

Audience 站台1.3使用說明

RD 2 黃彥鈞

大綱

1. 版本更新簡介
2. 更新功能
3. 操作簡介
4. 注意事項

版本更新簡介

- 版本開發背景:
 - 因應接下來的深度學習模型應用資料標註需求，因此開發新的資料整備模組用來儲存模型資料，以及串接資料標註平台。
- 1.3 版站台
 - 將停用原資料準備模組功能
 - 模型任務資料改由串接新資料整備模組
 - 資料整備模組提供資料標註(doccano)、資料維護(CRUD)功能
 - CRUD: create, retrieve, update, delete
 - 原有貼標與模型訓練流程不變
 - 使用者需更新舊資料至新資料整備模組 (洽 RD2 Audience 負責人)

版本更新簡介

- 更新功能
 - ~~資料準備任務~~
 - 資料整備任務
 - 資料標註平台
 - 資料維護模組

操作簡介-資料整備模組！

- 資料整備任務流程：



操作簡介-資料整備模組2

AUDIENCE
TOOLKITS

說明書

資料整備任務

模型建立任務

族群貼標任務

說明書

Audience Toolkits 族群訓練工具包

藉由此工具，您可以從0開始訓練一個屬於你的機器學習模型，並且進一步應用至OpView的族群標籤。

什麼是Audience族群標籤

Audience，aka族群，藉由透過網路上內容作者發布的內容，判斷其背後可能代表的族群類型。這邊假設網路內容可用的資訊如發文頻道、發文作者名稱、發文內容等資訊，會透露出一些某些族群的特有資訊。目前支援可用的欄位名稱有 標題、內容、來源、來源網站、作者。

舉個例子

在dcard討論區中，作者名稱有個固定的格式，可以拿來判斷發文者性別。例如「台灣大學/F」，其代表的涵意為「台灣大學的某位女生」，可為其貼上女性的族群標籤。

再舉個例子

在bbs論壇中，男性的使用者在發文的時候，有一個習慣是會以「小弟我...」作為句子開頭，以示禮貌。這個時候可藉由「小弟我...」為開頭的關鍵字或規則，為其貼上男性的族群標籤。

以上為藉由欄位的選擇與規則即可判斷的狀況，但若不是光靠規則可以判斷的狀況時，我們也可以藉由「監督式學習」的機器學習模型進行判斷，只需提供足夠的「內容」與其代表的「族群標記」，即可讓機器自動尋找文章中的有用資訊，學習如何判斷文章可能帶有的族群標籤。

只要準備好模型，並可設定族群貼標任務，應用於您希望生效的資料範圍上。

可用工具

此工具包將製作流程分為三種任務工具：

資料整備任務

在這裡您可以建立、描述，並管理您想進行的資料標記任務，完成的標記任務可以在「模型訓練任務」中使用。

模型建立任務

在這裡您可以使用「資料標記任務」中完成的資料來建立、訓練，並驗證您的機器學習模型。

族群貼標任務

在這裡您可以藉由選擇「模型訓練任務」中您的模型，並挑選應用的資料範圍，建立族群貼標任務。

2022/5/17

eLand

6

操作簡介-資料整備模組3

Doccano 標註平台

建立任務儲存
已標記資料集

上傳格式範例資料(csv) 下載

AUDIENCE
TOOLKITS

說明書

資料整備任務

模型建立任務

族群貼標任務

資料整備任務

任務列表

Show 10 entries

Search:

任務名稱	任務類型	創建者	建立時間	更新時間
機器學習資料測試	machine_learning_task	ychuang	2022-04-22 15:24:31	
測試	rule_task	ychuang	2022-05-05 13:58:08	
規則資料建立測試	rule_task	ychuang	2022-04-28 14:30:56	

Showing 1 to 3 of 3 entries

Previous 1 Next

任務範例資料

資料標註平台

建立新任務

資料整備流程

先至 **doccano** 資料標註平台 進行資料標註，再建立新任務將 結果資料(csv) 上傳至任務頁面。

[What is doccano?](#)

