1 文本

目標

在完成這章後，你將能夠

- 了解文本由代碼組成
- 注意到文件的字元、段落及頁格式
- 知道製作文本不同的方法
- 學會如何從掃描影像抽取文本

1.1 多媒體元素

多媒體元素 (Multimedia elements) 包括文本、圖形、動畫、聲音及視像。在演示及網頁中，多媒體元素扮演重要的角色。

各種多媒體元素的優點，摘錄如下：

多媒體元素	優點
文本	文本提供資訊，並連繫其他媒體元素。
圖形	圖形輔助文本，提供視覺效果，幫助用戶在最短時間內了解材料，「一張圖片可值千言。」
動畫	動畫有效地示範想法及說明抽像概念；當用戶對文字或靜態圖片感到厭倦時，動畫可以提高他們的興趣。
聲音	聲音提供除視覺外的另一種媒體，音響效果可提高用戶的注意力；錄下的聲音可強加內容的可信性。
視像	視像可帶給聽眾最大的震撼，「一個視像可值千張圖片。」你所付出的努力，必定會得到欣賞。

表 1　多媒體元素的優點

代碼 Code　　演示 Presentation　　網頁 Web pages

1.2　文本的優點

文本 (Text) 由字元組成，可準確無誤地傳達信息，並常用於內文及標題。

與其他多媒體元素比較，文本要求最少量的存貯空間，只需很少的資源，便可建立或顯示文本。

文本的傳輸只需很短的時間，因此有些網頁包含「純文字版本」，目的是顧及互聯網連接速度較慢的用戶。

1.3　字元代碼

文本均以**代碼 (Code)** 存貯於電腦中。

表 2　中文及英文的編碼

語言	代碼	代碼長度	可能的組合
英文	ASCII	8 位元 (1 字節)	256
中文	大五碼、國標碼、漢字碼	16 位元 (2 字節)	65,536

英文字元以「**美國訊息交換標準碼**」**(ASCII Code)** 來存貯。每個 ASCII 代碼代表一個字元，其長度為八**位元 (Bit)** 或一**字節 (Byte)**。利用 ASCII 代碼，可為 $2^8 = 256$ 字元編碼，這足以代表所有英文字母、阿拉伯數字及常用的符號。

由於中文字元超過 256 個，中文字的編碼需要兩個字節，並能代表高達 2^{16}=65,536 個中文字。今天，中文的編碼方式有多個，包括大五碼、國標碼、漢字碼等。

在編寫網頁時，為確保文本能準確地在用戶的瀏覽器上顯示，必須指定所用的字元集，例如，若網頁只有英文，字元集可設定為 "iso-8559-1" 如下：

```
<head>
<meta http-equiv="Content-Type" content="text/html;
        charset= iso-8859-1">
</head>
```

表 3　中文的編碼方式

編碼方式	使用語言	網頁上設定
大五碼 Big-5 codes	繁體中文	charset=big5
國標碼 GB (GuoBiao) codes	簡體中文	charset=gb2312
漢字碼 HZ (Hanzi) codes	簡體中文	charset=hz

字元 Character　　標題 Headline　　內文 Contents　　編碼 Encode
美國訊息交換標準碼 ASCII, American Standard Codes for Information Interchange

1.4 格式特徵

文本的格式可分類如下：

- 字元格式
- 段落格式
- 頁格式

A. 字元格式

字元格式 (Character formatting) 指定文本的外觀，包括顏色、字體、字型大小、字型款式。表 4 說明一些字元格式及例子：

表 4　字元格式

字體	字型大小	字型款式
這是細明體 **這是粗黑** 這是中圓 這是顏體	6點, 8點, 10點, 12點, 14 點, 16 點, 20 點	這是**粗體** 這是*斜體* 這是加底線 這是***粗斜體*** 這是**粗體加底線**

點數 (Point) 量度字元的高度，每點代表 1/72 吋，因此 72 點字元的高度大約是一吋。

圖 1　字型大小。一個有72點的英文字高度大約為一吋

沒有**襯線**的英文字體稱為 **sans serif**（**sans**，法語，解「沒有」）。

若文件以網頁等電子形式來發表，應避免使用罕見的字體。原因是，若所用的字體並沒有在用戶的電腦上安裝，便會被其他字體取代，以致文本不能如預期般顯示。

瀏覽器的字型大小有獨特的量度方式，其大小級別由 1 變化到 7，預設值是 3。

字體 Typeface　字型大小 Font size　字型款式 Font Style　襯線 Finishing stroke
粗體 Bold　斜體 Italics　加底線 Underline　預設值 Default value

