國 立 交 通 大 學

資訊科學與工程研究所

碩 士 論 文

近體詩主題辨識系統研製

Jintishi Processing and Categorization

研 究 生：王笙權

指導教授：梁　婷　教授

中 華 民 國　一百零一 年 八 月

近體詩分類系統研製

Jintishi Processing And Categorization

研 究 生：王笙權　　Student：Sheng-Chiuan Wang

指導教授：梁　婷　　Advisor：Dr. Tyne Liang

國 立 交 通 大 學
資 訊 科 學 與 工 程 研 究 所
碩 士 論 文

A Thesis
Submitted to Institute of Computer Science and Engineering
College of Computer Science
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
Master
in

Computer Science

August 2011

Hsinchu, Taiwan, Republic of China

中華民國 一百零一年 八月

近體詩主題辨識系統研製

研究生：王笙權　　指導教授：梁　婷　博士

國立交通大學

資訊科學與工程研究所

摘要

近體詩是中國文學的精粹之一，以精簡的文字表達豐富的情感與思想。此外詩作也可能包含大量的典故與對仗，因此近體詩對於一般人而言在理解與創作上存在著一定程度的困難。有鑑於此，本論文利用文本分類技術，以進行近體詩處理研究，並建立一個近體詩主題辨識系統。此系統提供詩作相關查詢及詩作處理功能包括斷詞、概念標記、情感辨識、及內容主題辨識等。本研究將主題辨識歸類成詠物述志、山水田園、情愛閨怨、贈別思友、邊塞征戰、社會民生等六項；情感辨識標註為喜愛、怨怒、哀愁等三項。在主題辨識的實驗中我們以 992 首七言律詩作為實驗語料，萃取詩作的八種詞彙與概念特徵，以支援向量機(SVM)模組進行辨識。經過 tenth-fold cross-validation 檢驗，主題辨識的平均正確率為 69.12%。以同樣的模組，在情感辨識的實驗中我們以 492 首七言律詩作為實驗語料，得到 70.7%的辨識正確率。

Jintishi Processing and Categorization

Student：Sheng-Chuan Wang Advisor：Dr. Tyne Liang

Institute of Computer Science and Engineering

National Chiao Tung University

ABSTRACT

Jintishi is one of the Chinese literature classics. Jintishi reveals rich emotion and thoughts in few words. Jintishi may contain allusions and follows syntactic and semantic parallelisms making them difficult to be understood. Therefore, we used text classification techniques to analyze Jintishi and built up a Jintishi topic identification system. The system provides poem search and poem analysis including word segmentation, semantic tagging, topic identification and emotion identification. We classified Jintishi into six topic categories, namely, Chanting Object, Landscape, Desperate Wife, Farewell, Frontier and Social Poem. Additionally, our system supports emotion categorization, namely, happiness, sadness or anger. We used 992 seven-character Lushi in topic identification labeling experiment. We extracted eight lexical and concept Jintishi features and used support vector machine to identify topics for each poem. We get 69.12% accuracy after ten-fold validation. The emotion identification method was performed and tested too. Using 492 seven-character Lushi as test corpus, we get 70.7% accuracy.

誌謝

首先最感謝的是我的指導教授梁婷老師，感謝老師在這兩年的研究所生涯中啟發我對於論文研究的方向，並指導我對於論文寫作上的技巧，而且除了學術理論上的教導外，老師也經常將自己的人生經驗和我們分享，在老師的啟蒙之下著實讓我獲益良多。另外也要感謝口試委員張俊盛教授、楊武教授、葉慶隆教授對於論文的寶貴建議，讓本論文更加完善。

其次要感謝楊哲青學長，總是能將我零散的想法整理成完整的架構，每一次的討論都能將論文的內容向前推進一步，真的十分的感謝。還有實驗室的冠熙學長，每次總是不辭辛勞的處理實驗室大小事務，解決我們各式各樣的疑難雜症，你的樂觀態度也總是讓我在低潮時給我力量，我們實驗室能有你這名大將真好。同時俊樺學長和家棋學長當我程式有問題時問你們總是能夠迎刃而解。羿賢與鴻達總是主動邀請我每日的籃球運動，讓我適時的抒解生活上的壓力。晨輝一整個就是情報王，總是可以找到有趣的話題一起聊天。以及晉榮在競爭激烈的課業勇奪書卷獎。上還有其他無法一一詳述的朋友們，感謝你們在我研究生涯中的一路相伴。

最後要感謝我的家人，感謝你們在我求學路上徬徨無助時適時的拉我一把，陪我分享著每一刻的快樂與憂傷，有了這個強力的精神支柱，讓我更有勇氣在人生道路上大步的向前邁進。

目錄

表目錄

圖目錄

第一章 緒論

研究背景與目的

興起於唐宋時代的近體詩，是中國文學的精粹之一，使用精簡的文字表達複雜的情感與思想。從勉人勵志、感時諷世、親友離別、思鄉之苦等詩作流傳至今，仍為人們所朗誦應用。這些詩作不僅能提升人們使用語言的能力，也豐富人們生活內涵。創作近體詩有嚴謹的格律要求[12]，如符合字數的規定、平仄聲律的安排、對仗的要求如表 1。因此在有限的字句裡，要表達豐富的意境，還要兼顧形式、內容、聲調的美感，誠屬不易，對一般人而言在理解與創作上有一定程度的困難。有鑒於此，本論文利用文本分類技術，進行近體詩處理研究，建立一個近體詩主題與情感辨識系統。藉此協助使用者對詩作內容及情感表達的理解，進而增加學習與欣賞的樂趣。

表 1 近體詩格律

字數的規定	絕句：一首共 4 句，又分五言一句或七言一句，稱五言絕句或七言絕句 律詩：一首共 8 句，又分五言一句或七言一句，稱五言律詩或七言律詩
平仄的安排[1]	以七言律詩為例有四種類型： 1. 平起首句不押韻 2. 平起首句押韻 3. 仄起首句不押韻 4. 仄起首句押韻
對仗的要求	1. 絕句可對仗，也可以不對仗 2. 律詩的第二聯、第三聯必須對仗；第一、第四聯可對可不對 3. 對仗的兩句，句型相同、詞性相同、平仄相反

目前線上詩詞系統大多著重於內容查詢的功能，對於詩作理解上幫助甚少。因此我們要解決的問題在於建構一個系統，處理項目包含斷詞、概念標記、情感辨識和主題辨識等功能，幫助使學習者更容易了解詩詞內容。

在自然語言處理中，斷詞的正確與否往往影響後續的實驗。目前的斷詞系統大多著重於白話文的應用，且斷詞的處理經常隨著不同的領域切詞的規則也不盡相同。因此我們必須針對近體詩文本，開發新的斷詞系統。另外，一詞多義的問題也是自然語言處理中重要的課題，以近體詩的七言律詩為例，詩人僅使用 56 個字表達複雜的情感、思想。為了幫助使用者理解詩作，我們將每首詩作的每個詞彙標記概念，藉此找出詩作所要表達的情感及思想。並利用這些詞彙與概念，透過文本分類技術辨識詩作的主題與情感。本研究依據朱我芯的定義將詩作歸類成詠物述志、山水田園、情愛閨怨、贈別思友、邊塞征戰與社會民生等六項主題[14]

[1] http://www.360doc.com/content/10/1216/18/3966739_78751830.shtml

如表 2。另外在詩作情感方面，依照顏崑陽所定義的，將詩作歸類成喜樂、怨怒、哀愁等三項[11]如表 3。

我們收集了 7117 首近體詩，以 992 首七言律詩作為實驗語料，萃取詩作的九種詞彙與概念特徵，以支援向量機(SVM)模組進行辨識，完成近體詩主題辨識系統的建置。本論文結構：第二章為論文的相關研究部分；第三章介紹詩作處理流程及方法；第四章討論詩作特徵與分類實驗；第五章介紹並展示我們的系統；第六章敘述結論及未來的發展方向。

表 2 詩作主題定義

主題	定義
詠物述志	借萬物寄託詩人自己的感情，詠物詩中的寄托往往跟詩人的經歷際遇、人生態度、生活作風、價值取向等有關係，以表現手法上說是借物來抒發志向。
山水田園	描寫清新的自然景色，山水草木，都富含詩人獨特的審美情趣，或歌詠閑適恬淡的田園生活、田間勞作為題材的詩歌。
情愛閨怨	描寫男女愛慕之情和愛情生活，或抒發離別相思之苦，大多是用第一人稱來直敘自己的愛情，也有些是以第三人稱觀點來寫。
贈別思友	表現朋友之間的摯愛深情、離情別緒，一般為即景抒情，詩的開頭是敘事，或寫景，然後是抒情表意。
邊塞征戰	描寫邊塞風光、反映邊疆將士生活為基本內容，抒發報效國家、渴望建功立業的豪情，或狀寫將士的鄉愁、邊塞征戰的殘酷、描寫塞上絕域的奇異風光等。
社會民生	利用嘲諷或勸喻手法，揭露社會黑暗、世態炎涼。或以憑弔古跡、歷史故事、古人事跡為題材，借此抒發情懷，諷刺時事。也有懷才不遇時，詩人抒發情感，或是感嘆年華老去仍無所做為。

表 3 詩作情感分類定義

情感	定義	歸納
喜樂	「喜」:亢奮的情緒狀態 「樂」:寧和、積極、愉快的心態沒有一時起伏翻騰的情緒	科場得志、親人朋友同堂之歡、 歸隱之樂、國家振興之喜、男女相愛之喜、生命或自然冥合之樂
怨怒	「怨」:一種夾在怒與哀之間的情緒 「怒」:最具衝動性的，發動時會引起強烈活動且具有相當程度的破壞性	大丈夫之怒、俠客之怒、反戰之怒、貧士之怒、烈女之怒、謫臣之怨、農臣之怨
哀愁	「哀」:悲傷的情緒 「愁」:是介於憂慮與悲哀的情緒	個人境遇之悲、愛情的悲哀、鄉愁的悲哀、生命的悲哀、歷史的悲哀

