Experiment No.5

Title: Applying similarity measures on the numeric datasets

Batch: **B4** **Roll No.: 16010420133** **Experiment No.: 5**

Aim: Applying similarity measures on the numeric datasets

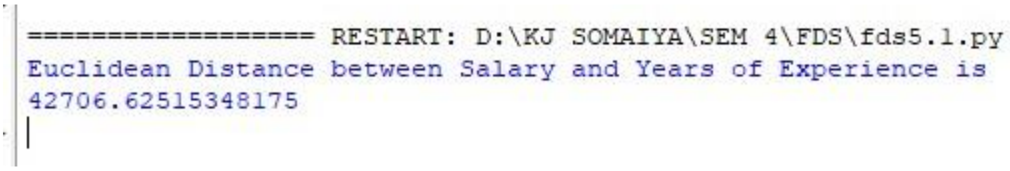
Resources needed: DataSet(csv), Jupyter Notebook

Results:

Identify the suitable attributes to apply the numeric similarity measures and write python code to calculate Euclidean, Manhattan similarity measures on it. Euclidean Distance

```
 from pandas import *  import pandas as pd
import numpy as np  sum=0  data =
pd.read_excel("HRDataset_v2.xlsx")
sal=data['Salary'].tolist()

years=data['YearsOfExperience'].tolist(
)  sal_arr=np.array(sal)
years_arr=np.array(years)
length=len(sal)  #print(length)  diff=[]
diff=sal_arr-years_arr
sq=np.square(diff)  sumsq=np.sum(sq)
squareroot=np.sqrt(sumsq)
 print("Euclidean Distance between Salary and Years of Experience is ")
print(squareroot)
```



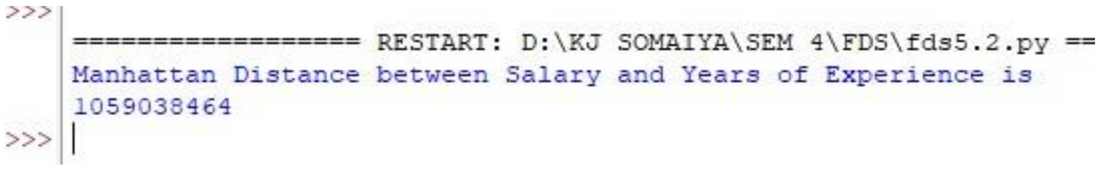
```
================== RESTART: D:\KJ SOMAIYA\SEM 4\FDS\fds5.1.py
Euclidean Distance between Salary and Years of Experience is
42706.62515348175
```

Manhattan Distance

```
def distancesum (x, y, n):
        sum = 0  for i in range(n):
       for j in range(i+1,n):
                        sum += (abs(x[i] - x[j]) +  abs(y[i] -
                                           y[j]))

        return sum

from pandas import *  import pandas as pd  import numpy as np
sum=0  data = pd.read_excel("HRDataset_v2.xlsx")
x=data['Salary'].tolist()  y=data['YearsOfExperience'].tolist()  n = len(x)
print("Manhattan Distance between Salary and Years of Experience is
")  print(distancesum(x, y, n) )
```



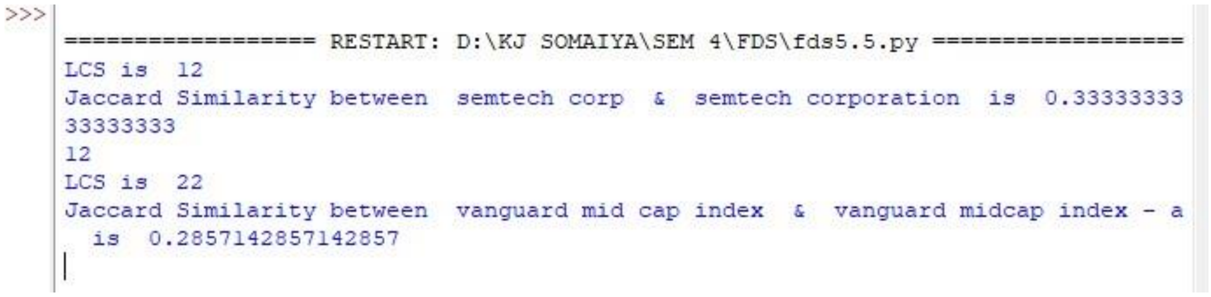
```
>>>
    ================== RESTART: D:\KJ SOMAIYA\SEM 4\FDS\fds5.2.py ==
    Manhattan Distance between Salary and Years of Experience is
    1059038464
>>> |
```

