Data mining final repor

主題: 通識課程的資料關聯分析與分群分析

專題實作目的：

這項研究將以通識課資料集，目的分為兩個部分，第一個是希望透過關聯分析（apriori algorithm），找出通識課程的各項屬性中有沒有什麼有趣的關聯性。第二部分則是希望透過分群演算法 (Kmeans, Agglomerative Clustering, DBSCAN)，分析課程的課程名稱、描述、學程是否有明顯的區隔在不同的通識領域上，希望發現在通識課程中，是否有課程分類較為不合適的課程，應重新分類。

採用data mining 模組

- apriori_python
- sklearn
- transformers
- scipy
- numpy

程式/環境設定, 執行方式說明

python version: python3.11.4
environment: 請安裝上方的模組，可使用 pip 進行安裝
執行方式：可直接執行 `python3 main.py` 便會列出執行的結果

改變控制參數/技術說明(須說明為何想改變控制的想法)

資料前處理

1. 過濾出所有通識課的資料
2. 將有復數資料的資料切分（授課教師, 教學方法, 學程...）

- **Associate**:
  將所有資料轉換成字串形式（類似於購買的物品），選課率（選課人數/(可選人數 + 授權碼人數)）則轉換成低中高等離散資料，以每0.25切一級距
- **Clustering**:
  使用 bert toknizer 將課程名稱、課程說明、學程轉為 token_id ，並轉換成 one-hot vector ，並將所有的 vector 相加，並過濾掉 人文藝術、社會科學、自然科學、邏輯運算、大學入門 這五類以外的課程

改變控制參數

- **Associate**:
  使用 apriori algorithm 套件，主要改變 minsup 及 minconf，希望觀察在不同的參數下課程間會有什麼關聯性，為避免產生過多關聯式，我盡可能地由大至小調整，在此實驗後，我又增加了必須含有課程領域的條件，希望觀察是否有課程領域與某些特徵有相關性。
- **Clustering**:
  主要使用 sklearn 的 Kmeans, Agglomerative Clustering DBSCAN。
  Kmeans 及 Agglomerative Clustering 將 n_clusters 皆設定為 5（人文藝術, 社會科學, 自然科學, 邏輯運算, 大學入門 5類），DBSCAN 則讓他自己分群。
  Agglomerative Clustering 將嘗試四種不同的演算法（ward, single, complete, average）去找尋最佳模型。
  針對 DBSCAN 做 model selection（eps 將嘗試 5~ 10、min_samples 將嘗試 3~6) 去找尋最佳模型。

評估方法

- **Associate**:
  主要透過調整參數，並希望看產生了多少關聯式去調整參數，避免產生過多關聯式，越高 support 及 confident 的將優先考慮。
- **Clustering**:
  主要使用 Entropy 及 Purity 進行分析，比較各模型的優劣程度。在 model selection 時，以 Puriry 為主要的參考依據。針對 DBSCAN 則額外要求雜訊的數量要在 0.2 以下，

結果及討論

- **Associate**:
  - 結果
    - **minsup = 0.8 minconf = 0.9**
      {選課率高} -> {講述法} conf = 0.981
      {講述法} -> {選課率高} conf = 0.987
    - **minsup = 0.76 minconf = 0.9**
      {上課地點本部} -> {選課率高} conf = 0.992
    - **minSup = 0.3, minConf = 0.9 限縮必須含有課程領域**
      {人文藝術} -> {選課率高} conf = 0.980
      {人文藝術} -> {選課率高, 講述法} conf = 0.980
      {人文藝術, 講述法} -> {選課率高} conf = 0.980
      {人文藝術} -> {講述法} conf = 1.0
      {人文藝術, 選課率高} -> {講述法} conf = 1.0

- 討論

  1. 通常使用講述法的課程，選課率通常會有較高的傾象
  2. 而在本部授課的課程，雖然 support 較低，但通常一開出來也會有很高的選課率
  3. 人文藝術的通識課通常會有選課率高及使用講述法等情形

- **Clustering**:

  - 結果

    - Agglomerative Clustering 最佳演算法：ward

	ward	**single**	**complete**	**average**
Purity	**0.560**	0.352	0.346	0.352

    - DBSCAN 最佳參數 eps=7.0 min_samples=3
      此僅顯示 noise 在 0.2內的

      在不同參數下的 Purity

min_sample\eps	**7.0**	**8.0**	**9.0**
3	**0.346**	0.289	0.308
4	0.327	0.289	0.308
5		0.264	0.283
6		0.264	0.283

    - 各分群演算法比較:

	Kmeans	**Agglomerative Clustering (ward)**	**DBSCAN (eps=7.0 min_samples=3)**
Entropy	**0.831**	1.087	1.107
Purity	**0.591**	0.560	0.346

  - 討論

    1. 在此資料集中 Agglomerative Clustering 使用 ward 演算法效果最好
    2. 在此資料集中 DBSCAN 使用 eps=7.0 min_samples=3 演算法效果最好

3. 在此資料集中應用各類分群演算法，Kmean 分群演算法有最好的分群效果
4. 然而真實分類下來，過多的相同名稱及描述課程會大大影響分群結果，並且只使用 tokenizer 做成 one-hot vector，對於文字的理解可能是不夠的，需要對文字作更深度的 embadding ，才能讓語意或描述相近的課程分類相近一些。

部分分類結果：