中華民國第 56 屆中小學科學展覽會

作品說明書

高級中等學校組　電腦與資訊學科

第二名

052502

英文篇章難易度自動分級之研究

學校名稱：臺北市立第一女子高級中學

作者：	指導老師：
高二 任恬儀 高二 許湘鈴	黃芳蘭

關鍵詞：語法分析、程度分級、語言學習

摘要

以製作適合高中生的英文篇章難易度自動分級為初衷，本研究採高中英文課文為語料，針對「如何分級」，意即從文章萃取哪些特徵、利用何工具或語料協助萃取特徵、以何工具分級等因素，進行研究與實驗，並建立一套新方法。首先進行前處理，再嘗試以單字、句型的數量或比例、句長、音節長、整合以上分析等各式特徵，支持向量機(Support Vector Machines)、隨機森林分類器(Random Forest Classifier)、決策樹分類器(Decision Tree Classifier)、卷積神經網路句分類器(Convolutional Neural Networks for Sentence Classification)等工具，進行將篇章分為高中一、二、三年級等三個難易度等級的測試，建立自動分級模型。最後製作成可供大眾使用的自動分級網頁。各項測試之中，最佳分類效能為整合各項特徵時得到的分類正確率 65.04%，經模擬得知，此效能確實優於過去研究。

壹、 研究動機

處在國際交流頻繁的現今，英文能力非常關鍵。於是，我們時常找尋各類英文文章來閱讀。然而，對於像我們這樣的高中生來說，找到適合自己程度的英文文章卻是一大問題——市面上雖有難易度分級的雜誌和參考書，但文章有限，還得付費購買；網路上的英文文章雖然比比皆是，其難易度卻十分駁雜。因此，我們想設計一套能夠「輔助學習者篩選出難易度合適的文章來閱讀」的判別方式。

貳、 研究目的

製作判定高中英文文章難易度的的機器學習模型：

一、由單字判定文章難易度

二、由句型判定文章難易度

三、由表面特徵判定文章難易度

四、整合單字、句型、表面特徵，判定文章難易度

並撰寫可供大眾利用的英文文章難易度自動分級網頁。

參、　研究設備及器材

一、 硬體

（一） 筆記型電腦（寫程式。CPU: Intel Core i5-4210U with Turbo Boost up to 2.7 GHz）

（二） 工作站電腦（執行程式。CPU: Intel(R) Core(TM) i7-2600 CPU @3.40GHz 四核心）

二、 軟體及工具

（一） Python（程式語言）

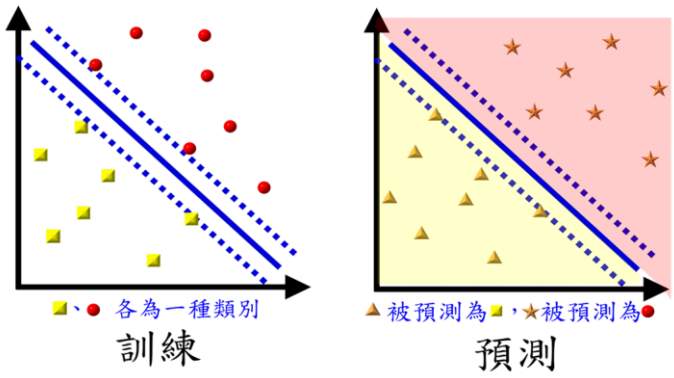
（二） Django（Web 應用框架）

（三） Stanford Parser（句型剖析工具）

（四） Natural Language Toolkit （NLTK，自然語言工具）

（五） Scikit-learn（**監督式學習**領域工具）

1．**支持向量機**（Support Vector Classification, **SVM**）[2]（主要以此實驗）



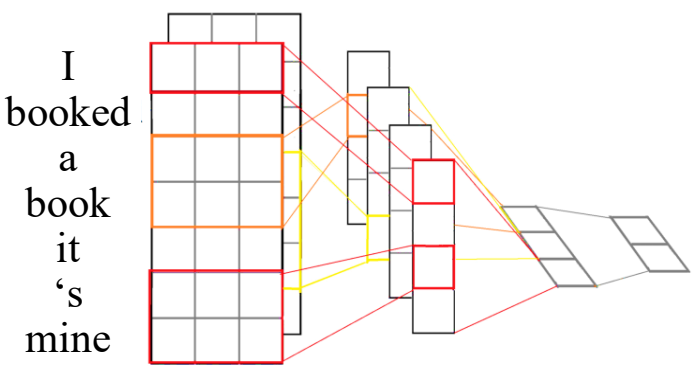
圖一：SVM 原理示意圖

建構高維「超平面」分類資料點。如圖，以二維示意 SVM：得到特徵，以機器學習在類別間劃界線（圖中藍實線）、建立模型，用以預測類別。SVM 有 C、γ 兩影響分類效果之參數，須逐個嘗試 2 的次方以求最佳參數組合。

2．隨機森林分類器（Random Forest Classifier, RFC）

3．決策樹分類器（Decision Tree Classifier, DTC）

（六） **卷積神經網路**句分類器（Convolutional Neural Networks for Sentence Classification, **CNN** sentence）[1]（**深度學習**領域工具）



圖二：CNN 原理示意圖

卷積神經網路(CNN)基於試圖使用多個處理層，對數據進行高維抽象。本圖以 I booked a book. It’s mine.為例。CNN sentence 利用此概念，使用 word2vec 的向量表，以 300 個維度表示每個向量，判斷句意。技術較新穎。

