

**Batch: A2 Roll No.: 16010421075 Experiment No.: 5**

**Aim:** Applying similarity measures on the numeric datasets

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Resources needed:** Any programming language, any data source

(RDBMS/Excel/CSV)

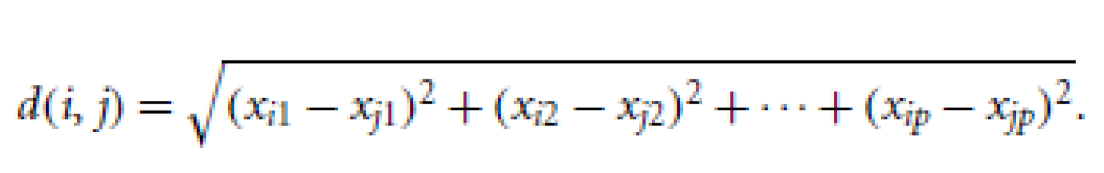
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Theory:**

**Similarity measures:**

Similarity measures for numeric attributes include the *Euclidean, Manhattan*, and *Minkowski distances*.

The most popular distance measure is Euclidean distance (i.e., straight line or “as the crow flies”).  be two objects described by *p* numeric attributes. The Euclidean distance between objects *i* and *j* is defined as,

……………….…..……..(1)

Another well-known measure is the Manhattan (or city block) distance, named so because it is the distance in blocks between any two points in a city (such as 2 blocks down and 3 blocks over for a total of 5 blocks). It is

defined as,

……………..........................(2)

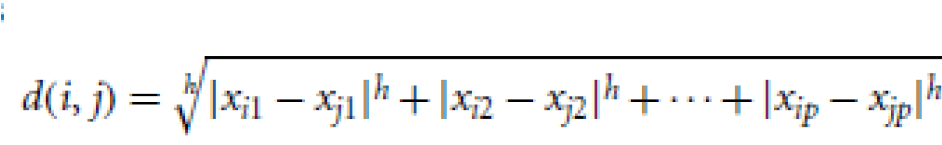
Both the Euclidean and the Manhattan distance satisfy the following mathematical properties:

Non-negativity: Distance is a non-negative number.

Identity of indiscernible: The distance of an object to itself is 0.

Minkowski distance is a generalization of the Euclidean and Manhattan

distances. It is defined as,

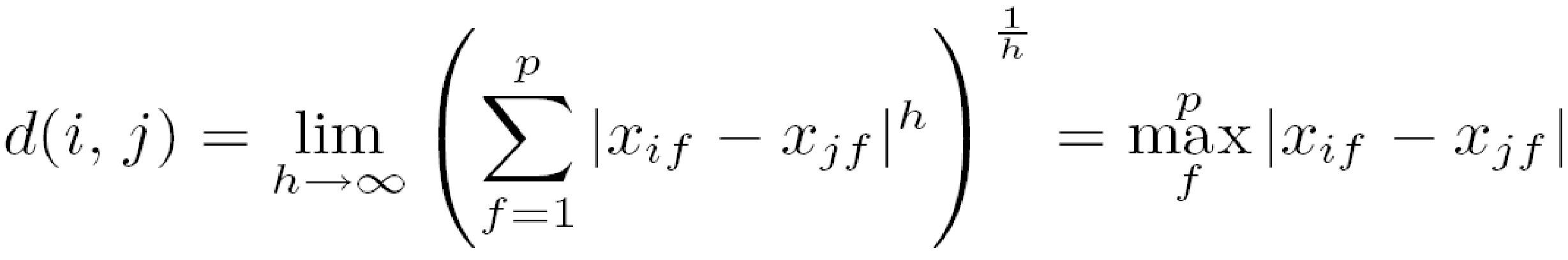
………………………………….(3)

Where *h* is a real number such that *h* >= 1. It represents the Manhattan distance when *h* = 1 and Euclidean distance when *h* = 2.

*When h* →∞., its a “supremum” (Lmaxnorm, L∞norm) distance.

– This is the maximum difference between any component

(attribute) of the vectors



\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Procedure / Approach /Algorithm / Activity Diagram:**

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

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Results: (Program printout with output / Document printout as per the format)**

import pandas as pd

import math

df = pd.read\_csv("Fire-Incidents.csv")

d1 = df["Incident\_Station\_Area"].tolist()

d2 = df["Civilian\_Casualties"].tolist()

a, b = 0, 0

for i in range(0, len(d1)):

  x = d1[i]

  y = d2[i]

  n1 = x-y

  a = a + abs(n1)

print("manhattan distance: ", a)

for i in range(0, len(d2)):

  a = d1[i]

  b = d2[i]

