Instructor: LCK & WJ

Lab 4: Data Preparation and Cleansing

Submission

DUE: 11:59 AM — April 1 (Sat), 2023

Section 0 -

Your name: 廖泓傑

Your student ID: b07703001

Task 1 – Cleaning a dataset with OpenRefine (1.5 point)

First things first:

1) Make sure you have OpenRefine
2) Download the Student_RPT_07_2015.csv @ COOL

For this section, please spend some time exploring the features in OpenRefine. Pastva_OpenRefine.pdf @COOL may help you to get familiar with OpenRefine to some degree.

Description of the original dataset: https://data.gov.tw/dataset/24730

Cleaning goals:

1. Using 國外郵政國名/地區名中英文對照表
   https://www.post.gov.tw/post/internet/Postal/sz_a_e_info.jsp
   [If failed, simply Google it] to clean country names in the dataset, e.g., use 韓國 for 大韓民國(南韓)
   Accomplished
2. The uppercases and lowercases in "對方學校(機構)英文名稱"should be consistent, e.g., USE University of Warsaw for UNIVERSITY OF WARSAW
   Accomplished
3. Since there is no "夜間班" in 學制, please reconstruct this column into records like "學士班"、"碩士班" and "博士班"
   Accomplished
4. Any idea of value like "TATUNG-OKUMA CO.,LTD" or university names recorded in Italian? How will you clean it? Discuss your strategies and try apply some of it into your dataset. Tell us about your attempts. Note: You don't have to check every single name.
   對於 "TATUNG-OKUMA CO.,LTD" ，我會在觀察到如 LTD 的縮寫時，用之前 Of 改成 of 的 case sensitive 方法，保留全大寫。
   另外 ，若大學名稱為無法顯示之外語，在 OpenRefine 中會呈現"?"，考量到不確定原文為何，所以保留原本之內容，若有需要使用這些大學名稱進行統計，則盡量使用中文的欄位。
5. If we would like to group similar departments ("系所名稱") together for a further analysis (e.g., 中國文學、中國語文、國文學系), how OpenRefine can accomplish this task? Is there any other resource that we need? Discuss the possible strategies (required) and try it yourself.

參考 108 學年度大學校院系所彙整表(Retrieved from: https://data.gov.tw/dataset/27932) 中的系所代碼可以將不同學校名稱不同，但性質相似的系所合併

第一步: 先用 OpenRefine 內建的 cluster 將表示方法不同的整併，如空格數量、是否有括號等

第二步: 用 Nearest Neighber 演算法將相近的學系整併，需人工篩選避免有誤，多為「學系」和「系」，以及一些相似字的整併

第三步: 用 n-gram fingerprint 將 size 設為 1，可以整併大多數倒置的學系名稱，如「社會與區域發展學系」和「區域與社會發展學系」

若對整併學系之內容有疑問則查詢 108 學年度大學校院系所彙整表的系所代碼是否相同

N.B. It is **not necessary to clean the entire dataset** in response to this request, as mentioned during the lecture. Cleaning a dataset can be time-consuming. Instead, you are encouraged to share your research findings and any new insights you have gained.

6. **[Deliverable]** Please upload your processed and reasonably clean dataset (it doesn't need to be perfectly clean) to Google Drive. Then, paste a shareable URL here. Note: Before submitting, ensure that your link is publicly accessible to the Teaching Assistants. Failure to do so may result in point deductions. https://docs.google.com/spreadsheets/d/1BMap5ymCgqQReq2to-RpXIv4M1vxxupU/edit?usp=sharing&ouid=116250462944821545397&rtpof=true&sd=true
7. **[Bonus]** At your choice, try to explore OpenRefine a bit more and try some new features that we don't cover in the lecture or Lab. Please tell us what you discovered with a few screenshots. (up to +0.25)

在第五步時，發現學系名稱的 cluster 數離我的目標，也就是各學系代碼只有一個 cluster 有一段差距，在網路上查詢後，用一個方式解決。也就是新增一個欄位，是學系代碼後面加上學系名稱，這樣有同樣的學系代碼的資料就會在 cluster 的演算法更接近，最後成功將 cluster 數量降至 1004，和學系的 969 很接近，做完 cluster 後再用 return value[6:] 就可以去掉學系代碼，我認為比我原本在第五步的方法更有效。