字樣	例子	HTML 碼
Sans Serif	Example for sans-serif	**** Example for sans-serif ****
Serif	Example for serif	**** Example for Serif ****

表 5　Serif 及 Sans Serif

B. 段落格式

段落格式 (Paragraph formatting) 指定**行距 (Line spacing)**、段落**寬度 (Paragraph width)**、**文本對齊 (Text alignment)** 等。

在文字處理中，按一下**換行 (Enter)** 鍵便可建立新的段落。然而，在網頁上，每個段落是建立在標籤 和 之間。

對齊決定段落的邊緣形狀。對齊的格式有靠左、靠右、置中及左右的對齊（見表 6）。

	例子	HTML 碼
靠左	靠左的文本在左邊的邊緣上是平順的但在右邊是破碎的。靠左排列是桌上型出版被推薦的方法。較容易讀而且字留間隔是較平均的。	 ....
置中	每行字是置中的; 因此，左邊的和右邊的邊緣是破碎的。排列置中時常作為大標題，正式的邀請或公告。	 ....
靠右	靠右的文本在右邊的邊緣上是平順的但在左邊是破碎的。靠左排列常用作圖片說明或廣告，但是不被推薦為內文。 我們習慣於從左邊到右邊讀。 如果本文的左邊邊緣不平順，對閱讀是困難的。	 ....
左右對齊	右邊和左邊的邊緣上 均是平順。 因為要維持左右 邊緣都平順，字和字之間的正常空間被改變為了 ，會造成不平均的白色空間。	 ....

表 6　文本對齊

邊緣 Edge　　靠左 Aligned left　　靠右 Aligned right　　置中 Centered
左右對齊 Justified

C. 頁格式

在文字處理中，**頁格式 (Page formatting)** 指定紙張的大小、方位 (風景/肖像)、邊緣、頁首和頁尾等。

在網頁上，頁格式指定背景的顏色，影響整個網頁的背景。

```
<body bgcolor="#CCFF99">
```

1.5 建立文本

A. 鍵入文本

文本通常是透過鍵盤，輸入電腦。以鍵盤來輸入中文有多個方法，下列是較常用的中文輸入法：

- 倉頡輸入法 (Changjei)
- 速成 (Quick)
- 注音輸入法 (Phonetic)
- 九方輸入法 (Q9)

專業打字員較喜歡使用倉頡輸入法，因為幾乎所有的中文字都可按獨特的序列輸入，毋需作出選擇，較為便捷。

B. 其他方法

除鍵盤輸入外，建立文本還有其他的方法。印刷文件上的文字可以透過掃描器及**「光符識別」(OCR)** 軟件，直接輸入電腦：首先，掃描器產生黑白的數碼影像，然後光符識別軟件對影像進行分析，產生可編輯的文本。識別後所產生的文字檔可由文字處理器編輯，而檔案較原本的數碼影像小。

手寫文本亦可透過**「手寫識別」**軟件，轉換成文本。除軟件外，你還需要一塊手寫板。同樣地，口語亦可直接轉換成文本，這便需要麥克風、音效卡和**「話音識別」**軟件。

	文本的創造方法	工具
1.	鍵盤輸入	鍵盤
2.	光符識別	掃描器、光符識別軟件
3.	手寫識別	手寫板、手寫識別軟件
4.	話音識別	麥克風、音效卡、話音識別軟件

表 7　文本創造方法

風景 Landscape　肖像 Portrait　輸入法 Input method
光符識別 OCR, Optical character recognition　手寫識別 Handwriting recognition
話音識別 Voice recognition

1.6　減少文字檔的大小

圖 2　檔案壓縮實用程式

文字檔的大小可利用**壓縮實用程式 (Compression utility)** 來減少。壓縮的過程必須是**無損耗 (Lossless)** 的，意謂當經壓縮的檔案被**解壓縮 (Decompressed)** 時，最初的檔案可完全地回復過來。