第二章 相關研究

2.1 斷詞處理

詞是最小有意義的語言單位，任何語言處理的研究都必須先能分辨文本中的詞才能進行進一步的處理。此外處理不同領域的文本時，領域相關的特殊詞彙或專有名詞，常常造成斷詞系統錯誤的切分詞彙。

俞士汶與胡俊峰[2003]觀察到詩詞文本常出現合成詞，這種情況下單使用互訊息不能顯示較好的結果，藉由詞彙的結合強度找出這一詞彙是否該被切分。首先經請專家對詩詞人工切詞建立語料庫，對語料庫中在同一句詩中同現的所有二字組統計相鄰與不相鄰的頻率，利用這兩個字相鄰的次數除以全部出現的次數得到此二字組的結合強度。其結果在出現頻率大於 20 且其結合強度大於 1 時，有 90% 的機率是一個詞。

陳紹宜[2010]考量近體詩特有句法結構幫助詩作斷詞。五言詩的句法結構{2/2/1, 2/3, …}共 9 種句型規則、七言詩的句法結構{4/3, 2/2/3, …}共 14 種句型規則。以黃鶴樓的「白日依山盡」這句，可以切分「白日/依山/盡」、「白日/依山盡」的斷詞，再根據詞彙出自的詞彙庫給予權重，選擇最高分的句型作為最後斷詞結果。斷詞成果在 70 首五言絕句 F-score 為 69.15%。

2.2 詞義處理

一詞多義的處理在自然語言裡是重要的研究議題，目前常用的解歧義的方法可分成監督式消解歧義與非監督式消解歧義。

監督式消解歧義技術，主要是透過人工標記的語料當作系統的訓練語料，因此建立相當耗時。柯淑津[2004] 提出以詞彙為主，結合配搭訊息與機率模式的方式來處理語意自動標記工作。首先由人工進行標記部分工作，再透過以 bootstrap 技巧來搭配詞彙來決定目標詞彙的詞義，並且決定出詞義的句子擴充訓練語料，再利用機率模式設法標註更多句子。經過實驗在可接受的應用率 74.9%下能獲得高正確率 92.6%。

非監督式消解歧義不需要有人工標記好的語料，因此有較好的通用性與便利性。Rada Mihalcea[2007]使用 wikipedia 及 WordNet 解決一詞多義的問題。收集 wikipedia 中含有短文解釋的詞彙以及 WordNet 當中定義的短文，然後這些短文經過 POS tag 後，使用資料探勘的方式找出每個語意所搭配的特徵，完成訓練語料。當發生一詞多義時，會根據前後文的特徵來解決詞彙歧義，其消解正確率可達 84.65%。

2.3 詩作分類

詩作的分類有很多種包含主題應用、風格派系、情感分類等等。目前以分析詩作的詞彙、概念、詞性及標題為主要特徵，最透過 Naïve Bayes、SVM 等分類器分類結果。但在近體詩詞彙屬於文言文，使得詞彙概念辨識較為困難，同時也沒有一個標準語料庫做為系統測試基準。

Gamon[2004]針對 1441 篇勃朗特文學作品依照作者類別進行分類。其主要方法是利用微軟所開發的 NLPWin 對每一篇作品進行語意、詞性以及句型架構的分析，並統計這三種特徵的頻率，按照頻率高低篩選出較佳的特徵之後，使用 SVM 分類器找出此篇作者。其分成三類的正確率達到 97.5%。

Yi Yong et al.[2005]針對 398 首宋詞依照風格分類成豪放及婉約。此篇最大的特點是未經過處理斷詞，他們把每一個字單視為一個特徵，透過基因演算法找出最佳的特徵集，最後透過 Naïve Bayes 判斷其風格分類。實驗結果顯示正確率達到 88.5%。

Zhong-Shi et al[2007]透過專家對 413 首宋詞做切詞和標記語意。利用詞連接的自然語言分析方法作為斷詞依據，將詞彙之間的連接建立成最佳搜尋樹，透過此搜尋樹決定句子的中心語以及句子的語意傾向，利用 Information gain 找出最佳的詞彙及語意特徵，透過 SVM 分類成豪放或婉約的風格。此研究分類正確率達 88.6%。

劉博榮[2010]把五言絕句分成六大類。斷詞部分利用句法結構{2/2/1, 2/3, …}共 9 種句型規則結合詞庫匹配的方法，計算出每一種句型的分數並取其最高分的句型完成斷詞。並提出啟發式規則將每首詩的每一詞彙標記一個概念，再利用 Chi Square 篩選出較具有分類依據的特徵，最後使用 SVM 分類器完成分類。其研究正確率達 72.4%。

表 4 詩作分類相關研究

	英文	中文			
	Gamon [2004]	Yi Yong et al. [2005]	Zhong-Shi et al. [2007]	劉博榮 [2010]	本研究
外部支援辭典	無	無	專家制定詞集	典故、CKIP、舊版 TYCCL	典故、CKIP、新版 TYCCL
斷詞處理	無	單字詞	詞彙連結	句型規則	複合規則
概念標記	NLP win	無	詞彙連結	啟發式規則	啟發式規則
實驗語料	1441 篇 勃朗特作品	398 首 宋詞	413 首 宋詞	1080 首 五言絕句	992 首 七言律詩
訓練:測試	8:2	5:5	9:1	8:2	8:2
分類模組	SVM	Naïve Bayes	SVM	SVM	SVM
特徵選取	Frequency	Genetic Algorithms	Information gain	Chi Square Test	Chi Square Test
分類特徵	語意、詞性、句型架構	詞彙	詞彙、語意	詞彙、語意	詞彙、語意
類別數	3 類	2 類	2 類	6 類	6 類
正確率	97.5%	88.5%	88.6%	72.4%	七言律詩69.12% 五言絕句72.81%

第三章 研究方法及步驟

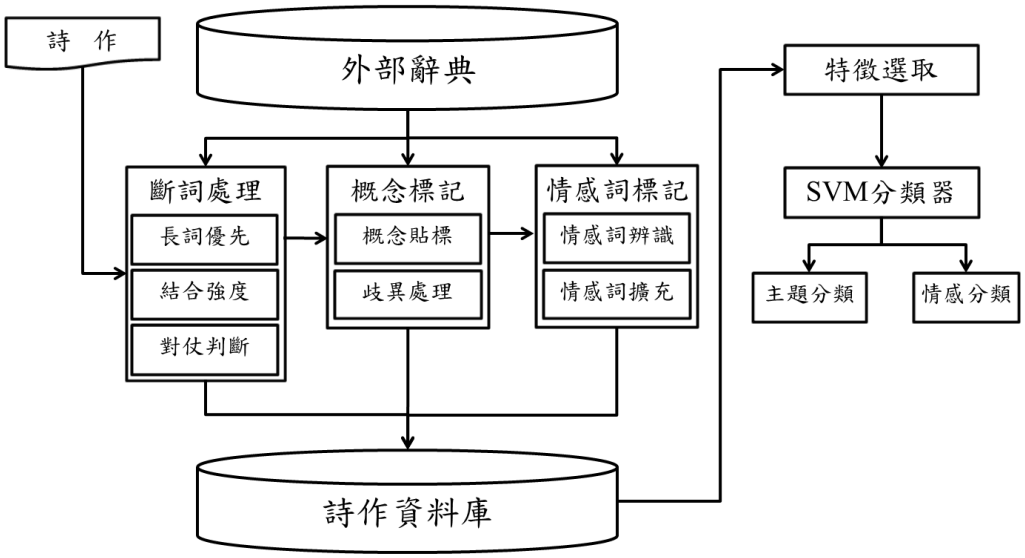

圖 1 系統流程圖

本研究的處理流程如圖 1 所示，輸入一首詩作，經過斷詞處理切分詞彙，再對每一個詞彙標記同義詞詞林之概念，同時辨識這些詞彙是否屬於情感詞彙，最後將這些詞彙、概念、情感詞存入詩作資料庫。接著根據我們定義的特徵於詩作資料庫選取特徵資訊，透過支援向量機模組對詩作進行主題辨識與情感辨識。後面將詳細介紹本論文使用的語料來源、外部辭典、斷詞處理、概念標記及情感詞標記之步驟。

3.1 語料來源

詩作語料庫是來自「維基文庫[2]」，由於簡體轉正體關係使得資料常有亂碼或缺字的現象，故我們藉由「新詩改罷自長吟 全唐詩檢索系統[3]」幫助修正其正確性詩作。另外維基文庫中也發現詩作重覆的情形，我們在語料庫中也將其他重覆的部分刪除，並建立近體詩語料庫共 7117 首詩作如表 5，且記錄每首詩作的詩題、作者、類型、詩文。

表 5 近體詩語料詩作數量

詩作類型	詩作數量
五言絕句	1080
五言律詩	3362
七言絕句	1683
七言律詩	992

[2] http://zh.wikisource.org/
[3] http://cls.hs.yzu.edu.tw/tang/database/index.html

3.2 外部辭典建立

隨著文本領域的不同專有名詞也不盡相同。目前現有的中文辭典，同義詞詞林、E-Hownet、中研院八萬目詞等詞庫，這些詞庫收錄了詞彙及語意。雖然這些詞庫都是用於現代白話文，但測試 unigram、bigram 與 trigram 的比對，我們發現還是有部分的詞彙被收錄在這些詞庫中。在測試中發現同義詞詞林匹配到的詞彙多於中研院八萬目詞與 E-Hownet。另外，根據我們的觀察結果，詩人在創作近體詩時常使用典故、地名與風景名勝等等詞彙。有鑒於此，我們以同義詞詞林作為主要詞庫，並擴增典故詞彙、專有名詞與地名詞彙，建立符合近體詩文本的辭典，來源與收錄詞條數如表 6。