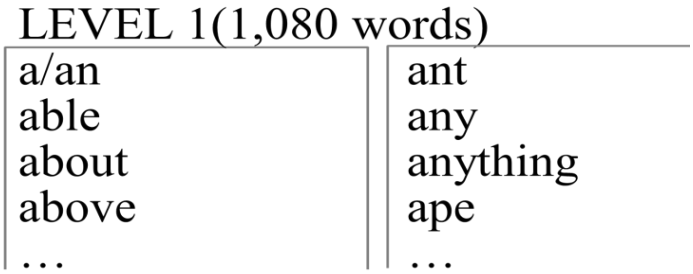
三、 語料：

（一） 美國國家語料庫（American National Corpus, ANC）

一個美式英語的大型電子彙集庫，對於任何用途完全公開而無限制。本實驗使用書面（Written）部分的頻率資料（Frequency Data），以在 ANC 語料庫中出現過的次數（Count，以下簡稱計數）排序前 24,521 行的資料。

圖三：ANC 節錄
如圖，每行有四項資料（綠字），包括詞和其原形（如橘框），詞性的代號，以及計數。採依計數排序（籃框）的版本。

（二） 高中英文詞彙參考表（The English Reference Word List for Senior High School, ERWL）



圖四：ERWL 節錄
將高中英文詞彙分為六個等級，每個等級分別包含 1080 個詞彙，並依字典序排序。

（三） 各版本 103 學年度高中英文課本電子檔

	三民	龍騰	南一	遠東施	遠東陳	總和
高一	24	24	24	24	23	119
高二	24	24	24	30	24	126
高三	22	20	23	26	20	111
總和	70	68	71	80	67	356

表一：課文來源分布（單位：篇）
採三民乙版，每冊比甲版多兩課，其他課文內容無異；遠東施版係由施玉惠、林茂松、黃崇術、Sarah Brooks 編著；遠東陳版則由陳純音主編。

肆、　研究過程與方法

本研究首先探討相關研究，接著進行前處理、特徵萃取、機器學習、網頁撰寫等。其中，前處理為其他部分之基礎；特徵萃取部分，我們發想、嘗試並探討以各式特徵以進行自動分級；機器學習部分，我們嘗試了包括監督式學習和深度學習兩個領域的幾種方法，監督式學習領域測試 SVM、RFC、DTC 等分類工具，深度學習則測試 CNN sentence，而由於 CNN sentence 性質特殊，和前述實驗分開進行；最後完成整合模型，並撰寫成方便大眾使用的網頁。

