

拼音文本詞彙統計工具 使用說明

(以哈利波特為例)

1. 載入文本：提供三種文本格式與來源
 - a. 從 DocuSky 匯入：已在 DocuSky 建庫
 - b. 上傳 xml 檔：本地 .xml 檔案
 - c. 上傳 txt 檔：本地 .txt 檔案

拼音文本詞彙統計工具

使用說明English

載入文本

從 DocuSky 載入

上傳 xml 檔

上傳 txt 檔

Corpus Name - Document Name

2. 載入後，內文可於黑框中檢視，並使用頁籤切換文件

拼音文本詞彙統計工具

使用說明English

載入文本

從 DocuSky 載入

上傳 xml 檔

上傳 txt 檔

哈利波特英文版 - 01Harry-Potter-and-the-Philosophers-Stone_00001

« 1 2 3 4 5 ... 198 »

go to the th document

Go

Chapter 1 The Boy Who Lived Mr. and Mrs. Dursley, of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much. They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense. Mr. Dursley was the director of a firm called Grunnings, which made drills. He was a big, beefy man with hardly any neck, although he did have a very large mustache. Mrs. Dursley was thin and blonde and had nearly twice the usual amount of neck, which came in very useful as she spent so much of her time craning over garden fences, spying on the neighbors. The Dursleys had a small son called Dudley and in their opinion there was no finer boy anywhere. The Dursleys had everything they wanted, but they also had a secret, and their greatest fear was that somebody would discover it. They didn't think they could bear it if anyone found out about the Potters. Mrs. Potter was Mrs. Dursley's sister, but they hadn't met for several years; in fact, Mrs. Dursley pretended she didn't have a sister, because her sister and her good-for-nothing husband were as unDursleyish as it was possible to be. The Dursleys shuddered to think what the neighbors would say if the Potters arrived in the street. The Dursleys knew that the Potters had a small son, too, but they had never even seen him. This boy was another good reason for keeping the Potters away; they didn't want Dudley mixing with a child like that. When Mr. and Mrs. Dursley woke up on the dull, gray Tuesday our story starts, there was nothing about the cloudy sky outside to suggest that strange and mysterious things would soon be happening all over the country. Mr. Dursley hummed as he picked out his most boring tie for work, and Mrs. Dursley gossiped away happily as she wrestled a screaming Dudley into his high chair. None of them noticed a large, tawny owl flutter past the window. At half

3. 載入後，出現分析選項，有三種模式
- a. 詞頻：n-gram 分析
 - b. 詞彙清單：給定詞彙清單進行統計
 - c. 標記：標記統計

拼音文本詞彙統計工具

使用說明English

somebody would discover it. They didn't think they could bear it if anyone found out about the Potters. Mrs. Potter was Mrs. Dursley's sister, but they hadn't met for several years; in fact, Mrs. Dursley pretended she didn't have a sister, because her sister and her good-for-nothing husband were as unDursleyish as it was possible to be. The Dursleys shuddered to think what the neighbors would say if the Potters arrived in the street. The Dursleys knew that the Potters had a small son, too, but they had never even seen him. This boy was another good reason for keeping the Potters away; they didn't want Dudley mixing with a child like that. When Mr. and Mrs. Dursley woke up on the dull, gray Tuesday our story starts, there was nothing about the cloudy sky outside to suggest that strange and mysterious things would soon be happening all over the country. Mr. Dursley hummed as he picked out his most boring tie for work, and Mrs. Dursley gossiped away happily as she wrestled a screaming Dudley into his high chair. None of them noticed a large, tawny owl flutter past the window. At half

詮釋資料

filenamecorpustopictitlebook_code

詞彙統計

詞頻

詞彙清單

標記

N-gram 大小1

是否使用停用詞Off

編輯停用詞清單

文件範圍e.g. 1,5,6-10

是否使用詞形還原Off

開始統計

詞頻

詞彙清單

標記

文件範圍e.g. 1,5,6-10

是否使用詞形還原Off

詞彙清單編輯上傳 (.txt)

開始統計

詞頻

詞彙清單

標記

文件範圍e.g. 1,5,6-10

標記Udef_TeacherUdef_sorcery

開始統計

4. 調整參數後，按「開始統計」進行分析（第一次跑會套用 NLP 套件，需要較長時間）

詞彙統計

| 詞頻 | 詞彙清單 | 標記 |
|--|------|----|
| <div>N-gram 大小: 1</div> <div>文件範圍: e.g. 1,5,6-10</div> <div>是否使用停用詞: Off</div> <div>是否使用詞形還原: Off</div> <div>編輯停用詞清單</div> <div>開始統計</div> | | |