表 6 外部辭典來源

	來源	詞條數量
典故	詩詞典故[4]	2807
	詩詞曲典故[5]	25490
專有名詞	中研院八萬目詞[6]且屬性為+countries 與 +nomenclature	708
地方詞	八萬目詞且屬性為+districts	1029
同義詞詞林	哈工大資訊檢索研究室同義詞詞林擴展版[7]	77303

除此之外我們還觀察到，中研院八萬目詞有專含有名詞與地方詞的貼標(+countries, +nomenclature, +districts)，因此我們從中研院八萬目擴增專有名詞與地方詞。同時為了後續概念標記的處理，於擴增的同時將這些詞彙標記同義詞詞林的概念的代碼(Di02A01_國家、Dd15B02_姓氏、Cb25A11_洲縣)如表 7。

表 7 中研院八萬目詞與相對應的概念

詞彙	詞性	定義	對應到同義詞詞林的概念
七雄	Naea	+countries	Di02A01_國家
三晉	Naea	+countries	Di02A01_國家
張	Nbc	+nomenclature	Dd15B02_姓氏
段	Nbc	+nomenclature	Dd15B02_姓氏
河南	Nca	+districts	Cb25A11_洲縣
江西	Nca	+districts	Cb25A11_洲縣

[4] http://ch.eywedu.com/Story/Untitled-2.htm
[5] http://cls.hs.yzu.edu.tw:88/CM/query/orig_source.htm
[6] http://www.aclclp.org.tw/use_ced_c.php
[7] http://ir.hit.edu.cn/

典故詞彙方面，我們利用網路收尋詩詞典故，找到兩個網站含有典故詞彙。第一個詩詞典故網站除了有典故詞彙之外，還有典故的概念。如圖 2 中「亡羊路」與「西州路」的概念為「城建」。因此我們只需要將「城建」轉換成同義詞詞林的概念代碼，其轉換的步驟如表 8 所示。

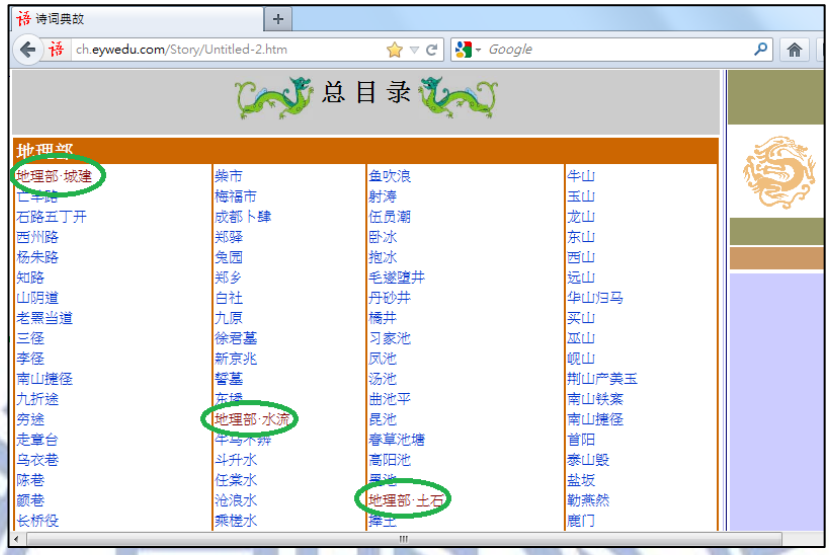

圖 2 詩詞典故網站

表 8 詩詞典故概念標記之方法

步驟一：擷取冒號後的片段，視為一個詞彙

步驟二：將詞彙比對同義詞詞林

若有存在，則標記其概念代碼(可以有多個概念)

若不存在，則人工判斷並標記代碼(可以有多個概念)

另一個詩詞曲典故網站如圖 3 所示，記錄著從漢朝到宋朝的典故。不像前一個網站，只有同義典故的資訊可以利用。只能使用當前的典故與同義典故，比對前一網站所建置的典故辭典，完成概念標記其方法如表 9。我們僅找到 273 個典故的概念。

典故	一丘之貉	
相關人物		
關鍵詞	相關典故	
	同義典故	一丘; 一丘貉; 同一丘; 貉一丘;
	參見典故	
典故出處	出處內容	
1.《漢書》卷六十六〈公孫劉田王楊蔡陳鄭列傳・楊敞・〈子〉楊惲〉~2891~	惲聞匈奴降者道單于見殺，惲曰：「得不肖君，大臣為畫善計不用，自令身無處所。若秦時但任小臣，誅殺忠良，竟以滅亡；令親任大臣，即至今耳。古與今如一丘之貉。」	

圖 3 詞曲典故網站

表 9 詩詞曲典故概念標記之方法

步驟一：當前的典故詞彙比對典故辭典

若有存在，則標記其概念代碼(可以有多個概念)

若不存在，則將同義典故的詞彙比對典故詞彙

若有存在，則標記其概念代碼

若不存在，則放棄標記

3.3 斷詞處理

詞是最小有意義的語言單位，任何語言處理的研究都必須先能分辨文本中的詞才能進行進一步的處理。本論文的斷詞處理分成三個部分，分別是詞庫匹配、詞彙結合強度與對仗處理。詳細步驟描述如下。

3.3.1 詞庫匹配：

使用我們收集的外部辭典，利用 N-gram 將詩作詞彙與詞庫詞彙做匹配，由 4-gram 遞減至 1-gram 如果詞彙有出現在詞庫出現即可視為一個詞；在匹配過程中我們有設定詞庫優先權,其順序典故 > 專有名詞 > 地方詞 > 同義詞詞林。

以杜甫的屬相為例：

直接比對詞庫，並標記詞彙的來源其結果如表 10。其標記符號如下:

a:表示此詞彙來自典故

c:表示此詞彙來自 CKIP

t:表示此詞彙來自同義詞詞林

s:表示此詞彙是透過結合強度計算而來

u:表示剩下來的單字詞彙

d:表示對仗詞彙

表 10 詞庫匹配結果

輸入	輸出
丞相祠堂何處尋	丞相(t) 祠堂(t) 何處(t) 尋(u)
錦官城外柏森森	錦官城(a) 外(u) 柏(u) 森森(t)
映階碧草自春色	映(u) 階(u) 碧(u) 草(u) 自(u) 春色(t)
隔葉黃鸝空好音	隔(u) 葉(u) 黃鸝(t) 空(u) 好(u) 音(u)
三顧頻繁天下計	三顧(a) 頻繁(t) 天下(t) 計(u)
兩朝開濟老臣心	兩(u) 朝(u) 開(u) 濟(u) 老臣心(a)
出師未捷身先死	出師(t) 未(u) 捷(u) 身(u) 先(u) 死(u)
長使英雄淚滿襟	長(u) 使(u) 英雄(t) 淚滿(a) 襟(u)

3.3.2 詞彙結合強度：

根據我們的觀察，完成詞庫匹配後的結果會偏向一個字例如「樓/前/海/月/伴/潮/生」，但實際應該分成「樓前/海月/伴/潮生」。這錯誤是因為詞庫裡沒有這些複合字「樓前、海月、潮生」。複合字判斷我們使用俞士汶與胡俊峰[13]提出的「結合強度」來計算兩兩單一詞是否能成為複合字。

我們利用七言律詩、七言絕句、五言律詩、五言絕句等四種體裁作為結合強度的訓練資料集，總共 7717 首近體詩；258,980 個字。主要計算方法，將相鄰的次數除以同一句中所有的配對如公式(1)。根據我們的觀察當字串的相鄰次數大於 5 且其結合強度大於 1 時可以確定該字串是一個詞。

$$結合強度 = \left(\frac{freq(C_iC_{i+1})}{freq(C_i, C_{i+1})}\right)^2 \times ln(freq(C_iC_{i+1})) \quad (1)$$

範例：

承接上一步驟的結果，計算兩兩單字的結合強度，若大於門檻值就將這兩個單字合併為雙字詞。經過計算後我們發現整首詩作只有「碧草」是一個詞彙而非兩個單字，所以將「碧草」合併成一個雙字詞如表 11。

表 11 詞彙結合強度結果

輸入	輸出
丞相(t) 祠堂(t) 何處(t) 尋(u)	丞相(t) 祠堂(t) 何處(t) 尋(u)
錦官城(a) 外(u) 柏(u) 森森(t)	錦官城(a) 外(u) 柏(u) 森森(t)
映(u) 階(u) 碧(u) 草(u) 自(u) 春色(t)	映(u) 階(u) 碧草(s) 自(u) 春色(t)
隔(u) 葉(u) 黃鸝(t) 空(u) 好(u) 音(u)	隔(u) 葉(u) 黃鸝(t) 空(u) 好(u) 音(u)
三顧(a) 頻繁(t) 天下(t) 計(u)	三顧(a) 頻繁(t) 天下(t) 計(u)
兩(u) 朝(u) 開(u) 濟(u) 老臣心(a)	兩(u) 朝(u) 開(u) 濟(u) 老臣心(a)
出師(t) 未(u) 捷(u) 身(u) 先(u) 死(u)	出師(t) 未(u) 捷(u) 身(u) 先(u) 死(u)
長(u) 使(u) 英雄(t) 淚滿(a) 襟(u)	長(u) 使(u) 英雄(t) 淚滿(a) 襟(u)

3.3.3 對仗的處理：

近體詩當中的律詩最大特徵在於對仗的要求，律詩除了首尾兩句可對仗，也可不對仗，沒有硬性要求，但中間兩聯卻有要求必須上下句對仗；絕句則沒有要求，因此在律詩中處理對仗是不可或缺的。對仗的規則詞性相同、字數相同、概念相似，我們就是利用「字數相同」這項特性來處理律詩的對仗。

完成詞彙結合強度後取得的結果，先給予二字詞以上的詞彙權重分數，典故 2 分、來自 CKIP 的詞彙 2 分、由結合強度統計出來的詞彙給予 2 分最後出自同義詞詞林的詞彙給定 1 分。最後將上下句分別做加總，取出分數較高者，其分數底的句子會依照分數高者重新斷詞。