一、 相關研究探討

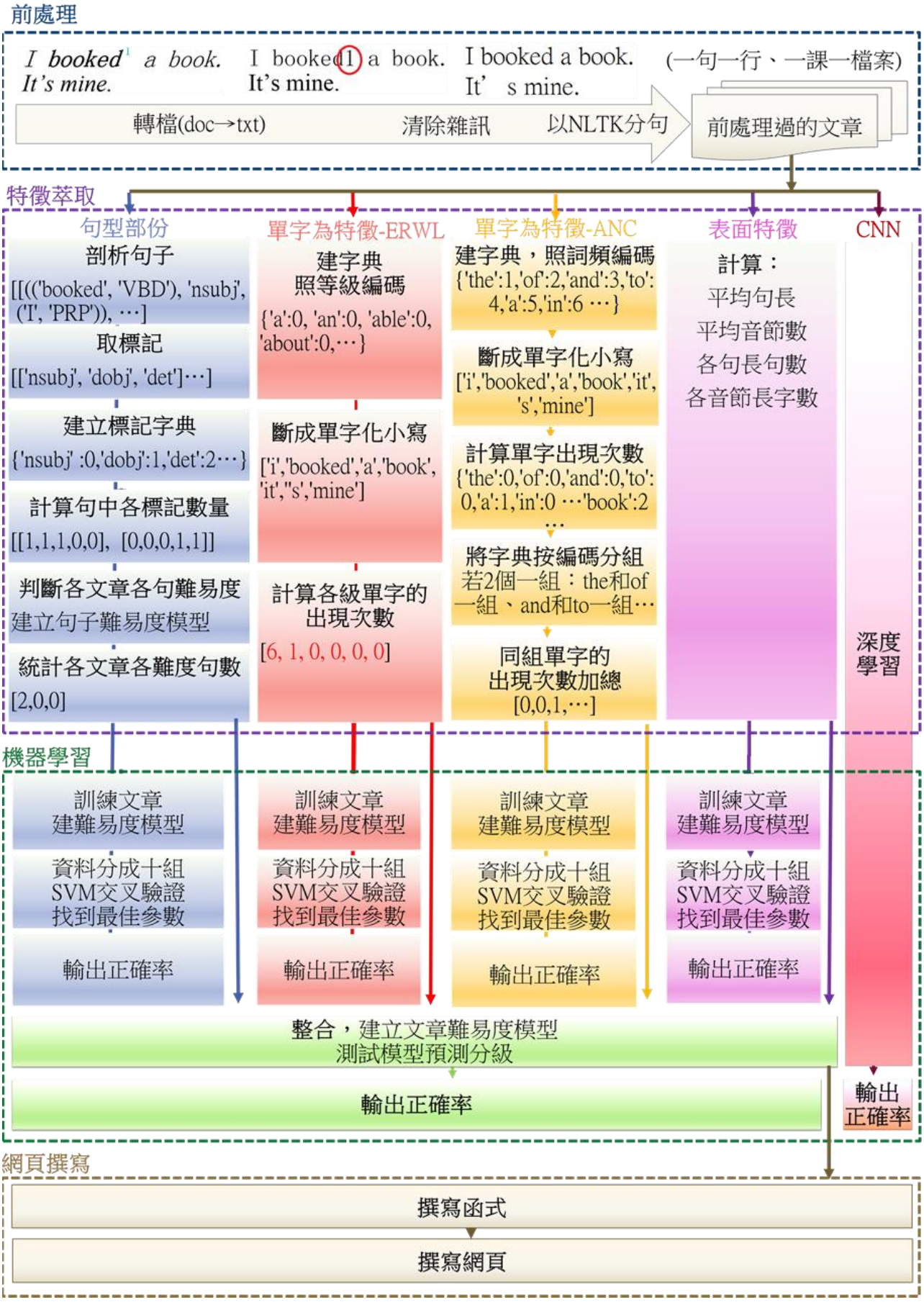
表二：相關研究之探討與整理

我們整理了過去試圖以類似難易度的標準分類文章的幾個研究。由於研究多是基於過往之成果，故在此以「新特徵」表示該研究新使用的特徵。

研究	語料／新特徵／方法	探討
適讀性公式[早期國外研究、5]	語料：英文讀本等	1.學理上有可議之處[4] 2.未考量是否以英文為母語
	新特徵：文長、句數等	
	方法：適讀性公式、Microsoft Word[5]	
全民英檢(GEPT)[3,4,6]	語料：全民英檢(GEPT)	僅概分為初、中、中高級，無法對應國內學生實況。
	新特徵：子句結構等	
	方法：C5.0 決策樹、KNN 算法	
我們的構想	**語料：高中英文課文** **新特徵：單字向量、句型結構、單字頻率等** **方法：支持向量機(SVM)、卷積神經網路(CNN)**	

二、流程圖

圖五：實驗整體流程圖 以”I booked a book. It’s mine.”為例。機器學習部分以 SVM 示意。



三、前處理

(一) 撰寫程式，處理各版本高中英文課本電子檔(doc)

1．將**課文**部份取出，儲存為純文字文件(txt)

2．去除上、下標（部分課文中有標註生字的下標編號，轉成 txt 時會變成普通數字，影響判讀）、亂碼（doc 排版符號等）等雜訊

3．使用 NLTK 中的 Punkt Sentence Tokenizer 將課文分句，儲存（一個純文字文件放一課、一行放一句）

(二) 撰寫程式，處理欲使用之**字彙庫**

1．整理字彙內容（如：將”a (n)”分成”a”和”an”兩字）

2．去除雜訊（避免編碼問題，去除含非英文語言字母之單字）

3．改寫成程式容易讀取之格式（如：加入跳格、換行符號）

四、特徵萃取

相較於過去研究使用過的各式特徵，我們企圖**萃取出更簡單卻更有效的特徵**，因此，我們試圖從生活經驗發想，並使用更具效益的語料、資料、工具等實作，得到，如單字、句型之相關特徵；此外，我們亦**基於舊有想法，發想出新方法**，如表面特徵的部分。

中學階段課文之學習重點在於單字、句型，應是辨別難易度的重要特徵，因此我們構思了針對單字、句型，及較淺顯的各式表面特徵，進行量化的一系列方式，以萃取特徵、進行機器學習。

為了應用於處理長文（如：長篇小說等），避免文長影響模型判別其難易度，**我們對於以下「計算出現次數」的實驗，進行了「計算出現頻率」的版本**，意即每個數據都除以該數據總和。**這也是我們首創的想法之一**。

以下為了方便舉例，我們對所有的例子做一個簡單的假設：

假設”My dog also likes eating sausage.”是一篇高一課文。

(一) 以單字為特徵

1. 字彙集來源

若欲以各個單字出現過的次數為特徵，可先取得一個字彙集，以字彙集建立字典（dic）。我們測試了兩個字彙集。較容易取得、屬於完全公開資源的字彙集是**美國國家語料庫（ANC）**。ANC 有以詞頻排序過的資料，並且包含單字變化形-原形-詞性-詞頻的對應關係，但以「英語為母語者產出之書面資料」的詞頻排序，和學生最熟悉的英文課文略有出入（如表三）。我國大學入學考試中心公布之**《高中英文詞彙參考表》（ERWL）**的分級可能相對貼近我國學生，但 ERWL 僅分將單字分為六級。

表三：ANC 與課文之詞頻序比較節錄

我們計算英文課文的詞頻序高低，和 ANC 略有出入。

單字	ANC	課文
of	2	3
and	3	4
to	4	2
is	7	8
for	8	10

2. **轉換成原形**

許多單字是以複數形、過去式等形態出現在文章中，可能造成影響。因此，我們嘗試兩種方式，一是直接以單字的各式形態為索引，利用 ANC 建立字典，計算在課文中出現的次數；二是利用 ANC 中的單字－原形詞對應關係，將文章中出現的詞全部轉換成原形詞，再計算在課文中出現的次數。

3. 依詞頻排序**分組、加總**出現次數

一篇課文僅數百字，但建立能夠容下所有課文中之多數單字的字典，其單字數量達上萬，而將數百字的課文，分散進數萬個元素的特徵矩陣中，數據相當稀疏。若詞頻高低和該詞的難易度具有一定程度的關聯性，應可將單字以詞頻排序過後，每一定數量的單字的出現次數加總，放進一個新的矩陣，再進行機器學習。至於究竟是否分組、加總，或每多少個字詞加總一次，即探討之處所在。

4. **頻率版本**

我們也試了改計詞頻的版本，將所有「字的出現次數」數據除以總字數。

(二) 以單字為特徵、ANC 為語料之步驟

1. 從課文讀出單字。

斷字處理，得['my', 'dog', 'also', **'likes'**, 'eating', 'sausage']

2. （**轉換成原形**才須進行）用 ANC 的「變化形-原形」對應關係，建「變化形-原形」之對應字典，將課文中讀出的單字都轉成原形。

{'the': 'the', 'of': 'of', 'a': 'a',…, **'likes': 'like'**,…}（表示 likes 的原形是 like）
轉換得['my', 'dog', 'also', **'like'**, 'eat', 'sausage']

3. 建立單字和 ANC 詞頻序對應的字典，以及和其字典等長之零矩陣。計算**單字在課文中出現的次數（即「計數」）**於詞頻序對應之矩陣位置，完成其新計數矩陣。

{'the':0, 'of':1, 'a':2,…, **'like':**,…}（表示 likes 的詞頻序 62）
[0, 0, 0, …, **1,**…] （表示詞頻序 62 的單字 **like 總共出現 1 次**）

