透過實際案例學習如何撰寫好的求職履歷 - 使用文字探勘技術之初步研究

李昀潔, 資管博一
呂孟芸, 國企四
張譯心, 國企四
陳韋霖, 資工碩一
詹雅安, 會計四

對於學生而言，除了努力完成學業或完成研究以取得畢業之外，如何透過簡歷 (Resume) 的撰寫去應徵企業的招募也是一件重要的事。然而如何撰寫簡歷不是一件簡單的。因此我們小組從網際網路上找尋到一個人簡歷的開放資料集，透過分群方法（Clustering）來對該資料集進行研究，我們檢驗履歷的寫法是否能對應到公司招聘的職位，另外我們也找出一些能夠申請多種工作類別的簡歷寫法，此次研究希望透過該開放資料集所作的初步研究，希望能對學生們製作個人簡歷上能有所幫助。

CCS Concepts: • **Applied computing → Document management and text processing**;

ACM Reference Format:
李昀潔, 呂孟芸, 張譯心, 陳韋霖, and 詹雅安. 2022. 透過實際案例學習如何撰寫好的求職履歷 - 使用文字探勘技術之初步研究. 1, 1 (January 2022), 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 研究目標

投履歷找工作時，如何撰寫一份符合業主要求的履歷，尤其對於高度專業的工作而言，更顯重要，然如何撰寫履歷等事宜，雖坊間有許多教條方法供大眾學習，但寫法上需考慮求職者的狀況，並無法直接使用法則來套用。有鑑於此，我們使用網路上公開之求職簡歷資料，運用課堂所學習之文字探勘技術來針對該些資料進行分析，藉此探求有趣的訊息以幫助求職者來面對如何撰寫履歷之問題，於本專案中，我們設定了兩個主題進行探討，其一為驗測該些資料是否各對應之履歷內容是否符合該求職之工作，其二為找出部份具有跨領域寫法之履歷。

Authors' addresses: 李昀潔, 資管博一; 呂孟芸, 國企四; 張譯心, 國企四; 陳韋霖, 資工碩一; 詹雅安, 會計四.

XXXX-XXXX/2022/1-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 研究方法

本專案中，我們所探討之資料，為從網路上所抓下來之公開資料，而該些資料除了各履歷的內容外，也包含了各別履歷所對應之求職職位。

以文字探勘而言，雖然求職職位與其履歷資料這兩種資料型可視為一種監督式學習技術之應用（Supervised Learning）[6]。然於本專案中，欲探討之主體非以分類問題之面象而進行之，而是希望從資料本身特性出發，藉由資料分析方法，試圖找出有趣的履歷範本，意即透過非監督式學習技術（Unsupervised Learning）[6]。

關於分析內容，共分為兩項工作，第一我們先檢驗資料集內容是符合常識，意即我們想確認是否各求職類別是否對應該些求職履歷，透過分群的方法去檢驗是否各群集（Cluster）內是否有內聚，第二透過分析資料內容，找出部份具有跨領域之履歷資料並作為範本供讀者參考，其方法上我們視此該些資料視為資料探勘問題 [4]，以找出適合之履歷資料。

2.1 資料集

針對履歷資料，我使用了公開資料集作為本專案之資料，其以 Kaggle 網站公開之實際之工作求職者履歷資料進行分析 [1]。

該資料共擁有三個欄位，分別為 ID，Resume_str，Resume_html 及 Category，ID 欄位表示流水號，中間兩欄位分別為 Resume 的純文字及其原始抓下來的格式（內含 HTML 標籤），Category 表示對應所求職之職位名，此資料集共有 24 個求職職位，求職種類及數量表可參照. 1。資料集共有 2484 筆資料，意即共有 2484 筆 resume。

2.2 資料前處理

於本專案中，我們主要進行兩項工作，然而在進行之前，我們需要進行資料前處理工作，才能將履歷資料轉換成我們可以分析的資料。

我們透過使用 Sentence-Transformers[8] 工具，將資料集之真實履歷資料，透過 BERT 方法 [3]，將該些履歷內的文字，轉成合適之 word embeddings。該些 word embeddings 為資料集中，所有履歷之文字，經過轉換後之新資料表示法，其原因為，若不經過轉換，原始資料中所有文字所表示之方法，其通常表示法可能為詞袋（bag of word）或是 one-hot encoding 編碼等，但這些表示法都有空間浪費的問題，因為所有文字所產生之矩陣（Matrix）都具有高度稀疏性 [9]，因此透過 BERT 轉換 [3]，將原有的矩陣維度進行降維（dimension reduction），可在保存原有資料的特徵外，減少了原本高度稀疏所造成的問題。

