Assignment 2 Report

Aditya Jain

1 Part A

1.1 Part A.1

Figure 1 shows the code output. The count of even numbers is **514** and **496** for odd numbers.

```
(gbif_species_trainer) Adityas-MacBook-Air:Assignment2 adityajain$ python partA1.py
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/02/08 14:32:06 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... u
sing builtin-java classes where applicable
Count of even numbers is 514
Count of odd numbers is 496
```

Figure 1: Result of Part A.1

1.2 Part A.2

Figure 2 shows the output for salary sum per department. Following is the answer:

- Sales: 3,488,491
- Research: 3,328,284
- Developer: 3,221,394
- QA: 3,360,624
- Marketing: 3,158,450

```
(gbif_species_trainer) Adityas-MacBook-Air:Assignment2 adityajain$ /Users/adityajain/miniconda3/envs/gbif_species_trainer/bin/pytho
n /Users/adityajain/Dropbox/UofT_Studies/MIE1628/Assignment2/partA2.py
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/02/09 11:24:54 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
 applicable
/Users/adityajain/miniconda3/envs/gbif_species_trainer/lib/python3.8/site-packages/pyspark/python/lib/pyspark.zip/pyspark/shuffle.p
y:65: UserWarning: Please install psutil to have better support with spilling
/Users/adityajain/miniconda3/envs/gbif_species_trainer/lib/python3.8/site-packages/pyspark/python/lib/pyspark.zip/pyspark/shuffle.p
y:65: UserWarning: Please install psutil to have better support with spilling
Salary sum per department is [('Sales', 3488491), ('Research', 3328284), ('Developer', 3221394), ('QA', 3360624), ('Marketing', 315
8450)]
```

Figure 2: Result of Part A.2

1.3 Part A.3

Figure 3 shows the count result.

The count is as follows:

- *Shakespeare*: 22
- *GUTENBERG*: 100

```
(gbif_species_trainer) Adityas-MacBook-Air:Assignment2 adityajain$ python partA3.py
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/02/10 14:43:13 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
/Users/adityajain/miniconda3/envs/gbif_species_trainer/lib/python3.8/site-packages/pyspark/python/lib/pyspark.zip/pyspark/shuffle.py:65: UserWarning: Please install psutil to have better support with spilling
/Users/adityajain/miniconda3/envs/gbif_species_trainer/lib/python3.8/site-packages/pyspark/python/lib/pyspark.zip/pyspark/shuffle.py:65: UserWarning: Please install psutil to have better support with spilling
[('Shakespeare', 22), ('GUTENBERG', 100), ('WILLIAM', 128), ('WORLD', 98), ('COLLEGE', 98), ('why', 114), ('Lord', 402), ('Library', 4)]
(gbif_species_trainer) Adityas-MacBook-Air:Assignment2 adityajain$
```

Figure 3: Result of Part A.3

- *WILLIAM*: 128
- *WORLD*: 98
- *COLLEGE*: 98
- *why*: 114
- *Lord*: 402
- *Library*: 4