另一個方法是將文件匯出到**純文字檔 (Plain text file)**，藉著除去格式屬性，減少檔案的大小，但是這會導致某些資訊遺失。

活動 1　使用光符識別軟件捕捉文本

所需硬體	掃描器連接到一部電腦
所需軟體	SimpleOCR -- 免費軟件 http://www.simpleocr.com/

活動目的

經過這個活動後，你將能夠

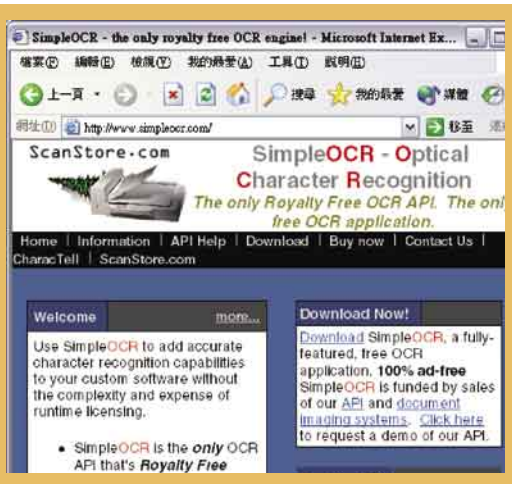
- 使用掃描器掃描一份文件
- 為光符識別選擇適當的圖像格式
- 欣賞光符識別軟件如何識別來自掃描影像的文本
- 知道光符識別軟件的限制

在這個活動中，你將使用掃描器去掃描一份文件，並轉換成可編輯的文本。

SimpleOCR®

SimpleOCR® 是能識別英文印刷文本的免費軟件，收費版本更能識別手寫文本。

步驟 1　準備工作



1. 準備一份有文本和圖形的印刷文件。

2. 下載來自 SimpleOCR 的 simpleocr.exe
   - 到 http://www.simpleocr.com

   或，
   - 到 http://www.download.com
   - 然後以關鍵字 "SimpleOCR" 進行搜尋

3. 安裝光符識別軟件：
   - 按兩下 simpleocr.exe，跟隨指令

文字檔 Text file　　匯出 Export　　可編輯 Editable　　免費軟件 Freeware

步驟 2 掃描文件

1. 執行程式：
   - 按一下 開始 ▶ 所有程式 ▶ SimpleOCR ▶ SimpleOCR
   - 按一下 機器印刷 (Machine Print)

2. 你可參看如何使用該程式的示範：
   - 按一下 示範 (Demo)

3. 開始使用軟件：
   - 按一下 選擇 (Select)

4. 製作一份新的文件：
   - 按一下 加頁 (Add Page)

