

大雅語文智能

目錄

1. 計劃簡介

- 1.1 現況
- 1.2 計劃目標:釐定篇章的難度等級、主題和運用的語文知識
- 1.3 未來方向:字卡、默書、建立文庫、推薦文章、學生個人檔案

2. 預期功能

- 2.1 字:字數統計
- 2.2 字:按學習階段認字
- 2.3 詞:分詞
- 2.4 字和詞:統計字詞的出現頻率
- 2.5 字和詞:按字詞分析主題或文體
- 2.6 句:語文知識(句式、複句、標點符號和部份修辭)

3. 文本的分析方法

- 3.1 情況1:有特定文字和詞語作條件
- 3.2 情況2:以規則(Rules)來作條件
- 3.3 情況3:需要編輯者判斷
- 3.4 情況4:由用家(文章上載者)提供資料

4. 附件

- 附件1 字表
- 附件2 小學文章篇幅表
- 附件3 詞語表 (臺灣、香港)

1. 計劃簡介

1.1 現況

語文課本和補充練習五花八門，坊間對於來自不同出版社、不同地區、不同難度的課本通常只能靠主觀感覺或自身經驗來評價和分辨。市面一直缺少了一套能夠量度語文課本篇章程度的標準和系統。是次計劃正是希望能建立統一量度標準，用來比較和檢視不同出版社、不同地區、不同難度的語文課程的準則和指標。

1.2 目標

計劃目標為開發一個利用人工智能科技分析中文篇章的網上平台，透過評定篇章中的潛在語文知識(字／詞／句／段／篇)和主題，比對不同地區的語文課程來測量篇章程度。

1.3 發展方向

學校

教師可利用此平台分析語文課程中的篇章，以圖解儀表板的形式查看篇章中的語文知識、主題和難度等級，以不同地區的課程比對難度等級。根據資料分析，教師可檢視和修訂現有課程，分配適當的篇章、補充練習、題目和學習資源來自訂校本課程。

家長和學生

收集學生研習過的篇章後，平台可為學生建立個人學習檔案，評估學生在語文知識、主題和難度等級的學習成果，進而推薦適當的篇章作個人化學習。

出版、教育和研究

網上平台收集不同地區語文課程的篇章後，可建立語文學習的文庫。文庫的潛在價值在於，出版社能夠依據本計劃數據審視作品；教育者又可安排不同的教學策略；研究者亦可以利用文庫所提供的素材，研究不同年代、地區的中國語文科教材。

作者

作者可利用此平台評鑑作品，安排銷售策略，透過撰寫作品獲取利潤。

2. 預期功能

2.1 字:字數統計

大雅語文智能，為上載文章統計字數，然後根據附件2(下表)，以年級為文章分類，評定文章所屬的難度等級。

中文科閱讀理解文章字數一覽

年級	教育局 TSA	教科書			新領域
		啟思	現代	朗文	
一	/	107-117	38-81	39-67	≤150
二	/	166-197	130-195	133-181	151-250
三	450-550	313-402	281-318	207-298	251-400
四	/	302-494	318-340	395-510	401-550
五	/	620-689	470-557	491-540	551-650
六	700-1000	639-661	570-639	456-620	≤651

圖I:參考教科書後得出文章篇幅數字

2.2 字:按學習階段認字

按照附件1(〈小學中文科常用字表〉)，識別文章中每個文字，並按「學習年級」、「常用度」、「學習難度」統計文字分佈。最後按不同「學習年級」字詞分佈的百分率，評定文章所屬的難度等級。

編號	生字	學習年級	常用度	學習難度
2891	風	1	194	1
280	功	1	424	2
2412	要	1	21	3
1748	田	1	784	4
1723	現	1	68	5
1473	洋	1	731	6
2845	青	1	459	8
2470	誰	1	617	9
608	姐	1	945	10
44	五	1	154	11
1396	歌	1	475	12
986	才	1	199	15
1852	知	1	164	17
1265	朋	1	518	19
283	助	1	552	20

圖II:〈小學中文科常用字表〉(別稱〈三千字表〉)節錄

2.3 詞:分詞

本計劃在分析自然語言時，均需要先運用中文分詞，將中文字句預先分拆成語義獨立的組合後，才能深入分析字、詞、句等部分。處理中文自然語言時，可按照詞性或語法，將篇章中的詞語區分出來，成為文本掘挖的第一步。

2.4 字和詞:統計字詞的出現頻率

透過輸入文章，統計文章內所有字或詞語出現的頻率。記錄各種字詞的出現頻率，用來評估文章的特色、文體或文章主題。同時成為中文教科書的基本研究素材。

編號	字詞	頻率	頻率百分率
18	的	30	5.79%
12	問	18	3.47%
11	學	17	3.28%
63	為	10	1.93%
15	是	8	1.54%
48	他	8	1.54%
59	媽	8	1.54%
65	麼	8	1.54%
10	「	7	1.35%
13	」	7	1.35%
14	這	7	1.35%
67	從	7	1.35%
142	你	7	1.35%
5	有	6	1.16%
23	開	6	1.16%
64	甚	6	1.16%
105	個	6	1.16%

