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

台灣火車站距離和飲料資料集

Taiwan Railway Station Distance & Drink 資料集

楊欣蓓、黃裕鳴、游棨翔

指導老師：許中川 博士

Advisor: Chung-Chian Hsu， Ph.D.

 中華民國113年6月

June 2024

**摘要**

本研究旨在利用Python探索MDS在台灣火車站地理距離可視化中的應用，並比較One-Hot Encoding和Word2Vec在飲料數據集處理上的效果差異。

在車站資料集中的一部分，本研究定義了台北、新竹、台中、斗六、高雄、花蓮玉里、台東知本等七個火車站的經緯度。然後，考量到地球的曲率，使用了測地線方法來測量火車站之間的地理距離，生成距離矩陣。接下來，使用MDS方法，將這些距離數據降維至二維空間，並繪製降維後的二維散點圖。此外，還有用Google地圖標記這些火車站的位置，以便對比。

資料集二的部分則比較了One-Hot Encoding和Word2Vec在飲料資料處理上的效果，發現Word2Vec能夠更有效的捕捉飲料之間的語意關係，並透過熱圖和t-SNE視覺化展示其優勢。實驗結果表明，相較於One-Hot Encoding，Word2Vec能更好的捕捉飲料之間的語意關係，在飲料推薦或市場區隔上具有更大的潛力。

關鍵字：降維、MDS、One-Hot Encoding、Word2Vec、t-SNE

**一、緒論**

1. **研究動機**

隨著台灣交通日益發達，跨縣市通勤、洽公與旅行的需求也日益增加。然而，民眾對於火車站之間的距離與相對位置，往往缺乏直觀且易於理解的認知。因此，本研究希望透過資料視覺化技術，將台灣主要火車站的地理位置與相對距離以更直觀的方式呈現，為旅客提供便捷的出行參考。

同時，飲料市場競爭激烈，消費者口味多變。如何從眾多產品中脫穎而出，並精確掌握消費者偏好，成為各大飲料廠商欲解決的難題。本研究將聚焦於常見的碳酸飲料與咖啡兩大類別，透過分析其相似度，期望能發掘消費者的偏好趨勢，為產品研發與營銷策略提供有價值的參考依據。例如，相似度高的飲料可能存在市場競爭關係，而相似度低的飲料則可能具有市場區隔的潛力。

1. **研究目的**

首先，本研究聚焦於台灣主要火車站之間的地理關係。本研究將收集台北、新竹、台中、斗六、高雄、花蓮玉里、台東知本等七個主要火車站的經緯度數據，計算其地理距離，並透過MDS方法將距離數據降維至二維平面。最後，結合Google地圖標記，繪製直觀的二維散點圖，協助民眾更有效地規劃行程。

其次，本研究將分析飲料市場中碳酸飲料與咖啡的相似度。本研究將使用包含Drink、Rank、Amount、Quantity等特徵的飲料資料集，透過t-SNE方法降維並視覺化。同時，本研究將比較1-of-k編碼與屬性值相似度計算兩種方法在處理名目屬性上的效果，期望能揭示消費者偏好趨勢，為廠商提供產品開發與市場策略的參考依據。

**二、實驗方法**

1. **實作說明**

首先，本研究收集了台北、新竹、台中、斗六、高雄、花蓮玉里、台東知本等七個主要火車站的經緯度資料。根據這些經緯度資訊，本研究利用測地線方法計算火車站之間的地理距離，生成距離矩陣。測地線方法能更精確地反映地球曲率對距離的影響，因此相較於平面距離計算更準確。接著，本研究使用MDS方法將距離矩陣降維至二維平面，以便更直觀地呈現火車站之間的相對位置。同時，本研究使用Google地圖標記這些火車站，將理論結果與實際地理位置進行對比驗證。

在飲料市場相似度分析實驗中，本研究使用包含Drink、Rank、Amount、Quantity等特徵的飲料資料集。為便於視覺化與分析，本研究將透過t-SNE方法將高維資料降維至二維平面。

為了比較不同編碼方式對相似度分析的影響，本研究將對飲料資料集中的名目欄位（如Drink）進行One-Hot Encoding和Word2Vec兩種處理方式，比較其在反映飲料相似度上的效果。接著，利用t-SNE將處理後的資料降維並繪製散點圖，以視覺化方式呈現飲料之間的相似度，進一步探討消費者偏好。

1. **操作說明**

本研究採用Python 3.8作為編寫語言，並使用Visual Studio Code作為開發環境。使用MDS對火車站經緯度降維，且利用One-Hot Encoding與Word2Vec處理Drink欄位，並利用t-SNE降維。

