

男女歌詞用詞分析與差異

作者:古貿昌,余興祐
許家誠,施文捷

簡介

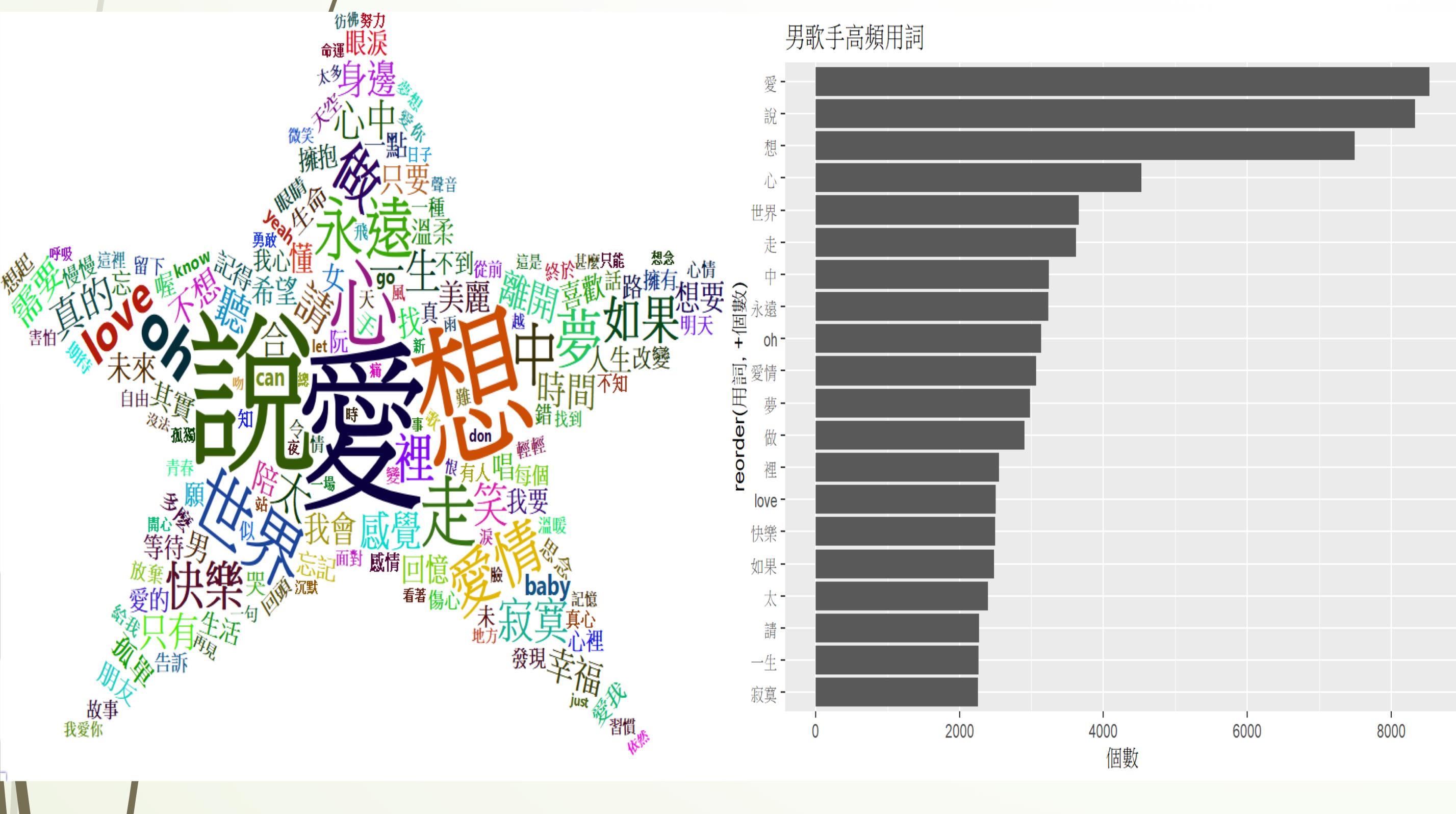
歌詞是語言中非常有趣的一部份，情感表述的方式比起日常用語來說更為豐富多變，我們對於用資料科學處理歌詞會有甚麼樣的結果感到好奇，決定了主題要與歌詞相關。社會的壓力或是期望，歌詞都會反映出來，不管內容是順從還是反抗社會。雖然歌詞的表現形式很容易讓意思模糊不清，或是充滿晦澀的象徵，有可能正是這種沒有清楚界線的文字更能表現出平常隱藏在表象下的真實。我們認為性別是一個分類歌手簡單直觀的標籤，也蠻可能有明顯的表現差異，所以最後決定蒐集男女歌詞並做出分析。

