Emoji recommender

表情圖示推薦系統

張鎧 R09725057

張智鈞 R09725051

吳延東 R09725049

沈宏穎 R09725043

1. Purpose

表情圖示（emoji），是使用在網頁和聊天中的形意符號，最初是日本在無線通信中所使用的視覺情感符號（圖畫文字），可用來代表多種表情，如笑臉表示笑、蛋糕表示食物等，而現在聊天軟體盛行，在聊天過程中人們往往需要用文字以外的方式去表達自己的情緒，像是貼圖、動圖，而其中表情圖示 emoji 的使用也自然而然漸趨頻繁，但每次使用 emoji 時都得自己去一個一個找來用其實累積起來已經花去了大量時間，同時，每個人其實都有自己特定幾個喜愛用的 emoji，或是有些比較常見的表情像是開心、難過等，大家選的表情都大同小異，此外現今手機的打字功能已經普遍能推薦使用者下個可能會使用到的字，且準確率十分的高，這時若有個能夠推薦使用者貼圖的系統就能讓使用者方便許多。

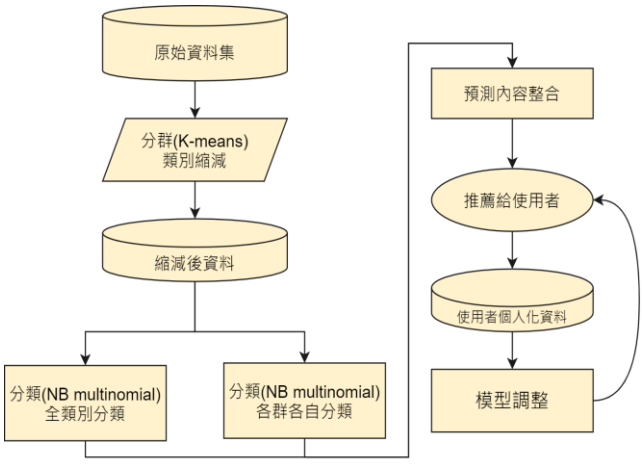
其實現在大家的手機中都有選擇 emoji 替代文字的功能，像是輸入什麼注音符號會跑出什麼 emoji，但此功能本身並不方便，在花了大量時間進行手動設定之後，還得輸入與當初設定正確無誤的注音才能夠跑出想要的貼圖，所以在很不實用的情況下也自然沒有什麼人知道此這功能，因此使用這真正需要的是在文字輸入完之後，系統能夠自動推薦結尾的 emoji，此時使用者就可以不用顧慮自身一定得打什麼字才能得到想要的貼圖，又能夠省去許多找 emoji 的時間。

2. Methodology

2.1 System Overview

資料來源是從 2020 年在 KDD conference 發表的一篇論文 *Emoji Prediction: Extensions and Benchmarking*，總共有 2,183,418 筆資料，每筆資料都有文字以及其對應的 emoji，總共有 221 種 emoji 的類別。

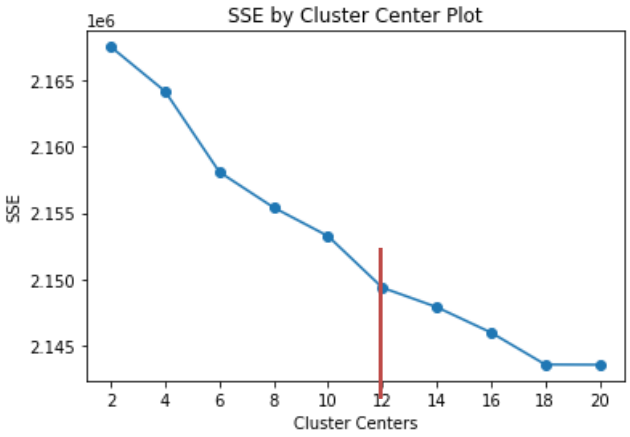
本研究方法將過程分為兩個 stage 加上一個 re-train 的機制，第一個 stage 將原始資料集的資料進行 k-means 分群，將分群完的資料丟進去第二個 stage 中進行 classification，此步驟有兩個分類器，分別是使用 NB multinomial 分類器為所有的類別進行分類，以及同樣使用 NB multinomial 分類器分別為每一群各自進行分類，之後藉由兩個分類器投票產生預測結果，將預測結果推薦給使用者後，系統會將使用者每次的選擇紀錄到使用者的個人化資料，以方便我們日後利用這些個人化資料進行預測模型的 re-train，使我們的推薦結果更加個人化。