**三、實驗設計**

1. **資料集**

**3.1.1 火車站經緯度**

此表格展示了台北火車站、新竹火車站、台中火車站、斗六火車站、高雄火車站、花蓮玉里及台東知本的經緯度座標。

**表1**

*各個火車站的經緯度座標*

|  |  |  |
| --- | --- | --- |
| **車站** | **緯度** | **經度** |
| 台北火車站 | 25.047637204053995 | 121.5171273798563 |
| 新竹火車站 | 24.801750331885113 | 120.9716203981461 |
| 台中火車站 | 24.137523918186687 | 120.68683583244295 |
| 斗六火車站 | 23.712201175089422 | 120.54104266971164 |
| 高雄火車站 | 22.6396706176077 | 120.30261518317573 |
| 花蓮玉里 | 23.727506999999997 | 120.29999999999999 |
| 台東知本 | 23.712201175089422 | 120.54104266971164 |

**3.1.2 Drink 資料集**

表2展示了Drink 資料集的原始型態，記錄了7種不同飲料的Class、名稱、Rank、容量(常態分佈)和數量(隨機分佈)，並指定了每種飲料要生成的模擬資料筆數。表3展示了依照Amount常態分佈及Quantity隨機分佈並依照Count筆數生成的模擬資料。

**表2**

*Drink 資料集(處理前)*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Class** | **Drink** | **Rank** | **Amount(N (μσ))** | **Quantity** | **Count** |
| A | 7up | 7 | (100, 200) | Random(500, 1000) | 100 |
| B | Sprite | 6 | (200, 10) | Random(500, 1000) | 200 |
| C | Pepsi | 5 | (200, 10) | Random(500, 1000) | 100 |
| D | Coke | 4 | (400, 100) | Random(500, 1000) | 400 |
| E | Cappuccino | 3 | (800, 10) | Random(1, 500) | 400 |
| F | Expresso | 2 | (800, 10) | Random(1, 500) | 200 |
| G | Latte | 1 | (900, 400) | Random(1, 500) | 100 |

**表3**

*Drink 資料集(處理後)*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Class** | **Drink** | **Rank** | **Amount** | **Quantity** |
| 0 | A | 7Up | 7 | -55.983657 | 919.387694 |
| 1 | A | 7Up | 7 | 278.756034 | 559.495568 |
| 2 | A | 7Up | 7 | 213.574166 | 858.120052 |
| 3 | A | 7Up | 7 | 115.276056 | 675.560194 |
| 4 | A | 7Up | 7 | 113.712519 | 856.110844 |
| … | … | … | … | … | … |
| 1495 | G | Latte | 1 | 1342.372717 | 368.955640 |
| 1496 | G | Latte | 1 | 833.656659 | 223.793875 |
| 1497 | G | Latte | 1 | 1034.502112 | 425.712543 |
| 1498 | G | Latte | 1 | 504.050802 | 141.633063 |
| 1499 | G | Latte | 1 | 1085.908829 | 99.448849 |

1. **資料前處理**
2. **火車站經緯度**

根據火車站經緯度資料，計算每一對火車站之間的地理距離(km)。考慮到地球的球面特性，不直接採用平面歐式距離，而是利用大圓距離，透過輸入的火車站經緯度資料，計算出每對火車站之間沿地球表面測量所得的最短距離（大圓距離）。此過程將結果儲存為一個二維矩陣，為後續分析提供基礎。

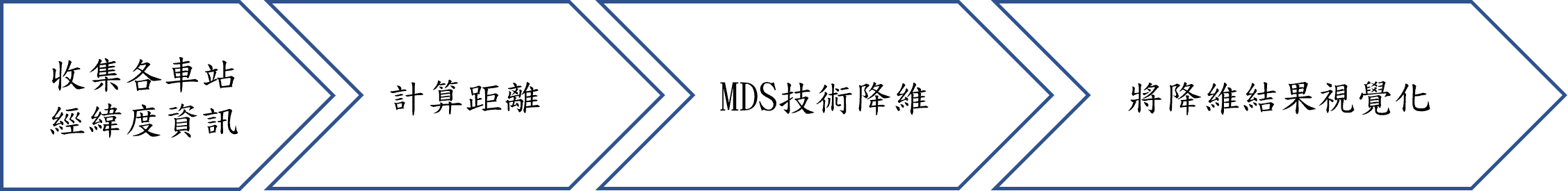
|  |  |
| --- | --- |
|  |  |

1. **Drink資料集**

在Drink資料集中，對'Drink'欄位分別進行兩種不同的處理方式：

* One-Hot Encoding：將每種獨特的飲料名稱轉換為一個二元向量，這個過程將每種獨特的飲料名稱(如：'7up'、 'Sprite')轉換為一個二維向量。
* Word2Vec：利用gensim庫中的Word2Vec模型學習每種飲料名稱的詞向量表示，不同於One-Hot Encoding，Word2Vec會將單詞映射到一個連續的向量空間中，使得在該向量空間中，語意相似的詞彙會靠的更近。