可設定參數：

- 文件範圍(全)：要統計的文件頁數，預設全部
- N-gram 大小(詞頻)：要統計的 gram 數，選項 1-5
- 是否使用停用詞(詞頻)：去除停用詞再進行統計，stopwords(英文)由 python NLP 套件提供，按「編輯停用詞清單」可刪減、增加停用詞，按「儲存」更新停用詞清單

是否使用停用詞 ☒ On 編輯停用詞清單

拼音文本詞彙統計工具

停用詞清單

預設停用詞清單取自：自然語言處理 python 套件 NLTK。

Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. O'Reilly Media Inc.

i

i'm

me

my

myself

we

our

ours

ourselves

you

you're

you've

you'll

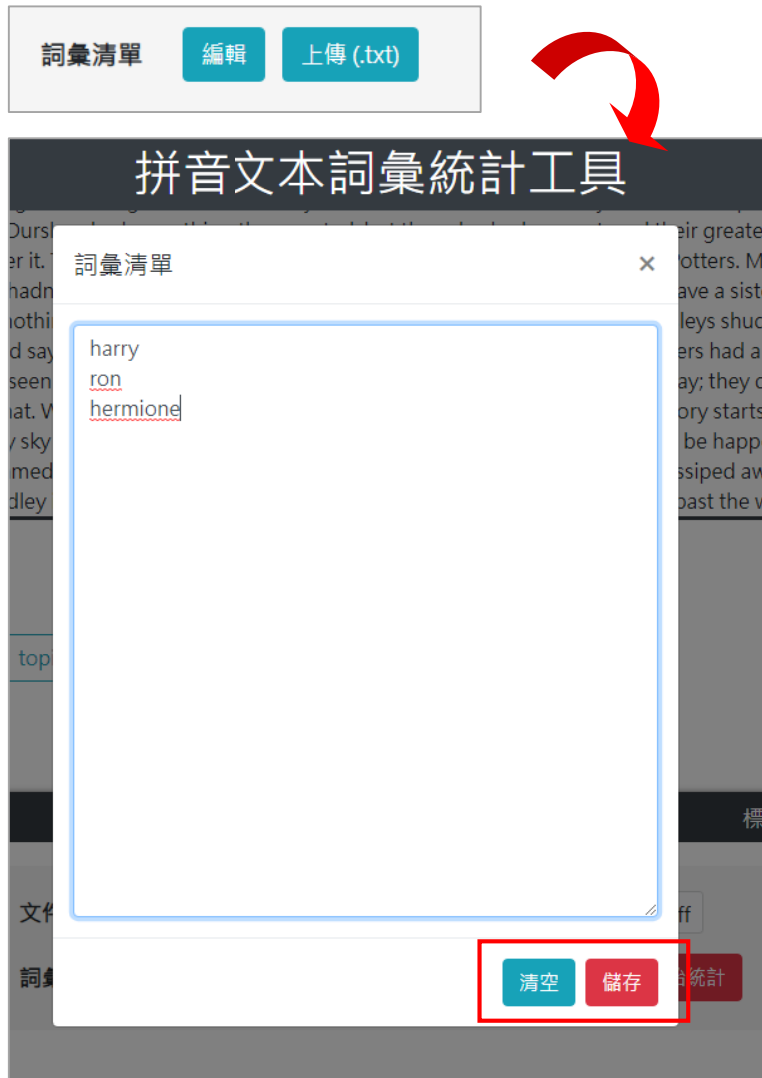
you'd

your

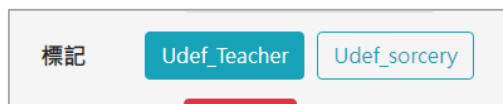
yours

儲存

- 是否使用詞形還原(詞頻、詞彙清單)(限英文)：還原動詞變化、名詞單複數再進行統計，例：say, says, said, saying => say
- 詞彙清單：欲進行統計的詞彙清單，可按「編輯」手動輸入詞彙，也可直接「上傳(.txt)」的詞彙清單，按「清空」刪除所有詞彙，按「儲存」更新詞彙清單