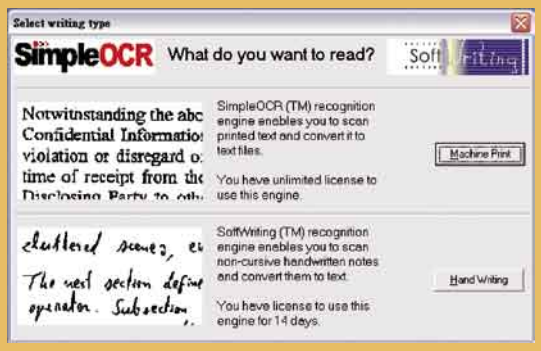
   「選擇來源 (Select Source) 視窗」彈出。

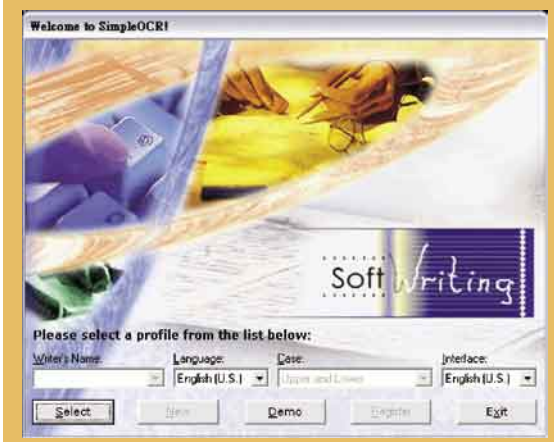
   確定掃描器已連接到你的電腦，並將一份印刷文件放在掃描器上。

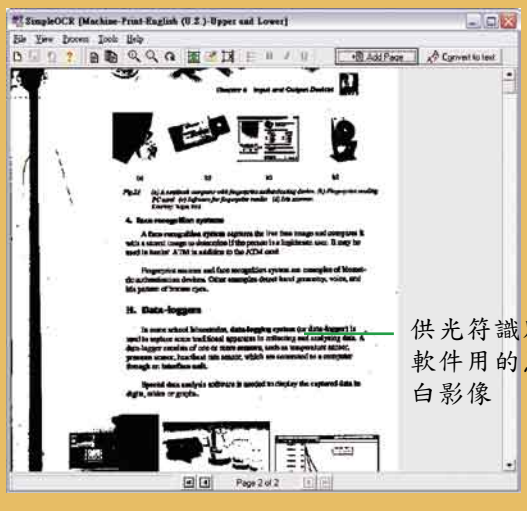
5. 掃描文件：
   - 選擇 掃描器 (Scanner)
   - 按一下 確定 (OK)

   若掃描程序成功，掃描後的影像將出現在屏幕上，如右圖所示。

注意： 掃描影像應該是黑白的，這類影像最能配合光符識別的使用。







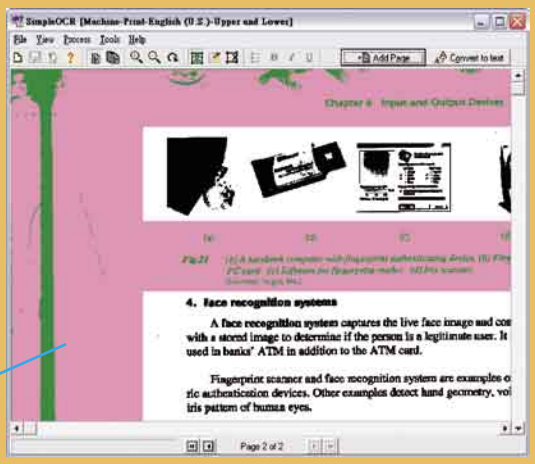
步驟 3　轉換成文本

1. **標示不想要的區域：**

- 按一下 **忽略指定區域** (Ignore Region) ( )
- 拖曳滑鼠以標示不想要的區域

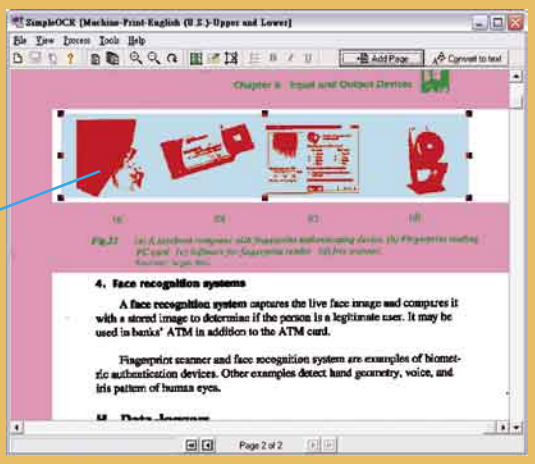
可使用 **縮放** (Zoom) (或) 改變你的視野，作較精確的區域選擇。

- 按一下已標示的區域，調整其大小



2. **選擇要保留的圖片：**

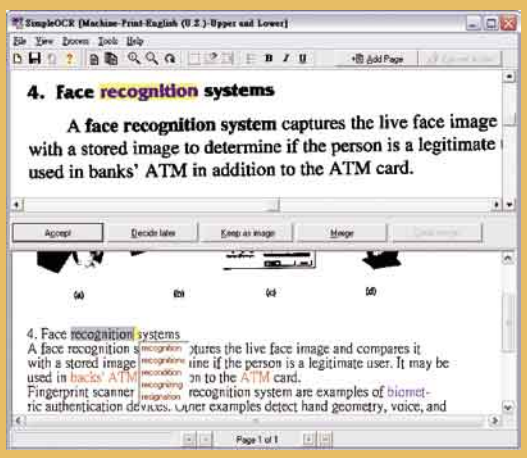
- 按一下 **圖片區域** (Picture Region) ( )
- 標示你想 SimpleOCR 匯入到文書文件的圖片



3. **將影像轉換成文本：**

- 按一下 **轉換成文本** (Convert to text)