1. **實驗設計**



**圖1**

*火車距離資料集實驗設計流程圖*

本研究於收集七個火車站的各個經緯度，並利用利用側地線方法計算各個火車站之間的距離，並生成矩陣，接著使用MDS方法將距離矩陣降維至二為平面，可更簡易地觀察火車站之間的相對位置以及大致上的距離差。此外使用Google地圖標記這些火車站，可更直觀地做出對比驗證。



**圖2**

*Drink資料集實驗設計流程圖*

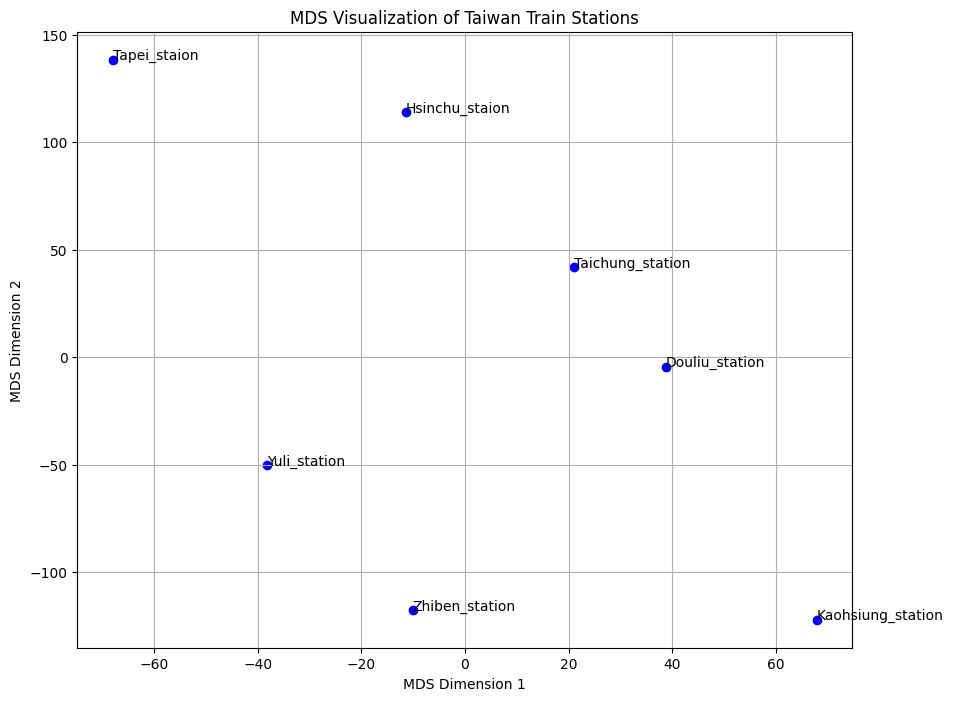
本研究於Drink資料集先行依照Amount常態分佈及Quantity隨機分佈並依照Count筆數生成的模擬資料，並對Drink欄位做One-Hot Encoding和Word2Vec資料前處理，接著利用t-SNE進行降維，欲觀察兩者之間在視覺化方面的差異，並探討何者較容易觀測。

