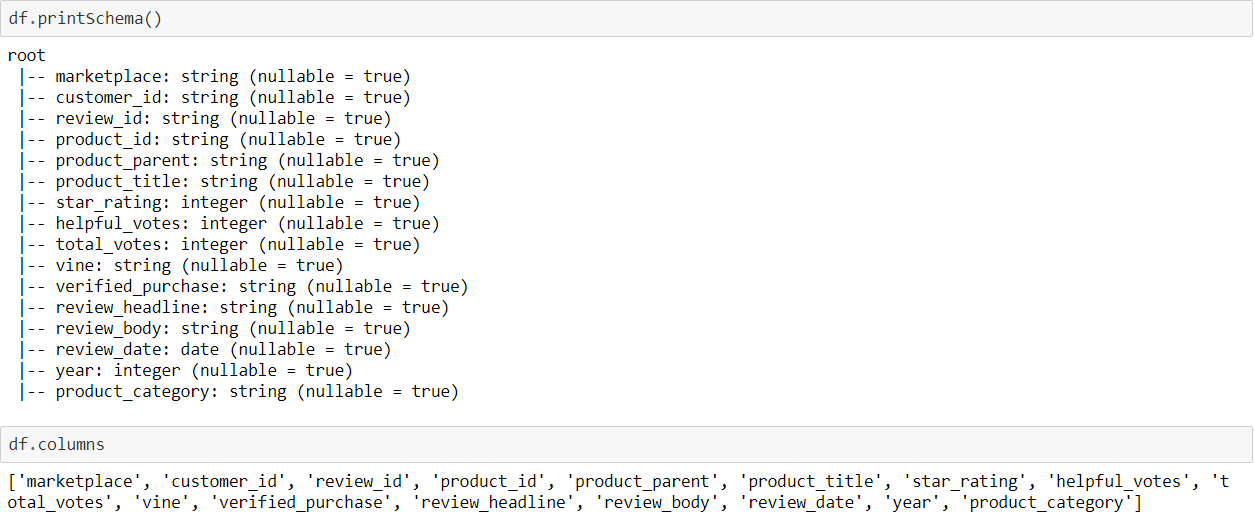
**Initial start to Spark Session :**



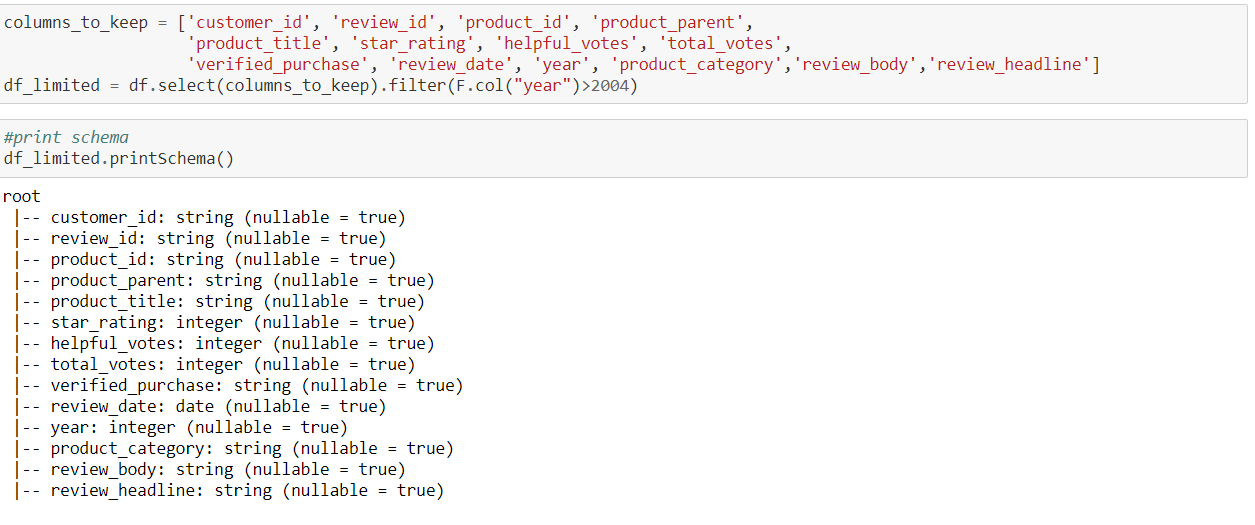
**Load the initial Dataset :**



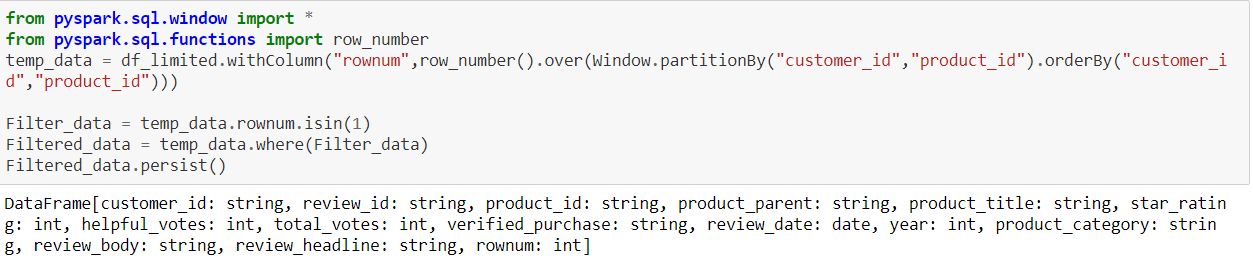
**Print Schema and Number of Columns :**