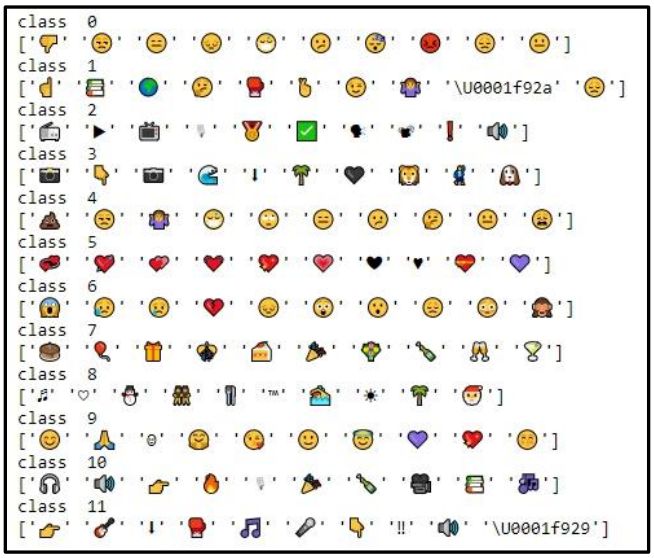
圖一、系統流程圖

2.2 Stage1：k-means 分群

在第一階段的模型中，我們使用 k-means clustering 作為分群方法，k-means 的目的是把 n 個點劃分到 k 個群中，使得每個點都屬於離他最近的均值（此即群的中心）對應的群，以之作為分群的標準，而對分群問題來說，要將資料點分為幾群便成為了最主要的問題，為了得知適合的分群數目，我們使用手肘法，得以判斷以 12 群作為基準來進行分群較為合適。分群完之後依據 emoji 在各群內的量，刪減 emoji 至 158 個，並準備送到下一個階段的模型進行預測。



圖二、手肘法決定以 12 群作為基準分群



圖三、第一階段 k-means 分群的結果

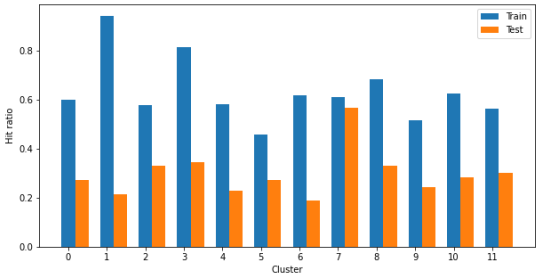
2.3 Stage2：分類

2.3.1 全類別分類

第一個分類器將刪減部分 emoji 過後的資料併群，而在分類器的選擇上我們選擇 NB Multinomial 作為我們的分類器進行分類，在全類別分類的分類器中，得出的每筆預測會直接挑選出前 20 名 emoji 作為預測結果。

2.3.2 各群各自分類

第二個分類器將第一階段得出的分群結果進行 emoji 刪減過後，針對 12 群分別訓練 12 個 NB Multinomial 分類器，接著計算文字的 tf-idf 值與 12 群各自的中心點的距離，使用最接近的兩個分類器分別預測 10 個來作為此分類器的預測結果。



圖四、12 個群各自分類器的預測表現

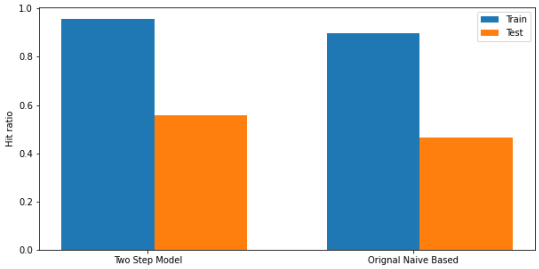
2.4 預測內容整合

將上述兩種分類器（共三個分類器）產生出的分類結果進行整合，整合方式為三個分類器各自選出的排名結果加總，有重複則刪去，最後取出至少前 20 名的 emoji 作為該次推薦結果。

