# 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 maximum occuring pairs for a 0.04 *min_support* is :

```
FreqItemset(items=[u'spaghetti', u'mineral water'], freq=448)
FreqItemset(items=[u'chocolate', u'mineral water'], freq=395)
FreqItemset(items=[u'eggs', u'mineral water'], freq=382)
FreqItemset(items=[u'milk', u'mineral water'], freq=360)
FreqItemset(items=[u'ground beef', u'mineral water'], freq=307)
```

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 maximum occuring pairs (given as their ID's) for a 0.0236 *min_support* is :

```
FreqItemset(items=[u'22386', u'85099B'], freq=554)
FreqItemset(items=[u'22697', u'22699'], freq=557)
FreqItemset(items=[u'22726', u'22727'], freq=536)
FreqItemset(items=[u'22384', u'20725'], freq=530)
FreqItemset(items=[u'22383', u'20725'], freq=526)
```

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