**Columns to Keep:**



**Filter Data before further Analysis :**

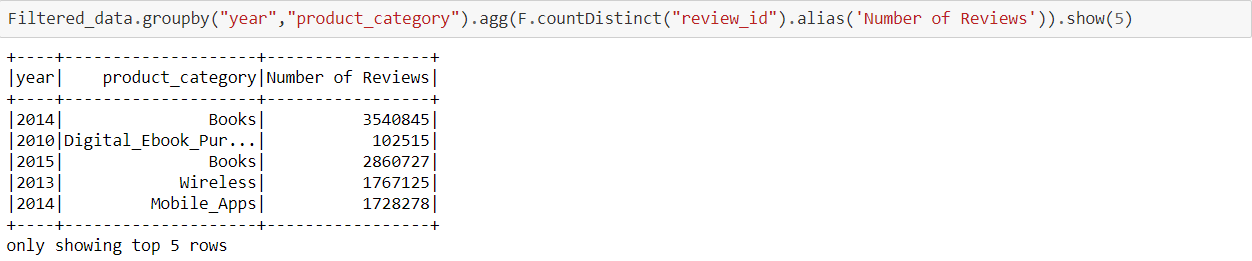


**Question 1) Explore the dataset and provide analysis by product-category and year**

**1) Number of reviews**

Query : Filtered\_data.groupby("year","product\_category").agg(F.countDistinct("review\_id").alias('Number of Reviews')).show(5)

Output :

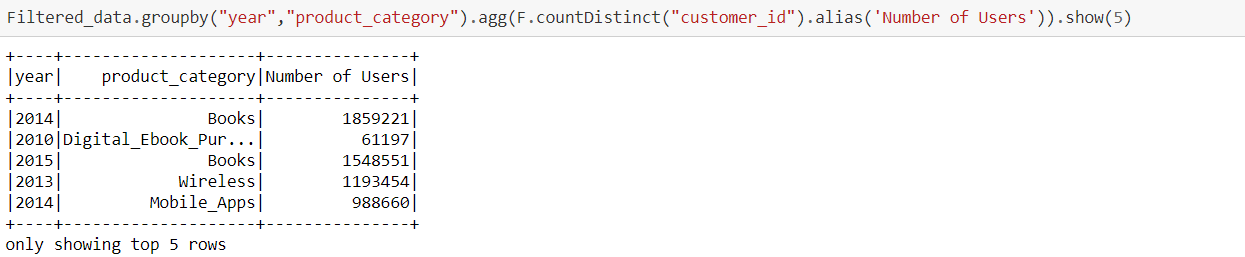


**2) Number of users**

**Query :**

Filtered\_data.groupby("year","product\_category").agg(F.countDistinct("customer\_id").alias('Number of Users')).show(5)

**Output:**



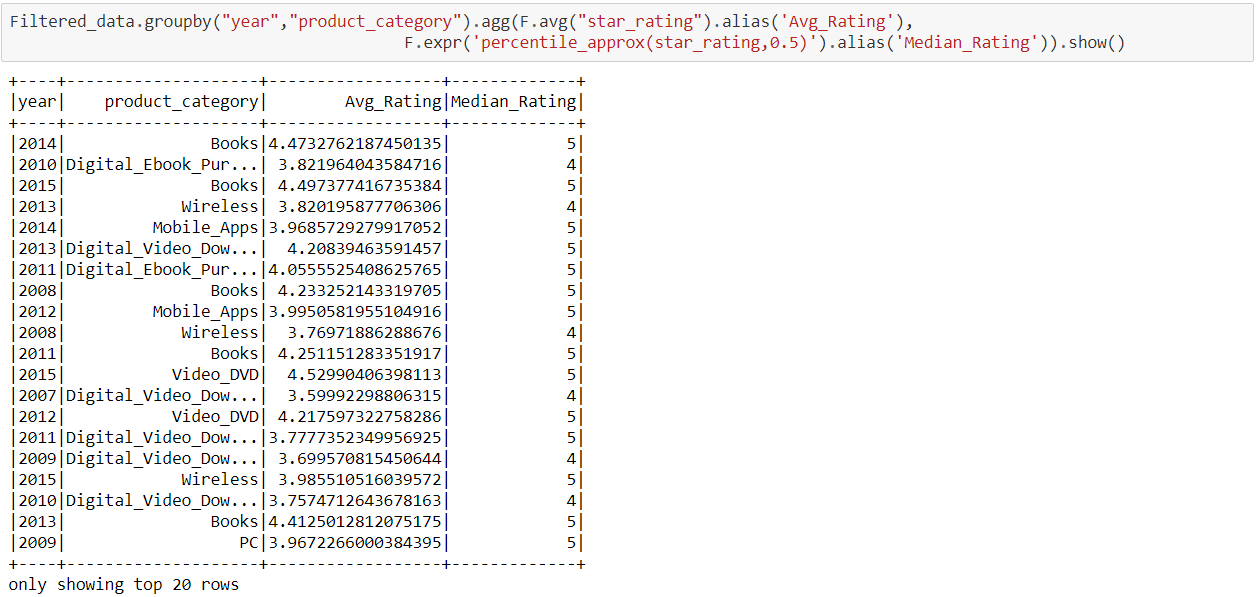
**3) Average and Median review stars**

**Query :**

Filtered\_data.groupby("year","product\_category").agg(F.avg("star\_rating").alias('Avg\_Rating'),

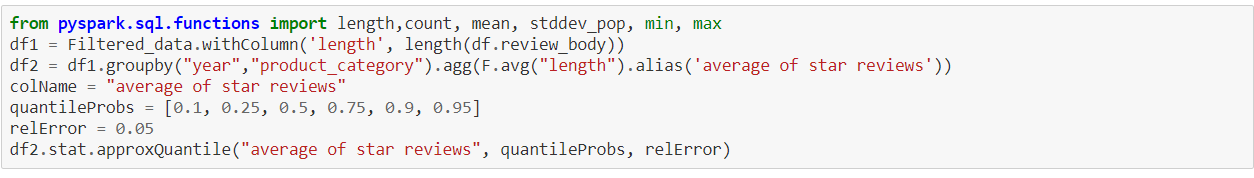
F.expr('percentile\_approx(star\_rating,0.5)').alias('Median\_Rating')).show()