  n2 = a-b

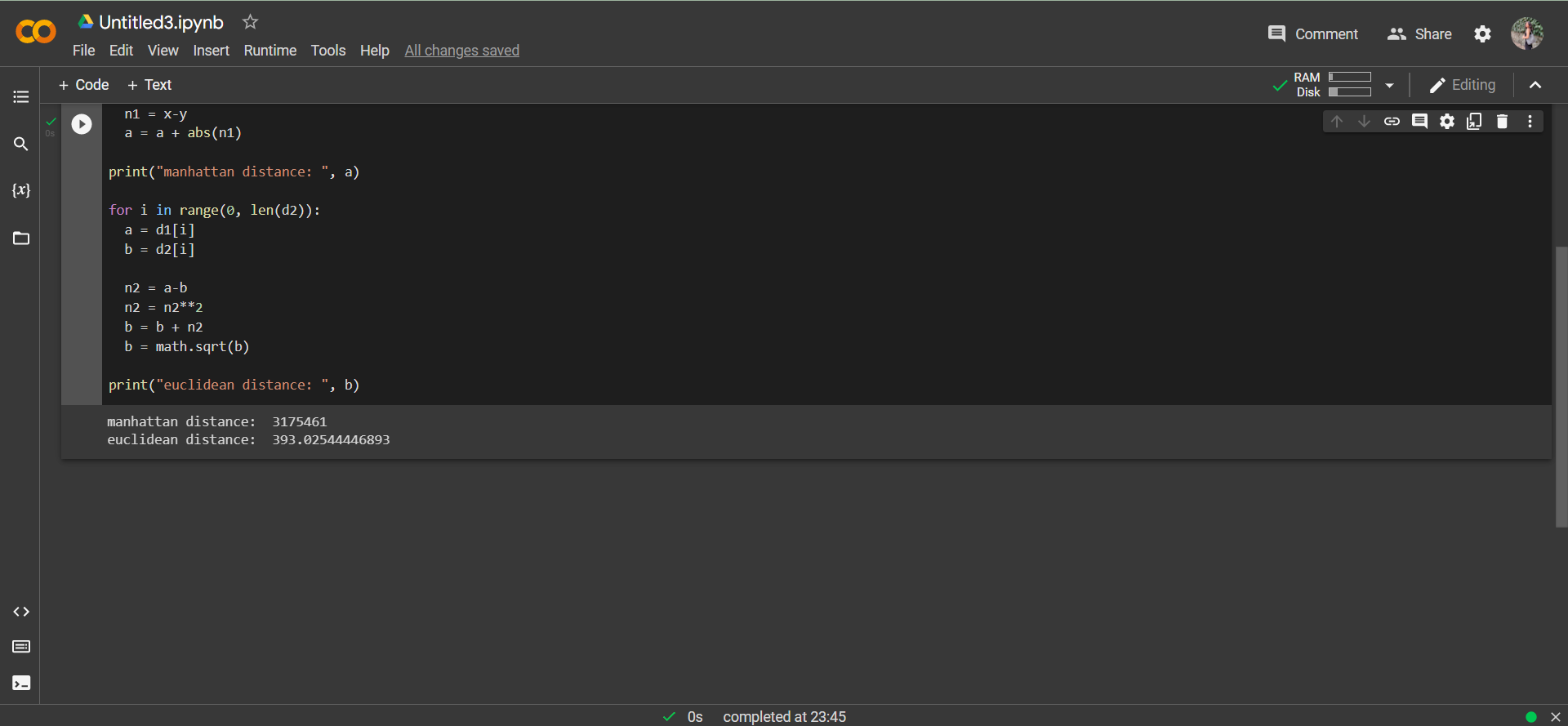
  n2 = n2\*\*2

  b = b + n2

  b = math.sqrt(b)

print("euclidean distance: ", b)

Output:





**Questions:**

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

Answer: The different applications of numeric similarity measures are:

 Euclidian distance: It is useful to use this formula whenever you want to

compute the distance between 2 points in the absence of obstacles on the

pathway.

 SED: SED can reduce computational work while calculating distances

between observations. For instance, it can be used in clustering, classification,

image processing, and other domains.

 Canberra distance: It is a weighted version of Manhattan distance used in

Clustering, like Fuzzy Clustering, classification, computer security, and

ham/spam detection systems. It is more robust to outliers in contrast to the

previous metric.

 Cosine distance: This metric is widely used in text mining, natural language

processing, and information retrieval systems.

 Standardized Euclidian distance: Standardization or normalization is a

technique used in the pre-processing stage when building a machine learning

model.

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

Answer: There are several applications or areas where we use the text similarity; these areas are Information retrieval, clustering, text categorization, topic detection, question answer session, machine translation, text summarization etc. These text similarity measures play an increasingly vital role in text related research and applications in several tasks such as text classification, information retrieval, topic tracking, document clustering, questions generation, question answering, short answer scoring, machine translation, essay scoring.

**Outcomes:** CO2: Comprehend descriptive and proximity measures of data.

**Conclusion:** In completing this experiment, I understood and implemented the concepts of Minkowski distance (Manhattan and Euclidean). I also learnt their importance in data analytics.

**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.