Identify the suitable attributes to apply the textual similarity measures and write python code to calculate Longest common subsequence, edit distance similarity measures on it. LCS & Jaccard Similarity

```
from pandas import *  import
pandas as pd  import numpy as np
data = pd.read_csv("test.csv")
data1=data['description_x'].tolist()
data2=data['description_y'].tolist()


def lcs(X,Y,m,n):
   if m==0 or n==0:
       return 0  elif X[m-
   1]==Y[n-1]:  return
   1+lcs(X,Y,m-1,n-1)
   else:
       return max(lcs(X,Y,m,n-1),lcs(X,Y,m-1,n))


for i in range(len(data1)):
    print("LCS is ",len(data1[i]))
    X=data1[i]
   Y=data2[i]
   m=len(data1[i])
   n=len(data2[i])
    X_tokens=set(X.lower().split())      Y_tokens=set(Y.lower().split())      jaccard_similarity=
   len(X_tokens.intersection(Y_tokens))/len(X_tokens.union(Y_tokens))          print("Jaccard
   Similarity between ",X, " & ", Y, " is ",jaccard_similarity)  if m==0 or n==0:  print('0')  elif
   X[m-1]==Y[n-1]:
       print(1+lcs(X,Y,m-1,n-1))
   else:
       print(max(lcs(X,Y,m,n-1),lcs(X,Y,m-1,n)))
```



```
>>>
================== RESTART: D:\KJ SOMAIYA\SEM 4\FDS\fds5.5.py ==================
LCS is  12
Jaccard Similarity between  semtech corp  &  semtech corporation  is  0.33333333
33333333
12
LCS is  22
Jaccard Similarity between  vanguard mid cap index  &  vanguard midcap index - a
  is  0.2857142857142857
```

Questions:

1. What are the different applications of Numeric similarity measure?

Many real-world applications make use of similarity measures to see how two objects are related together. We can use these measures in the applications involving Computer vision and Natural Language Processing, for example, to find and map similar documents. One important use case here for the business would be to match resumes with the Job Description saving a considerable amount of time for the recruiter. Another important use case would be to segment different customers for marketing campaigns using the K Means Clustering algorithm which also uses similarity measures.

2. What are the different applications of finding similarity between textual attributes?

Biomedical Informatics: To develop the biomedical ontologies namely the Gene Ontology we used the semantic similarity. Similarity methods are mainly used to compare the genes and they can also used in other bio-entities

Geo-Informatics: Similarity measure also used to find the similarities between geographical feature type ontologies. Several tools are available to do this task such as (i) The OSM Semantic Network used to compute the semantic similarity of tags in OpenStreetMap . (ii) Similarity Calculator is used to find the similarity between two geographical concepts in the Geo-Net-PT ontology and (iii) SIM-DL similarity server computes the similarity between geographical feature type ontologies.

Natural Language Processing: It is field of Computer Science and linguistics. There are several fields where STS can play an important role directly or indirectly such as sentiment analysis, natural language understanding and machine translation.

Outcomes:

CO 2 Comprehend descriptive and proximity measures of data

Conclusion: (Conclusion to be based on the objectives and outcomes achieved)

I have implemented the python code for finding the similarity between textual and numerical data.

Grade: AA / AB / BB / BC / CC / CD /DD

Signature of faculty in-charge with date

References:

Books/ Journals/ Websites:

1. Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3nd Edition
2. Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. Introduction to data mining.
   Pearson Education India, 2016.