**Output :**



**4) Percentiles of length of the review. Use the following percentiles:**

**Query :**

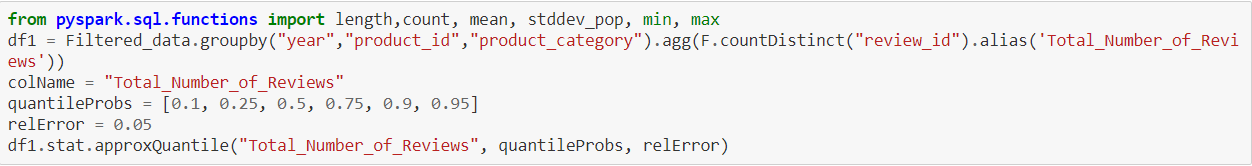


**Output :**

[205.48028507600662, 349.16350523270023, 586.2441338149835, 853.2765772362093, 945.7590082915988, 2207.5789473684213]

**5) Percentiles for number of reviews per product. For example, 10% of books got 5 or less reviews**

**Query :**



**Output :**

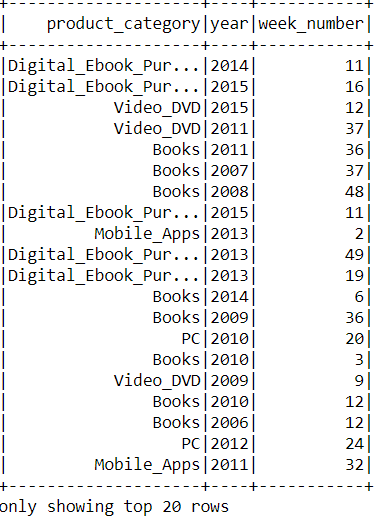
[1.0, 1.0, 2.0, 4.0, 5108.0, 31128.0]

**6) Identify week number (each year has 52 weeks) for each year and product category with most positive reviews (4 and 5 star)**

**Query :**



**Output :**



**Q2) Provide detailed analysis of "Digital eBook Purchase" versus Books.**

1. Using Spark Pivot functionality, produce DataFrame with following columns:
   1. Year
   2. Month
   3. Total number of reviews for "Digital eBook Purchase" category
   4. Total number of reviews for "Books" category
   5. Average stars for reviews for "Digital eBook Purchase" category
   6. Average stars for reviews for "Books" category

**Query :**

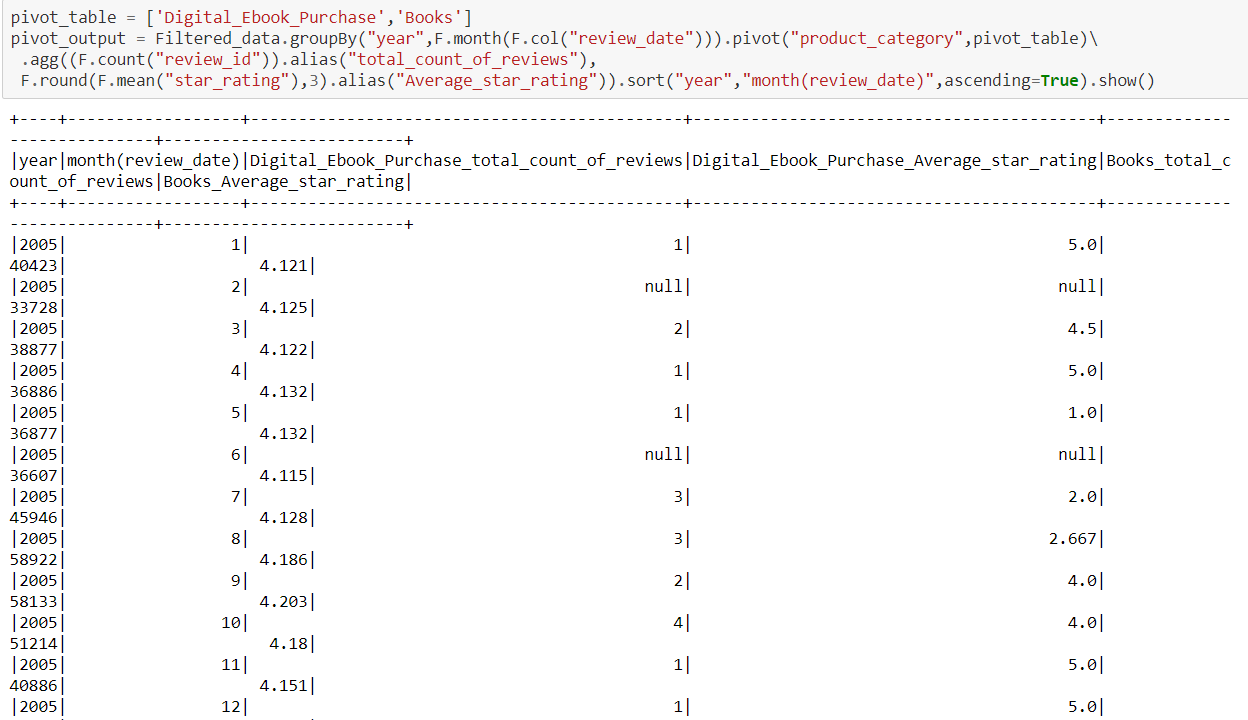
pivot\_table = ['Digital\_Ebook\_Purchase','Books']

pivot\_output = Filtered\_data.groupBy("year",F.month(F.col("review\_date"))).pivot("product\_category",pivot\_table)\