1.4 Part A.4

Figure 4 shows the output for the most and least count words. The least count words are (1 each): *anyone*, *restrictions*, *online*, *www*, *gutenberg*, *org*, *COPYRIGHTED*, *Details*, *guidelines*, and *Posting*. The max count words are: *the* (11410), *I* (9712), *and* (8941), *of* (7967), *to* (7742), *a* (5796), *you* (5360), *my* (4919), *in* (4803), and *that* (3864).

```
(gbif_species_trainer) Adityas-MacBook-Air:Assignment2 adityajain$ /Users/adityajain/miniconda3/envs/gbif_species_trainer/bin/python /Users/adityajain/Dropbox/UofT_Studies/MIE1628/Assignment2/partA4.py
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/02/10 11:12:48 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
/Users/adityajain/miniconda3/envs/gbif_species_trainer/lib/python3.8/site-packages/pyspark/python/lib/pyspark.zip/pyspark/shuffle.py:65: UserWarning: Please install psutil to have better support with spilling
/Users/adityajain/miniconda3/envs/gbif_species_trainer/lib/python3.8/site-packages/pyspark/python/lib/pyspark.zip/pyspark/shuffle.py:65: UserWarning: Please install psutil to have better support with spilling
The 10 words with least count are [('anyone', 1), ('restrictions', 1), ('online', 1), ('www', 1), ('gutenberg', 1), ('org', 1), ('COPYRIGHTED', 1), ('Details', 1), ('guidelines', 1), ('Posting', 1)]
The 10 words with max count are [('the', 11410), ('I', 9712), ('and', 8941), ('of', 7967), ('to', 7742), ('a', 5796), ('you', 5360), ('my', 4919), ('in', 4803), ('d', 4365), ('that', 3864)]
(gbif_species_trainer) Adityas-MacBook-Air:Assignment2 adityajain$
```

Figure 4: Result of Part A.4

2 Part B

2.1 Part B.1

The data consists of ratings given to 100 movies by 30 unique users. A total of 1501 ratings are given in the range of 1-5. The top 20 movies with the highest average rating are (in descending order): *32*, *90*, *30*, *23*, *94*, *49*, *29*, *18*, *52*, *62*, *53*, *92*, *46*, *68*, *87*, *2*, *69*, *27*, *88*, and *22*. The top 15 users who provided the most number of highest (5) ratings are (in descending order): *11*, *22*, *23*, *26*, *2*, *17*, *24*, *8*, *12*, *14*, *16*, *18*, *28*, *9*, and *21*. Figure 5 shows the code output.

```
(gbif_species_trainer) Adityas-MacBook-Air:Assignment2 adityajain$ /Users/adityajain/miniconda3/envs/gbif_species_trainer/bin/python /Users/adityajain/Dropbox/UofT_Studies/MIE1628/Assignment2/partB1.py
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/02/12 15:50:36 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
The total number of ratings given are 1501
/Users/adityajain/miniconda3/envs/gbif_species_trainer/lib/python3.8/site-packages/pyspark/python/lib/pyspark.zip/pyspark/shuffle.py:65: UserWarning: Please install psutil to have better support with sp
illing
The total number of movies are 100
The total number of unique users are 30
The range of movies ratings is [('3', 179), ('1', 941), ('2', 207), ('4', 99), ('5', 75)]
Top 20 movies with the highest average ratings are [('32', 2.92), ('90', 2.81), ('30', 2.5), ('23', 2.47), ('94', 2.47), ('49', 2.44), ('29', 2.4), ('18', 2.4), ('52', 2.36), ('62', 2.25), ('53', 2.25),
 ('92', 2.21), ('46', 2.2), ('68', 2.16), ('87', 2.13), ('2', 2.11), ('69', 2.08), ('27', 2.07), ('88', 2.06), ('22', 2.05)]
Top 15 users who provided the highest ratings are [('11', 8), ('22', 6), ('23', 6), ('26', 6), ('2', 5), ('17', 5), ('24', 5), ('8', 4), ('12', 4), ('14', 4), ('16', 4), ('18', 4), ('28', 3), ('9', 2),
('21', 2)]
(gbif_species_trainer) Adityas-MacBook-Air:Assignment2 adityajain$
```