4. （**分組、加總**才須進行）將單字按照 ANC 計數排序後分組，再利用前一步完成之新計數矩陣，加總各組的新計數，成為一個新的、濃縮過的矩陣。

若 500 個單字一組。第 0~499 個單字的**出現次數總和為 3**、第 500~999 個為 1、第 1000~1499 個為 0…，則得特徵矩陣[3, 1, 0, …]

5. （**頻率版本**才須進行）將前一步驟所得矩陣除以總字數。

原矩陣[3, 1, 0, …]除以總字數 6 字，得新矩陣[**0.5**, 0.1667, 0, …]

6. 得 x、y 矩陣，放進分類器訓練、測試。

特徵矩陣[**0.5**, 0.1667, 0, …]屬高一難易度，記為[1]，則：
輸入資料：x 矩陣（特徵）為 [0.5, 0.1667, 0, …]，y 矩陣（難易度）為 [1]

(三) 以單字為特徵、ERWL 為語料之步驟

1. 以 ERWL 製作「單字-難度等級」對應字典。

{'a':1, 'an':1, 'able':1, 'about':1, 'above':1,…, 'like':1,…}（這些字皆屬**難度 1**）

2. 從課文讀出單字、建立字典並計算數量。和 ANC 為語料之步驟類似，但改計「各級單字出現的次數」。詳見以單字為特徵、ANC 為語料。

[6,0,0,0,0,0]（**難度 1** 的字記在第 1 格，共出現 6 次）

3. （**頻率版本**才須進行）將前一步驟所得矩陣除以總字數。

原矩陣[6,0,0,0,0,0]除以總字數 6 字，得新矩陣[1,0,0,0,0,0]

4. 得 x、y 矩陣，放進分類器訓練、測試。

特徵矩陣[1,0,0,0,0,0]屬高一難易度，記為[1]，則： **輸入資料**：x 矩陣（特徵）為 [1,0,0,0,0,0]，y 矩陣（難易度）為 [1]

(四) 以句型為特徵

句型特徵我們採用詞之間的**關係（Dependency）資料**，也就是各詞之間的關係。亦是一**前所未見之首創**作法。

1. 以句為單位

以句型為特徵時，必須將資料斷句，以句子為單位剖析。對每篇課文，我們對當中的各個句子剖析，得到第一層特徵——關係資料。各個句子依其「各式關係出現過的次數」為特徵，以分類器進行第一層分類，分成三個難易度。此次分類，得到以句子為單位的分級結果。

2. 以篇為單位

再以各篇課文「第一層分類時，被分成各個難易度的句子分別有幾句」為特徵，以分類器進行第二層分類，得到以篇章為單位的分級結果。

(五) 以句型為特徵——以句為單位之步驟

1. 將標籤編碼，製成字典。

> {"advmod": 0, "advcl": 1, "mark": 2, "nsubj": 3, …}（表示 adbmod 編碼 0）

2. 使用 Stanford Parser 剖析每一個句子，得到其中的關係資料。

> (('likes', 'VBZ'), 'nsubj', ('dog', 'NN'))
>
> (('dog', 'NN'), 'nmod:poss', ('My', 'PRP$'))
>
> (('likes', 'VBZ'), 'advmod', ('also', 'RB')) ……

3. 取這之中表示兩者關係的**關係標記**，組成句型矩陣，得到**句子架構**。

> ['nsubj', 'nmod:poss', 'advmod',……]

4. 統計每個句子**各標記出現的次數**，儲存於矩陣中該標籤的字典編碼位置。

> [1,0,0,1,1,0,….]（表示編碼 0 的 adbmod 共出現 1 次）

5. 得 x、y 矩陣，分別在 SVM、RFC 、DTC 訓練、測試。

> 特徵矩陣[1,0,0,1,1,0,….]屬高一難易度，記為[1]，則：
>
> **輸入資料**：x 矩陣（特徵）為 [1,0,0,1,1,0,….]，y 矩陣（難易度）為 [1]

(六) 以句型為特徵——以篇為單位之步驟

1. 分別讀出每篇文章、計算文章中分屬三個年級難易度的句子各有幾句，作為特徵。

> [1,0,0]（表示有 1 個句子被判成了**記在第 1 格的高一難度**）

2. （**頻率版本**才須進行）將前一步驟所得矩陣除以總句數。

> 原矩陣[1,0,0]除以總句數 1 句，得新矩陣[1,0,0]

3. 各文章輪流當作測試資料，其他文章當作訓練資料，計算正確率。

> 特徵矩陣[1,0,0]屬高一難易度，記為[1]，則：
>
> **輸入資料**：x 矩陣（特徵）為 [1,0,0]，y 矩陣（難易度）為 [1]

(七) 表面特徵

表面特徵當中，為了實用考量，我們依舊秉持著「**避免受文長影響**」的原則。因此單純的總字數、總句數等會被文長影響的特徵，我們皆不採用。不受影響的特徵如平均句長、平均音節數被過去諸多研究不斷使用，故我們將不多加探討。而我們**首創**的**各句長句數、各音節長字數**兩特徵雖不複雜，過去研究卻未見到。

1. 平均句長

$$\text{平均句長} = \frac{\text{文章總字數}}{\text{文章總句數}}$$

2. 平均音節數

$$\text{字平均音節數} = \frac{\text{文章總音節數}}{\text{文章總字數}}$$

$$\text{句平均音節數} = \frac{\text{文章總音節數}}{\text{文章總句數}}$$

3. 各句長句數

$$\text{i 句長句數} = \text{文中句長 i 的句子數} \quad \forall 1 \le \text{i} \le 40, \text{i} \in N$$

但考量短文句數不多，資料較分散，將 n 個句長的句子數加總，達到類似平滑的效果，形成新的特徵矩陣。

$$\text{i 句長句數特徵} = \sum_{k=i}^{i+n} \text{文中句長 k 的句子數} \quad \forall 1 \le \text{i} \le 40 - \text{n}, \text{i} \in N$$

4. 各音節長字數

$$\text{i 音節長字數} = \text{文中音節長 i 的字數} \quad \forall 1 \le \text{i} \le 10, \text{i} \in N$$