上傳格式請參考 任務範例資料

machine_learning_task

建立任務請選擇 機器學習模型資料

每項任務請至少準備以下類型資料:

- train 訓練資料
- dev 驗證資料
- test 測試資料

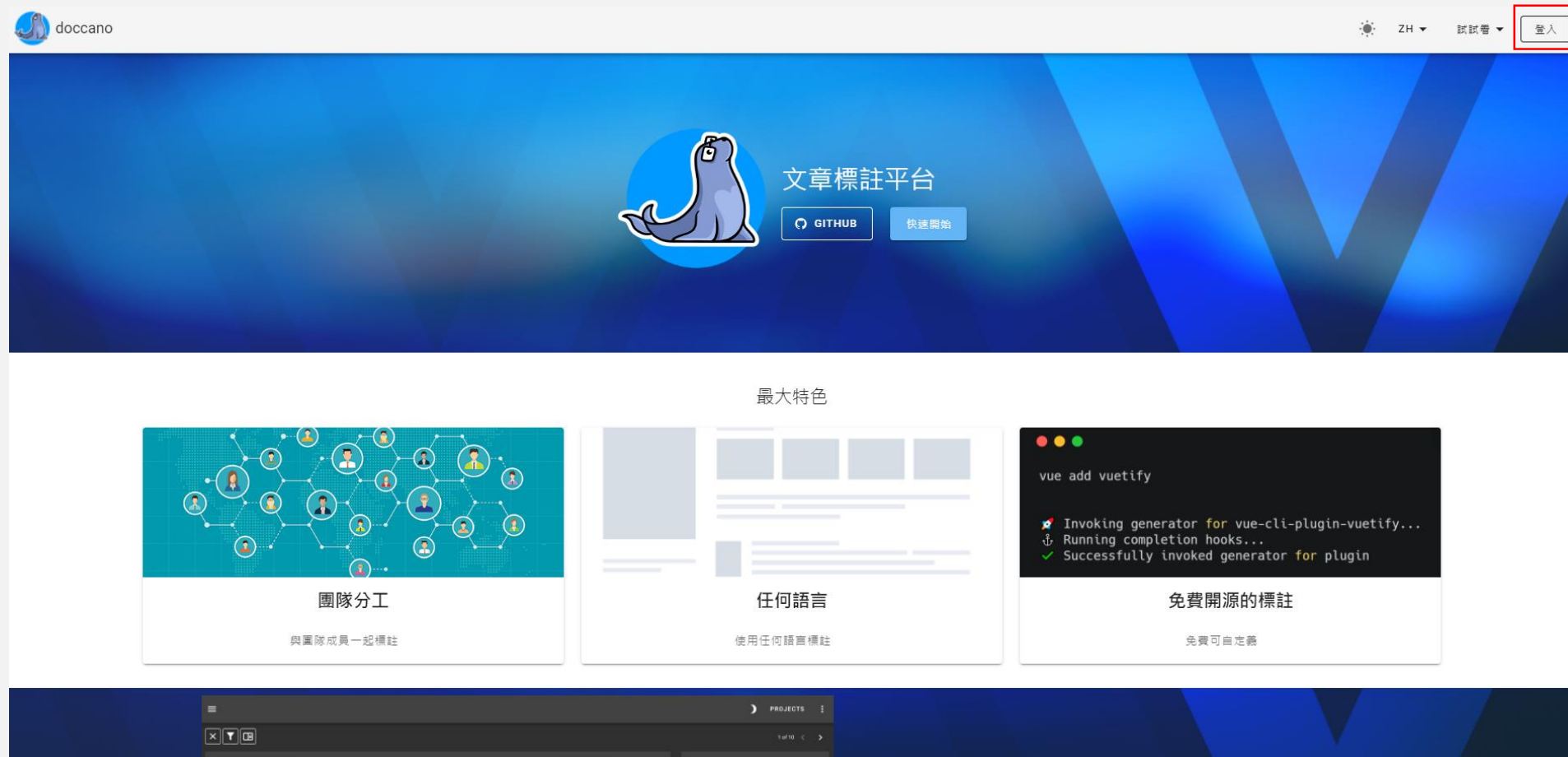
上傳資料欄位，*為必要欄位:

- title
- author

任務頁連結與任務資訊

任務流程說明

操作簡介-doccoano資料標註平台！



操作簡介-doccoano資料標註平台2

背景主題與語言

建立與刪除任務

創建

刪除

查詢

<input type="checkbox"/>	名稱	描述	類型	Updated	Tags
<input type="checkbox"/>	test2	test multi label	DocumentClassification	06/04/2022 09:53	
<input type="checkbox"/>	testing	test	DocumentClassification	15/04/2022 09:08	



每頁最多可顯示 10 1-2 of 2

任務頁面連結與任務列表

操作簡介-doccoano資料標註平台3



Audience 任務請選擇
文章分類任務標註


建立專案項目

✓ Positive  Negative 

Fair drama/love story movie that focuses on the lives of blue collar people finding new life thru new love. The acting here is good but the film fails in cinematography, screenplay, directing and editing. The story/script is only average at best. This film will be enjoyed by Fonda and De Niro fans and by people who love middle age love stories where in the courtship is on a more wiser and cautious

✓文章分類任務標註

✓ Cat  Dog 



影像分類


After bowling Somerset out for 83 on the opening morning at Grace Road, Leicestershire extended their first innings by 94 runs before being bowled out for 296 with England discard Andy Caddick taking three


•ORG
•LOC •ORG
•LOC •PER

序列資料標註




If it had not been for his help, I would have failed.

New text

S'il ne m'avait pas aidé, j'aurais échoué. 

S'il ne m'avait pas aidée, j'aurais échoué. 

序列生成

✓ Flight  FlightTime  Airfare 

I want to fly from Boston at 8:38 am and arrive in Denver at 11:10 in the morning.