範例：

律詩的格律有要求第二聯、第三聯一定要對仗，因此只處理此兩聯。

第二聯上句：映(u) 階(u) 碧草(s) 自(u) 春色(t)

其分數：0+0+2+0+2=4

第二聯下句：隔(u) 葉(u) 黃鸝(t) 空(u) 好(u) 音(u)

其分數：0+0+2+0+0+0=2

第二聯最後結果為：映 階 碧草 自 春色；隔 葉 黃鸝 空 好音

第三聯上句：三顧(a) 頻繁(t) 天下(t) 計(u)

其分數：2+1+1+0=4

第三聯下句：兩(u) 朝(u) 開(u) 濟(u) 老臣心(a)

其分數：2+0+0+1=3

第三聯最後結果為：三顧 頻繁 天下 計；兩朝 開濟 老臣 心

表 12 對仗處理結果

輸入	輸出
丞相(t) 祠堂(t) 何處(t) 尋(u)	丞相(t) 祠堂(t) 何處(t) 尋(u)
錦官城(a) 外(u) 柏(u) 森森(t)	錦官城(a) 外(u) 柏(u) 森森(t)
映(u) 階(u) 碧草(s) 自(u) 春色(t)	映(u) 階(u) 碧草(s) 自(u) 春色(t)
隔(u) 葉(u) 黃鸝(t) 空(u) 好(u) 音(u)	隔(u) 葉(u) 黃鸝(t) 空(u) 好音(d)
三顧(a) 頻繁(t) 天下(t) 計(u)	三顧(a) 頻繁(t) 天下(t) 計(u)
兩(u) 朝(u) 開(u) 濟(u) 老臣心(a)	兩朝(d) 開濟(d) 老臣(d) 心(d)
出師(t) 未(u) 捷(u) 身(u) 先(u) 死(u)	出師(t) 未(u) 捷(u) 身(u) 先(u) 死(u)
長(u) 使(u) 英雄(t) 淚滿(a) 襟(u)	長(u) 使(u) 英雄(t) 淚滿(a) 襟(u)

3.4 詞彙概念的標記與歧異消解

3.4.1 概念標記:

研究指出概念是有助於分類[3][4][5][18][19]，因此我們需要把每個詞彙標記概念。首先每個詞彙都進行詞庫匹配比對外部辭典，並標記所有回傳結果的概念。如果沒有回傳任何概念即是未知詞，這些未知詞是由單一字詞、結合強度與對仗處理所切分出來的詞彙。在未知詞的處理中，首先判斷詞長，若詞長等於 1，則標記”unknown”；如大於 1，則將詞彙切分成單一字的集合。此集合的所有單一字都需要比對詞庫，並標記所有回傳的概念；所有單一字都不存在辭典中，則標記”unknown”。

表 13 概念標記結果

斷詞結果	概念標記
三顧(a)	皇帝_Af05A01
頻繁(t)	一再_Ka10A02
天下(t)	地方_Cb08A01
計(u)	儀錶_Bo18A01,方法_Db09A01,計畫_Df09A01,打算_Ga05B01,計算_Hj29C01
兩朝(d)	unknow@雙方_Dd05B06,二_Dn04A03,幾_Dn05B02,絲_Dn10A15,少_Eb01B01,朝代_Ca02B01,早晨_Ca27B01,政府_Dm01A01,朝廷_Dm01A05,向_Kb01A01
開濟(d)	unknow@挖_Fa10A01,開_Fa31A01,射擊_Hb06A01,設立_Hc05B01,開除_Hc22A02,支付_He10C01,駕駛_Hf01A01,揮筆_Hg11A01,沸騰_Ia10A01,解凍_Ia11C01,開花_Ib21A01,舉行_Ie13B01,開始_Ig01A01,周濟_Hi36A03
老臣(d)	unknow@老人_Ab02A01,公公_Ab02A04,慣例_Da03A04,老_Eb15D01,長久_Eb24A01,舊_Eb29A01,年老_Eb36A01,深_Ec05A01,本來_Ed51B01,老練_Ee21A01,非常_Ka01A01,一直_Ka11B01,官吏_Af08A01
心(d)	心_Bc05F01,心_Bk14B01,中_Cb04C01,點_Cb23A01,內心_Df02A01

最後發現 1044 個單字詞(unique 426)找不到概念，這些詞彙中有部分的頻率偏高，此時我們利用「教育部國語辭典」查詢頻率大於 10 的未知詞的概念，並以人工方式標記概念。例如: 圖 4"殢"在教育部國語辭典裡有三種解釋，再分別將詞彙「滯留、逗留、纏綿、糾纏、沉迷、沉溺」於詞庫做概念查詢。查詢後得到「滯留_Hj02A01、逗留_Hj02A01、纏綿_Ed32A05、糾纏_Hi51A01、沉迷_Ga14A01、沉溺_Ga14A01」，因此"殢"將會被標記成"殢_Hj02A01、Ed32A05、Hi51A01、Ga14A01"。雖然經過人工查詢、標記但仍然有 89 個(unique 83)在教育部國語辭典找不到其資訊，這些詞彙被標記為"unknow"。

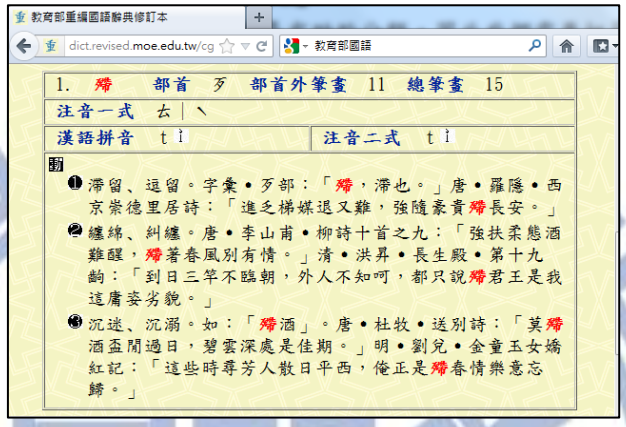

圖 4 教育部國語辭典

3.4.2 詞彙歧義消解

前一步驟詞彙標記概念，產生了 30521 個詞彙擁有多個概念，但我們期望一個詞彙只能擁有一個概念，為此我們改良劉博榮[19]提出的啟發示規則概念歧義處理消解多義詞之概念。啟發示規則概念歧義處理，是由共同出現詞彙、共同出現概念、最短概念距離三個規則循序漸近處理歧義消解。

建立共現資料庫：

「共同出現詞彙、共同出現概念」其主要方法是藉由人工標記產生的共同出現詞彙及共同出現概念幫助消解歧義詞彙。此資料庫由人工對詩作進行斷詞並標記其概念，接著 window size 以一句為限，建立共同出現詞彙的配對及共同出現概念的配對，並且記錄配對出現的次數；另外也記錄每個詞彙和每個概念出現的次數。範例如下表 14：

表 14 以「日日/思歸/勤/理/鬢」為例建立共同出現詞彙資料庫之表示

人工標記	共現詞彙表	共現概念表	詞彙頻率表	概念頻率表
< 日日, (C_1,2) >	< (日日, C_1), (思歸, C_2), 3>	< (C_1, C_2), 5>	<日日, 5>	< C_1, 7>
< 思歸, (C_2,1) >	< (日日, C_1), (勤, C_3), 2>	< (C_1, C_3), 7>	<思歸, 3>	< C_2, 3>
< 勤, (C_3,2) >	< (日日, C_1), (理, C_4), 1>	< (C_1, C_4), 4>	<勤, 6>	< C_3, 4>
< 理, (C_4,2) >	< (日日, C_1), (鬢, C_5), 1>	< (C_1, C_5), 4>	<理, 2>	< C_4, 4>
< 鬢, (C_5,2) >	< (思歸, C_2), (勤, C_3), 1>	< (C_2, C_1), 1>	<鬢, 7>	< C_5, 6>

共同出現詞彙：

首先建立每一詩句詞彙組合，例如詩句被斷詞成「日日/思歸/勤/理/鬢」，五個詞彙則產生的詞彙配對集合為 $Pair\{$(日日,思歸), (日日,勤), (日日,理), (日日,鬢), (思歸,勤), (思歸,理), …$\}$，然後查詢人工標記資料庫，看這些配對是否有存在共同出現資料庫。如果有出現，代表這兩個詞彙的概念有可能與資料庫相同，因此標記其概念。

此方法有可能發生多個配對同時被查詢到，且同一詞彙有重複但概念不一樣的情況。例如兩個配對$Pair_1\{(W_1, C_1), (W_2, C_2), 3\}$與$Pair_2\{(W_1, C_3), (W_3, C_4), 2\}$，表示$Pair_1$和$Pair_2$同時出現資料庫中，但$W_1$卻是不同概念，此時較比兩配對的頻率並標記較高者。如果兩個配對的頻率一樣，則比較詞彙W_2和W_3的頻率並標記較高者的配對概念。

共同出現概念：

當詩句進行共同出現概念處理時，先將詞彙分類已消歧義之詞彙$Wu = \{C_1, C_2, \dots\}$及待消歧義之詞彙$Wa = \{(C_1, C_2, \dots, C_k), (C_1, C_2, \dots, C_k), \dots\}$兩集合。再透過已知概念配搭候選概念所組成的概念配對，尋找人工標記的共同出現概念資料庫。

因此當 Wu 集合不是空集合時，將會產生概念配對$Pair(Wu_i, Wa_{j_k})$並查詢資料庫。如果命中多個配對，例如$Pair_1(Wu_i, Wa4_{j_k})$與$Pair_2(Wu_i, Wa_{j_{k+1}})$之狀況，則比較其出現頻率並標記較常出現之概念配對。倘若多個配對其頻率一樣時，則比較單一概念的頻率，並標記較高頻率者之配對。但發生詩句全部詞彙都有多概念，即 Wu 集合為空時，則產生$Pair(Wa_{j_k}, Wa_{j+1_k})$的概念配對，並重復查詢資料庫之動作。