五、 機器學習

本實驗以 **10 次交叉驗證**（10-fold cross validation）方式，將前特徵萃取得到之 x 矩陣（特徵）及 y 矩陣（難易度）輸入分類器訓練和測試。主要以 SVM 分級，其有 C、γ 兩參數，須從 $2^{-5}, 2^{-4}, 2^{-3}, \cdots, 2^{5}$ 開始，逐步縮小範圍暴力嘗試，尋找能夠得到最佳效果的參數組合。以 SVM 分析各項特徵之效能，並測試 RFC 和 DTC 等有別於 SVM 之分類工具。

六、 CNN sentence

由於該工具使用的資料為句子，所以和其他實驗分開介紹。

(一) 測資

1. 工具內部會有自動分成訓練與測試資料，因此只將資料分成一、二、三年級。

2. 後因為考量來自同一篇文章的句子可能會有相近的文意，故將各句子依文章分類。

(二) 程式流程

1. process_data.py

(1) 功能：以 Google Word2Vec 及測試資料的每一個句子，建立名為 mr.p 的 pickle 檔。

(2) 修改部份：設定路徑將原本的「反向句子設定為編號 0，正向句子設定為編號 1」，改成「一、二、三年級分別設定為 0、1、2」。原為第幾份資料的 0~9 隨機數字，改至每篇文章隨機一次，確保文章不會同時出現在訓練資料和測試資料中。

2. Conv_sentence.py

(1) 功能：讀取 mr.p 的資料，及模式的引數，整理出形式，及依 Word2Vec 整理出多維向量，引用 Theano 模組，進行深度學習。

(2) 修改部分：max_len 為一個句子的最大長度，init 為分類的數量，由正面負面兩項，改為一、二、三年級三項。

七、**整合**

整合前，我們將程式**重新撰寫**，力求**模型未來的維護與擴增更加容易**。如：將前處理程式碼嵌進主程式、所有特徵萃取的程式重寫成函式等等。模型流程簡述如下：

（一） 輸入分句、經過前處理的文章。

（二） 分析**單字特徵**：ANC 各個單字、ERWL 各級單字出現次數。

（三） 分析**句型特徵**：被判成一、二、三年級難度的句子個數。

（四） 分析**表面特徵**。

（五） 將分析得到之特徵矩陣相接，以先前訓練出之 SVM 模型預測分級並輸出。

八、網頁撰寫

為了應用於長文的判別，我們亦建置了不會被文長影響的**頻率版本之整合模型**，並以 Django **撰寫網頁**，**直接應用**此模型。

伍、 研究結果

一、 效能評估方式

下表為一般判斷時的情況。其中，「人工判斷」相當於我們現有的「課文實際上是高一、二、三」，而「系統判斷」則相當於「分類器預測為高一、二、三」。

若以「從高中課文中判斷高一課文」為例，「相關」相當於「被認為是高一課文」，「不相關」相當於「被認為不是高一課文」。因此有了判斷「正確」、判斷「錯誤」，以及「正例」、「負例」之分。

表四：判斷情形

		系統判斷	
		相關	不相關
人工判斷	相關	正確正例（true positive, TP）	錯誤負例（false negative, FN）
	不相關	錯誤正例（false positive, FP）	正確負例（true negative, TN）

（一） 正確率（Precision）

$$\text{正確率(Precision)} = \frac{\text{正確正例(TP)}}{\text{正確正例(TP)} + \text{錯誤正例(FP)}}$$

（二） 召回率（Recall）

$$\text{召回率(Recall)} = \frac{\text{正確正例(TP)}}{\text{正確正例(TP)} + \text{錯誤負例(FN)}}$$

（三） F1 度量（F1-measure）

$$\text{F1 度量} = \frac{2 * \text{正確率} * \text{召回率}}{\text{正確率} + \text{召回率}}$$

二、 以單字為特徵

（一） 以 ANC 為字表

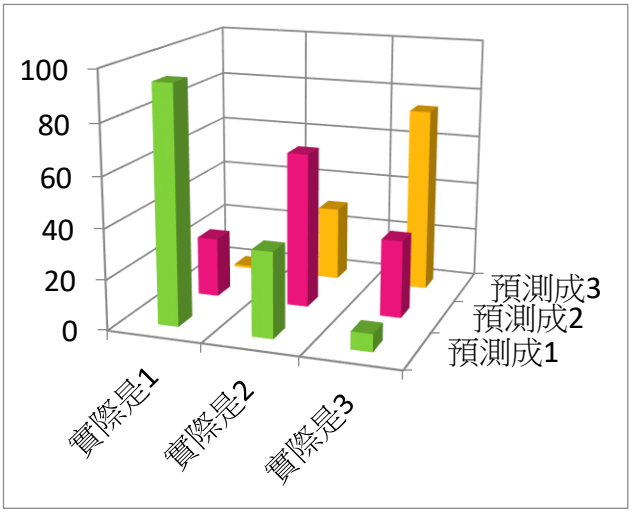
1． 轉原形正確率（表五）

	轉換原形	不轉換原形
排序分組	64.30%	63.25%
不排序分組	63.80%	64.42%

機器學習工具：SVM＼特徵：不轉換成原形、不排序分組

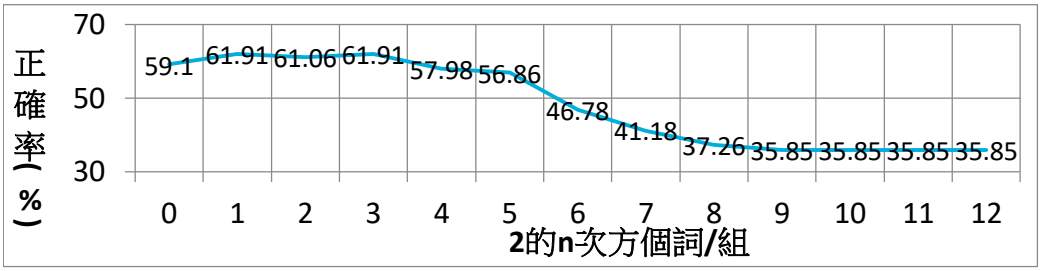
（表六，單位：篇）F-measure =0.644158。

預測 / 實際	1	2	3	正確率(%)
1	**94**	24	1	78.33
2	34	**62**	30	48.44
3	7	31	**74**	64.35
正確率(%)	69.12	52.1	68.52	**64.43**



