國立雲林科技大學資訊管理系

機器學習-作業四

Department of Information Management

National Yunlin University of Science & Technology

Assignment

應用降維技術於資料集

Dataset Dimensionality Reduction

巫宇哲、鄭皓名、翁振洋

指導老師：許中川 博士

Advisor: Chung-Chian Hsu, Ph.D.

中華民國 112 年 6 月

June 2023

摘要

本研究的動機是為了研究和展示台灣高鐵站點之間的距離關係以及各式品牌飲料之間的關係，目的是探索和視覺化台灣高鐵站點之間的地理距離關係以及各式品牌飲料之間的關係。在資料集一的研究主要使用 python 中的 haversine 套件來計算任意站點兩點間的距離，以及使用多維尺度繪圖(Multi-Dimensional Scaling,MDS)的方法將這些距離轉換成 2D 平面上的視覺化表示。在資料集二則透過 MDS、t-SNE 等降維技術演算法，讓高維的維度可以降到三維以下以便人們可以看到資料分布，並使用 word embedding 將名目型資料轉換成向量和 Matplotlib 將降維後的維度以 2D 或 3D 的方式呈現出來。在資料集一的實驗結果表明，MDS 技術能夠更好地保持資料點之間的距離關係，與 Google 地圖上的座標點呈現較高的一致性。然而，t-SNE 的效果相對較差。這種差異可能源於 t-SNE 和 MDS 之間的基本原理和優化目標的不同。t-SNE 旨在在低維空間中保留資料點之間的局部關係，而 MDS 則更關注保持資料點間的全局距離。在資料集二(Drink Dataset)中的實驗結果表明，分為兩部分來探討，第一部分是使用 One-hot Encoding 做前處理對於咖啡類飲品跟汽水類飲品之間的區隔較使用 Word2Vec 做前處理的降維結果還要不明顯，而且使用 Word2Vec 後做 t-SNE 降維，同類別的飲料品項資料點的分布較為靠近，意味著組內變異較小。第二部分是使用 t-SNE 做為降維的結果相較 MDS 更能看出咖啡類飲品跟汽水類飲品之間的區別。根據兩個資料集使用 MDS 與 t-SNE 演算法降維後的結果，可以得出不同資料集使用不同的降維演算法，結果也不相同，採用哪一種降維技術依舊需要考量到資料集的特性以及降維後所要觀察的內容。

關鍵字：MDS、t-SNE、Word2Vec、word embedding

一、 緒論

1.1 動機

本研究的第一項動機是為了研究和展示台灣高鐵站點之間的距離關係。台灣高鐵系統是一個重要的交通基礎設施，連接了台灣的各個城市，為人們提供了快速、便捷的交通選擇。透過這個研究，可以更深入地了解台灣高鐵站點之間的地理距離和空間關係，並透過多維尺度繪圖(Multi-Dimensional Scaling)的方法將這些距離轉換成 2D 平面上的視覺化表示。

本研究的第二項動機是探索各式品牌飲料之間的關係。在現代社會中，有許多品牌的飲料可供消費者選擇，而這些品牌通常在市場上有著不同的地位、知名度和消費者偏好。研究各式品牌飲料之間的關係有助於更深入地了解飲料市場的現狀和消費者行為。

1.2 目的

本研究的目的是探索和視覺化台灣高鐵站點之間的地理距離關係，這項研究可以提供高鐵站點之間的距離資訊，對於需要規劃交通路線的人們來說，這將提供有價值的參考。無論是商務旅行、觀光旅遊還是其他交通需求，這些距離資訊和視覺化表示將幫助使用者更好地理解各個高鐵站點之間的距離和時間成本，從而做出更明智的交通安排。

本研究的目的是探索和視覺化各式品牌飲料之間的關係，這項研究可以比較不同品牌之間的特點，像是品牌知名度、品質聲譽、價格定位、市場定位等方面。此研究還可以理解品牌間的競爭關係，這有助於品牌管理者制定策略以應對市場競爭，並提供消費者更多選擇和優質的產品。同時還可以了解消費者對不同品牌的態度和行為模式。