•City •Time
•City •Time

文意識別和槽位填充

項目名稱

請輸入項目名稱

描述

Tags

☐ Allow single label

☐ 隨機文檔序列

☐ 所有使用者共享標註

創建

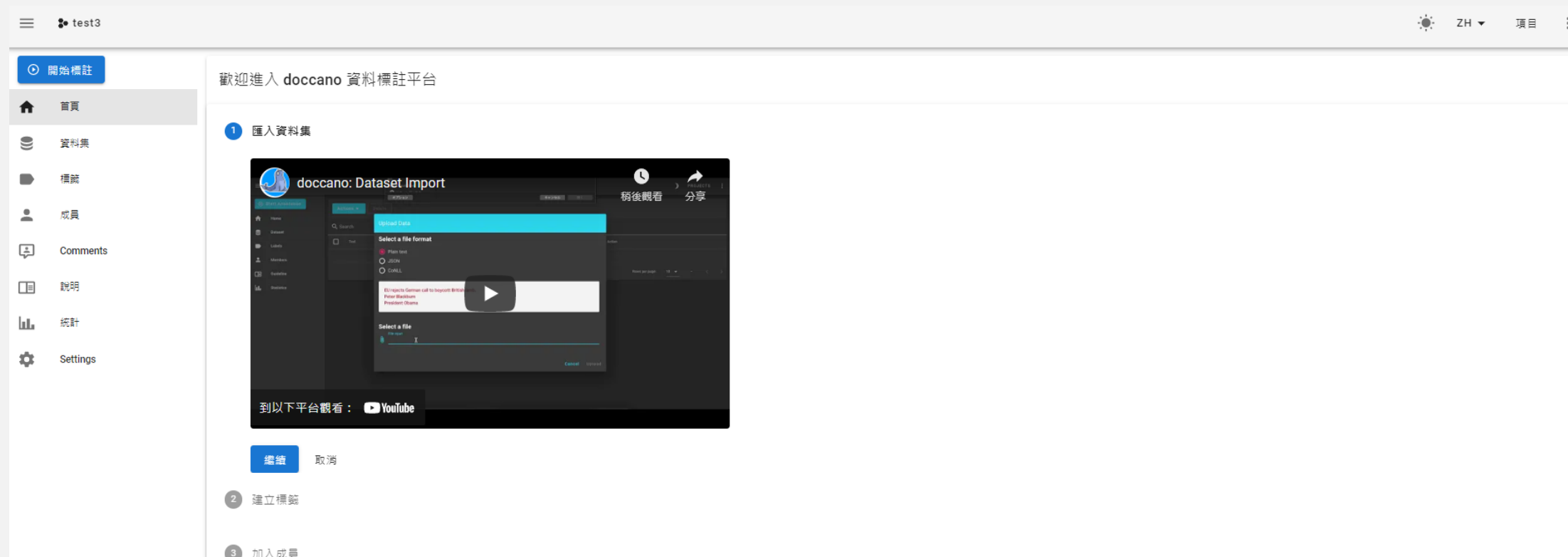
選擇是否為單標籤標註

2022/5/17

操作簡介-doccoano資料標註平台4

先行觀看標註教學，再開始標註資料 ...

下載標註完成之資料集，返回 Audience 站台資料整備任務頁面



操作簡介-資料整備模組4

The screenshot shows the 'AUDIENCE TOOLKITS' interface. On the left is a sidebar with links: '說明書', '資料整備任務', '模型建立任務', and '族群貼標任務'. The main area is titled '資料整備任務' and '任務列表'. It features a table with columns '任務名稱' and '任務類型'. The table lists three tasks: '機器學習資料測試' (machine_learning_task), '測試' (rule_task), and '規則資料建立測試' (rule_task). A modal window titled '建立任務' is open in the center, containing fields for '任務名稱', '任務簡述', a '任務類型' dropdown menu (currently set to '機器學習模型資料'), and a checkbox for '是否為多標籤'. A '送出' button is at the bottom right of the modal. In the background, the top right of the interface shows a user profile 'ychuang' and a navigation bar with buttons: '任務範例資料', '資料標註平台', and '建立新任務' (highlighted with a red box). Below the navigation bar, there is a '資料整備流程' section with instructions on how to create a task using the 'doccano' platform and a list of required data types: 'train 訓練資料', 'dev 驗證資料', and 'test 測試資料'.

AUDIENCE TOOLKITS

資料整備任務

任務列表

Show 10 entries

任務名稱	任務類型
機器學習資料測試	machine_learning_task
測試	rule_task
規則資料建立測試	rule_task

Showing 1 to 3 of 3 entries

建立任務

任務名稱:

任務簡述:

任務類型: 機器學習模型資料

是否為多標籤: ☐

送出

任務範例資料 資料標註平台 建立新任務

資料整備流程

先至 **doccano** 資料標註平台 進行資料標註, 再建立新任務將 結果資料(csv) 上傳至任務頁面。

What is doccano?

上傳格式請參考 任務範例資料

machine_learning_task

建立任務請選擇 機器學習模型資料

每項任務請至少準備以下類型資料:

- train 訓練資料
- dev 驗證資料
- test 測試資料

上傳資料欄位, *為必要欄位:

操作簡介-資料整備模組5

AUDIENCE
TOOLKITS

說明書

資料整備任務

模型建立任務

族群貼標任務

資料整備任務

ychuang

規則資料建立測試

single-label rule_task

由 ychuang 建立

規則資料

上傳資料

下載資料

任務說明

操作

Search:

新增規則

label	rule_type	match_type
-------	-----------	------------

上傳資料

檔案：
選擇檔案 未選擇任何檔案
• 此為必需欄位。

是否覆寫資料? ☐

Close 上傳

請上傳標註資料，資料集格式
請參考 任務說明-任務範例資料
請務必包含必要欄位。

請上傳 .csv 文字檔格式
並選擇是否覆寫資料(無勾選是新增資料)