.agg((F.count("review\_id")).alias("total\_count\_of\_reviews"),

F.round(F.mean("star\_rating"),3).alias("Average\_star\_rating")).sort("year","month(review\_date)",ascending=**True**).show()

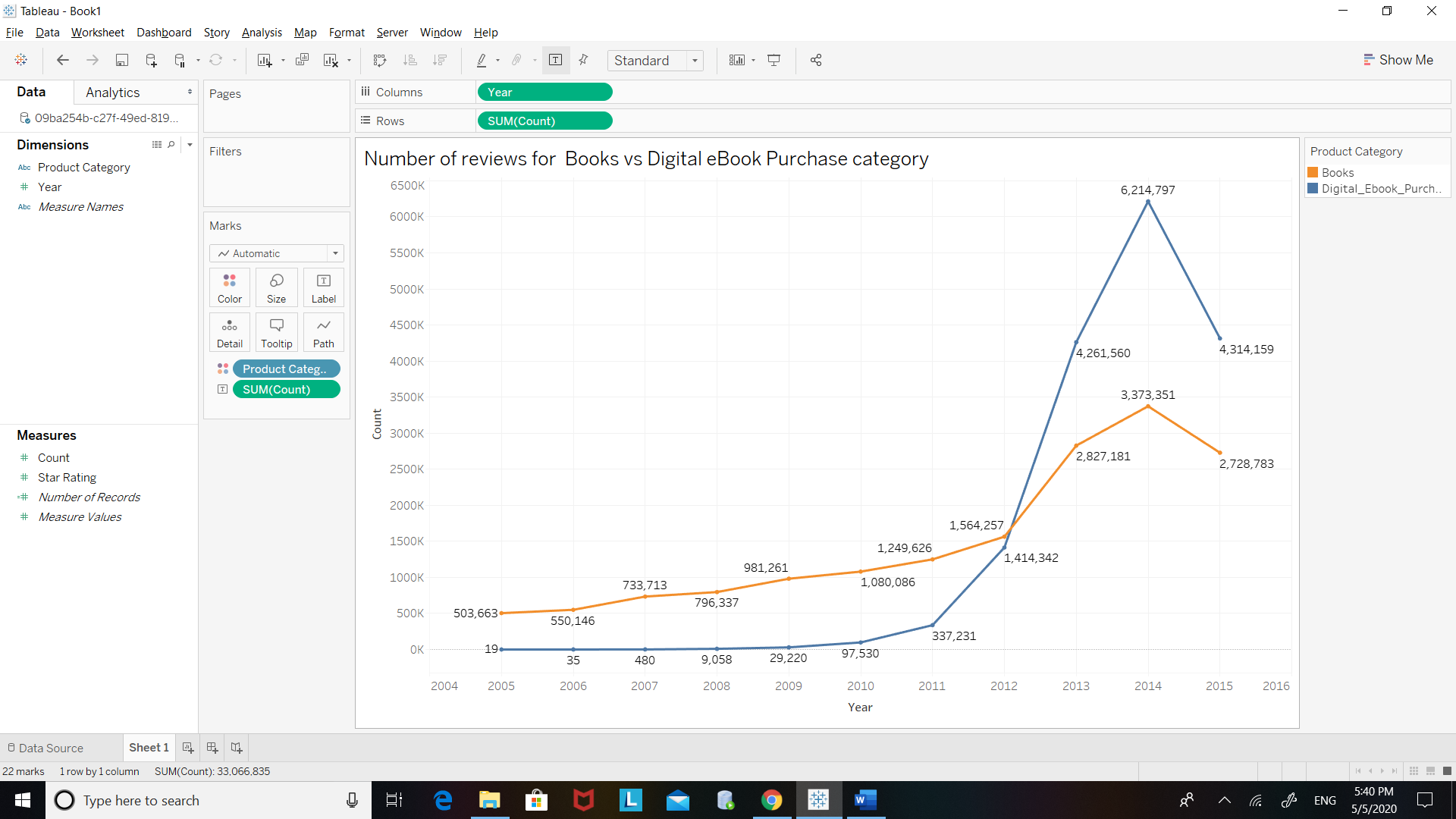
**Output :**



**Q2) 2)**

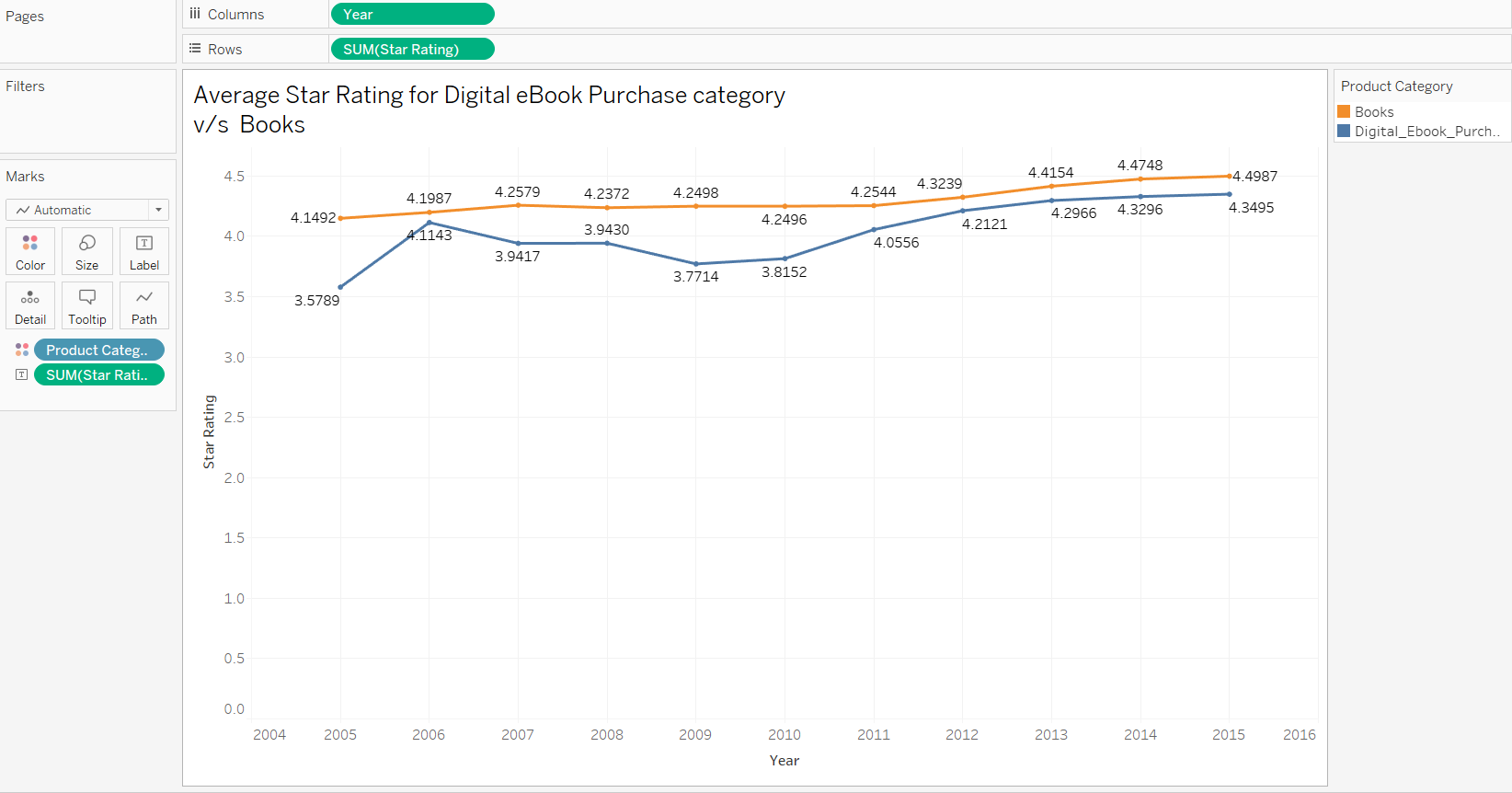
**2. Produce two graphs to demonstrate aggregations from #1: 1. Number of reviews 2. Average stars**

**1 – Number of reviews for Books Vs