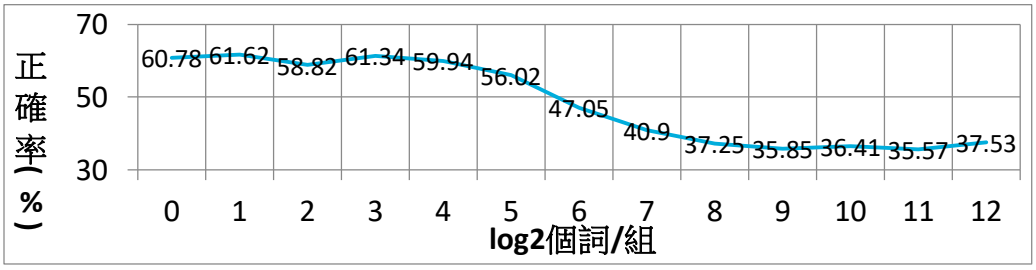
2． 分組

(1) 不轉原形，嘗試 2 的各個指數，以將不同數量的詞劃為一組加總。不調整參數，得到正確率如下圖。2^0~2^5 個詞為一組加總，正確率在 60%上下波動。進一步測試 2^0~2^5 中的每個整數測試，不調整參數。最高正確率為每 5 個一組時的 62.75%。調整參數後，不轉換成原形、排序後每 5 個詞為一組，若取 C=2^0，γ =2^{-11}，可得到 63.31%的正確率。（圖六 ）



(2) 轉原形，2^0~2^4 個詞一組時，正確率在 60%上下波動。進一步以 2^0 到 2^4 的每個整數逐一測試，不調整參數，每 7 個詞為一組時，正確率可達 63.03%。

i. 調整參數後，轉換成原形、不排序分組，若取 C=4.18，γ =$2^{-14.6}$，可以得到最高正確率 64.15%。(圖七)



3． 其他工具的嘗試（表七，正確率）

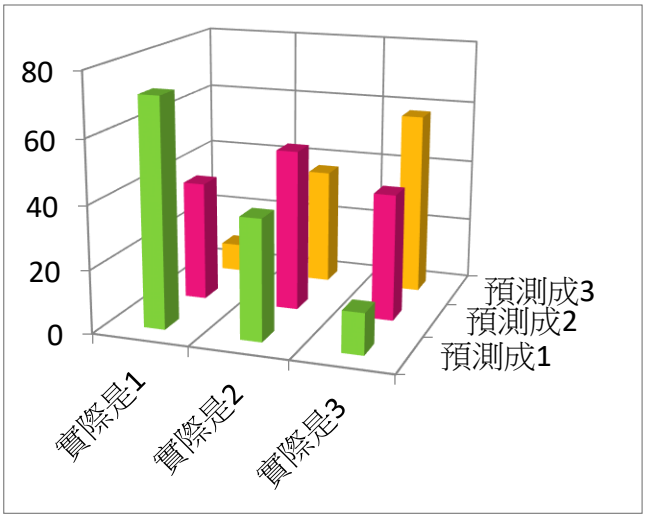
RFC	DTC
46.80%	51.37%

（二） 以大考中心(ERWL)為字表（表八，正確率）

計次數	計頻率
35.67%	50.84%

(表九，單位：篇)機器學習工具：SVM\ 特徵：大考中心單字頻率。F-measure=0.511520。

預測 / 實際	1	2	3	正確率(%)
1	**72**	38	9	60.50
2	38	**51**	37	40.47
3	13	40	**58**	52.25
正確率(%)	58.54	39.53	55.77	50.84



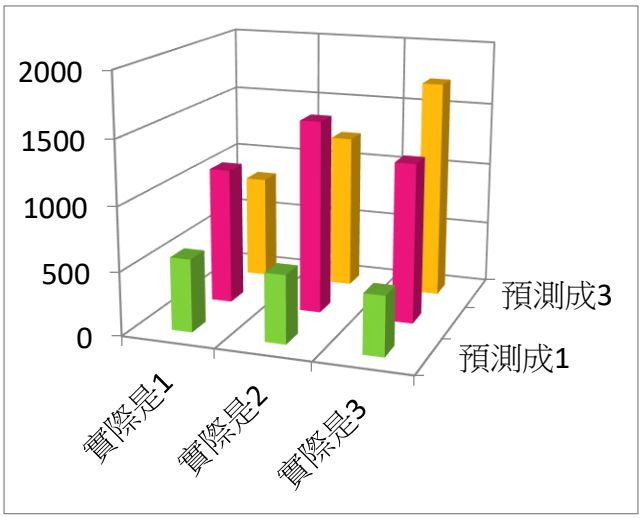
三、 以句型為特徵

(一) 以句為單位（表十，正確率）

使用工具	SVM	RFC	DTC
F-measure	0.395043	0.378382	0.35829

(表十一，單位：篇)機器學習工具：SVM\ 特徵：各句型標記數量 F-measure=0.395043。

預測 / 實際	1	2	3	正確率(%)
1	**573**	1078	828	23.11
2	536	**1518**	1231	46.21
3	466	1253	**1722**	50.04
正確率(%)	36.38	39.44	45.54	**41.42**



