Subject: 21DS602

Lab Session Date: 29-10-2021

Main Section (Mandatory):

Please refer to the "***Sheet1***" worksheet of ***Lab Session2 Data.xlsx*** file provided. The data contains text reports from an assignment from 37 students (S1 to S37).

Following nomenclature to be used for this assignment

1. Report text available against each student → document or instance
2. Words in the report text -> tokens
3. Splitting a sentence or document into constituent words → tokenization
4. Stop-words → commonly used words; meaning could still be retained without them
5. White spaces, commas and full stops to act as token splitter delimiters.

M1. Load the documents for all the 37 students. Tokenize all the documents and store the tokens. You may:

- use the NLTK package for tokenization, or
  (Suggestion: Python users →
  `import nltk`
  `from nltk.tokenize import word_tokenize`*).*
- write your own code for splitting the document into tokens (words)

M2. Merging the tokens from all documents, create a master list of distinct tokens available across all documents. Let us call this as ***"token population"***

M3. Load the ***stop-words*** using NLTK package. Study these ***stop-words***. What do you think they represent?

(Suggestion: Python users → ***`from nltk.corpus import stopwords`****).*

M4. Create a ***"bag-of-words"*** from the ***"token population"*** by removing the ***stop-words***.

M5. For each document / instance, create 2 feature vectors as follows

- First vector attributes indicate the presence / absence of tokens from ***bag-of-words*** in that document
- Second vector attributes indicate the count of presence of each word from ***bag-of-words*** in the document
- Do this for all documents. After this is done, you should have 37 vectors each of first and second kind.

M6. Calculate the Jaccard Coefficient between the document vectors. Use first vector for each document for this. Jaccard coefficient between two binary vectors is defined as:

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

$f11$ = number of attributes where $\mathbf{x}$ was 1 and $\mathbf{y}$ was 1; $\mathbf{x}$ & $\mathbf{y}$ are vectors

M7. 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

M8. Plot the Jaccard and cosine matrices and see the results.

Suggestion to Python users →

```
import seaborn as sns
sns.heatmap(data, annot = True)
```

Optional Section:

O1. Load the "**Purchase Data**" worksheet of **Lab Session1 Data.xlsx** (provided last week). Take the first 2 purchase records provided. Calculate the following distances between these purchase records:

- Manhattan or city-block distance
- Euclidean distance

Observe the values and draw your inferences.

O2. For the same two above records, calculate the Minkowski distance (formula provided below) for orders (r values) 1 to 10. Make a plot of the Minkowski distances against the r values and observe the nature of the graph.

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}$$

O3. Load the "**IRCTC Stock Price**" worksheet of **Lab Session1 Data.xlsx**. Consider the "**High**" and "**Low**" columns as 2 vectors (249-dimensional vector space). Repeat exercises O1 & O2 on these vectors.

Report Assignment:

Q1. What are your observations from the similarity study conducted in this experiment? Please design a plagiarism detection system based on your learnings in this exercise. Mention the advantages and disadvantages of your designed system with respect to performance, efficiency and accuracy. Please highlight failure scenarios of your designed system.

Q2. Mention a few use cases where similarity measure is useful for classification / categorization and may be used instead of distance measure.