屏幕的下半部顯示 SimpleOCR 已成功地轉換的本文。



步驟 4 校對

在存貯轉換後的文本前，SimpleOCR 需要你作出校對。轉換後的文本 (在視窗下半部) 以不同的顏色標示，並可依下表詮釋：

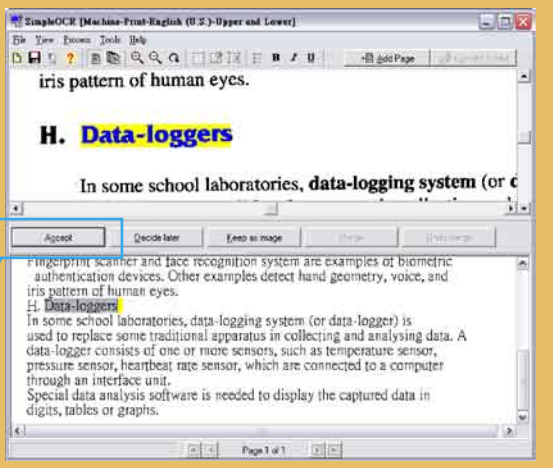
	文本的顏色	意義
1.	黑色	軟件假定轉換是正確的。 正常情況下，你不用修改這些字。
2.	黃色的背景	正在編輯中的字
3.	紅色	字典中找不著的字
4.	藍色	SimpleOCR 不能確定的字
5.	綠色	從縮寫轉換過來的字

1. 校對文件：

為每個可疑 (有黃色背景) 的字執行以下步驟：

- 選擇來自下拉式選單的建議或直接鍵入你自己的文字
- 按一下 **接受** (Accept)

顏色將會變回常態。

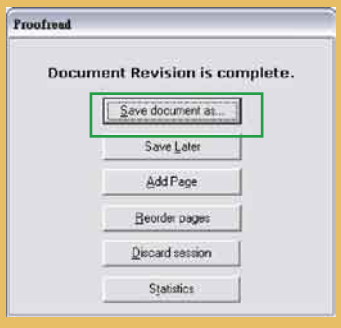


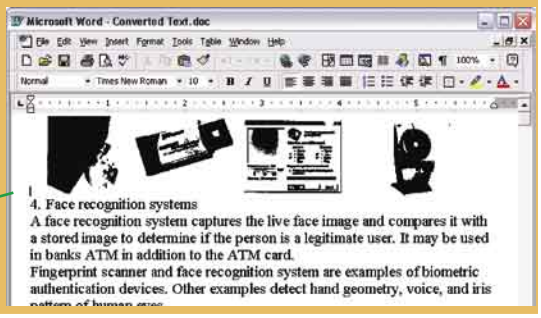
處理後按此按鈕

2. 確認更改：

在按最後一次 **接受** (Accept) 後，「**校對**(Proofread) **視窗**」便會彈出。

- 按一下 **存檔成 ...** (Save document as ...)
- 以 MS Word® 文件形式把文本存貯起來





給轉換後本文和圖形出現在MS Word中

摘要

1. 文本均以代碼存貯在電腦上。
2. 美國信息交換標準代碼(ASCII)應用於英文字元，每個ASCII代碼的長度是一字節。
3. 中文字元需要兩個字節編碼。
4. 字元格式決定字元的外觀，例如顏色、字體、字型大小、字型款式等。
5. 段落格式包括行高、段落寬度、及文本對齊。
6. 頁格式記載紙張的大小、方位 (風景/肖像)、邊緣、頁首和頁尾。
7. 文本能透過打字、光符識別、手寫識別及話音識別而建立。
8. 無損耗壓縮：當壓縮後的檔案被解壓縮時，最初的檔案可完全回復。

練習

多項選擇題

1. 代表英文字 "Computer." 的位元數量(不包括引號)是
   A. 8
   B. 9
   C. 64
   D. 72

2. 下列哪個字元格式不適用於強調網頁內的文本?
   A. 粗體字
   B. 斜體字
   C. 加底線
   D. 粗體-斜體字

3. 下列哪項不是中文編碼方法?
   A. Big-5
   B. Changjei
   C. GB Code
   D. HZ Code

4. 下列哪項不是純文字檔?
   A. HTML 檔
   B. XML 檔
   C. TXT 檔
   D. Word 文件

5. 文字檔的大小取決於
   A. 字元的數量
   B. 字的數量
   C. 頁的數量
   D. 字體大小

引號 Quotation mark

問答題

注意：部分題目可能涉及其他章節的內容。

1. 彼得在某非政府青少年協會中工作。協會將要建立一個網站，幫助青少年了解使用政府服務的程序。

網站的語言將同時包括中、英文。中文有繁、簡體之分，但是彼得發現將兩者都同時包括在他的網站是相當困難的，最後，他決定只使用繁體中文。

(a) 寫出在技術上建立一個網站包括繁、簡體的困難。 (3 分)