Digital eBook Purchase category**



From the above graph we can conclude that the total count for the number of reviews for digital books is greater than that for the printed books.

**2 – Avearage Star rating for Digital eBook Purchase category vs Books**

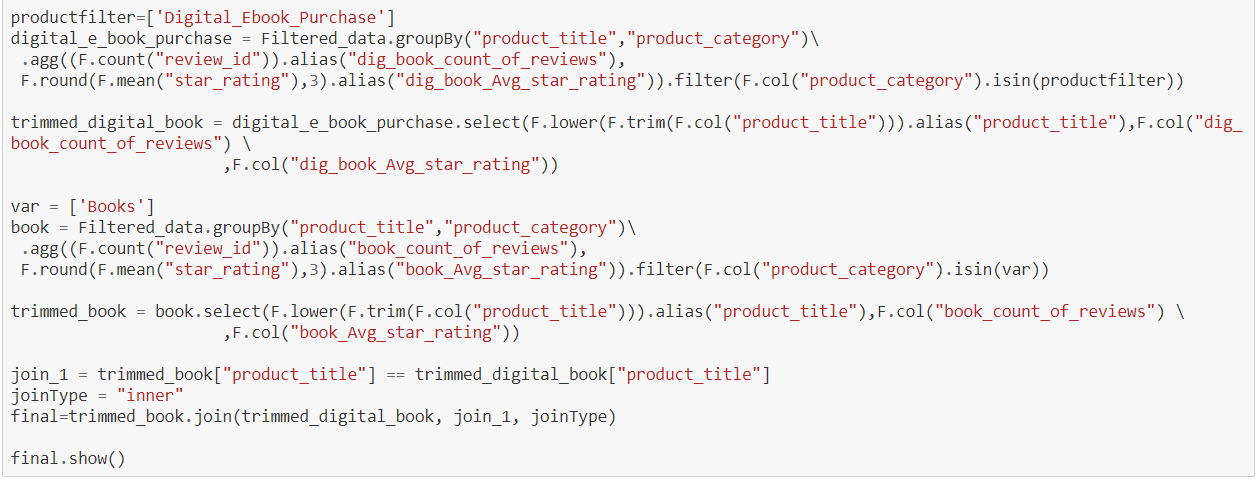


From the above graph we can conclude that the average rating for printed books is more than the average star rating for Digital eBook category.