圖III:模擬詞頻統計

2.5 字和詞:按字詞分析主題或文體

承接2.4部份，利用詞頻統計，配合不同主題的字詞表，期望能成為了解文章文體的跳板。從某類型字詞出現頻率的百分比，可推測文章的主旨或主題，從而推斷文章的體裁。

2.6 句:語文知識(句式、複句、標點符號和部份修辭)

按照大雅圖畫中國語文科語文知識重點分類，利用關聯詞等，通過掃瞄和搜尋，記錄並統計文章中所出現的語文知識。再者，根據語文知識的分層，透過句語文知識的分佈和困難程度，來判斷文章的難度等級。

3. 文本的分析方法

3.1 情況1：有特定文字和詞語作條件

以「並列複句」為例，用以區分「並列複句」的字詞有如下：

區分「並列複句」的字詞(關聯詞)	
1	又
2	也
3	還
4	一會兒……一會兒
5	有時……有時
6	一方面……一方面
7	一方面……另一方面
8	一邊……一邊
9	既……又
10	既……也
11	又……又
12	是……不是
13	不是……而是……

以附件1文章《學與問》為例，以下標記的句子，因吻合上表第9項「既……又」，所以歸入「並列複句」。

學問學問，既要學又要問。學與問是相輔相成的，只有在學中間，在問中學，才能求得真知。我們從小養成了勤學好問的習慣，就好比插上了兩隻強健有力的翅

Microsoft Office User
句：複句：並列複句（9. 既……又）

〈學與問〉 —— 並列複句

人們常把有知識說成「有學問」，這是很冇道理的。知識是學來的，也是問來的。
「問」常常是打開知識殿堂的金鑰匙是通向成功之門的鋪路石。

波蘭偉大的天文學家哥白尼，小時候就非常喜歡問。他對世界充滿了好奇，經常纏着爸爸媽媽問這問那：太陽為甚麼總是從東方升起，從西邊落下？晴朗的夜空有那麼多星星，為甚麼到了白天卻無影無蹤了？小雞為甚麼要從雞蛋裏出來，而不從母雞的肚子裏出來？哥白尼對科學奧祕的不懈探求，正是從這些稀奇古怪的「為甚麼」開始的。

我們面對的是一個五彩繽紛的世界。這個世界日新月異，瞬息萬變。作為新一代的小學生，我們更應當像哥白尼那樣，遇事多問幾個「為甚麼」，學會從平常的事物中發現問題。有了問題，可隨時隨地請教別人。你可以請教父母和老師，也可以請教同學和朋友。只要他確實能給你啟發，給你幫助，不管他年長年幼，地位高低，都可以成為你的老師，都應該向他請教。古人說的「能者為師」就是這個道理。

在求知的過程中，我們還要善於把勤學好問和觀察思考結合起來。北宋有個大科學家，名叫沈括。他小時候讀白居易的詩《大林寺桃花》：「人間四月芳菲盡，山寺桃花始盛開。」他想：為甚麼同是桃花，開花的時間相差這麼遠呢？

他去問媽媽，媽媽說：「興許是花開花落，有早有遲吧！」媽媽的回答沒能解開沈括的疑團，他仍然把這個問題放在心上。有一次，他隨大人到深山的寺廟裏去，發現那裏的溫度要比山下低得多，才明白了其中的道理。

學問學問，既要學又要問。學與問是相輔相成的，只有在學中問，在問中學，才能求得真知。我們從小養成了勤學好問的習慣，就好比插上了兩隻強健有力的翅膀。到那時，知識的天空將任你翱翔，宇宙的奧秘將任你探求，你將真正成為學習的主人。

Commented [MOU1]: 句：複句：並列複句（9. 既……又）

3.2 情況2：以規則(Rules)來作條件

以「引號」為例，用以區分「引號」的條件(Rules)有如下：

區分「引號」的條件(Rules)		
<u>分為：</u>		
I. 特殊含義		
II. 引用		
III. 對話		
IV. [需要編輯者判斷]		
歸類為	條件	
1 I. 特殊含義	若……少於4個字元，且獨立見於字詞表 「……」	
2 II. 引用	若……多於4個字元，不獨立見於字詞表 「……」	
3 III. 對話	配合冒號出現 ：「……」	
4	配合冒號(:)和「說」出現 A) 說：「……」或 B) 「……」XXXX說，「」	
5	配合冒號(:)和「道」出現 A) 道：「……」或 B) 「……」XXXX道，「」	
6	配合冒號(:)和「曰」出現 A) 曰：「……」或 B) 「……」XXXX曰，「」	
7	配合冒號(:)和「講」出現 A) 講：「……」或 B) 「……」XXXX講，「」	
8	配合冒號(:)和「喊」出現 A) 喊：「……」或 B) 「……」XXXX喊，「」	
9	配合冒號(:)和「叫」出現 A) 叫：「……」或 B) 「……」XXXX叫，「」	
10 IV. [需要編輯者判斷]	若……少於4個字元，但不獨立見於字詞表 「……」	