二、 方法

2.1 實作說明

本研究的第一項工作是從 Google 地圖上尋找台北高鐵站、苗栗高鐵站、雲林高鐵站、台南高鐵站、高雄高鐵站、花蓮豐濱、台東鹿野各個地方的緯度和經度，然後藉由經緯度去計算出之間的距離，使用 python 中的 haversine 套件計算任 2 點間的距離，haversine 是一種根據兩點的經度和緯度來確定大圓上兩點之間距離的計算方法，計算後將得到的任兩點距離轉成距離矩陣，最後本研究將多維尺度繪圖(Multi-Dimensional Scaling)的方法將這些距離轉換成 2D 平面上的視覺化表示。本研究的第二項工作是生成資料，透過常態分配設定不同的平均數與標準差來生成資料，以及使用隨機生成的資料並設定相對應的區間來生成需要的資料筆數。此時 Drink Dataset 之中的 Amount 與 Quantity 欄位的資料已

生成完畢，接下來是資料前處理，將 Drink Dataset 之中的 Class 與 Drink 欄位使用 One-hot Encoding 與 Word2Vec 的技術將名目尺度的資料轉換成向量資料以便後續做訓練，最終使用 t-SNE 與 One-hot Encoding、Word2Vec 這兩項技術來比較其結果，並查看其降維的效果。

2.2 操作說明

本研究執行環境採用 Python3.7.16，以 Visual Studio Code 作為開發工具，利用 sklearn 裡面多維縮放(MDS)、t-隨機鄰近嵌入法(t-SNE)等降維技術工具讓高維的維度可以降到 3 維以下以便人們可以看到資料分布，並使用 haversine 套件來計算 2 個座標點的距離、word embedding 將名目的資料轉換成向量和 Matplotlib 將降維後的維度以 2D 或 3D 的方式呈現出來。

三、 實驗

3.1 資料集

3.1.1 資料集一

從 Google 地圖上面尋找台北高鐵站、苗栗高鐵站、雲林高鐵站、台南高鐵站、高雄高鐵站、花蓮豐濱、台東鹿野各個地方的緯度和經度，如表 3 所示。

表 1 各區域車站經緯度

	緯度	經度
台北高鐵站	25.04733097	121.517619097881
苗栗高鐵站	24.6057177	120.8254565
雲林高鐵站	23.73631869	120.4165046
台南高鐵站	22.9246529245525	120.285653801857
高雄高鐵站	22.68785475	120.309129169351
花蓮豐濱	23.58521358	121.5027313
花蓮豐濱	22.95515334	121.1577754

3.1.2 資料集二

資料集名稱: Drink Dataset

A 類別的資料使用常態分配生成資料 300 筆，平均數設定 100、標準差設定 200，以及使用隨機生成資料 300 筆，區間設定 500~1000 。

B 類別的資料使用常態分配生成資料 150 筆，平均數設定 200、標準差設定 10，以及使用隨機生成資料 150 筆，區間設定 500~1000 。

C 類別的資料使用常態分配生成資料 150 筆，平均數設定 200、標準差設定 10，以及使用隨機生成資料 150 筆，區間設定 500~1000 。

D 類別的資料使用常態分配生成資料 300 筆，平均數設定 400、標準差設定 100，以及使用隨機生成資料 300 筆，區間設定 500~1000 。

E 類別的資料使用常態分配生成資料 150 筆，平均數設定 800、標準差設定 10，以及使用隨機生成資料 150 筆，區間設定 1~500 。
F 類別的資料使用常態分配生成資料 150 筆，平均數設定 800、標準差設定 10，以及使用隨機生成資料 150 筆，區間設定 1~500 。
G 類別的資料使用常態分配生成資料 300 筆，平均數設定 900、標準差設定 400，以及使用隨機生成資料 300 筆，區間設定 1~500 。

