MLSD: Assignment 1
Frequent itemsets and association rules
Similar items

– Due date: April 23, 2021 –

For each of the following exercises, you should implement and test the code locally using the provided Spark VM. You may use the provided sample datasets (or create your own) for developing.
Run the final code over the full dataset using an EMR cluster, through AWS Educate / Vocareum account. As a suggestion, use the CLI to launch the cluster, run the code, and collect the results.

Data for this assignment are available here: `https://bit.ly/2HUrW4T`

What to submit
In your final submission, include the code, with explanatory comments where necessary, and the result of the algorithms or a download link if the files are too large.

1. The file 'conditions.csv.gz' lists conditions for a large set of patients. The file contains the following fields, with multiple non-consecutive entries for each patient:

   START,STOP,PATIENT,ENCOUNTER,CODE,DESCRIPTION

   PATIENT is the patient identifier
   CODE is a condition identifier
   DESCRIPTION is the name of the condition

   You will need to reorganize the data before applying the algorithms.

   2.1. Using the A-Priori algorithm, obtain the 10 most frequent itemsets for sizes $k = 2$ and $k = 3$. Set a support threshold of 1000.

   2.2. Obtain associations between conditions by extracting rules of the forms $(X) \rightarrow Y$ and $(X, Y) \rightarrow Z$, with minimum standardised lift of 0.2.
   Write the rules to a text file, showing the standardised lift, lift, confidence and interest values and sorted by standardised lift.

2. Implement and apply LSH to identify similar movies based on their plots.
   Use the dataset available here: `http://www.cs.cmu.edu/~ark/personas/`