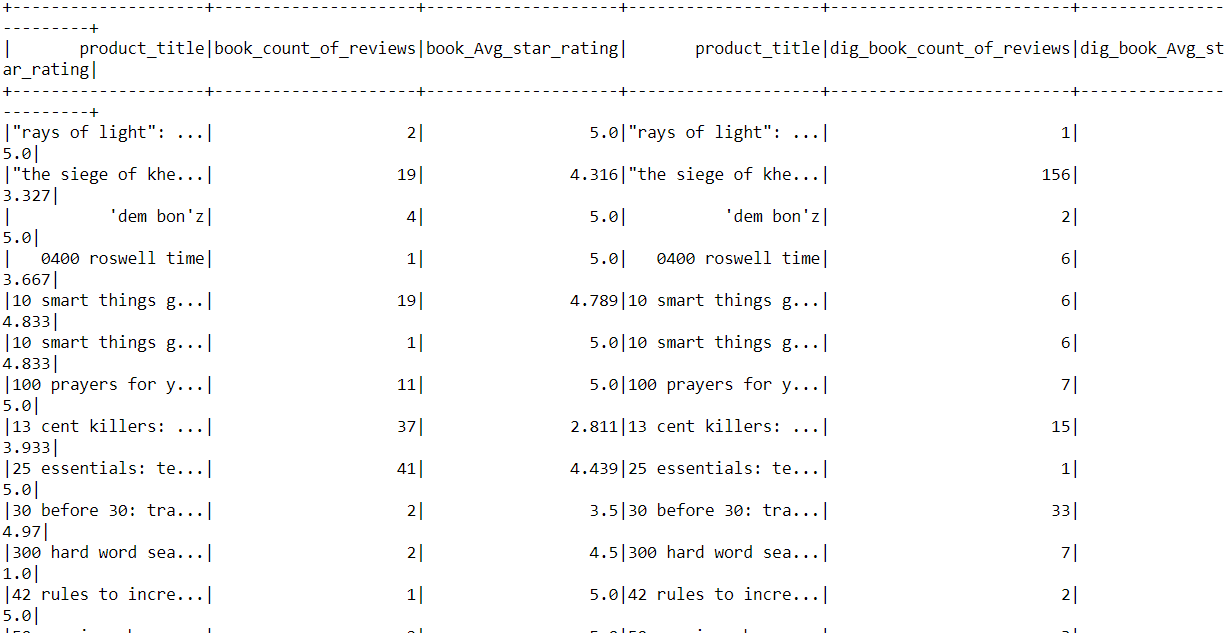
**Q3) Identify similar products (books) in both categories. Use "product\_title" to matchproducts. To account for potential differences in naming of products, compare titlesafter stripping spaces and converting to lower case.**

**Is there a difference in average rating for the similar books in digital and printed form?**

**Query :**



**Output :**



The above analysis shows that Printed Books has more higher rated stars as compared to the digital form.

**Q2) 3) 2) To answer #1, you may calculate number of items with high stars in digital form versus printed form, and vise versa. Alternatively, you can make the conclusion by using appropriate pairwise statistic**

**Query :**

high\_stars = F.col("book\_Avg\_star\_rating")>4

final.where(high\_stars).count()

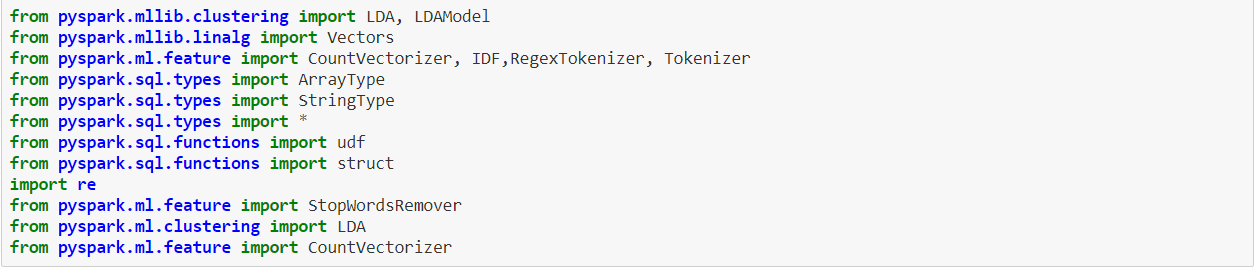
high\_stars\_1 = F.col("dig\_book\_Avg\_star\_rating")>4

final.where(high\_stars\_1).count()

**Output :**

245526

**Importing ML Libraries :**



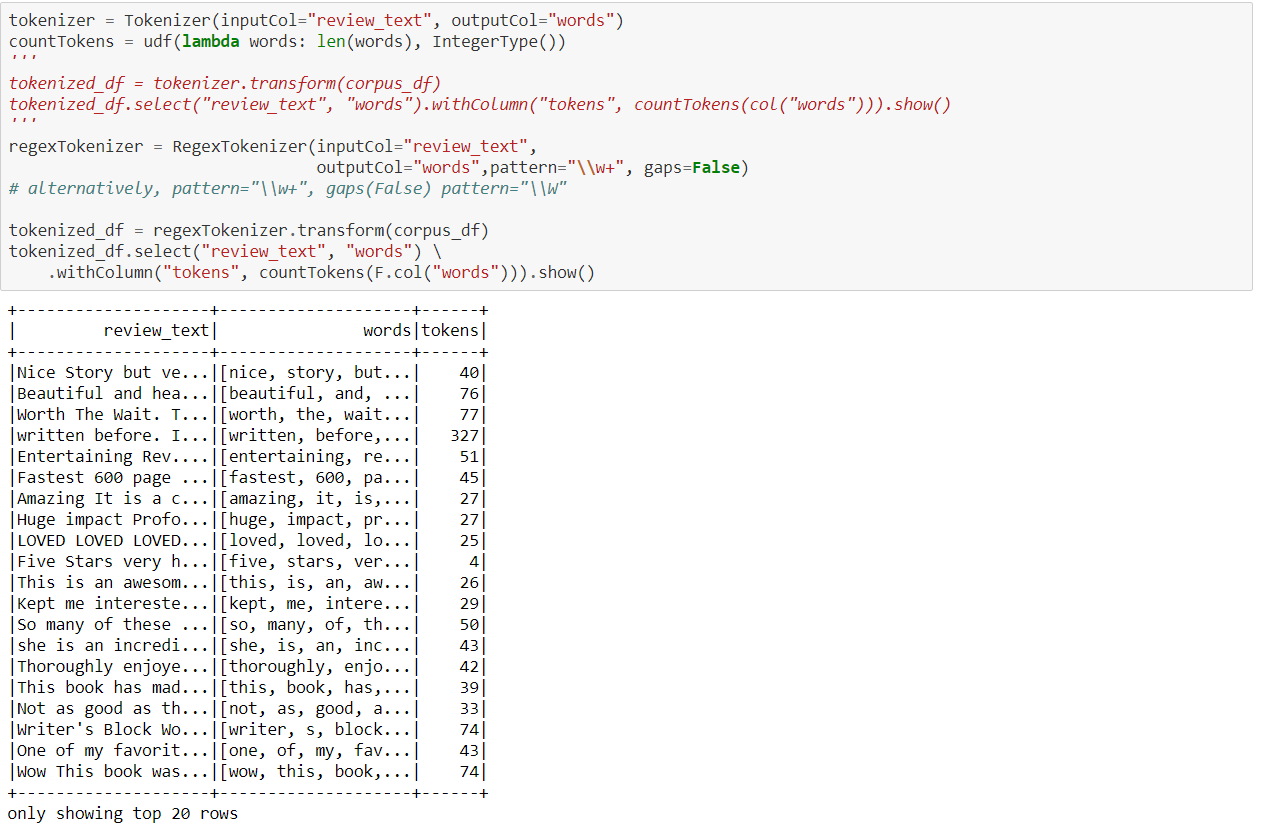
**Q2) 4) Using provided LDA starter notebook, perform LDA topic modeling for the reviews in Digital\_Ebook\_Purchase and Books categories. Consider reviews for the January of 2015 only.**

