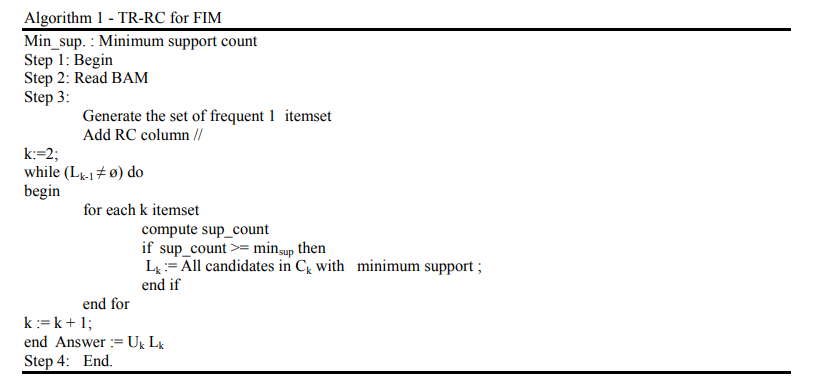
Association Rule Mining Project

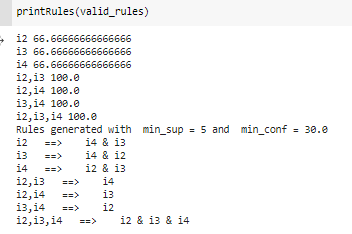
# Apriori Algorithm

Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently a itemset occurs in a transaction. Given a set of transactions, the goal is to find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction.

## Transaction Reduction

A transaction that does not contain any frequent k-itemsets cannot contain any frequent (k+1)-itemsets. Therefore, such a transaction can be marked or removed from further consideration. Implemented the below algorithm from scratch per the [paper](https://www.sciencedirect.com/science/article/pii/S1877050915004512). Refer transaction\_reduction.py for implementation.



Rules are generated for toy data as below with minimum support of 2 and confidence of 30%.

## Hash based Technique

When scanning each transaction in the database to generate the frequent 1-itemsets, L1, we can generate all the 2-itemsets for each transaction, hash (i.e., map) them into the different buckets of a hash table structure, and increase the corresponding bucket counts. A 2-itemset with a corresponding bucket count in the hash table that is below the support threshold cannot be frequent and thus should be removed from the candidate set.

We used itemset of size 2 for frequent set generation using the below hash function.

**H(x,y)= ((Order of first)\*10 + (Order of second)) mod 7**

Refer hashing.py for implementation details.

## Rule generation using WEKA

**Data Source:**

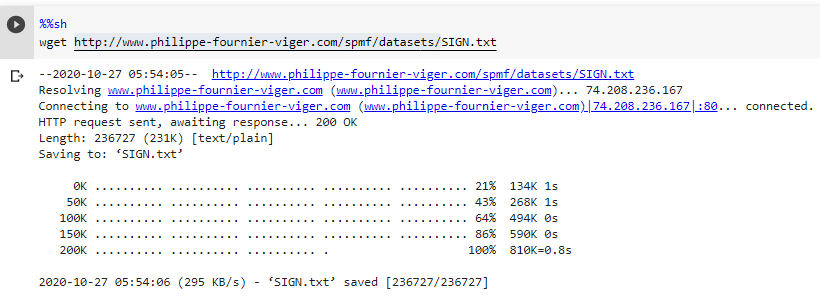
We used transactions data available at **SPFM** website<http://www.philippe-fournier-viger.com/spmf/index.php> **.** SPMF is an open-source softwareanddata mining librarywritten in Java, specialized in pattern mining(the discovery of patterns in data)

**Library Used:** WEKA library is used for rule generation using SPFM data.

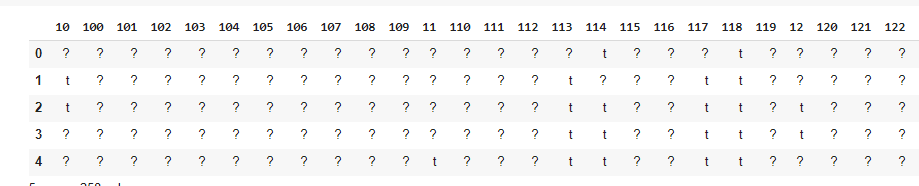
**Script**: SPFM2WEKA parser is written to convert data in SPFM format to WEKA format. Refer the script convertspfm2weka.py

Example#1: - SIGN data

A dataset of sign language utterance containing approximately 800 sequences and 267 items. The original dataset file in another format can be obtained [here](http://cs-people.bu.edu/panagpap/Research/asl_mining.htm) with more details on this dataset.



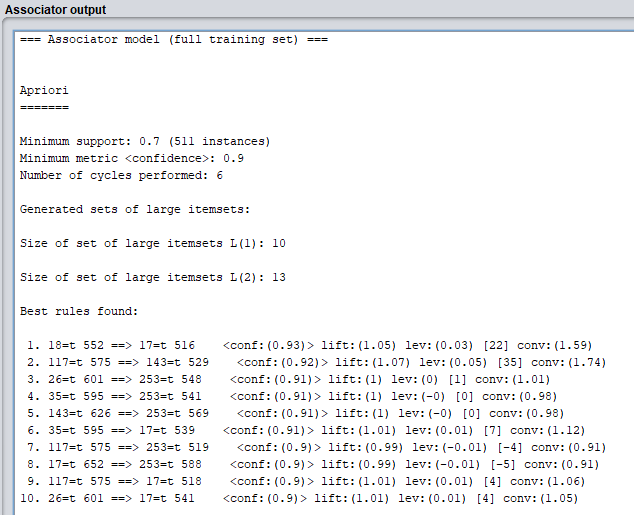
This raw data is converted to WEKA format, where Binary matrix of transactions and items is created.



Here each cell[i,j] represents if the transaction ‘i’ has item ‘j’ . Presence is represented as ‘t’ and absence as ‘?’.

**Rules**:

Below Rules generated in the WEKA tool with given support and confidence.



# FP-Growth Algorithm

Apriori Algorithm has slow performance and has below drawbacks

* At each step, candidate sets must be built.
* To build the candidate sets, the algorithm must repeatedly scan the database.

These two properties inevitably make the algorithm slower. To overcome these redundant steps, a new association-rule mining algorithm was developed named Frequent Pattern Growth Algorithm. It overcomes the disadvantages of the Apriori algorithm by storing all the transactions in a Tree Data Structure.

**Rules**:

Below Rules generated in the WEKA tool with given support and confidence.