研究方法

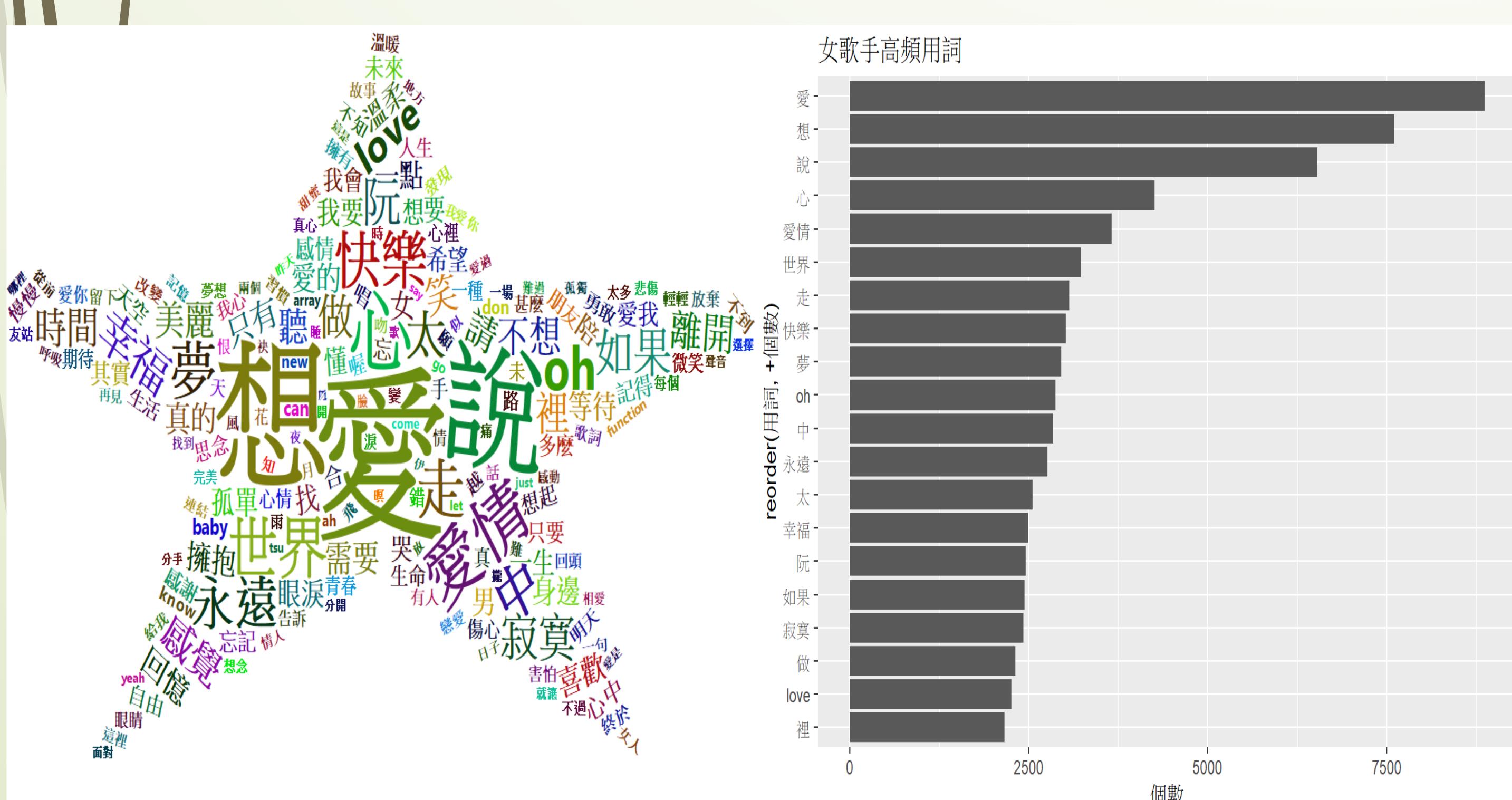
我們稍微修改了網路上的爬蟲，用python從魔鏡歌詞網上爬取華語熱門排行前80名歌手的所有歌詞，最後取得了男女各約14000首歌的歌詞。碰到的問題幾個是魔鏡歌詞的格式有點混亂，還包含許多意義不明的空白頁，讓我們在整理資料時遇到了一些麻煩。