3. System Outcome

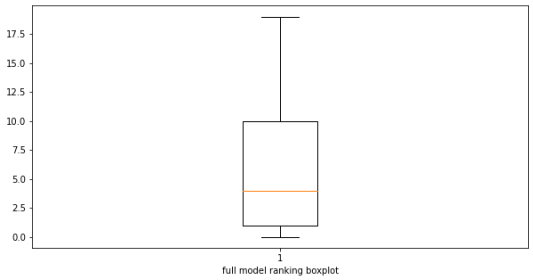
3.1 模型比較

為了比較本次研究中 two-step 模型的表現，我們將 two-step 模型的預測表現，與只使用 NB Multinomial 模型對整個資料集進行預測的表現進行比較。

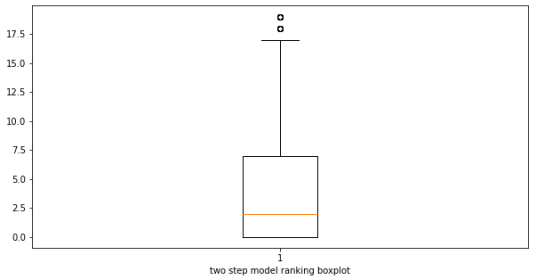


圖五、兩模型預測結果比較

我們以 Hit-ratio 作為比較基準，Hit 表示預測的 20 個 emoji 中有命中實際需求，Miss 表示 20 個 emoji 中沒有命中，數值（0-1）越大表示預測越準確，從圖五可以看出不論在 train 的結果以及 test 的結果，two-step model 比只使用 NB Multinomial 分類器的分類結果還好。



圖六、NB Classifier 預測分布



圖七、two-step model 預測分布

從圖六、七可以得知兩個模型的預測分布情形，y 軸代表命中的預測排名，越低代表排名越好，代表越適合使用者的 emoji 的排名會被排到更前面。如圖所示，可以看出 two-step model 的預測表現分布比只使用 NB Multinomial Classifier 進行分類的模型還要好。總結來說，除了模型的準確率以外，本研究中亦考量到排名結果來衡量模型的效果。

3.2 預測結果

下圖為模型預測的範例，可以看出預測結果還算不錯，也出現了一些有趣的結果，比方說因為我們的資料集來自 2020 年，因此在輸入 Covid-19 時，可以發現預測的結果會出現口罩及全球的 emoji ，又或者說輸入 China 或 Trumph 的時候，出現的 emoji 大都分布在負面情緒或者是無奈等，根據結果推測，應是因為資料集的蒐集來源有包含 Twitter、Facebook 等社群，因此反映了該社群中大多數人的使用習慣。從結果而言，我們認為除了能根據訊息推測 emoji 以外，推測出來的 emoji 的結果亦可以反映出時事、社群內對於不同事物的反應情緒。

Message	Emoji
Covid-19 is spreading around the world	
I got a cold in this cold day	
She get a champion in the game	
Donald Trump is a president of USA	
China is a greatest country	
I love my dog	
Let's have a party tonight	
Happy halloween!	
Would you marry me	
Happy winter vacation!	

圖八、推薦結果範例

4. Conclusion

本研究提供的系統可以藉由輸入不同的句子，利用分群的方法刪減較不常用的 emoji 後，再透過兩個不同類型的模型進行預測並且整合，即時產生不同的 emoji 推薦序列，並能夠將使用者點選的結果回傳至系統內部儲存，將來更能夠對於這些使用者資料進行個人化調整，使推薦結果更貼近每個人日常所需。

4.1 未來方向

上述有提到 two-step 之後的 re-train 階段，可以使用先前蒐集到的個人化資料，重新訓練 stage 2 的分類器，使得預測結果更貼近每一個使用者的習慣。另一方面，我們希望減少模型的大小，並且提升預測的反應時間，在模型的訓練上，因為訓練資料包含了 200 多萬筆的資料，並且採用了分群及分類等兩種模型，我們希望再藉由個人化 re-train 的同時，進一步縮小模型。

另外可以藉由解讀文字得來的 emoji 列表，使得平台更進一步了解使用者對於新議題的想法，讓本系統不只對使用者有所助益，也對平台方有所貢獻。

5. Reference

1. Dataset：https://github.com/hikari-NYU/Emoji_Prediction_Datasets_MMS
2. https://arxiv.org/ftp/arxiv/papers/2007/2007.07389.pdf