- 標記：列出抓取到的所有標記名稱，選擇欲進行統計的標記，例子如下圖：選擇 Udef_Teacher，Udef_sorcery 不選



5. 結果顯示在下方，一次分析會產生一個表格

例子一：用不同的參數跑了三次詞頻分析的結果

拼音文本詞彙統計工具

使用說明English

詞頻

詞彙清單

標記

N-gram 大小2

是否使用停用詞On

編輯停用詞清單

文件範圍e.g. 1,5,6-10

是否使用詞形還原On

開始統計

#TermFreq

1of the4896

2in the3566

3said harry2623

4he was2483

5at the2442

6to the2382

#TermFreq

1said harry2623

2said ron1537

3said hermione1265

4said dumbledore662

5professor mcgonagall594

6mrs weasley592

#TermFreq

1say harry2619

2say ron1539

3say hermione1265

4say dumbledore662

5mrs weasley599

6professor mcgonagall598

例子二：使用詞彙清單內的詞進行統計

詞頻

詞彙清單

標記

文件範圍e.g. 1,5,6-10

是否使用詞形還原Off

詞彙清單

編輯上傳 (.txt)

開始統計

#TermFreq

1harry18234

2ron8182

3hermione5356

例子三：分別跑了兩個標記的統計

詞類

詞彙清單

標記

文件範圍

e.g. 1,5,6-10

標記

Udef_Teacher

Udef_sorcery

開始統計

• 文件範圍：

• 標記：Udef_Teacher

| # | Term | Freq |
|---|------------|------|
| 1 | Dumbledore | 3351 |
| 2 | Hagrid | 2036 |
| 3 | Snape | 1826 |
| 4 | Lupin | 810 |



• 文件範圍：

• 標記：Udef_sorcery


| # | Term | Freq |
|---|--------------|------|
| 1 | Occlumency | 48 |
| 2 | Accio | 33 |
| 3 | Expelliarmus | 26 |
| 4 | Stupefy | 22 |

6. 結果表格：詞彙照出現次數排序，上方帶有工具列

| # | Term | Freq |
|---|------------|------|
| 1 | of the | 4896 |
| 2 | in the | 3566 |
| 3 | said harry | 2623 |
| 4 | he was | 2483 |
| 5 | at the | 2442 |
| 6 | to the | 2382 |

-  : 刪除此次統計結果
-  : 顯示統計參數，可自行設定顯示的詞彙數量



-  : CSV 格式的統計結果(以文件為單位)，可以「複製」到剪貼簿，或者直接「下載 CSV 檔」



若有選擇詮釋資料，會一併將資料輸出。

詮釋資料

filename

corpus

topic

title

book_code

統計結果 (CSV 格式)

"corpus","title","term","frequency"

"哈利波特英文版","01.01 The Boy Who Lived","privet drive","8"

"哈利波特英文版","01.01 The Boy Who Lived","never seen","3"

"哈利波特英文版","01.01 The Boy Who Lived","said mr","2"

"哈利波特英文版","01.01 The Boy Who Lived","said mrs","1"

"哈利波特英文版","01.01 The Boy Who Lived","albus dumbledore","2"

"哈利波特英文版","01.01 The Boy Who Lived","professor mcgonagall","25"

"哈利波特英文版","01.01 The Boy Who Lived","said professor","6"

"哈利波特英文版","01.01 The Boy Who Lived","said dumbledore","11"

"哈利波特英文版","01.01 The Boy Who Lived","madam pomfrey","1"

"哈利波特英文版","01.01 The Boy Who Lived","last night","1"

"哈利波特英文版","01.01 The Boy Who Lived","harry potter","5"

"哈利波特英文版","01.01 The Boy Who Lived","professor dumbledore","2"

"哈利波特英文版","01.01 The Boy Who Lived","sirius black","1"

"哈利波特英文版","01.01 The Boy Who Lived","i've got","1"

"哈利波特英文版","01.01 The Boy Who Lived","could see","1"

"哈利波特英文版","01.01 The Boy Who Lived","front door","2"

"哈利波特英文版","01.01 The Boy Who Lived","said hagrid","1"

"哈利波特英文版","01.02 The Vanishing Glass","privet drive","1"

複製

下載 CSV 檔

7. 可以切換中、英文介面

英文文本詞彙統計工具

English

English Term Statistics

中文