表 2 Drink Dataset

Class	**Drink**	**Rank**	**Amount(*N*(*m*,*s*))**	**Quantity**	**Count**
A	7Up	7	(100, 200)	Random(500, 1000)	300
B	Sprite	6	(200, 10)	Random(500, 1000)	150
C	Pepsi	5	(200, 10)	Random(500,1000)	150
D	Coke	4	(400, 100)	Random(500, 1000)	300
E	Cappuccino	3	(800, 10)	Random(1, 500)	150
F	Espresso	2	(800, 10)	Random(1, 500)	150
G	Latte	1	(900, 400)	Random(1, 500)	300

N(*m*,*s*)：Normal Distribution

3.2 前處理

- 資料前處理
  - 資料集一

將台北高鐵站、苗栗高鐵站、雲林高鐵站、台南高鐵站、高雄高鐵站、花蓮豐濱、台東鹿野，這 7 個地方透過 google 地圖找到他們緯度和經度，然後用 python 中的 haversine 套件計算任 2 點間的距離，haversine 是一種根據兩點的經度和緯度來確定大圓上兩點之間距離的計算方法，最後將得到的任兩點距離轉成距離矩陣，如下表 3。

表 3 任兩點間的距離矩陣

	台北高鐵站	苗栗高鐵站	雲林高鐵站	台南高鐵站	高雄高鐵站	花蓮豐濱	台東鹿野
台北高鐵站	0.000000	85.385034	183.534166	267.156463	289.709359	162.587259	235.493531
苗栗高鐵站	85.385034	0.000000	105.198560	194.830040	219.645320	132.674836	186.623374
雲林高鐵站	183.534166	105.198560	0.000000	91.236726	117.099278	111.898097	115.203717

台南高鐵站	267.156463	194.830040	91.236726	0.000000	26.440513	144.411544	89.370476
高雄高鐵站	289.709359	219.645320	117.099278	26.440513	0.000000	157.643816	91.916108
花蓮豐濱	162.587259	132.674836	111.898097	144.411544	157.643816	0.000000	78.421814
台東鹿野	235.493531	186.623374	115.203717	89.370476	91.916108	78.421814	0.000000

■ 資料集二(Drink Dataset)

➢ 名目型資料

I. One-hot Encoding：針對名目屬性的欄位，如：Class、Drink，以 One-hot Encoding 的方式轉換成只由 0、1 構成的向量。這邊以欄位 Drink 使用 One-hot Encoding 技術為例，如下表 4。

表 4 欄位 Drink 使用 One-hot Encoding 後的結果

	Drink		7Up	Cappuccino	Coke	Espresso	Latte	Pepsi	Sprite
0	7Up		1	0	0	0	0	0	0
1	7Up		1	0	0	0	0	0	0
2	7Up		1	0	0	0	0	0	0
…	…	→	…	…	…	…	…	…	…
1497	Latte		0	0	0	0	1	0	0
1498	Latte		0	0	0	0	1	0	0
1499	Latte		0	0	0	0	1	0	0

II. Word2Vec：使用 Gensim 套件匯入 Word2Vec、FastText 等模型，將欄位 Drink 的飲料名稱轉換成向量形式表示。

➢ 數值型資料

I. 對於欄位 Amount 與 Quantity 分別讓資料點在區間內進行常態分佈與亂數分配。

II. 針對數值屬性的欄位，如：Amount、Quantity，使用 Minmax Scaler 正規化技術，將數值限縮於 0~1 之間，以提升降維後不同類別的資料點的區分可以更加明確。正規化處理結果如下表 5。

表 5 使用 Minmax Scalar 正規化後的結果

	Amount	Quantity		Amount	Quantity
0	0.234211	749		0.23421083	0.74924774
1	0.172499	557		0.17249898	0.55667001
2	0.290095	756		0.29009506	0.75626881
…	…	…	→	…	…
1497	0.596732	193		0.59673223	0.19157472
1498	0.428290	308		0.42829027	0.30692076
1499	0.598219	276		0.59821866	0.27482447

