Assignment No. 1

* Problem Statement:
  For an organization, choose a set of business Processes. design Star/snowflake Schema for analyzing these Processes. Create a fact constellation Schema by combining them. Extract data from diff. data sources, apply suitable transations and load into designation table using ETL tool

* Objective : i.) understanding the basics of Star/ snowflake/ fact constellation Schema.
ii.) Learn how to install & use Pentaho tool.

* Outcome : i.) Pentaho was successfully installed
ii.) Process Schema set was decided & Schema
iii.) ETL was performed on dataset in Pentaho

* Theory :

- ETL Stands for Extract, Transform & load. It extracts data from diff. RDBMS source system, transform data by applying calculation & concatination then load data into data warehouse system. It is loaded in the form of dimention & fact table.

- Extraction

A staging area is req. during ETL load for various reasons.

(1.) The source system are only available for a specific period to extract data. this period of time is less than data load time. this Staging area allows to extract data from source system & keep it in staging area before the time slot.

(2.) Staging area is also req. when you want to get data from multiple source together

- Transformation

In data transformation, you apply a set of functions on extracted data to load it into the target system.

Data which does not req. any transform is known as direct move pass thrw data

Diff transformations on extracted data from the source system can apply, customised to the business decided.

- Load

In this phase, data is loaded into the end target system, either a flat file or a Data warehouse.

⊙ Star Schema ⇒

- Every dimention represented with only one dimention table.
- dimention table should contain a set of attributes
- dimention table joined to fact table using foreign k
- dimention table are not joined to each other
- fact table contains key & measure
- easy to understand, widely supported by BI tools
- optimal disc usage, denormalized structure, faster queries
- dimention tables are not normilized
- cube processing is faster

⊙ Snowflake Schema is an extension of Star Schema, & it adds additional dimentions

- dimention table normilized, which splits data into additional tables
- uses smaller disc space
- reduced query performance due to additional table
- complex database design
- low level data redundancy.

⊙ Pentaho data integration provides the ETL capabilities that facilitates the process of consuming, cleaning & storing data using a uniform & consistent format that is accessable & relevent to end users.

* This tool enables users to -
1) load huge datasets into databases
2) clean data with steps ranging from ver
simple to very powerful complex transformati
3) data warehouse populations, and more

* The PDI client 'Spoon' is a visual, drag
& drop, no-code interface that is used
to perform ETL.

* Conclusion:
1) Pentaho Data integration tool was
successfully installed & used on a Sales
dataset to successfully Extract, transfor
& load in MySQL.
2) A business was picked & a set of
business processes developed & represented
using Snowflack Schema.

Assignment No. 2

* Problem Statement :
Consider a suitable dataset, for clustering of data instances in diff. groups, apply diff. clustering techniques (min. 2). Visualize the clusters using suitable tools.

* Objective : i) understand various clustering types & how to implement the same.
ii) use Python libraries & appropriate data sets to perform clustering & visualize the same.

* Outcome : Understood K-means, Hierarchical clustering & perform it on dataset cars data by brand.

* Theory :

⊙ Clustering or cluster analysis is a task of grouping a set of objects in such a way that objects in the same group (cluster) are more similar to each other than to those in each other.
⊙ Clustering is the main task of exploratory data mining & a common

technique for statistical data analysis, and is used in many fields.

⊙ K-means clustering is a type of unsupervised learning, which is used when you unlabeled data (i.e. data without defined categories or groups). the goal of this algorithm is to find groups in the data, with the number of groups represented by variable K. the algorithm works iteratively to assign each data point to one of K groups based on the feature provided. data points are clustered based on feature similarity. the result of this algorithm is -

⇒ the centroid of K-clusters, can be used to label new data

⇒ labels from the training data (each data point is assigned to single cluster)

⊙ Rather than defining groups, before looking at the data, clustering allows you to find & analyze the groups that have formed organically.

⊙ Hierarchical clustering is an algorithm that groups similar objects into groups called cluster. the end point is a set of cluster, where each cluster is distinct from each other, and the object

within each cluster are broadly similar to each other.

⊙ Given a set of N items to be clustered and an N×N distance metrix, the basic process of hierarchical clustering is:

→ assign each item to its own cluster, So now there are N clusters; let the dist. between the cluster equal the dist. (Similarities) between the items they contain.

→ Find the closest (most similar) pair of clusters and merge them into a single cluster

→ compute the distances (similarities) between the new cluster & each of the old cluster

→ Repeat above 2 steps until all clusters are clustered into a single cluster of size N.

⊙ The default distance measured in the Euclidean dist. which is the Sq. root of the Sum of the differences.

⊙ In the agglomerative clustering approach, there are 4 possible methods, ward's method being one of them. It says that the dist. between two clusters, A & B, is how much the Sum of Squares will increase when they are merged

- The dataset used was 'mall-customer' which contains various parameters.
- these entries were clustered using the algorithm.

* Conclusion:

K-means and hierarchical (agglomerative) clustering techiques were understood, successfully implemented, & the req. output was obtained.

Assignment No. 3

* Problem Statement:
  Apply a-Priori algorithm to find freq. occuring items from given data and generate strong association rules using support & confidence thresholds.

