Subject: 19CSE305

Lab Session: 02

Notes:

1. Please read the assignment notes carefully and comply to the guidelines provided.
2. If you have not completed the prerequisite assignments, please complete them before starting these assignments.
3. Refer to your class notes on encoding, normalization, distance and similarity metrics
4. Data for the experiment is provided in excel file "***19CSE305_LabData_Set3.xlsx***". This file has 2 worksheets containing 2 sets of data.
    a. "**thyroid0387_UCI**" contains thyroid data obtained from UCI repository. A question mark in the value means the value is unknown / missing.
    b. "**marketing_campaign**" contains market data for customer behavior analysis

Main Section (Mandatory):

A1. **Data Exploration:** Load the data available in "**thyroid0387_UCI"** worksheet. Perform the following tasks:

- Study each attribute and associated values present. Identify the datatype (nominal etc.) for the attribute.
- For categorical attributes, identify the encoding scheme to be employed. (Guidance: employ label encoding for ordinal variables while One-Hot encoding may be employed for nominal variables).
- Study the data range for numeric variables.
- Study the presence of missing values in each attribute.
- Study presence of outliers in data.
- For numeric variables, calculate the mean and variance (or standard deviation).

A2. **Data Imputation:** employ appropriate central tendencies to fill the missing values in the data variables. Employ following guidance.

- Mean may be used when the attribute is numeric with no outliers
- Median may be employed for attributes which are numeric and contain outliers
- Mode may be employed for categorical attributes

A3. **Data Normalization / Scaling:** from the data study, identify the attributes which may need normalization. Employ appropriate normalization techniques to create normalized set of data.

A4. **Similarity Measure:** Take the first 2 observation vectors from the dataset. Consider only the attributes (direct or derived) with binary values for these vectors (ignore other attributes). Calculate the Jaccard Coefficient (JC) and Simple Matching Coefficient (SMC) between the document vectors. Use first vector for each document for this. Compare the values for JC and SMC and judge the appropriateness of each of them.

$$JC = (f11) / (f01 + f10 + f11)$$
$$SMC = (f11 + f00) / (f00 + f01 + f10 + f11)$$

$f11$= number of attributes where **the attribute carries value of 1 in both the vectors.**

A5. **Cosine Similarity Measure:** Now take the complete vectors for these two observations (including all the attributes). Calculate the Cosine similarity between the documents by using the second feature vector for each document.

If **A** and **B** are two document vectors, then

$$\cos(\mathbf{A}, \mathbf{B}) = \langle \mathbf{A}, \mathbf{B} \rangle / \|\mathbf{A}\| \, \|\mathbf{B}\|$$

$$< A, B >= \sum_{k=1}^{n} a_k * bk$$

$\|\mathbf{A}\|$ and $\|\mathbf{B}\|$ are lengths of vectors A & B

A6. **Heatmap Plot:** Consider the first 20 observation vectors. Calculate the JC, SMC and COS between the pairs of vectors for these 20 vectors. Employ similar strategies for coefficient calculation as in A4 & A5. Employ a heatmap plot to visualize the similarities.

Suggestion to Python users →

```
import seaborn as sns
```

```
sns.heatmap(data, annot = True)
```

Optional Section:

O1. Repeat experiments (A4 to A6) by taking observation samples from different regions of the data. You may try a random sampling to pick 20 vectors randomly form the set.

O2. Try the same exercise on data available on "**marketing_campaign**" worksheet.

Report Assignment:

1. Write your understanding of your project in the introduction section of the report.
2. Download at least 10 published papers (from IEEE Xplore, Springer, Elsevier or Science Direct) for your project. Study these papers use them for literature survey section of your report.
3. Using the learnings so far, design a system that could be used for customer / patient segmentation. Enrich your answer with:
    a. Flow diagram to depict the data flow. Example: input handling, preprocessing, similarity scoring, output.
    b. Architecture diagram for the system should be in methodology or system description section. Detail what happens in each block.
    c. define parameters to be used in the system; assign values for these parameters and justify them.

Since this is a Design work, the solution should be provided in the methodology section of the IEEE format of report. The results may be taken from above experiments and discussed to conclude the paper.