Figure 5: Result of Part B.1

2.2 Part B.2

RMSE for (80, 20) split is 1.41 (figure 6) and for (70, 30) split is 1.5 (figure 7). Since the former has more training data, hence it has a lower test error.

```
(gbif_species_trainer) Adityas-MacBook-Air:Assignment2 adityajain$ /Users/adityajain/miniconda3/envs/gbif_spec
ies_trainer/bin/python /Users/adityajain/Dropbox/UofT_Studies/MIE1628/Assignment2/partB2.py
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/02/12 17:07:41 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using built
in-java classes where applicable
23/02/12 17:07:58 WARN InstanceBuilder$NativeBLAS: Failed to load implementation from:dev.ludovic.netlib.blas.
JNIBLAS
23/02/12 17:07:58 WARN InstanceBuilder$NativeBLAS: Failed to load implementation from:dev.ludovic.netlib.blas.
ForeignLinkerBLAS
23/02/12 17:07:59 WARN InstanceBuilder$NativeLAPACK: Failed to load implementation from:dev.ludovic.netlib.lap
ack.JNILAPACK
Root-mean-square error for (80.0, 20.0) is 1.41
(gbif_species_trainer) Adityas-MacBook-Air:Assignment2 adityajain$
```

Figure 6: Result of Part B.2.1

```
(gbif_species_trainer) Adityas-MacBook-Air:Assignment2 adityajain$ /Users/adityajain/miniconda3/envs/gbif_spec
ies_trainer/bin/python /Users/adityajain/Dropbox/UofT_Studies/MIE1628/Assignment2/partB2.py
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/02/12 17:06:49 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using built
in-java classes where applicable
23/02/12 17:07:14 WARN InstanceBuilder$NativeBLAS: Failed to load implementation from:dev.ludovic.netlib.blas.
JNIBLAS
23/02/12 17:07:14 WARN InstanceBuilder$NativeBLAS: Failed to load implementation from:dev.ludovic.netlib.blas.
ForeignLinkerBLAS
23/02/12 17:07:14 WARN InstanceBuilder$NativeLAPACK: Failed to load implementation from:dev.ludovic.netlib.lap
ack.JNILAPACK
Root-mean-square error for (70.0, 30.0) is 1.5
(gbif_species_trainer) Adityas-MacBook-Air:Assignment2 adityajain$
```

Figure 7: Result of Part B.2.2

2.3 Part B.3

Mean absolute error (MAE) is the average of the sum of absolute differences between the true and predicted values, thus measuring average of the residuals. Mean squared error (MSE) is the average of the sum of squared differences between the true and predicted values, thus measuring the variance of the residuals. Root mean squared error (RMSE) is the square root of MSE, which gives a measure of the standard deviation of residuals.

RMSE is more sensitive to large error values than MAE, hence the former is a better evaluation metric. Since RMSE is already used in the previous part, training will be evaluated using MAE in this part. Figure 8 and figure 9 shows the MAE for (80, 20) and (70, 30) split respectively. MAE is same in both the cases as expected because MAE is not sensitive to large error values. But we have seen in the previous part that (80, 20) split performs better on RMSE because it has more training data.

```
(gbif_species_trainer) Adityas-MacBook-Air:Assignment2 adityajain$ /Users/adityajain/miniconda3/envs/gbif_species_trai
ner/bin/python /Users/adityajain/Dropbox/UofT_Studies/MIE1628/Assignment2/partB3.py
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/02/12 18:30:17 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
classes where applicable
23/02/12 18:30:33 WARN InstanceBuilder$NativeBLAS: Failed to load implementation from:dev.ludovic.netlib.blas.JNIBLAS
23/02/12 18:30:33 WARN InstanceBuilder$NativeBLAS: Failed to load implementation from:dev.ludovic.netlib.blas.ForeignL
inkerBLAS
23/02/12 18:30:34 WARN InstanceBuilder$NativeLAPACK: Failed to load implementation from:dev.ludovic.netlib.lapack.JNIL
APACK
Mean absolute error (MAE) for (80.0, 20.0) split is 0.98
(gbif_species_trainer) Adityas-MacBook-Air:Assignment2 adityajain$
```