3.3 實驗設計

圖 1 實驗設計流程

本實驗設計流程分為資料集一和二，兩個部分：

1. 將各個座標轉成距離矩陣之後，本研究使用了 MDS 和 t-SNE 兩種降維技術將資料維度降維成 2 維或是 3 維方便後續可視化以及觀察 2 種降維技術的效果如何。
2. 首先，針對名目屬性及數值屬性資料各別依照 3.2 節所提及的前處理方式進行處理，如：One-hot Encoding、Word2Vec 等。接著，依據不同的降維技術產生低維度的座標點，如： MDS 與 t-SNE 兩種降維度技術，最後，再可視化座標點為 2D 或 3D 圖。

3.4 實驗結果

3.4.1 資料集一

首先，本研究先使用了 MDS 降維技術，其基本的思想，是將高維坐標中的點投影到低維空間中，保持點彼此之間的相似性盡可能不變，圖表上，相似的數據，比不相似的更加靠近。如圖 2 所示為經過 MDS 降維技術後的可視化 2D 圖、3D 圖和 google 地圖標出來的實際點，由於降維後的 2D 圖跟在 google 地圖上看到的座標點有點相反，因此我們有將座標點進行旋轉方便我們更直觀的做對比。

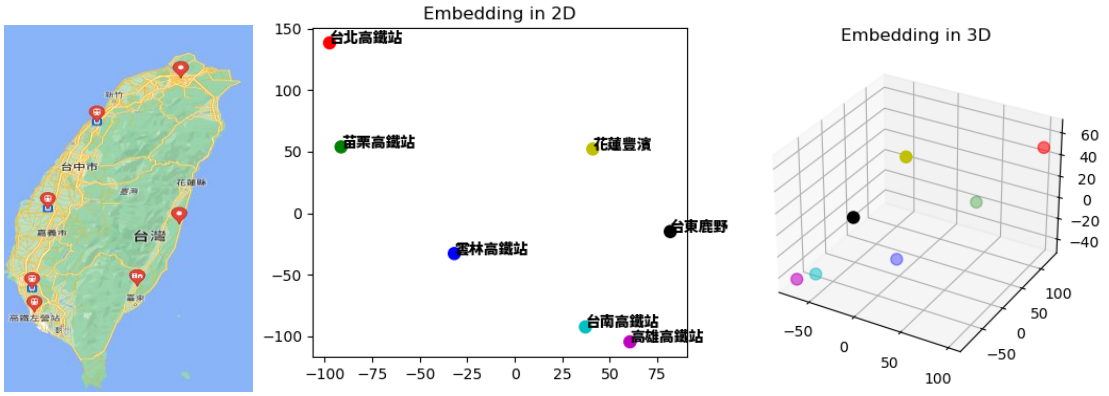

圖 2 實際座標點和降維後的 2D、3D 座標點

其次本研究也使用了降維的另一個技術 t-SNE，t-SNE 是非線性降維，用在非監督問題類類型中，流形還原的意義是將高維度上相近的點，對應到低維度上相近的點，盡量保持資料點之間的遠近關係，沒有資料點的地方，就不列列入考量範圍。如圖 3 所示為經過 t-SNE 降維技術後的可視化 2D 圖、3D 圖和 google 地圖標出來的實際點，由於 t-SNE 使用隨機初始化和隨機選擇鄰居的過程，以及優化過程中的隨機性，因此每次跑的結果可能都不盡相同，本研究測試了好幾次，找到比較符合 google 地圖上座標的相對位置，也將座標進行了旋轉。