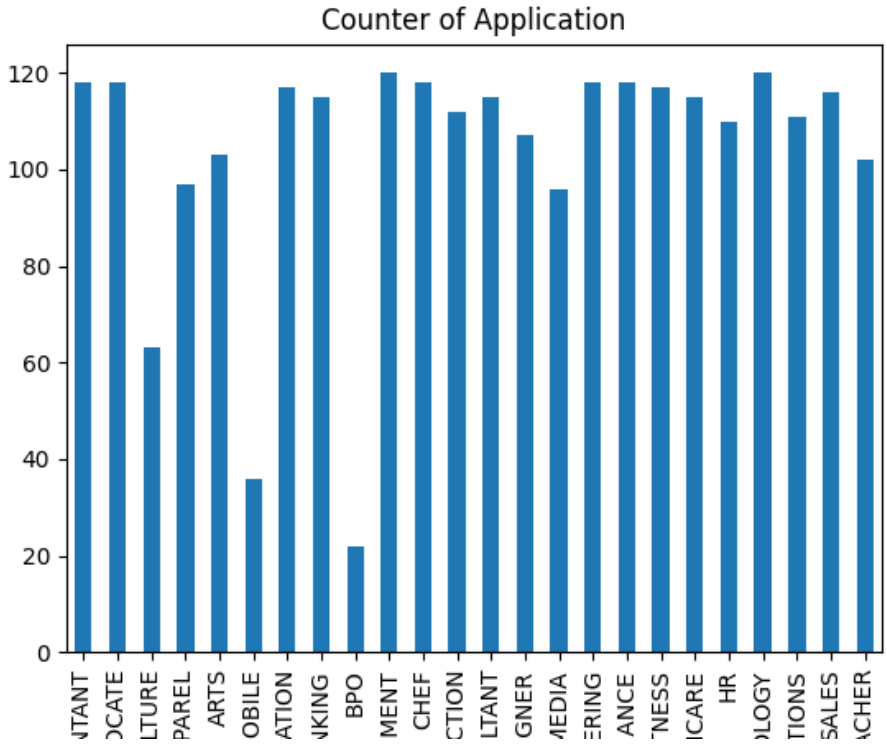

Fig. 1. The counting for each application catogery in resume

2.3 履歷一致性

在取得 word embeddings 後，我們會先針對此資料，使用 python 的 sklearn[2] 套件，使用主流之分群方法進行檢驗 [4]，其目的為確保該些履歷資料具有可信性，意即針對特定群體之履歷，其所對應之相同職缺，該些個別履歷資料之特性應具有相似性。

第一種分群方法，我們採用 k-means 分群方法 [5]，第二種分群方法，我們採用 PCA（Princial Component Analysis）[7]。

針對結果，我使用了 t-sne[10] 作圖像化。

2.4 跨領域範本

我們使用歐及里德距離公式，計算出各群之中存在較偏遠之履歷資料，並將其視為具有跨領域性質之履歷。

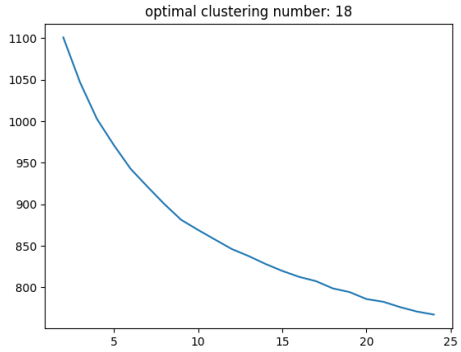

Fig. 2. Elbow method for optimal clustering number

3 研究結果

3.1 履歷一致性

為了要檢驗資料中的履歷是否與該申請之職缺相符，我們使用了 k-means[5] 作為分群方法，利用 elbow method 來取得最適分群數（Optimal clustering number），其結果如圖. 2，所得值為 18，

3.2 跨領域範本

4 結論

ACKNOWLEDGMENTS

謝謝老師上課所教授之知識，使得我們能得以完成此次報告研究。

REFERENCES

[1] Snehaan Bhawal. 2021. Resume Dataset. https://www.kaggle.com/snehaanbhawal/resume-dataset. Accessed: 2021-12-30.

[2] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. 108–122.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[4] Jiawei Han, Jian Pei, and Micheline Kamber. 2011. *Data mining: concepts and techniques*. Elsevier.

[5] James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Vol. 1. Oakland, CA, USA, 281–297.

[6] Kevin P Murphy. 2012. *Machine learning: a probabilistic perspective*. MIT press.

[7] Karl Pearson. 1901. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science* 2, 11 (1901), 559–572.

[8] Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. *arXiv preprint arXiv:2004.09813* (04 2020). http://arxiv.org/abs/2004.09813

[9] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*. Vol. 39. Cambridge University Press Cambridge.

[10] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).