Assignment 3

Om Shri Prasath, EE17B113

April 2020

Introduction

The assignment consists of running the Alternating Least Squares (ALS) algorithm and Frequent Pattern Mining (FP Mining) algorithm using the Spark **mllib** Library on the given set of data.

ALS Algorithm

The data is given as a **'::'**-split data. It is split into its components and passed to the model, which gives the output and Test-RMSE.

For different regularization parameters and iterations, we found the minimum RMSE for 20 iterations and regularization parameter of 0.1 For different train-test splits, we found the minimum RMSE for 0.9/0.1 split.

The outputs are present in *ALS_out_1.txt* and *ALS_out_2.txt*.

FP Growth Algorithm - 1

The data is already given in a proper format to use. We split the data using **','**-data.

The output for a 0.16 *min_support* is :

```
FreqItemset(items=[u'mineral water'], freq=1788)
FreqItemset(items=[u'eggs'], freq=1348)
FreqItemset(items=[u'spaghetti'], freq=1306)
FreqItemset(items=[u'french fries'], freq=1282)
FreqItemset(items=[u'chocolate'], freq=1229)
```

The output is present in *fp1_out.txt*.

FP Growth Algorithm - 2

The data needs to be prepossessed for use in FP Growth. It is done by grouping the object via their *Invoice Number* after removing NaN or empty rows. After that, it is run through the FP Growth Algorithm.

The output for a 0.06 *min_support* is :

```
FreqItemset(items=[u'85123A'], freq=2020)
FreqItemset(items=[u'22423'], freq=1884)
FreqItemset(items=[u'85099B'], freq=1643)
FreqItemset(items=[u'47566'], freq=1399)
FreqItemset(items=[u'84879'], freq=1385)
```

The output is present in *fp2_out.txt* and the code for cleaning is in *fp_convert_data.py*.