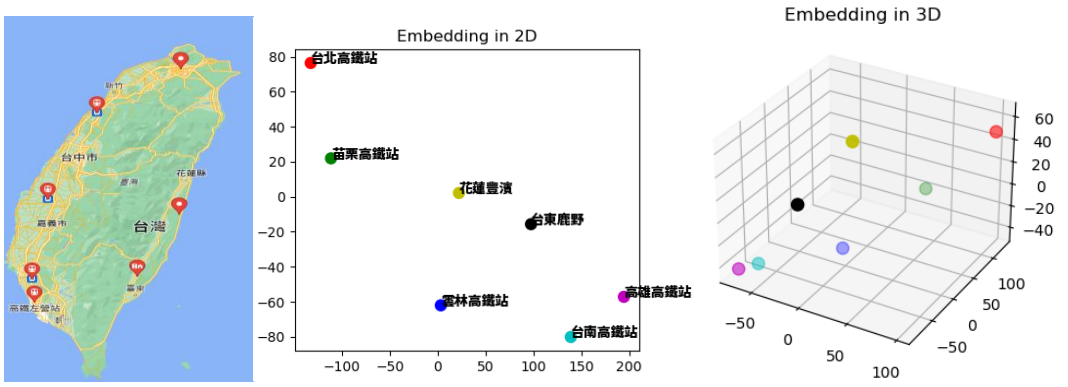

圖 3 實際座標點和降維後的 2D、3D 座標點

3.4.2 資料集二(Drink Dataset)

以下將以兩種降維技術(MDS、t-SNE)說明實驗結果並比較兩種技術對於飲料資料集降維的績效 。

1. 採用 t-SNE 降維技術

首先探討使用 t-SNE 對飲料資料集降維的結果。本研究對於該資料集內名目屬性的資料分別以 One-hot Encoding 與 Word2Vec 兩種技術處理，數值屬性的資料則是用 Minmax Scaler 的正規化技術處理，接著比較兩種對名目屬性資料的降維結果，以觀察降維後飲料種類間的相似程度。

對名目屬性資料使用 One-hot Encoding 技術，超參數設定如下表 6，learning rate 的部分讓模型自動配置適當的學習率，不同的超參數組合的降維結果，依照

組合順序，排列如下圖 4 至 7 所示。表 6 中的 perplexity 為困惑度，困惑度可解釋成有效鄰近樣本點數量，困惑度越大，近鄰越多，對小區域的敏感度就越小，大的資料集需要更大的困惑度，預設為 30；random_state 主要控制隨機數的生成，從結果可以看出，當 random_state 一致時，不同的訓練迭代次數會影響降維後的資料點的分布與相似度的關係。從結果可以看出，當調整較大的迭代次數確實能使模型在區分汽水類跟咖啡類飲品更加明確，而較大的困惑度可以使每個類別中的資料點差距越小，當 random_state 調整越大時，會使模型在區分汽水類跟咖啡類飲品更混亂。

表 6 t-SNE 超參數設定組合

	learning rate	perplexity	random_state	n_iter
1	auto	30	0	1000
2	auto	30	0	100000
3	auto	100	0	1000
4	auto	30	20	1000