最短概念距離：

同義詞詞林主要是收錄現代漢語詞彙的辭典，但也有少部分常見的古代用詞。它們把詞彙分成大、中、小三類，每個小類有很多詞，這些詞有根據詞義的遠近和相關性分成了若干個詞群如表 15。其中符號「=、#、@」分別代表「相似、不等、獨立」，在本研究考慮到方便仍視為相似，即「Da15B02=」與「Da15B02#」在本研究仍視為同義詞。根據同義詞詞林的概念表示，我們可以將「Da15B02」轉化成樹狀結構如圖 5，並且設定每個距離的值為 1。如此即可判斷兩個概念的相似度。

表 15 同義詞詞林概念符號

編碼位	1	2	3	4	5	6	7	8
符號舉例	D	a	1	5	B	0	2	=\#\@
符號性質	大類	中類	小類		詞群	園子詞群		
級別	第 1 級	第 2 級	第 3 級		第 4 級	第 5 級		

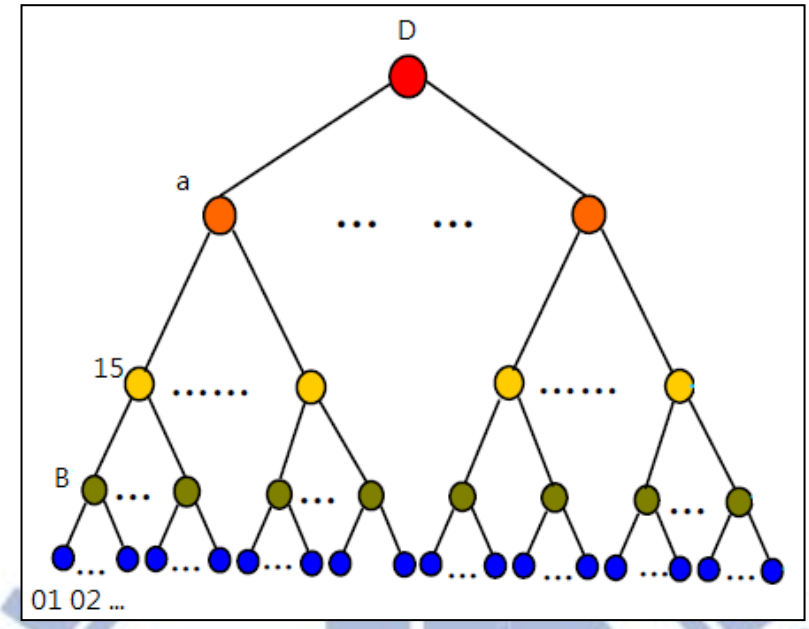

圖 5 同義詞詞林概念樹狀圖

我們觀察到近體詩經常看到在詞彙與詞彙之間有著相似的概念。以律詩來說詩人為了講求對仗，在第二聯與第三聯出現相似概念的次數為最。例如《吳融 上巳日花下閒看》的第二聯「雲鬢照水和花重，羅袖抬風惹絮遲」，當中的”雲鬢、羅袖” 及 ”花、絮”都是對仗詞彙。另外也發現詩作會出現句中對的情況例如《溫庭筠 七夕》的第一句「鵲歸燕去兩悠悠」，當中的” 鵲、燕” 與 ”歸、去”也屬於對仗詞。有鑒於此，我們將此方法分為兩個子項目「隔句-最短概念距離」與「同句-最短概念距離」其兩者定義如表 16，這也是我們和[劉博榮 ‘10]的方法差異之處。最後透過同義詞詞林的階層式架構找出相似概念。

表 16 規則限制

	Window size	字數
隔句-最短概念距離	同一聯	位子相同且字數相同
同句-最短概念距離	同一句	不限制

3.5 情感詞標記

情感辨識中情感詞是一重要指標，因此情感詞標記是一關鍵步驟。從先前的研究大部分使用詞庫來辨識情感詞彙，我們仿造其方法嘗試使用 NTUSD[8]、E-hownet 作為情感詞庫且辨識詩作詞彙但其效果不彰。主要原因有二，經過我們的斷詞處理後其結果偏向一個字，但其 NTUSD、E-hownet 單一字的情感詞寥寥無幾；另一個原因是 NTUSD、E-hownet 主要用於現代語的詞庫並不適用於詩作。因此選擇採用統計的方法辨識詩作中的情感詞彙。此方法主要是計算詞彙在每個情

[8] http://nlg18.csie.ntu.edu.tw:8080/opinion/pub1.html

感類別出現之機率如公式(2)，由於情感類別詩作數量分佈不均(232 首喜樂詩、73 首怨怒詩、388 首哀愁詩)，需要做正規劃如公式(3)，然後挑選其正規劃後的權重大於 0.7 之詞彙作為情感詞彙。另外有些詞彙字雖然不是情感詞，但其概念含有情感，因此希望可以找出含有情感概念的詞彙作為擴充。使用同樣的機率公式，然後挑選其正規劃後的權重大於 0.8 之概念。

$$P(term) = \frac{freq(term)}{|class_i|} \tag{2}$$

$$normalize = \frac{P(term)}{\sum_{i=1}^{3} \frac{freq(term)}{|class_i|}} \tag{3}$$

第四章 實驗語實驗分析

4.1 斷詞結果與分析

檢驗 120 首詩斷詞結果其七言律詩的 F-score 為 70.43%。先前提到同義詞詞林可以匹配較多的詞彙，但根據表 18 看到一個字的詞彙占了 69.26%，主要原因同義詞詞林是用於現代語料的工具，表示我們所擴增的詞彙數量是有待提升。

表 17 斷詞結果比較

	劉博榮 [2010]	本論文
字典	舊版 Tyccl、CKIP 專有名詞、典故	新版 Tyccl、CKIP 專有名詞、典故
規則	Lexicon base、句型規則	Lexicon base、詞彙結合強度、對仗
F-score	70 首五言絕句 69.15%	70 首五言絕句 77.18% 120 首七言律詩的 F-score 為 70.43%

表 18 字數統計

	詞彙個數	詞彙不重覆個數
一字詞	29104(69.26%)	2978
二字詞	12441	6041
三字詞	352	312
四字詞	125	120
五字詞	2	2
小計	42024	9453

韋莊的”上元縣”斷詞結果：

南朝 三十 六英 雄，角 逐 興 亡 盡 此中
有國 有 家 皆 是夢，為龍 為 虎 亦 成空
殘花 舊宅 悲 江令，落日 青山 弔 謝公
止 竟 霸 圖 何 物 在，石 麟 無主 臥 秋風

這首詩的第一聯和第二聯明顯的錯誤，正確斷詞應為：

南朝 三十六 英雄，角逐 興亡 盡 此中
有國 有家 皆 是夢，為龍 為虎 亦 成空

第一聯的錯誤是詞庫權重造成的，根據我們的方法，典故詞彙優先於其他詞彙。”六英”是一個典故詞，所以他會最先被切分成一個詞，導致其他詞彙被切分錯誤。第二聯則是結合強度訓練語料不足，導致可以找到”為龍”是複合字，但是”為虎”卻無法被切分出來。

4.2 概念標記結果與分析

概念標記結果

其標記結果如表 19 所示，概念個數有很大的差距(1 個~34 個)，是因為在標記的方法，如果詞彙不存在詞庫中則標記此詞彙所有字的概念，所以導致一個詞彙最多有 34 個概念。我們期望一個詞彙只有一個概念，單一概念的詞彙占了 27%，所以剩下的 73%的詞彙需要解歧義。

表 19 概念標記結果以 992 首律詩為例

概念個數	詞彙個數
0	89
1	11414
2	6393
3	5101
4	4254
5	3838
6 個以上	10935
總計	42024

詞彙歧義消解

此實驗資料集共 240 首詩(每個主題各 40 首)，這些詩作都經過人工斷詞並標記概念。使用 120 首詩(每個主題各 20 首，總詞彙 5165 個)做訓練語料建立共現資料庫，再拿另外的 120 首詩(每個主題各 20 首，總詞彙 5127 個)做為實驗語料。實驗結果如表 20。其中「單一概念」代表詞彙在比對同義詞詞林只找到單一概念，故不計算其正確率及召回率。

表 20 詞彙歧義消解實驗結果

	單一概念	共同出現詞彙	共同出現概念	隔句-最短概念距離	同句-最短概念距離
比對詞數	5127	3757	2943	1787	1076
標記詞數	1370	814	1156	711	1076
涵蓋率	26.72%	15.88%	22.55%	13.87%	20.99%
正確標記詞數	1370	653	931	564	346
正確率	-	80.22%	80.54%	79.32%	32.16%
召回率	-	17.38%	31.63%	31.56%	32.16%

從實驗可看出 ”共同出現詞彙”、”共同出現概念”、”隔句-最短概念距離” 三者的正確率高於 ”同句-最短概念距離”，其錯誤標記的原因整理如下。

1. 例如詩句「垂簾幾度青春老，堪鎖千年白日長」的詞彙 ”老” 經過共同出現概念所標記的意思是 Ab02A01_老人，但實際上正確的概念是 Eb36A01_年老。在

“共同出現詞彙”與“共同出現概念”的標記容易被頻率影響，造成多者恆多的情況。

2. 前文有提到對仗可分詞性相對、語意相似，而我們只考慮語意相似的部分，此情況容易發生標記相近的概念但非正確的概念。由其 ”同句-最短概念距離”最常出現此錯誤。

4.3 情感詞標記結果與分析

我們標記情感詞的方法，是依照詞彙在此情感類別出現的機率，詳細公式在前一章節已經說明。總共標記 584 個情感詞彙，並擴充 60 個情感概念如表 21。

表 21 情緒詞彙統計表

情感類別	單一字詞	雙字詞	情感概念
喜樂	110	96	23
怨怒	64	139	19
哀愁	93	82	18

表 22 呈現每個情感類別的前 15 個情感詞，可以發現有些詞彙不屬於情感詞彙，例如”退”、”液”等等。這是因為只考慮頻率作為辨識情感詞的關係，但是這些詞彙搭配其他詞彙可以表現出情感意境，因此我們仍視為情感詞彙。