(二) 以篇為單位：

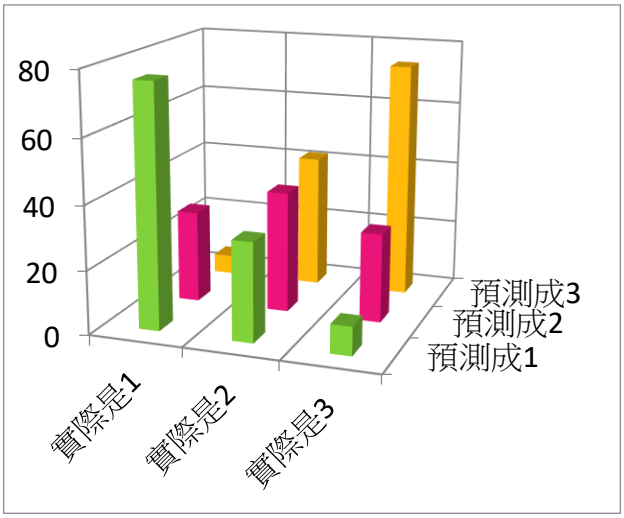
由於在句子的部份 SVM 的效果較佳，所以此部份的句子判斷皆使用 SVM 工具。

（表十二，正確率）

使用工具 / 特徵	SVM	RFC	DTC
各難易度句子數量	55.62%	53.67%	50.64%
各難易度句子比例	48.42%	41.32%	37.85%

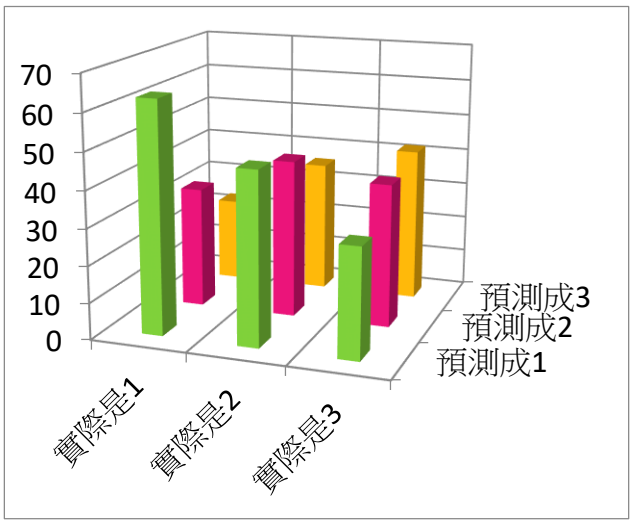
(表十三，單位：篇)機器學習工具：SVM\ 特徵：各難易度句子數量 F-measure=0.556167。

實際 \ 預測	1	2	3	正確率 (%)
1	**89**	26	4	74.79
2	44	**47**	35	37.30
3	9	39	**63**	56.76
正確率 (%)	62.68	41.96	61.76	55.90



(表十四，單位：篇)工具：RFC / 特徵：各難易度句子比例 F-measure=0.413186。

實際 \ 預測	1	2	3	正確率 (%)
1	**63**	33	23	52.94
2	47	**43**	36	34.13
3	30	39	**42**	37.84
正確率 (%)	45.00	37.39	41.58	41.57



四、 表面特徵

（表十五，正確率）

	表面特徵			
	平均句長	平均音節數	各句長句數	各音節長字數
計數量	44.38%	42.69%	56.74%	36.23%
計頻率			49.71%	37.64%

五、 CNN sentence

由 CNN sentence 測試結果：內部資料的測試（訓練資料中含有測試資料）的正確率是 98.98%，外部資料的測試（訓練資料中不含測試資料）正確率則是 38.25%。

六、 各項研究比較

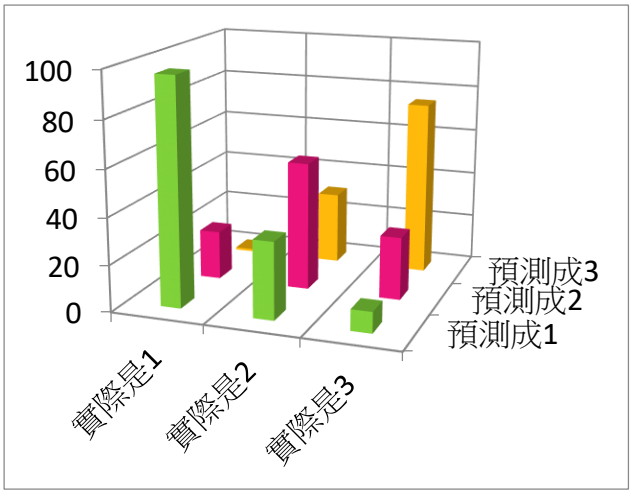
（表十六，正確率）			計數量	計頻率
句型為特徵	以句為單位		41.42%	
	以篇為單位		56.46%	48.42%
單字為特徵	ERWL		35.67%	50.84%
	ANC	分組*	34.83%	40.45%
		不分組	64.43%	52.25%
淺顯特徵	平均句長		44.38%	
	平均音節數		42.69%	
	各句長之句數		56.74%	49.71%
	各音節數之字數		36.23%	37.64%
CNN			38.25%	

七、 整合

本區結果皆為調整參數之後。

工具：SVM / 特徵：整合(**表十七，單位：篇**) F-measure=0.645169。

預測 / 實際	1	2	3	正確率 (%)
1	**97**	21	1	81.51
2	33	**55**	31	46.22
3	9	27	**75**	67.57
正確率 (%)	69.78	53.40	70.09	65.04



陸、 討論

依據我們的實驗過程與結果，我們將就討論分為以下幾個部分。

一、 特徵萃取結果

- **單字為特徵較易判別**：相較於句型為特徵，以單字為特徵較容易判別。因為：1.一篇文章中，字數通常較句數多，可以擷取的特徵數量較多，對於判別便較以句型為特徵有利。2. 高中階段難易度差異，單字大於句型。

- **句型為特徵，以篇為單位較易判別**：相較於以句為單位，以篇為單位較易判別。尤其是一年級，且一三年級有較理想的判斷分布，雖然一年級的句子，易被誤判成二三年級，但少了會被判斷為三年級的難句，所以會有較明顯的句子難易度分布特徵。