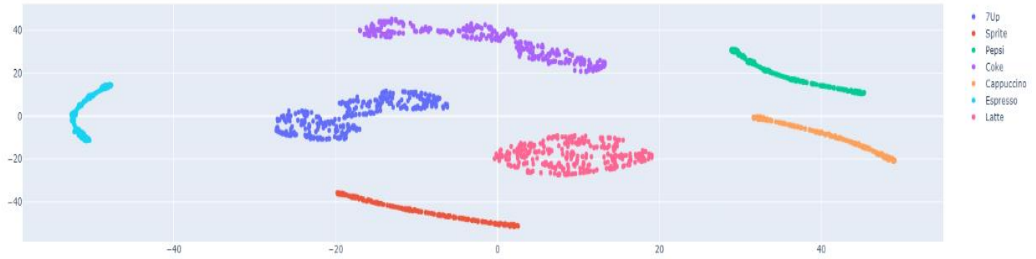

圖 4 降維結果以隨機性方式呈現

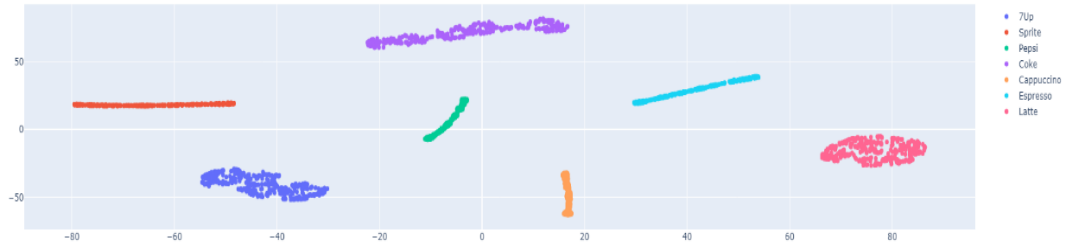

圖 5 訓練迭代次數較大的降維結果

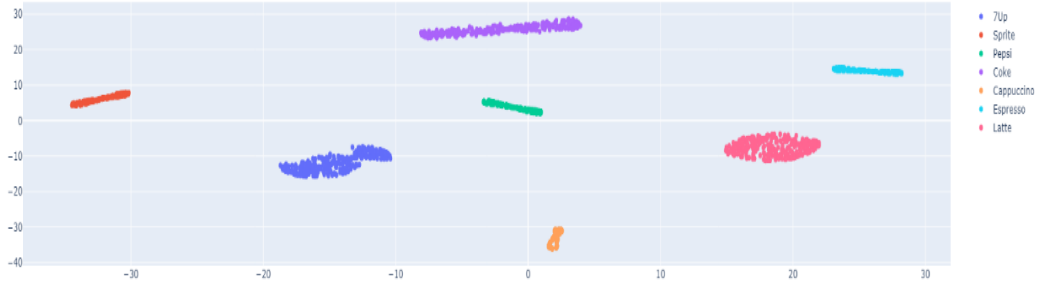

圖 6 較高困惑度的降維結果

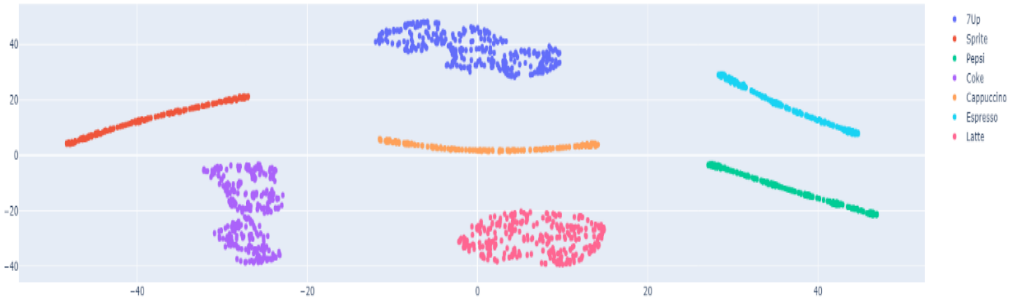

圖 7 t-SNE 設定較大的隨機性

接著是對名目屬性資料使用 Word2Vec 技術，這部分使用了 Gensim 套件以及在 Google 新聞數據集的一部分（約 1000 億個單詞）上訓練的預訓練向量，透過將飲品名稱轉換成各自對應的向量，再得出距離矩陣，最後使用 t-SNE 將資料點降維度至二維平面以觀察各種類飲品間的相似度。結果如下圖 8 所示，明顯可以看出咖啡類飲品跟汽水類飲品之間的相似度非常低。