表 22 情感詞辨識結果(前 15 個)

喜樂		怨怒		哀愁	
詞彙	權重	詞彙	權重	詞彙	權重
慶	0.944	夜長	0.941	淚	0.947
綵	0.938	家住	0.914	傷	0.9
上苑	0.93	胡天	0.914	趨	0.868
歡娛	0.93	長恨	0.914	相逢	0.868
韶	0.921	獨眠	0.914	風塵	0.843
奉	0.921	誰謂	0.914	渡	0.827
美	0.909	砧	0.914	首	0.827
歸來	0.893	免	0.914	忍	0.827
退	0.893	任他	0.914	不堪	0.827
乘	0.883	飄零	0.914	生涯	0.827
液	0.87	刑	0.914	滅	0.827
日暖	0.87	猶在	0.914	依依	0.807
精	0.87	無才	0.914	都	0.807
梅花	0.87	蓬山	0.914	茫茫	0.807
潭	0.87	穴	0.914	不勝	0.807

4.4 特徵定義與特徵選取

對於在詩作分類中所使用的特徵，我們將其列表如下：

1. 詩題單字詞：將詩題切分成單字詞作為特徵。
2. 詩題單字詞概念：利用詩題中的單字詞，其所標記的概念作為特徵。
3. 詩題雙字詞：將詩題切分成雙字詞作為特徵。
4. 詩題雙字詞概念：利用詩題中的雙字詞，其所標記的概念作為特徵。
5. 詩文單字詞彙：經過斷詞後所產生的單字詞彙。
6. 詩文單字詞概念：經過斷詞後所產生的單字詞彙，其所標記的概念作為特徵。
7. 詩文多字詞彙：經過斷詞後所產生的二字以上的詞彙。
8. 詩文多字詞概念：經過斷詞後所產生的二字以上的詞彙，其所標記的概念作為特徵。
9. 情感詞彙：使用情感詞彙當作特徵，期望能夠幫助分類。

決定使用的特徵後，接下來必需選取這些特徵中具代表性的部分，而有些在單一主題中出現頻率較高的詞彙或是概念，未必就是此主題具代表性的特徵，例如像「人」、「我」這類的詞在每一個主題都會出現的詞彙，若以這些詞彙當作關鍵字，對於整個訓練與分類過程幫助不大，且會降低分類時的正確率，在考量特徵選取時，需以能正確表示主題性質的特徵為主，常用的特徵選取方法有 TF-IDF、資訊增益(Information Gain)、卡方檢定(Chi-square test)…等，在 Yang et al. ['97]實驗中，卡方檢定和資訊增益相較於其他的方法有良好的分類正確度，經我們對於兩個方法的測試後考量所選出的特徵，所以卡方檢定來做為選取特徵值的方法。對於每一個特徵F_i，本研究所使用的卡方檢定公式如下：

$$\chi^2 = \frac{n(AD - BC)^2}{(A + C)(B + D)(A + B)(C + D)} \tag{6}$$

n：該主題所有的詩作數

A：屬於該主題且有特徵F_i的詩作數

B：屬於該主題且沒有特徵F_i的詩作數

C：不屬於該主題且有特徵F_i的詩作數

D：不屬於該主題且沒有特徵F_i的詩作數

除了使用卡方檢定外，我們另外訂定了一項特徵選取公式，稱為特徵類別比例(*Feature Class Ratio*)，這個公式用來計算單一特徵占不同主題之間的比例，藉以了解特徵對於該類別是否具代表性，特徵類別比例之公式如下：

$$Feature\ Class\ Ratio_c = \frac{Uni\ Class\ Frequeny_c}{All\ Class\ Frequeny} \tag{7}$$

表 23 Feature Class Ratio 參數說明

$Feature\ Class\ Ratio_c$	特徵於類別 c 所占之比例
$Uni\ Class\ Frequeny_c$	特徵在類別 c 出現的頻率
$All\ Class\ Frequeny$	特徵在全部的類別出現的頻率

整合以上兩個公式，我們特徵選取的方法如下：

首先，以特徵出現的詩作數為單位，計算卡方檢定的值，並刪除卡方檢定的值太低，對該類別缺乏鑑別度的特徵。然後，以特徵出現的頻率為單位，計算特徵類別比例的值，若特徵類別比例低於訂定的門檻就將該特徵刪除。

4.5 主題辨識與情感辨識實驗

在 992 首七言律詩中，我們依照朱我芯與顏崑陽所定義的資訊，詩作主題部分人工標記 160 首詠物述志、62 首山水田園、62 首情愛閨怨、284 首贈別思友、48 首邊塞征戰及 376 首社會民生。另外情感詩部分總共標記 232 首喜樂詩、76 首怨怒詩、385 首哀愁詩、轉換情感 20 首詩與中性 279 首詩，其中轉換情感和中性的詩作我們不做處理。

本論文中以支援向量機作為我們的理論基礎，支援向量機是由 Vapnik et al.等人所提出以統計學習理論(Statistical Learning Theory)為基礎，針對資料分類、迴歸與圖形辨識的機器學習工具，其應用領域包括影像辨識、資訊探勘、文件分類等。我們使用 Chang et al.所開發的 Libsvm 作為辨識詩作的分類器，其中的核心函數部分我們選用 RBF，參數 gamma 和 cost 的部分，則是利用 Libsvm 中的 grid 程式來反覆測試，找出最佳的 gamma 和 cost，再經由分類器來對訓練資料集來找出最佳的平面。

在特徵的組合方面，我們共有 9 個特徵，若直接以窮舉的方法來選取特徵組合相當的耗時，故我們使用的 Backward Sequential Selection Algorithm 來做特徵選取，這個方法大致上是先將全部的特徵放進 SF(Select Feature)集合，然後每一次減少一個特徵做分類實驗，挑一個具有最高分類正確率的特徵組合，放進 SF 中再做一次分類實驗，直到所有分類實驗的結果都小於 SF 的結果就停止此演算法。我們將演算法詳述如表 24。

表 24 Backward Sequential Selection Algorithm

Step1：將所有特徵放入 *SF* 集合中，執行詩作分類實驗。

Step1.1: 紀錄 *BestF=(SF)*，*BestEval=Eval(SF)*

Step2：對上一步驟的 *BestF* 集合每一次減少一種特徵 F，執行詩作分類實驗。

Step2.1: 如果 *Eval(SF)*大於 *BestEval*

則紀錄 *BestF=(SF)*，*BestEval=Eval(SF)*，並執行 Step2

Step3：若所有的 *Eval(SF)*小於 *BestEval* 時，停止演算法，可得對佳特徵組合 *BestF*

首先將所有的特徵視為最佳特徵組合，再利用 SVM 分類器找出其正確率為 64.62%。使用 Backward Sequential Selection Algorithm 找出最佳的特徵組合，第一回合所有組合的結果如表 25：

表 25 特徵代號

特徵代號	特徵描述
F1	詩題單字詞
F2	詩題單字詞概念
F3	詩題雙字詞
F4	詩題雙字詞概念
F5	詩作單字詞
F6	詩作單字詞概念
F7	詩作多字詞
F8	詩作多字詞概念
F9	情感詞彙

表 26 第一回合分類結果

特徵組合	正確數量/全部數量	正確率
F2+F3+F4+F5+F6+F7+F8+F9	124/195	63.59%
F1+F3+F4+F5+F6+F7+F8+F9	115/195	58.97%
F1+F2+F4+F5+F6+F7+F8+F9	125/195	64.10%
F1+F2+F3+F5+F6+F7+F8+F9	121/195	62.05%
F1+F2+F3+F4+F6+F7+F8+F9	121/195	62.05%
F1+F2+F3+F4+F5+F7+F8+F9	118/195	60.51%
F1+F2+F3+F4+F5+F6+F8+F9	127/195	65.12%
F1+F2+F3+F4+F5+F6+F7+F9	122/195	62.56%
F1+F2+F3+F4+F5+F6+F7+F8	134/195	68.72%

經過第一回合我們找出最高的特組合 SF={F1+F2+F3+F4+F5+F6+F7+F8}其正確率 68.72%，將此特徵組合繼續執行進行第二回合。

表 27 第二回合分類結果

特徵組合	正確數量/全部數量	正確率
F2+F3+F4+F5+F6+F7+F8	127/195	65.13%
F1+F3+F4+F5+F6+F7+F8	114/195	58.46%
F1+F2+F4+F5+F6+F7+F8	122/195	62.56%
F1+F2+F3+F5+F6+F7+F8	124/195	63.59%
F1+F2+F3+F4+F6+F7+F8	116/195	59.49%
F1+F2+F3+F4+F5+F7+F8	119/195	61.03%
F1+F2+F3+F4+F5+F6+F8	124/195	63.59%
F1+F2+F3+F4+F5+F6+F7	121/195	62.05%

經過第二回合的，我們發現所有結果都小於第一回合的結果，因此停止演算法，得到最佳特徵組合為 SF={F1+F2+F3+F4+F5+F6+F7+F8}。此特徵組合每類別辨識結果如表 28：

表 28 主題辨識結果

	C1	C2	C3	C4	C5	C6	小計
訓練集個數	129	50	50	228	39	301	797
測試集個數	31	12	12	56	9	75	195
正確辨識個數	23	5	5	49	3	49	134
錯誤辨識個數	8	7	7	7	6	26	61
辨識總數	33	6	5	86	5	60	195
Accuracy	74.19%	41.67%	41.67%	87.5%	33.33%	65.33%	68.72%