* Objective: Model associations between Products by determining sets of item freq. Purchased together & building association rules to drive recommendations.

* Outcome: Demonstrated market basket analysis using a-priori algorithm to find freq. occuring items from given data & generate strong association rules using support & confidence algo.

* Theory:
  Association rule mining finds interesting associations and relationships between large sets of data items. this rule shows how frequently an item occurs in a transaction. It is defined as an implementation expression of the form $X \rightarrow Y$ where X & Y are any 2 item sets.

Market Basket analysis is one of the key techniques used by large retailers to show association between items. It allows retailers to identify relationships between the items that people buy together freq, So that items can be strategically placed next to each other to boost sales.

Some definations :

↳ Support count : freq. of itemset occurence

↳ freq. Itemset : Itemset whose support >= minup threshold

The association rule evaluation metrix are

① Support (s) : The number of transactions that include items in the {x} & {y} parts of the rules as percentage of total number of transaction.

② confidence (c) : It is the ratio of number of transactions that includes all items in an itemset {A} as well as the transaction that include all items in {A} to the number of transaction that includ all items in {A}

③ Lift (l) : The lift of the rule $X \rightarrow Y$ is the confidence of the rule divided by

the expected confidence, assuming {X} & {Y} are indipendent of each other. expected confidence is confidence divided by freq. {Y}. Itemset occur together (i) as expected if lift = 1 (ii) more than expected if lift > 1 and (iii) less than expected if lift < 1

(A) Apriori Algorithm

→ It is an algorithm for freq. itemset mining & association rule learning over relational databases, first Possesed by Agrawal & Srikant in 1994.

It is designed to operate on databases containing transactions, with each transaction being a set of items (itemsets).

Given a threshold C, the algo identifies the itemsets, which are subsets of atleast C transactions in database

It uses a bottom-up approach, where frequent subsets are extended one at a time (called candidate generation), and groups of candidate are tasked against the data. the algo terminates when no other successful extension are found.

A-Priori used BFS & a hash tree to count candidate item sets efficiently. It generates K-itemsets of length k-1; then it prunes the candidate, which have

in-frequent sub pattern. After that it scans the transaction database, to determine freq. item sets among the candidate.

* Conclusion:

Thus frequently occuring items from given market basket dataset and strong association rules using support & confidence thresholds found using a-priori algorithm.

Assignment No. 4

* Problem Statement :
  Consider a suitable text dataset. Remove Stop words, apply Stemming & feature Selection techniques to represent documents as vectors classify documents & evaluate Precision, Recall.

* Objective : Learn how to tokenize & filter a document into its diff. words and then do word count for each word in doc. Apply stemming, feature Selection on doc (text)

* outcome : demonstrated text Processing using nltk, understood vectorizing & removing Stop words.

* Theory :
  - NLP is a subfield of linguistic computer Science & AI concerned with interaction b/w computer & Human language, in Particular how to Program computer to Process & Analyse large amt of Natural language dt
  - In computing Stop words are words that are filtered out before or after the natural language data (text) are

Processed. while 'stop words' typically refer to the most common wand, there is no universal list of stop words.

⊙ STEMMING ⇒ for grammitical reasons, doc are going to use diff. forms of a word (e.g. organized, organizing, etc). Additionally there are families of derivitavelly relative words with similar meaning like democracy democracy & democratization.

→ The goal of stemming (and lemmatization) is to reduce inflectional forms & sometimes derivationally linked forms of a word to its common base form

am, are, is → be

car, cars, cars', car's → car

→ when applied to a doc, result will be somewhat like this

the boy's cars are diff. color

the boy car be differ color

→ Stemming is a more coarse process, a heuristic that chops off the end of a word in the hope of achieving the goal correctly more of time, and often includes removal of derivational affixes

→ Feature selection is a process of selecting a subset of the terms occurring

in the remaining set and using only this subset as feature in text classification.

→ It serves 2 purposes. first, it makes training & applying a classifier more efficient by decreasing the size of V.cab. this is of particular importance to classifiers that are expensive to train.

→ second, feature selection often increases classification accuracy by eliminating noise feature these features when added to docs representation increases the classification error on new data.

→ Vectorization is a process of converting text into machine readable form. words are represented as 'vectors' (numerically)

→ Count vectorizer (one-hot encoding) involves counting the number of occurence for each word occuring in the doc.

→ Idea behind it is simple. vector is created having as many dimentions as there are distinct words in text/doc/collection of docs being used. each unique word has a unique dimention & is represented by a 1 in the dimention with 0s everywhere else. This results in huge & sparse vectors that capture no relational data.

→ If IDF vectors are related to one hot encoded vectors, but instead of just featuring a count, they feature numerical representation, where words aren't just present or not present instead, they are represented by their term freq. multiplied by inverse doc. freq.

→ In simpler terms, words that occur everywhere should be given very little weight or significance, coz they don't provide a large amt. of value, however if a word appears very little or freq. frequently but only in specific place, then they are possible of higher significance

→ the downside is there is no capture of semantic relatedness, this is solved with the co-occurence. matrix or a neural probabilistic model.

* Conclusion:

Text docs were tokenized, filtered, vectorized, & thus successfully processed basics of NLP were understood.