- **句型為特徵、以句為單位，難句較易判別**：相較於簡單的句子（如高一句子），難句反而較易判別。因為簡單的句子在簡單、困難的文章都會出現，沒有清楚特點，所以較難判別；反之，較難的句子有明顯「在簡單文章中不會出現」的特徵，因此高三的句子預測數量呈 3>2>1 的理想情形。

- **句型為特徵、篇章為單位，以 SVM 分類，加上各難易度句子的比例較易分類**：以「比例」作為特徵，相較於以「數量」作為特徵，相當於少了「文章總句數」的雜訊，分級時可以得到微小的改善。

- **表面特徵中，多個字為一級時，頻率版本效能較佳：**由於多個字為一組時，文長大幅干擾各級字在文中之出現次數。若採頻率版本，則可以看出各級字在文章中所佔比例，相較於次數，少了文長一變因，資料更能聚焦於「各難度字所佔比例」。

- **整合各個特徵，最易判別**：將特徵整合後，無論 SVM 或 RFC，得到的正確率都較分別測試時高。可推測各個特徵之間具有一定的互補性。

- **以篇為單位，高一最易判別**：以單字為例，高一課文不常出現高二、三的單字，會有清晰的「沒有高二高三難易度的單字」的特徵。句型亦是同理。

- **以篇為單位，高二最難判別**：高一、二，以及高二、三之間的文章，皆較容易混淆在一起，高一、三之間則不太容易混淆。因為課文的難易度往往循序漸進地增加，因此前後兩年級之間的難易度差異度較小。

二、 機器學習結果

- **SVM 的參數，對於分級效果影響甚大**：SVM 進行的實驗中，以預設參數分級，約略可得數據之間的相對結果。但參數選擇影響正確率至數十個百分點，若要得到最高分類正確率，調整、測試出最佳參數是必要的。

- **SVM 較適合篇章難易度分級**：就分類工具來說，若由 SVM 改用 RFC 或 DTC，正確率皆不如 SVM。因此，SVM 應比另兩項工具更適合篇章難易度分級。

三、 效能分析

為了評估效能，我們**模擬過去研究採用的方式，對我們的語料進行一樣的分級**（高中英文課文、分三級），其分類效能不盡理想（如表十八），可見**本研究實際上**相當具有挑戰性，而我們**提出新方法所得到的 65.04%，**確實較於過去研究**效能更佳**。

表十八：過去類似分級、研究之模擬與比較

模擬之研究	該研究所採方法／我們模擬的方式	(模擬)正確率
適讀性公式[5]	方法：Flesch 適讀性公式（該研究未分析分級效能）	42.22%
	模擬：Flesch 適讀性公式得特徵，SVM 分級	
全民英檢 GEPT)[3,4]	方法：表面特徵、子句結構等，C5.0 決策樹	57.31%
	模擬：同樣特徵，Scikit-learn 決策樹工具	
本研究	**方法：單字及句型之計數與頻率、改良表面特徵，支持向量機(SVM)、卷積神經網路(CNN)**	65.04%

柒、 結論

將篇章以單字為特徵的部分，和以句型為特徵的部分整合後，使用 SVM 和 RFC 兩項工具，皆能得到更高的正確率。可知單字、句型兩種特徵應具有互補性，相結合使得分類效果提升。未來可以考慮將 CNN sentence 分類納入整合，使自動分級多考量文意等更深入之資訊，對於提升正確率應該頗有幫助。

因為課文的難易度往往循序漸進地增加，因此前後兩年級之間的難易度差異度較小，高一、二，以及高二、三之間的文章，皆較容易混淆在一起，高一、三之間則不太容易混淆。高二文章最難判斷，而這樣的結果符合預期。

將多個字劃為一個難度等級、同等級出現次數加總，以建立濃縮過的特徵矩陣時，將整個特徵矩陣除以總字數，能去除文長的影響，使特徵更單純，正確率也大幅提升。

就篇章難易度的分級來說，RFC 或 DTC 的分級效果皆不及 SVM。因此，SVM 應比另兩項工具更適合篇章難易度分級。

篇章難度的自動判斷研究有一定的難度，和過去的研究相比，本研究提出了許多新的特徵萃取方式，最後也確實達到較佳的分級效果。未來將嘗試更多方向，找出能夠讓難易度更明顯的特徵，並和目前方法加以結合改善。

捌、　參考資料及其他

[1] Yoon Kim (2014). Convolutional Neural Networks for Sentence Classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746 – 1751. Retrieved from http://emnlp2014.org/

[2] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin (2010). A Practical Guide to Support Vector Classification. Retrieved from https://www.csie.ntu.edu.tw/~cjlin/

[3] 黃孝慈（2010）。利用字彙與子句結構進行全民英檢閱讀文章難易度分類之研究。長榮大學資訊管理學系碩士學位論文，未出版。

[4] 許珀豪（2013）。應用文字探勘技術於英文文章難易度分類。國立政治大學資訊管理學系碩士學位論文，未出版。

[5] 陳海泓（2013）。以適讀性公式挑選英文讀本之探究。教育資料與圖書館學，50(2)。

[6] 楊子儀（2009）。基於代理人計數之適性化英文閱讀文章推薦系統。長榮大學資訊管理學系碩士學位論文，未出版。

[7] 鄭恆雄、張郁慧、程玉秀、顧英秀、許秀玲、黃莉琪、劉欣潔、黃麗華（2001）。《大考中心高中英文參考詞彙表》編修研究計畫報告（第二期）。臺北市：大考中心高中英文參考詞彙表研究計畫小組。取自：http://www.ceec.edu.tw/

【評語】052502

1. 本作品具實用價值，且能採用適合的資料分析技術，完整度高。

2. 建議可多考慮實驗設計的合理性及完整性，用以驗證作品的實用價值。