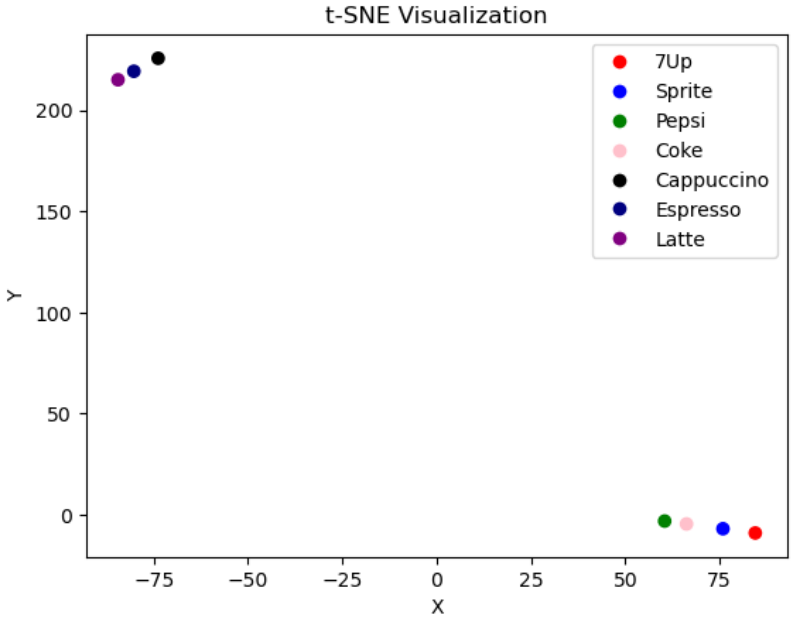

圖 8 t-SNE 降維後各類別間的相似度關係

從上述對名目屬性的資料的兩種不同的轉換方式，可以看出用 Word2Vec 技術對於 Drink Dataset 在降維方面比較清楚能劃分出咖啡類飲品跟汽水類飲品之間的不相似度。

2. 採用 MDS 降維技術

這部分在資料前處理階段，對名目型與數值型資料的處理方式皆與採用 t-SNE 降維技術所使用的方式一致，因此，這個段落主要以觀察使用 MDS 降維後資料點分布的狀況以及與採用 t-SNE 降維技術所產生的結果比較。

下圖 9 是對名目型資料使用 One-hot Encoding 技術做前處理後，再用 MDS 技術降維後的結果。圖 10 為使用 Word2Vec 做前處理後，再經 t-SNE 降維後資料點分布情形。比較使用 MDS 降維與 t-SNE 降維後對咖啡類飲品跟汽水類飲品

之間的區分情形，可以得知 t-SNE 做為降維的結果相較 MDS 更能看出咖啡類飲品跟汽水類飲品之間的不相似度，但在各飲品類別內，經 t-SNE 降維後的資料點分布較為分散，而這部分可以藉由提高 t-SNE 模型的困惑度，讓各飲品類別內的資料點分布較緊密。

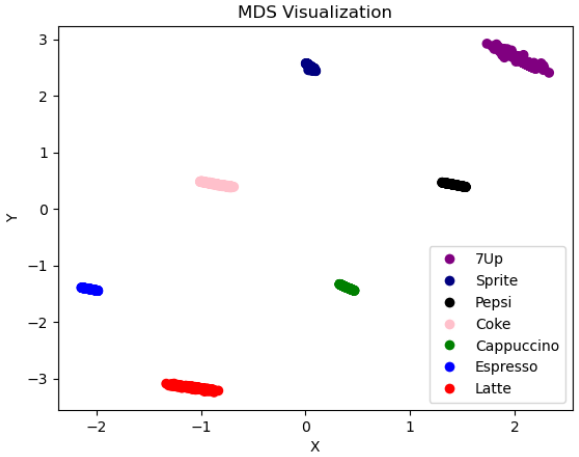

圖 9 MDS 降維後各類別間的相似度關係 (One-hot Encoding)