彼得將僱用一個能翻譯中、英文的助理。

(b) 該助理應該懂得如何輸入中文。試舉出**三個**中文輸入法的例子。 (3 分)

(c) 建議彼得如何設計網頁及安排檔案，以便用戶能選擇他們的語言。 (3 分)

(d) 有些中文字是電腦上沒有的。試向彼得提供**一個**建議，以便用戶的瀏覽器能顯示這些中文字。 (2 分)

彼得的網站相當繁忙。最近，他收到用戶的抱怨，投訴他的網站速度非常緩慢。

(e) 除替硬件升級外，有甚麼解決方案可加快用戶取得他們所需的資訊？ (1 分)

(f) 彼得建立了一個 FAQ 網頁。

i) 甚麼是 FAQ？ (1 分)

ii) FAQ 能如何幫助

(1) 網站管理員？

(2) 用戶？ (2 分)

2. 瀏覽某些網站時，當所有的文本已經下載後，有些圖形仍然在傳送中。

(a) 試提供**一個**理由。 (1 分)

(b) 檔案傳送的速度會否因文本字型的大小不同而有所分別？試解釋。 (2 分)

(c) 相同的文本可能會在兩部電腦中有不同的顯示。建議**兩個**理由。 (2 分)

(d) 某種情況下，網站因為太繁忙以致不能顯示某些圖形，怎樣才可將圖形上重要的資訊告知用戶？ (2 分)

(e) 建議**兩種**方法確保用戶的瀏覽器在顯示文本時，必定同時顯示圖形。 (4 分)

(f) 在網頁上，以加底線來強調某些文本**不是**恰當的做法。試解釋。 (1 分)

(g) 描述**三種**其他字元格式用以強調網頁上的文本。 (3 分)

3. 張先生是一位經濟科老師。他建造了一個網站，以便存放用於課堂上的演示文件 (例如 PowerPoint® 文件)。

(a) 張先生的一個學生剛購置了一部只安裝了瀏覽器的電腦，但不能成功地顯示張先生的演示。

i) 試建議**一個**原因。 (2 分)

ii) 試建議張先生應怎樣做，以便所有學生都能在家中看到網頁上的內容。 (1 分)

(b) 張先生亦決定上載他的筆記。他正在考慮使用 Word 還是 PDF 格式。寫出每種格式的**優點**及**缺點**。 (4 分)

(c) 在他的筆記上，有些文本是來自報章的。張先生首先把報章掃描，將數碼影像存檔，然後轉換成可匯入到他的筆記的文本。

i) 張先生用了甚麼軟件將掃描後的影像轉換成文本？ (1 分)

ii) 試比較掃描後影像的檔案大小與轉換後文本的檔案大小。 (1 分)

iii) 寫出影像內的文本和轉換後的文本之間的**一個**分別。 (1 分)

iv) 張先生應該以那種形式來存貯其掃描後的影像，以便促進轉換程序？試解釋。 (2 分)

v) 張先生在報章上發現一些有用的照片。但是，那些已存貯的圖形質素實在太差，無法使用。張先生應該怎樣做才可取得質素較佳的圖形？ (2 分)

4. 喬治建立了一個流行歌歌詞的網頁。該網頁是使用文本編輯器來編寫的，並包括許多圖形，例如橫幅、圖形連結和整合在圖形中的歌詞。

(a) 寫出**兩個**使用圖形來顯示文字的**優點**。 (4 分)

(b) 圖形的連結對一些用戶構成不便。
i) 討論造成這問題的原因。
ii) 建議如何解決這問題。 (4 分)

(c) 有些用戶不喜歡喬治把歌詞整合在圖形中。試列出**兩個**原因。 (2 分)

(d) 有些用戶希望喬治為每首歌加插 MP3 檔案，但被喬治拒絕了。提供**一個**支持他決定的理由。 (1 分)

(e) 解釋為何有些中文操作系統 (OS) 的電腦不能夠適當地顯示中文歌詞的字元。 (2 分)

(f) 喬治應如何避免歌詞裡包含誤拼的英文字? (2 分)