實驗結果顯示經過 Backward Sequential Selection Algorithm 以後使得正確率從 64.62%提升到 68.72%。實驗分析如下：

1. 根據演算法結果我們得到特徵"情感詞彙"對於主題辨識是負影響的，那是因為同一主題可以包含多種情感。例如「贈別思友」的詩作有敘述朋友之間因為離別感到悲傷，也有描寫他鄉遇故知的愉快心情；另外在「社會民生」類別中也有官場升遷與不得志的詩作。這些都是導致"情感詞彙"對於主題辨識是負影響的因素。
2. 「送耿拾遺歸上都：若為天畔獨歸秦，對水看山欲暮春。窮海別離無限路，隔河征戰幾歸人。長安萬里傳雙淚，建德千峰寄一身。想到郵亭愁駐馬，不堪西望見風塵。」這首詩系統辨識出來的類別是贈別思友，但正確的類別是邊塞征戰。錯誤的原因為詩題中的"送" 、"歸"，詩作內容的"獨"、"歸"、"別離"、"愁"等都是贈別思友的特徵，屬於邊塞征戰的特徵只有"征戰"、"駐馬"。

此外我們使用最佳特徵組合執行 tenth-fold cross-validation 測量分類，其正確率可達 69.12%。主題辨識實驗的語料分布如表 29，其結果如表 30。

表 29 tenth-fold cross-validation 測量之主題辨識語料分佈

主題類別	總資料量	亂數產生訓練數量	測試數量
詠物述志	160	144	16
山水田園	62	55	7
情愛閨怨	62	55	7
贈別思友	284	255	29
邊塞征戰	48	43	5
社會民生	376	338	38
總計	992	890	102

表 30 主題辨識 tenth-fold cross-validation 測量結果

正確分類/測試數量	正確率	平均正確率
74/102	72.55%	69.118%
70/102	68.63%	
73/102	71.57%	
69/102	67.65%	
67/102	65.67%	
73/102	71.57%	
70/102	68.63%	
68/102	66.67%	
69/102	67.65%	
72/102	70.59%	

雖然"情感詞彙"這一特徵無法幫助主題辨識，但單獨使用此特徵使用 SVM 分類模組進行情感辨識得到 70.7%正確率。情感辨識實驗的語料分布如表 31，其結果如表 32。

表 31 tenth-fold cross-validation 測量之情感辨識語料分佈

主題類別	總資料量	亂數產生訓練數量	測試數量
喜樂	232	208	24
怨怒	76	68	8
哀愁	385	346	39
總計	693	622	71

表 32 情感辨識 tenth-fold cross-validation 測量結果

正確率	正確分類/測試數量	平均正確率
50/71	70.42%	70.702%
51/71	71.83%	
50/71	70.42%	
49/71	69.01%	
51/71	71.83%	
50/71	70.42%	
50/71	70.42%	
49/71	69.01%	
52/71	73.24%	
50/71	70.42%	

第五章 系統展示與介紹

本系統從網路上擷取 7117 首近體詩，且記錄每首詩作的詩題、作者、類型、詩文。詩作辨識的實驗，以 992 首七言律詩建立分類模組。系統的平台為 Windows 7，並使用 Python 與 php 程式語言做為開發工具，建立近體詩主題辨識系統如圖 6 所示。此系統包含詩作查詢、主題定義、主題辨識與概念查詢等功能。

圖 6 近體詩主題辨識系統

詩作的分類有很多種，包含主題、情感、體裁、作者等等。因此為了讓使用者了解本系統分類的依據，我們提供主題定義的功能圖 7。本系統依據朱我芯的定義將詩作主題分成詠物述志、山水田園、情愛閨怨、贈別思友、邊塞征戰、社會民生等六項[14]；詩作情感方面，依照顏崑陽所定義的，將詩作歸類成喜樂、怨怒、哀愁等三項[11]。除此之外，同時提供每個類別常出現的詞彙與概念，讓使用者了解每個主題的特徵如圖 8。

圖 7 主題定義

圖 8 主題特徵

系統查詢功能方面，以詩作的標題、內容、作者、體裁以及主題做為查詢的指標。其中相同欄位可以使用布林檢索給定多項條件，不同的欄位以”AND”為條件判定。根據檢索的條件展示出原始資料，包含主題、情感、斷詞、概念與平仄如圖 9。為了讓使用者了解作者擅長的主題以及常使用什麼概念來作詩，本系統提供每位詩人在系統中的相關統計如下圖 10、圖 11。除此之外，本系統還提供使用者標註功能如圖 12，期望領域專家來幫助修正詩作資料，提高詩作資料正確性。

圖 9 詩作資料

圖 10 作者相關統計

圖 11 作者常用概念

圖 12 人工標記平台

有別於其他近體詩系統只有提供詩作查詢，主題辨識功能是我們的貢獻如圖13。辨識主題之前需要找出相關特徵，因此需要先進行斷詞及概念標記如下圖 14、圖 15，其方法已在第三章節描述，最後呈現系統辨識的結果圖 16。

圖 13 主題辨識首頁

圖 14 斷詞結果

圖 15 概念標記結果

圖 16 主題辨識結果

本系統的概念是源自於同義詞詞林，但是一般使用者可能不了解或不熟悉同義詞詞林，因此我們提供概念查詢的功能。此功能除了可以查詢概念，同時也能夠查詢詞彙如下圖 17、圖 18。

圖 17 概念查詢

圖 18 概念查詢解果

第六章 結論

本論文提出並實作一個近體詩主題辨識系統，當使用者輸入詩作，本系同能夠有效的對其詩作進行斷詞、概念標記、辨識詩作主題及辨識情感。本論文主要貢獻如下：

1. 收集 7117 首詩作，包含 1080 首五言絕句、3362 首五言律詩、1683 首七言絕句、992 首七言律詩。
2. 人工標記詩作主題及詩作情感 992 首七言律詩。
3. 收集 28297 個典故詞彙，包括 25490 的詩詞曲典故及 2807 的詩詞典故，其中 3080 個典故詞彙擁有概念標記。
4. 利用結合強度、律詩對仗的特性提高斷詞程序的效能。
5. 提出詩作詞彙概念的歧義解決策略。
6. 建置近體詩處理系統，提供查詢作者、詩作內容、主題、體裁、詩作標記平台、辨識詩作主題等多項功能。

本論文未來研究方向有下列幾點：

1. 擴充近體詩詞庫以提高近體詩作中的概念。
2. 擴充系統以提供詩作賞析之功能。
3. 針對詩作主題辨識將加強詩眼、詩句修辭等領域知識，提高辨識率。

參考文獻

[1] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines (2001). Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[2] Gerard Escudero and Lluís Màrquez and German Rigau (2004),"An Empirical Study of the Domain Dependence of Supervised Word Sense Disambiguation Systems."Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpor, Hong Kong.

[3] Ling-Yan Li, Zhong-Shi He, Yong Yi(2004), "Poetry Stylistic Analysis Technique Based on Term Connections." IEEE, 0-7803-8403-2.

[4] Keh-Jiann Chen, Shu-Ling Huang, Yueh-Yin Shih, Yi-Jun Chen(2005), "Extended-HowNet: A Representational Framework for Concepts" , In Proceedings of IJCNLP-05 Workshop on Lexical Semantic, Jeju Island, South Korea, p.p 1-6.

[5] Yong Yi, Zhong-Shi He, Liang-Yan Li, Tian Yu, Elaine Yi (2005), "Advanced studies on traditional Chinese Poetry style identification." In Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, China, vol.5, pp. 2936- 2939.

[6] Zhong-Shi He, Wen-Ting Liang, Liang-Yan Li, Yu-Fang Tian(2007), "SVM-Based Classification Method For Poetry Style." IEEE, 1-4244-0973-X

[7] Rada Mihalcea(2007), "Using Wikipedia for Automatic Word Sense Disambiguation." In Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

[8] Likun Qiu, Weishi Zhang, Changjian Hu, Kai Zhao(2009),"A Self-Supervised Model for Sentiment Classification." Proceedings of *CIK'09*, November2-6, Hong Kong, China.

[9] Xiaojun Wan (2009), “Co-Training for Cross-Lingual Sentiment Classification.” In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Suntec, Singapore, pp. 235–243.

[10] Mitesh M. Khapra, Salil Joshi, Arindam Chatterjee, Pushpak Bhattacharyya(2011), “Together We Can: Bilingual Bootstrapping for WSD.” Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, page 561-569.

[11] 顏崑陽，喜怒哀樂，台北市，月房子，1994。

[12] 許清雲，部編大學用書-近體詩創作理論，臺北市：洪葉文化，1997。

[13] 俞士汶，胡俊峰，“唐宋詩之詞匯自動分析及應用”，北京大學，博士論文，2003 年 8 月。

[14] 朱我芯，「深秋猿鳥來心上，夜靜松杉到眼前」—華文詩歌情境再現，第五屆全球華文網路教育國際研討會，台北，2007 年 6 月。

[15] 孫瑛澤，陳建良，劉峻杰，劉昭麟，蘇豐文，“中文短句之情緒分類”，第 22 屆自然語言與語音處理研討會，南投，台灣，2010 年 9 月。

[16] 柯淑津，黃居仁，洪嘉馡，劉詩音，簡卉伶，蘇依莉，“中文詞義全文標記語料庫之設計與雛形製作”，第十九屆自然語言與語音處理研討會，2007 年 9 月，台灣大學，台灣。