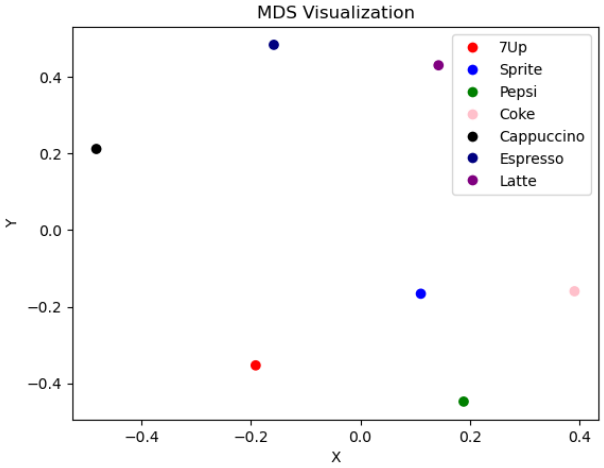

圖 10 MDS 降維後各類別間的相似度關係 (Word2Vec)

結論

在使用資料集一進行降維可視化時，我們將各個景點的緯度和經度轉換為對應的距離度量。我們採用了多維縮放（MDS）和 t-SNE 兩種降維技術進行比較。實驗結果表明，MDS 技術能夠更好地保持資料點之間的距離關係，與 Google 地圖上的座標點呈現較高的一致性。然而，t-SNE 的效果相對較差。這種差異可能源於 t-SNE 和 MDS 之間的基本原理和優化目標的不同。t-SNE 旨在在低維空間中保留資料點之間的局部關係，而 MDS 則更關注保持資料點間的全局距離。此外，t-SNE 在優化過程中的隨機性可能導致不同運行之間的結果變化較大，這些因素都對資料的呈現產生了影響，因此需要耗費時間找到適合的參數組合，以使得降維後的資料分布更接近 Google 地圖上的座標點。

在資料集二(Drink Dataset)中，對於欄位 Amount 與 Quantity 分別讓資料點在區間內進行常態分佈與亂數分配，接著對於名目型資料以兩種不同的技術轉換成向量方式表示，數值型資料則是以 Minmax Scalar 進行正規化讓數值縮在 0~1 之間，以提升降維可視化後，能明顯看出飲品種類間的相似度關係。對於該資料集，本研究分成兩個部分探討：在同一降維技術(t-SNE)中，比較使用 One-hot Encoding 與 Word2Vec 進行資料前處理後，對於降維的結果的差異性、比較兩種降維演算法對於 Drink Dataset 降維後的差異。

第一部分的探討，可以從 3.4 節實驗結果得知，使用 One-hot Encoding 做前處理對於咖啡類飲品跟汽水類飲品之間的區隔較使用 Word2Vec 做前處理的降維結果還要不明顯，而且使用 Word2Vec 後做 t-SNE 降維，同類別的飲料品項資料點的分布較為靠近，意味著組內變異較小。而第二部分，一樣能從 3.4 節實驗結果得知，使用 t-SNE 做為降維的結果相較 MDS 更能看出咖啡類飲品跟汽水類飲品之間的不相似度，但在各飲品類別內，經 t-SNE 降維後的資料點分布較為分散，雖然 t-SNE 有很強的捕捉局部特徵的能力，但由於 t-SNE 演算法具有隨機性，正如前段結論所述，t-SNE 在多次實驗可以產生不同的結果，跟 PCA 正好是相反，PCA 是確定性的（deterministic），且每次計算後的結果相同，對此隨機性，可以透過增加最大迭代次數(n_iter)來使不同種類的資料可以區分得更清楚。

參考文獻

方石劍(2022)，使用 Scikit-Learn 的 Python 多维缩放指南。
https://juejin.cn/post/7116381670801932302
資料降維與視覺化：t-SNE 理論與應用
https://www.mropengate.com/2019/06/t-sne.html
詞向量
https://huggingface.co/fse/word2vec-google-news-300