1. **實驗結果**

本實驗利用火車經緯度與飲料資料集，並透過MDS、t-SNE和Word2Vec技術，將高維數據降維並視覺化展示，揭示火車站之間的地理距離與飲料之間的相似度。

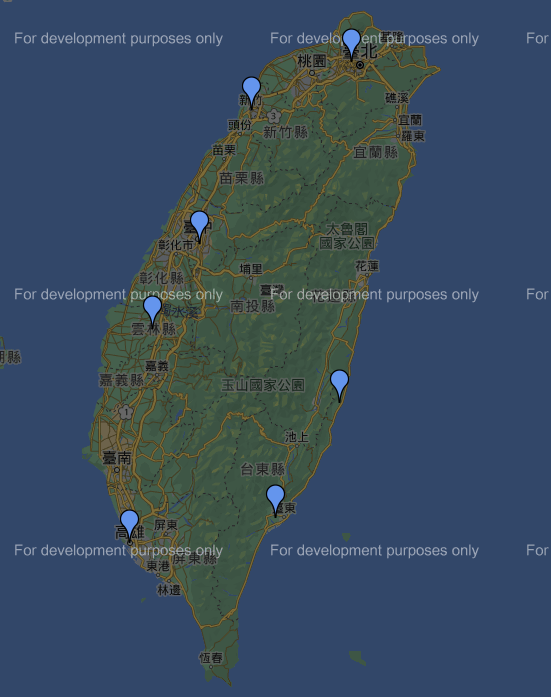
1. **火車站經緯度分析**

本實驗利用多維度縮放(MDS)技術，將台灣七個主要火車站的地理距離關係視覺化呈現。透過計算火車站間的大圓距離並進行降維，並繪製了散點圖(圖3)，清晰的展示了各站點的相對位置，與實際地理分佈高度吻合，通過與實際地圖的對比，進一步驗證了MDS方法在可是化地理距離上的有效性。



**圖3**

*台灣主要火車站距離之MDS視覺化*



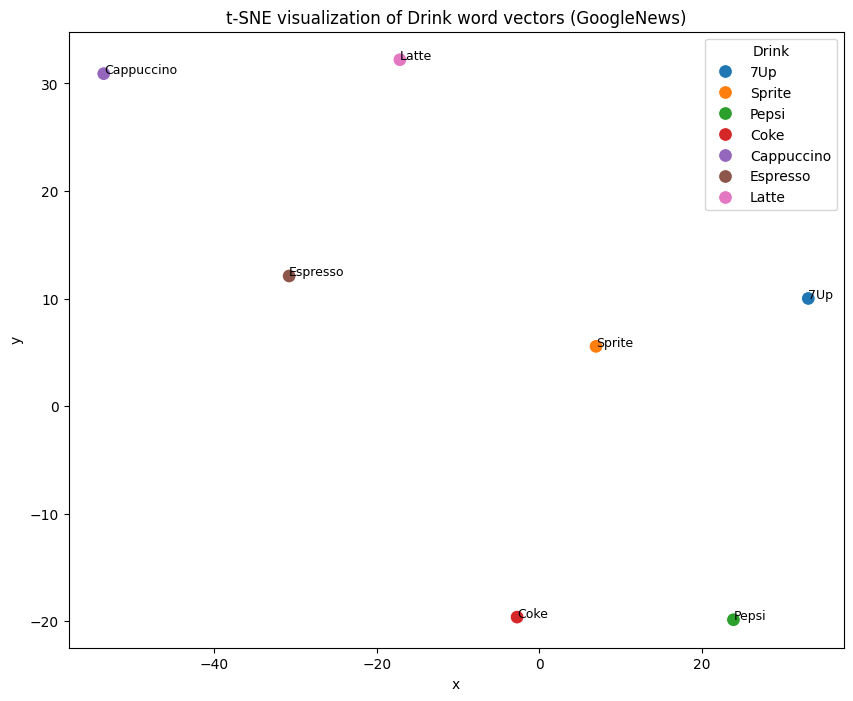
**圖4**

*台灣火車站座標真實地圖*

1. **Drink資料集分析**

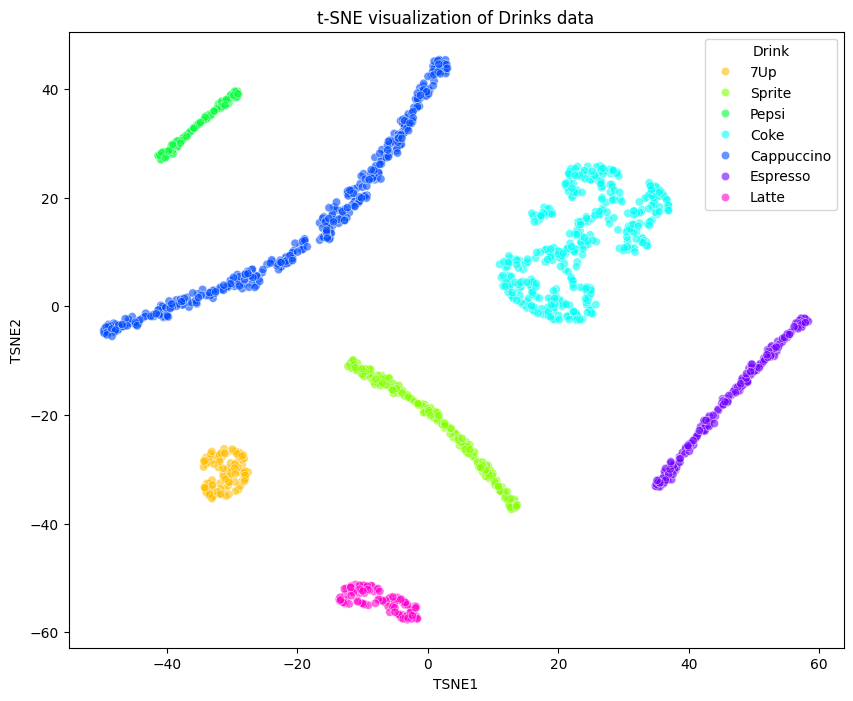
本研究旨在比較One-Hot Encoding和Word2Vec兩種方法在飲料資料處理上的效果。實驗結果表明，Word2Vec能夠更好的捕捉飲料之間的語意關係，相似的飲料在t-SNE視覺化圖中聚集在一起(圖5)。而One-Hot Encoding僅僅將飲料轉換為獨立的類別，無法捕捉飲料之間的相似性，因此在t-SNE視覺化圖中，不同飲料的分佈相對分散(圖6)。

