MLSD: Assignment 1
Frequent itemsets and association rules
Similar items

– Due date: March 23, 2023 –

For each of the following exercises, you should implement the solutions using Spark. Use small samples of the dataset for developing and initial testing, then run on the full data.

What to submit
For each exercise, submit a documented Jupyter notebook, a python script to run through spark-submit, and the results of the algorithm. If the results are too large, submit a download link instead.

The comments should explain the main steps of the solution with sufficient detail.

1. The file 'conditions.csv.gz' (available on the shared folder) 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 may prefer to reorganize the data before applying the algorithms.
   Try to use Spark for this as well.

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

   1.2. 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, sorted by standardised lift.

2. Implement and apply LSH to identify similar news articles.
   Use the dataset available on the shared folder.

   2.1. The number of bands and rows should be parameters. Select a combination that finds as candidates at least 90% of pairs with 85% similarity and less than 5% of pairs with 60% similarity.

   2.2. Implement a function that, given a news article, returns all other news articles that are at least 85% similar. You should make use of a pre-processed set of candidate pairs, obtained by LSH, and return only the ones that have Jaccard similarity–considering the shingles–above 85%.

   2.3. Using a sample of the dataset, evaluate the LSH method by calculating the Jaccard similarities and obtaining the percentage of false positives and false negatives.

   Note: You can average over multiple samples to get more robust values.