Task 2 – Tidy data (1.5 point)

First things first:

1) Download the 109 學年大教室修課人數_NTU.xlsx @ COOL

For this section, we would like you to identify a "non-tidy" dataset. Please answer the following questions:

1. Based on Wickham's tidy-data heuristics, is this a tidy dataset? Please explain your thoughts.
   Hint: any "five most common problems" identified?
   根據 Wickham 的 tidy-data heuristics，這並不是一個 tidy dataset，在某些地方這個 dataset 呈現資料的方式並不符合經驗法則。
   其中一個問題是同一類觀察單元的數值分散於多處，特別是在「星期一教室」、「星期二教室」等等欄位，以及「星期一時間」、「星期二時間」等等，更直覺的表現方式應將星期幾視為變數，這樣也可以避免空值欄位過多的問題。
   另一格問題是班次欄位，對於沒有多個班次的課程有缺值問題。
2. How would you "melt" or "clean" this dataset? Please outline your plan and strategy without actually performing the melt/clean process at this stage (Let's just focus on formulating a plan only).
   Hints: Consider the following sequence: database-level, schema-level, and instance-level.
   在 melting 的部分，我預計將星期、時間以及地點分為三個不同欄位，若一堂課有多天要上課，則會以多筆資料分別輸入。
3. Now, please go ahead and melt the data and briefly describe what/how you did.
   Hint: you are free to create multiple data tables (as a relational database).
   Accomplished
4. **[Deliverable]** Please upload your processed, relatively tidy dataset via Google Drive, paste a sharable URL here. N.B. Please recheck your URL, points will be deducted if your link is not publicly available to your TAs.
   https://docs.google.com/spreadsheets/d/1m72sAlEOL8xZKTIheK2X5LDBP53LxBSvhiypoqx1p5I/edit?usp=sharing
5. Briefly evaluate the effectiveness of your original plan as described in Step #2. Were there any interesting observations, unexpected situations, or lessons learned? (Remember, there is no perfect plan. Feel free to share negative results as well. ☺
   使用 Tableau Prep 可以很快速的達到 Melting data 的目的，達到在 step #2 的目標。我認為若要使資料更整齊，可能可以使用關聯式資料庫將課程資訊，以及課程的時間地點分開，不過我不太會建立關聯式資料庫。但在資料的處理上，如將空值以特定符號取代，或者字串的處理上，可能是因為不熟悉的關係，仍然是 excel 更加直覺，因此還有用 excel 稍微加工一下

Task 3 – Playing data view on Airtable (1 point)

First things first:

1) Register/Log in to https://www.airtable.com/
2) Download the Ghibli characters.csv @ COOL

For this section, we would like you to utilize Airtable based on a specific scenario. Please answer the following question.

1. What's the scenario you set to use Airtable? Why Airtable is a suitable tool for the scenario?
   想以視覺化的方式統整吉卜力角色的特徵時使用。Artable 的可以用更視覺化的方式去呈現各種統計資料以及其中的細節
2. How did you organize your data? Please share your steps and concerns.

把年齡跟身高的資料類型改為數字，方便作為圖表的統計量，加入各角色的圖片，來增加視覺化的效果，並把 special power 改為多選讓一個角色可以有多個超能力，最後將各筆資料用 release date 來 sort，更能看出吉卜力隨時間產生人物的變化。

3. Design two interfaces, dashboard and record review, based on the scenario. While designing the elements and layout of the interface, what's your goal?

[Deliverable] Please paste screenshots of your organized "data" and "interface", and paste sharable URL of your database. (please make sure you have published the interface, or we can't check any of your great work!)

[Bonus] At your choice, you can pick your own dataset to finish the task, or try to explore Airtable a bit more and try some new features that we don't cover in the lecture or Lab. Please tell us what you discovered with a few screenshots. (up to+0.25)