Figure 8: Result of Part B.3.1

```
(gbif_species_trainer) Adityas-MacBook-Air:Assignment2 adityajain$ /Users/adityajain/miniconda3/envs/gbif_species_trai
ner/bin/python /Users/adityajain/Dropbox/UofT_Studies/MIE1628/Assignment2/partB3.py
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/02/12 18:32:30 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
classes where applicable
23/02/12 18:32:48 WARN InstanceBuilder$NativeBLAS: Failed to load implementation from:dev.ludovic.netlib.blas.JNIBLAS
23/02/12 18:32:48 WARN InstanceBuilder$NativeBLAS: Failed to load implementation from:dev.ludovic.netlib.blas.ForeignL
inkerBLAS
23/02/12 18:32:48 WARN InstanceBuilder$NativeLAPACK: Failed to load implementation from:dev.ludovic.netlib.lapack.JNIL
APACK
Mean absolute error (MAE) for (70.0, 30.0) split is 0.98
(gbif_species_trainer) Adityas-MacBook-Air:Assignment2 adityajain$
```

Figure 9: Result of Part B.3.2

2.4 Part B.4

The parameters used for tuning the algorithm are: *maxIter*, *rank*, and *regParam*. *maxIter* is the maximum number of iterations the model is trained on the training set, *rank* is the rank of the factorization method, and *regParam* is the value of the regularization factor to prevent model overfitting. The range of values tested are: *maxIter* - [5, 10, 15], *rank* - [3, 6, 10, 14], and *regParam* - [0.1, 0.5, 2, 4]. This leads to a total of 48 combinations. Cross-validation is used to find the hyperparamters using number of folds k=3.

Figure 10 shows the output of the cross-validation and the best parameters found are: *maxIter* - 15, *rank* - 14, and *regParam* - 0.1. Training for more iterations leads to better learning of the underlying patterns in the data, while a small regularization parameter ensures that the model is not overfitting the training data. Increasing the rank in the ALS algorithm leads to putting more weight to graph connections in collaborative filtering.

```
(gbif_species_trainer) Adityas-MacBook-Air:Assignment2 adityajain$ python partB4.py
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/02/13 18:57:57 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
23/02/13 18:58:11 WARN BlockManager: Block rdd_22_0 already exists on this machine; not re-adding it
23/02/13 18:58:14 WARN InstanceBuilder$NativeBLAS: Failed to load implementation from:dev.ludovic.netlib.blas.JNIBLAS
23/02/13 18:58:15 WARN InstanceBuilder$NativeBLAS: Failed to load implementation from:dev.ludovic.netlib.blas.ForeignLinkerBLAS
23/02/13 18:58:15 WARN InstanceBuilder$NativeLAPACK: Failed to load implementation from:dev.ludovic.netlib.lapack.JNILAPACK
**Best Model**
  Rank: 14
  MaxIter: 15
  RegParam: 0.1
The least root mean squared error (RMSE) is 0.92
(gbif_species_trainer) Adityas-MacBook-Air:Assignment2 adityajain$
```

Figure 10: Result of Part B.4

2.5 Part B.5

The top 15 movie recommendations for user id 10 are: *92*, *12*, *19*, *71*, *81*, *91*, *34*, *46*, *93*, *32*, *95*, *82*, *64*, *65*, and *87*. The top 15 movie recommendations for user id 14 are: *43*, *85*, *58*, *90*, *2*, *41*, *70*, *30*, *60*, *77*, *87*, *61*, *18*, *74*, and *75*. Figure 11 shows the code output.

```
(gbif_species_trainer) Adityas-MacBook-Air:Assignment2 adityajain$ python partB5.py
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/02/14 15:08:00 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
The top 15 movie recommendations for user id 10 is [92, 12, 19, 71, 81, 91, 34, 46, 93, 32, 95, 82, 64, 65, 87]
The top 15 movie recommendations for user id 14 is [43, 85, 58, 90, 2, 41, 70, 30, 60, 77, 87, 61, 18, 74, 75]
(gbif_species_trainer) Adityas-MacBook-Air:Assignment2 adityajain$
```

Figure 11: Result of Part B.5