〈學與問〉——引號

人們常把有知識說成「有學問」，這是很冇道理的。知識是學來的，也是問來的。

「問」常常是打開知識殿堂的金鑰匙是通向成功之門的鋪路石。

波蘭偉大的天文學家哥白尼，小時候就非常喜歡問。他對世界充滿了好奇，經常纏着爸爸媽媽問這問那：太陽為甚麼總是從東方升起，從西邊落下？晴朗的夜空有那麼多星星，為甚麼到了白天卻無影無蹤了？小雞為甚麼要從雞蛋裏出來，而不從母雞的肚子裏出來？哥白尼對科學奧祕的不懈探求，正是從這些稀奇古怪的「為甚麼」開始的。

我們面對的是一個五彩繽紛的世界。這個世界日新月異，瞬息萬變。作為新一代的小學生，我們更應當像哥白尼那樣，遇事多問幾個「為甚麼」，學會從平常的事物中發現問題。有了問題，可隨時隨地請教別人。你可以請教父母和老師，也可以請教同學和朋友。只要他確實能給你啟發，給你幫助，不管他年長年幼，地位高低，都可以成為你的老師，都應該向他請教。古人說的「能者為師」就是這個道理。

在求知的過程中，我們還要善於把勤學好問和觀察思考結合起來。北宋有個大科學家，名叫沈括。他小時候讀白居易的詩《大林寺桃花》：「人間四月芳菲盡，山寺桃花始盛開。」他想：為甚麼同是桃花，開花的時間相差這麼遠呢？

他去問媽媽，媽媽說：「興許是花開花落，有早有遲吧！」媽媽的回答沒能解開沈括的疑團，他仍然把這個問題放在心上。有一次，他隨大人到深山的寺廟裏去，發現那裏的溫度要比山下低得多，才明白了其中的道理。

學問學問，既要學又要問。學與問是相輔相成的，只有在學中間，在問中學，才能求得真知。我們從小養成了勤學好問的習慣，就好比插上了兩隻強健有力的翅膀。到那時，知識的天空將任你翱翔，宇宙的奧秘將任你探求，你將真正成為學習的主人。

Commented [MOU2]: 句：標點符號：引號：[人工判斷]

(1. 若……少於4個字元，但不獨立見於字詞表「……」)

Commented [MOU3]: 句：標點符號：引號：特殊含義

(1. 若……少於4個字元，且獨立見於字詞表「……」)

Commented [MOU4]: 句：標點符號：引號：特殊含義

(1. 若……少於4個字元，且獨立見於字詞表「……」)

Commented [MOU5]: 句：標點符號：引號：特殊含義

(1. 若……少於4個字元，且獨立見於字詞表「……」)

Commented [MOU6]: 句：標點符號：引號：引用

(2. 若……不獨立見於字詞表「……」)

Commented [MOU7]: 句：標點符號：引號：引用

(3. 配合冒號出現)

Commented [MOU8]: 句：標點符號：引號：對話

(4A. 說：「……」或)

3.3 情況3：需要編輯者判斷

以上述例子為例：

[人們常把有知識說成「有學問」]一句中，
[有學問] 三字是由 [有] + [學問] 組成的詞組，整個詞組並不會獨立出現在字詞表內。

歸入 [需要編輯者判斷] 的詞條，代表需要編輯者額外判斷。所以，需要額外標記，提示編輯者。日後，當編輯者依靠人手歸類後，可考慮成為歸入「例外清單」或「採納清單」之中。

3.4 情況4：由用家(文章上載者)提供資料

類似於1.3，同樣是在分析資料之外，由用家輸入相關資料。在平台上提供標籤 (Tagging) 功能，讓用家在上載文章後，能夠以選擇題或直接輸入兩種方式，提供關於文章的資料，例如文章題目、作者、主題、體裁或出處等。

附件1〈小學中文科常用字表〉

<https://docs.google.com/spreadsheets/d/1GZ6WRFGM87qh7ltt5QosJeXOVBEIwGp-/edit?usp=sharing&ouid=112300334093909079389&rtpof=true&sd=true>

附件2〈小學文章篇幅表〉

中文科閱讀理解文章字數一覽

年級	教育局 TSA	教科書			
		啟思	現代	朗文	新領域
一	/	107-117	38-81	39-67	≤150
二	/	166-197	130-195	133-181	151-250
三	450-550	313-402	281-318	207-298	251-400
四	/	302-494	318-340	395-510	401-550
五	/	620-689	470-557	491-540	551-650
六	700-1000	639-661	570-639	456-620	≤651

附件3〈國教院三等七級詞表〉

https://drive.google.com/file/d/1Dfa34ydd_8sPzhwM0E9K3PqZYpDIRW-/view?usp=sharing