**Topic Modelling for reviews more than three stars**

**Query :**

In [11]:

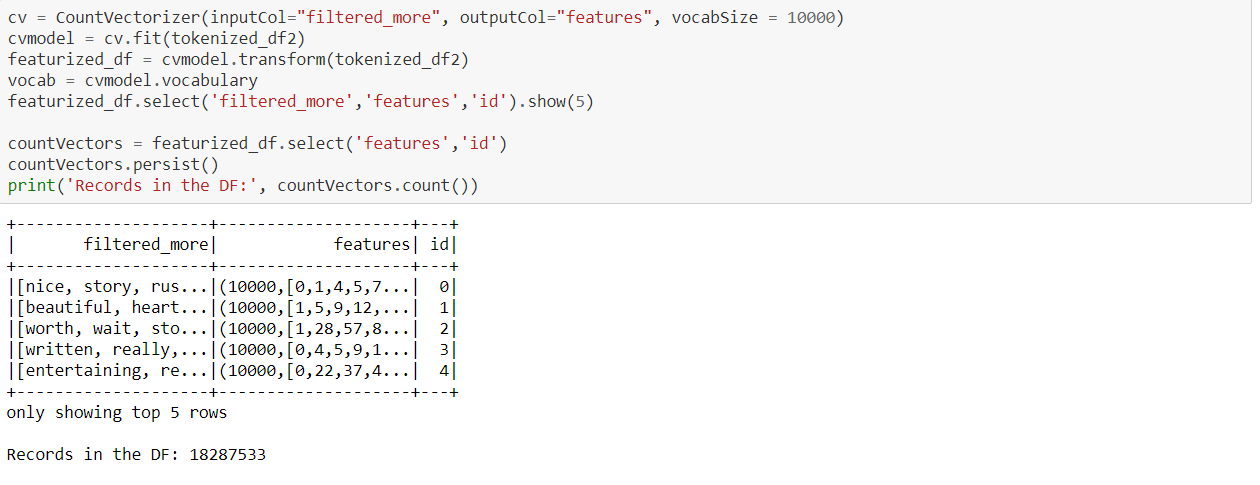
In [20]:



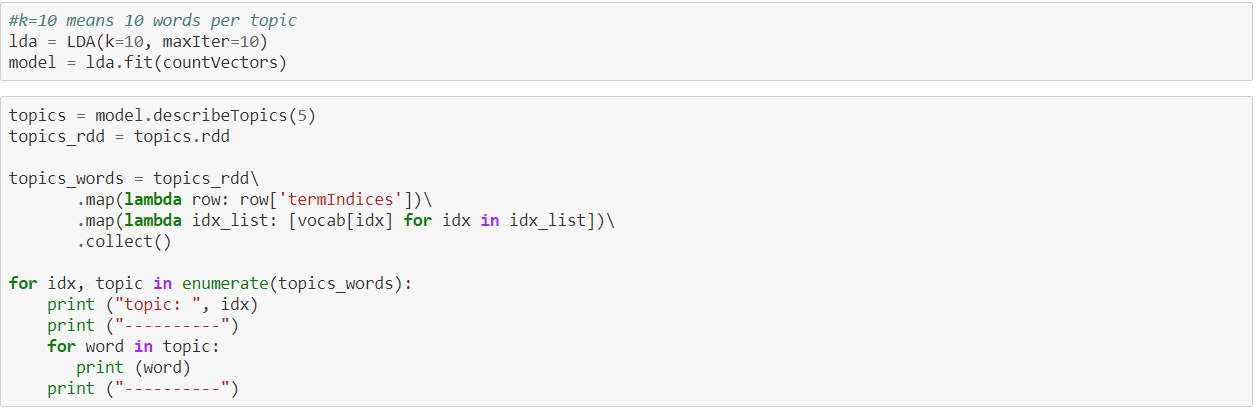
**Adding Stop Words**







**Performing LDA**

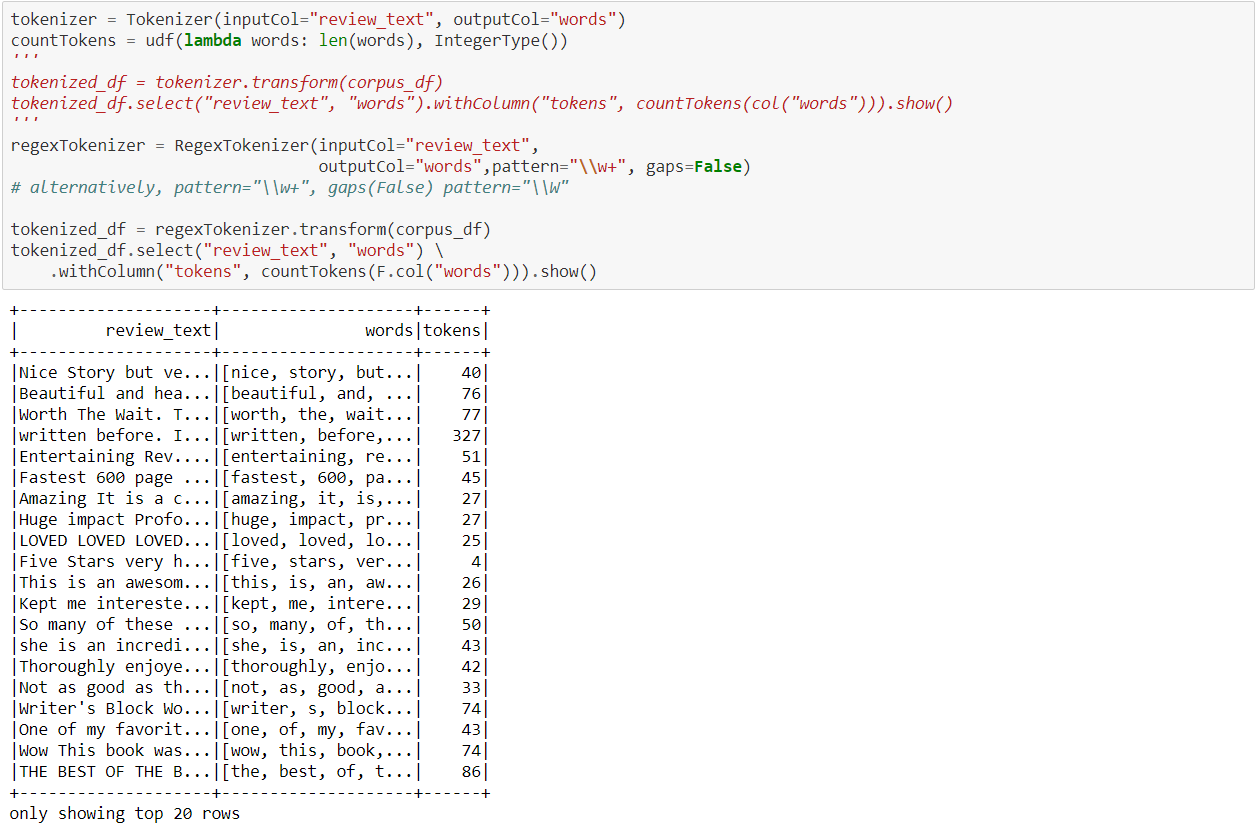


**Output :**



**Topic Modeling for Review Stars less than 3**

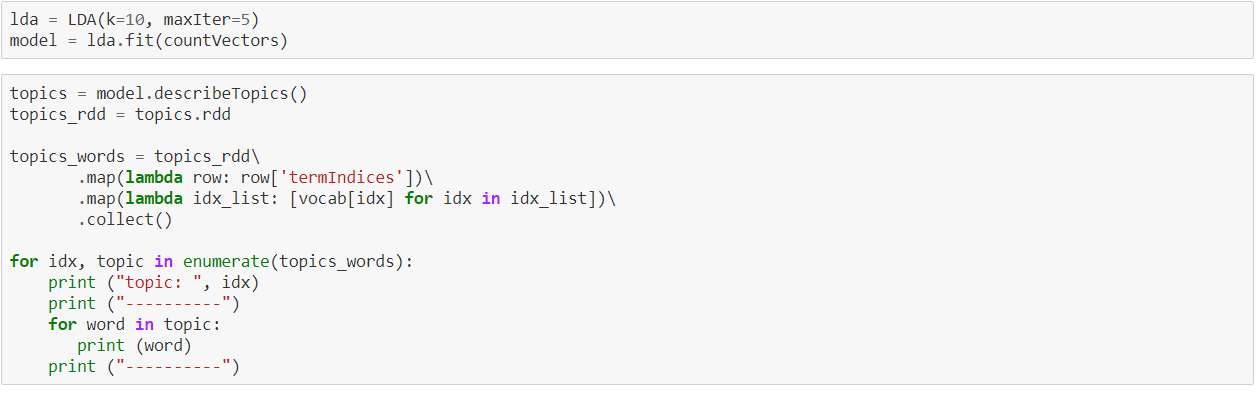








**Performing LDA :**



**Output:**



**Q2) 4) Does topic modeling provides good approximation to number of stars given in the review**

Ans : After running the LDAs for both the review stars greater than and less than 3 we can conclude that the topic modeling is does not give a good approximation to number of stars given in the review.

**Conclusion :**

Interpreting from the graph we could conclude that the count of reviews was more for Digital eBooks as compared to the Printed ones whereas the average star rating was more for the Printed Books as compared to the Digital eBooks.

Added to that we could also come to a conclusion that the topic modeling run for the LDA was not effective.

**References :**

Class Notes : Lectures 8,9,10,11

Class Mate : Prathamesh Namjoshi

Tableau

<https://spark.apache.org/docs/latest/api.html>