這種差異也在熱圖中得到了體現(圖7)，其中顏色越深表示相似度越高，可以明顯看出Word2Vec能夠有效捕捉飲料之間的語意關係，例：可樂和百事可樂的相似度較高，而One-Hot Encoding則無法呈現這樣的關係。



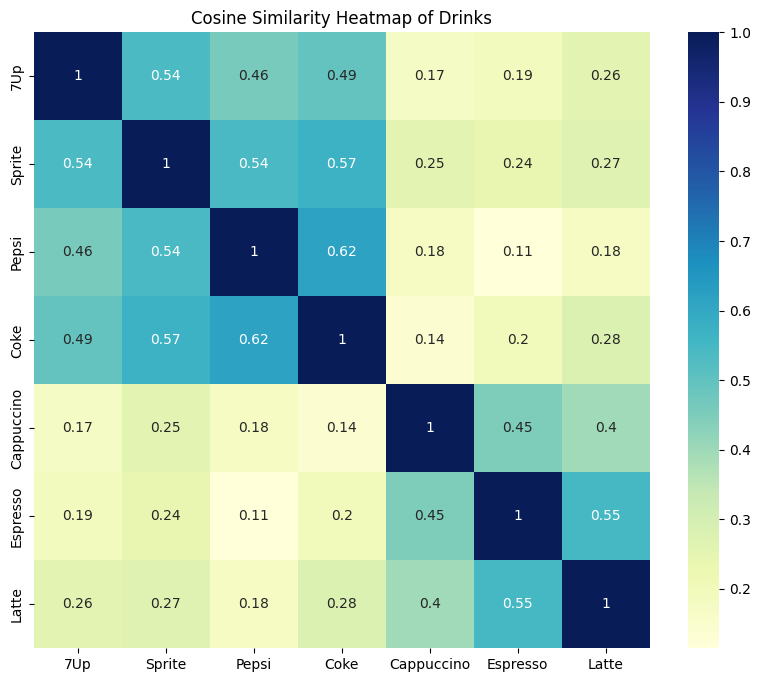
**圖5**

*Drink資料集使用Word2Vec之t-SNE視覺化圖*



**圖6**

*Drink資料集使用One-Hot Encoding之t-SNE視覺化圖*



**圖7**

*Drink資料集heatmap視覺化*

**四、結論**

本研究透過多尺度縮放(MDS)技術成功的將台灣七個主要火車站的地理距離關係視覺化，清楚呈現它們的相對位置，與實際地理分佈高度吻合。這種可視化方法能直觀的反映火車站之間的距離關係，更有助於旅客規劃行程。

此外，在飲料相似度分析方面，本研究比較了One-Hot Encoding和Word2Vec兩種方法在處理飲料資料上的結果。實驗結果顯示，Word2Vec在捕捉飲料之間的語意關係方面表現更為出色。相似的飲料在t-SNE視覺化圖中聚集在一起，而One-Hot Encoding則無法呈現這樣的關係。這項發現對於飲料市場的產品開發和行銷策略具有重要意義，因為Word2Vec能夠更精確地反映消費者對不同飲料的偏好與認知。

**參考文獻**

sklearn與機器學習系列專題之降維（三）一文弄清楚MDS特徵篩選&降維

https://blog.csdn.net/weixin\_45234485/article/details/109729659

[Day15] 文本/詞表示方式(五)-實作word2vec

<https://ithelp.ithome.com.tw/articles/10264523>

t-SNE實踐－sklearn教程翻譯

<https://blog.csdn.net/hustqb/article/details/80628721>

初學Pandas+Ploty+Dash大禮包

<https://weilihmen.medium.com/%E5%88%9D%E5%AD%B8pandas-ploty-dash%E5%A4%A7%E7%A6%AE%E5%8C%85-8661c04e67b7>