斷詞的部份我們使用jeibaR進行。停詞表的部分有點麻煩，一般的停用詞表會刪掉代名詞和連接詞等比較沒有具體意義的詞彙，然而一開始我們認為男女用詞上的差距可能就隱藏在這些詞中。我們先試了比較簡單的停詞表，但是結果難以看出所以然，最後決定修改了網路上比較齊全的停詞表自製了一份。計算斷完詞後的資料後，用結果製成詞頻表並比較男女歌手之間的分別。

研究結果



男歌手用詞頻率文字雲與詞頻表



女歌手用詞頻率文字雲與詞頻表

觀察文字雲與詞頻表可以發現，男女用詞在高頻率的字詞上相似度很高，而有些出現頻率稍低的詞則有出現明顯男女分別，例如「一生」只在男生的歌有較高的出現率，而「孤單」則幾乎只在女生的歌中出現（但是男生的歌詞中有不少「孤獨」）。

高頻用詞中，最高的幾項出現的總次數非常高，「愛」這個字在男女歌手的詞中都出現了超過8000次。一樣是高頻用詞，男女的「說」和「想」出現頻率剛好顛倒。

女生的歌詞中「阮」的出現頻率很高，應該是台語的「我們」，表示歌詞中混入了不少台語歌，不能確定這對數據會不會有很大的影響。

結論

有些平常使用的詞彙，本身可能帶有一些平時不容易察覺的性別意識在其中。例如歌詞中女性會使用「孤單」，而男性則偏好「孤獨」，就是很有趣的一點，雖然意思相近，孤單和孤獨乍聽之下也沒有明顯的男女氣質分別，歌手使用上卻仍各有偏好。

在差異之外有點令人訝異的是男女在大部分詞彙運用上保持非常高的一致性，出現明顯差異的地方都在整體出現率較低的詞彙上。一致性最明顯的地方，男女對「愛」的興趣都非常濃厚，如果去除「說」和「想」這兩個主題沒這麼明確的字，「愛」的出現頻率在男女用詞都足足領先了第二名一倍，28000條歌中出現了16000次。

總而言之，就大方向來說，男女歌手的用詞相當一致，而大方向之下有許多雖然較微小，但不該忽略的重要差異。

討論和貢獻

這次的結果應能作為研究社會性別差異的素材。例如結論中提到的「孤單」和「孤獨」的頻率差異，在辭典解釋上，「孤獨」偏向社會狀態，而「孤單」則是表達心理上的感受，我們推測在撰寫歌詞時，就算歌唱本身就是一種鼓勵情緒表達的藝術，仍會受到「男性不應該輕易表達（脆弱）情緒」這種印象所影響。歌詞的表現手法多變，直白的、含蓄的、熱情的、晦澀的，雖然意思常常因此變的不明確，但也因此容易規避社會框架的束縛，更接近內心真正想法。

我們這詞的研究也有不少不足之處。首先，停詞表並沒有很嚴謹的被決定，而停詞表示這個研究非常重要的環，會極大幅的影響研究結果，如果時間充足，應該對停詞表多做闡述。其次，只根據詞頻來分類男女歌手略顯粗糙，詞語之間的關聯性也很有可能有重要差別。最後，我們的資料本身不夠乾淨，有混到不少台語歌跟少量日語和英文，因為我們是根據歌手去爬歌詞。雖然有手動整理了一下，但或許有更好的資料爬取或整理方式。

參考資料

https://github.com/hhpslily/mojim_crawler