[17] 羅鳳珠，“植基於中國詩詞語言特性所建構之語意概念分類體系研究”，第九屆海峽兩岸圖書資訊學學術研討會，武漢大學，2008 年 7 月 3-6 日。

[18] 易勇，何中市，李良炎，周建勇，瞿義玻，“基於遺傳演算法改進詩詞風格判斷的研究”，重慶大學，2005 年，計算機科學，32 捲，7 號。

[19] 劉博榮，“近體詩自動分類研究”，國立交通大學，碩士論文，2010 年 8 月。

[20] 陳紹宜，“建構一個中文對聯創作的知識評價架構”，國立交通大學，碩士論文，2010 年 6 月。

附錄

特徵選取結果-詠物述志

詩題詞彙	牡丹、鸚鵡、成長、隱居、進士、海棠、刺史、揚子、和尚、香山、題、花、丹、牡、仙、廟、寺、香、竹、避
詩題概念	Bh02A02_牡丹、Bh02A44_香菊片、Am01B04_僧徒、Bi11B38_鸚鵡、Hj01B06_隱居、Bh07A55_海棠、Af09B05_狀元、Ib01C02_成長、Hj01B04_寄居、Bn24A03_寺、Hg11A01_揮筆、Ac03A01_美女、Eb35A01_雄、Ec01A01_紅、Dh01A01_神、Bn24A03_寺、Bn23B01_祠堂、Ih02A01_改變、Df10D01_夢幻、Bp22A01_香
詩文詞彙	牡丹、瓊瑤、芍藥、高樹、片片、美人、山色、自知、何用、幾日、白、艷、毛、瘦、栽、窮、緣、神、本、點
詩文概念	Bh02A02_牡丹、Bm16B04_漢白玉、Bp13A05_琴、Eb30A04_鮮豔、Ba06A04_香料、Ed59B01_繼、Ed49A01_恰當、Bi12D01_鶴、Da25A08_何用、Ah08B01_妻、Dn08A35_枝、Eb30A04_鮮豔、Hj38B01_躲藏、Hi36A01_幫助、De01C01_道德、Ia10B01_蒸發、Bp22A01_香、Ea02A04_高大、Bg01A03_洪水、Ea04C01_狹窄

特徵選取結果-山水田園

詩題詞彙	秋日、閒居、潤州、賓客、柳州、刺史、一二、知己、秋雨、曲江、巳、深、秋、霽、井、舟、登、連、溪、趙
詩題概念	Bd02A05_秋日、Hf03B01_划船、Hj01B09_家居、Hf06A01_經過、Bh01A75_栗樹、Bn01B33_套房、Eb01B01_少、Aj05B01_客人、Bf01A12_冰雨、Ed32A01_親密、Ca01B06_巳、Eb22B01_遲、Bn15B01_井、Ia02A04_放晴、Jd06C01_包含、Fb01A16_登高、Kc06A01_連、Be05C01_溪澗、Bo22A01_船、Be06A01_湖泊
詩文詞彙	忘機、十里、漠漠、踏青、海鷗、身外、神女、晴來、煙水、霽景、蠶、田、桑、種、袁、穗、傲、禾、跳、荇
詩文概念	Ca29C01_夜間、Bh09A30_桑、Df12A05_詩興、Bi12E01_鷗、Ea04A01_廣闊、Bh13B01_種子、Bf02A04_微風、Be05B19_鴨綠江、Bp36B02_煙火、Ab01B01_女人、Bi19B01_蠶、Ka26A01_恰巧、Ee34D01_驕傲、Ee36A01_嚴格、Bh09A30_桑、Bn12A01_田地、Br12B01_茶、Hj08D01_興、Bb02A02_沙、Hd16A01_疏浚

特徵選取結果-情愛閨怨

詩題詞彙	七夕、有人、海棠、櫻桃、春雪、鄰家、無、夕、代、意、賦、溪、蓮、羽、棠、娘
詩題概念	Ca25A14_七夕、Aa06B01_有人、Bh07A55_海棠、Bh07A09_櫻桃、Bf01B03_初雪、Aj02C01_鄰居、Ab01B01_女人、Ca02B01_朝代、Dk06A01_詞、Df08C01_意圖、Be05C01_溪澗、Bh01A36_棠梨、Je08B01_約束、Bk11A06_羽毛、Bh02A04_荷花、Ef14A01_富裕
詩文詞彙	鴛鴦、含情、雙飛、不語、珠簾、獨眠、青樓、蓬山、雲髻、弱柳、斂、妒、嬌、繡、眉、妝、字、勻、臉、邑
詩文概念	Ah01A01_親戚、Fc10B01_叫、Bi11B34_鴛鴦、Ka23C01_獨自、Bp27A01_床、Gb10B01_恨、Eb01A03_如林、Bp01B01_鏡子、Ab01B03_少婦、Id14A01_移動、Bk02B01_臉、Gb13B01_妒忌、Je08B01_約束、Eb30A06_嬌、Hj41A04_刺繡、Dj04B02_紅利、Bk12A01_眉毛、Ed37D01_均勻、Bo03A12_鎖鏈、Hc03C03_調撥

特徵選取結果-贈別思友

詩題詞彙	員外、郎中、處士、送客、友人、樂天、江南、長安、相公、湖南、寄、贈、送、州、別、員、酬、友、歸、少
詩題概念	Af03C03_豪紳、Ae15A01_醫生、Hi06A01_送別、Af11B01_隱士、Aj01A01_朋友、Cb08A22_西楚、Ee08A01_樂觀、Di02B23_阿肯色州、Af09A01_宰相、Cb25A11_營口、Hi26A01_贈送、Je05B01_寄予、Dd15B20_邵、Di02B16_道、Hj26C01_區分、Aa01A01_人、Aj01A01_朋友、Hi27D01_歸還、Eb25B01_暫時、Dn04A03_二
詩文詞彙	白髮、故人、草色、江上、別後、洞庭、青雲、千山、少年、相逢、君、老、好、腰、病、寄、公、別、滴、少
詩文概念	Bo22A01_船、Ah08B06_故人、Bk11B10_白髮、Dk27A01_詩、Eb21A01_遠、Gc03D01_何須、Ea01A01_長、Bo07B01_蠶簇、Ec01B01_黃、Gc02D01_不可、Aa03A01_你、Ed59A01_寄、Ib10A01_生病、Ib21C01_凋謝、Hi27C01_借、Ed24A01_光榮、Id18C01_相遇、Ab01A03_先生、Id10E01_滴、Kc01A01_和

特徵選取結果-邊塞征戰

詩題詞彙	將軍、八月、河南、司徒、河北、司馬、秀才、將、軍、六、八、諸、獵、作、河、寧、門
詩題概念	Hh06C02_將、Ca21A09_八月、Di02B19_西藏、Di02B28_花縣、Kb05C01_以、Di11A01_軍隊、Dn04A07_六、Dn04A09_八、Hd26A01_打獵、Ed61D01_各個、Ja01B01_當做、Ka17C01_難道、Di05B01_家庭、Bp29D01_帳幕
詩文詞彙	塞鴻、將軍、烽火、老將、征戰、已過、劍戟、金甲、詩書、戰馬、弓、箭、危、劍、殊、葬、軍、角、兵、旗
詩文概念	Be05A01_海洋、Hh06C02_將、Bn01B02_住房、Bc02A01_邊、Dk19A10_詩書、Ha02A01_鬥爭、Hh04C02_吹奏、Bo27A01_武器、Ae10A04_將領、Bk18A01_血液、Bo29B01_箭、Bo29E01_劍、Fa11B02_埋葬、Ca04C01_關頭、Bo18B09_弓、Da14B03_勳勞、Bg03A13_烽火、Ee13B01_怯懦、Hb04B01_防守、Di11A01_軍隊

特徵選取結果-社會民生

詩題詞彙	公主、春日、初春、恩賜、有感、曲江、早春、揚州、夏日、溫泉、應、制、奉、幸、春、和、聖、宮、主、莊
詩題概念	Af06A06_公主、Ca19B01_春、Ca19B02_初春、Hi26A11_敬獻、Df01B05_觀感、Cb08A16_南區、Bn01B18_故居、Bf01A12_冰雨、Be04A16_嵩山、Fb01A16_登高、Hd04A01_製造、Je14A01_接受、Ka26B01_幸虧、Hi18B01_答、Ca18A01_年、Bn01B08_宮殿、Ak03C01_聖賢、Kc01A01_和、Di23B01_宴會、Cb25C02_村莊
詩文詞彙	南山、上苑、帝城、旌旗、此日、蓬萊、春色、樓臺、天子、昔年、帝、聖、宴、綵、繞、御、買、殿、迎、出
詩文概念	Dh03B01_鳳、Bg05A01_氣、Bi14A01_魚、Bd01B03_銀河、Kb07C01_趁、Dn10A15_絲、Bg07A39_說話聲、Dc02A02_春光、Da11A01_恩惠、Bn23A01_皇宮、Df07A01_欲望、Bq02A01_絲綢、Di23B01_宴會、Ed15A01_猛烈、Fb01A11_閒逛、Bo21A23_車駕、Df07A02_情欲、Bh09A43_蘆葦、Bn13A01_圈、Af05A01_皇帝

情感詞彙辨識結果

喜樂		怨怒		哀愁	
詞彙	權重	詞彙	權重	詞彙	權重
慶	0.944	夜長	0.941	淚	0.947
綵	0.938	家住	0.914	傷	0.9
上苑	0.93	胡天	0.914	趨	0.868
歡娛	0.93	長恨	0.914	相逢	0.868
韶	0.921	獨眠	0.914	風塵	0.843
奉	0.921	誰謂	0.914	渡	0.827
美	0.909	砧	0.914	首	0.827
歸來	0.893	免	0.914	忍	0.827
退	0.893	任他	0.914	不堪	0.827
乘	0.883	飄零	0.914	生涯	0.827
液	0.87	刑	0.914	滅	0.827
日暖	0.87	猶在	0.914	依依	0.807
精	0.87	無才	0.914	都	0.807
梅花	0.87	蓬山	0.914	茫茫	0.807
潭	0.87	穴	0.914	不勝	0.807
陪	0.87	入夢	0.889	甘	0.782
賀	0.87	符	0.889	白頭	0.782
不是	0.87	殊	0.889	往事	0.782
殷	0.87	弓	0.888	埋	0.782
光輝	0.87	鋒	0.864	嗟	0.782