操作簡介-資料整備模組6（規則模型資料）

AUDIENCE
TOOLKITS

說明書

資料整備任務

模型建立任務

族群貼標任務

資料整備任務

規則資料建立測試

single-label rule_task

由 ychuang 建立

規則資料

上傳資料 下載資料 任務說明 操作

ychuang

新增規則

content	label	rule_type	match_type	
(如果{0,1})我是{1,2}女(?方)	男性	regex	partially	修改 刪除
我{0,1}(太太 老婆 妻子 岳母 岳父)	男性	regex	partially	修改 刪除
我{0,1}(當兵 退伍)	男性	regex	partially	修改 刪除
我[^男]{0,2}女[^男]{0,1}友	男性	regex	partially	修改 刪除
我是{1,2}男(?方)	男性	regex	partially	修改 刪除

Showing 1 to 5 of 5 entries

規則數量: 5

新增、修改 與 刪除規則資料

操作簡介-資料整備模組7 (機器學習資料)

機器學習資料測試

multi-label machine_learning_task
由 ychuang 建立
測試修改

上傳資料

Show 10 entries Search:

title	author	content	dataset_type	label	
		537 2 8/24 duncanga [問卦]會因為反中FB不敢發表在中國的炫耀文嗎538 2 8/24 duncanga [問卦]網遊、直播，儲值花錢算不算傻子	train	spam	<div>修改 刪除</div>
		momo好食堂等你來按讚>>https://momo.dm/fViMf3先到momo好食堂粉絲團按讚再到此貼文按讚+留言+tag2名好友來抽獎 張庭瓊寂寞又有了【Tefal法國特福】鈦金系列不沾刀具四件組，人人都是特級廚師。	train	spam	<div>修改 刪除</div>
		add some text data here	dev	spam	<div>修改 刪除</div>
		537 2 8/24 duncanga [問卦]會因為反中FB不敢發表在中國的炫耀文嗎538 2 8/24 duncanga [問卦]網遊、直播，儲值花錢算不算傻子？	train	spam	<div>修改 刪除</div>
		momo好食堂等你來按讚>>https://momo.dm/fViMf3先到momo好食堂粉絲團按讚再到此貼文按讚+留言+tag2名好友來抽獎 張庭瓊寂寞又有了【Tefal法國特福】鈦金系列不沾刀具四件組，人人都是特級廚師。	train	spam	<div>修改 刪除</div>

Showing 1 to 5 of 5 entries

Previous 1 Next

訓練集: 4 驗證集: 1

資料數量統計

新增、修改資料，
機器學習資料整備，不支援個別新增資料與個別標籤修改，不過可以修改標籤以外欄位資料，請在 **doccano** 標註平台控管。

操作簡介-資料整備模組8

機器學習資料測試

multi-label machine_learning_task
由 ychuang 建立
測試修改

上傳資料 下載資料 任務說明 操作 ▾

下載任務資料集(.csv)
與任務說明連結

Show 10 entries Search:

title	author	content	dataset_type	label	
		537 2 8/24 duncanga □ [問卦]會因為反中FB不敢發表在中國的炫耀文嗎538 2 8/24 duncanga □ [問卦]網遊、直播，儲值花錢算不算傻子	train	spam	修改 刪除
		momo好食堂等你來按讚>>https://momo.dm/fViMf3先到momo好食堂粉絲團按讚再到此貼文按讚+留言+tag2名好友來抽獎 張庭瓊寂寞乂有了【Tefal法國特福】鈦金系列不沾刀具四件組，人人都是特級廚師。	train	spam	修改 刪除
		add some text data here	dev	spam	修改 刪除
		537 2 8/24 duncanga □ [問卦]會因為反中FB不敢發表在中國的炫耀文嗎538 2 8/24 duncanga □ [問卦]網遊、直播，儲值花錢算不算傻子？	train	spam	修改 刪除
		momo好食堂等你來按讚>>https://momo.dm/fViMf3先到momo好食堂粉絲團按讚再到此貼文按讚+留言+tag2名好友來抽獎 張庭瓊寂寞乂有了【Tefal法國特福】鈦金系列不沾刀具四件組，人人都是特級廚師。	train	spam	修改 刪除

Showing 1 to 5 of 5 entries

Previous 1 Next

訓練集: 4 驗證集: 1

操作簡介-資料整備模組9(使用資料)

The screenshot displays the 'AUDIENCE TOOLKITS' interface. On the left sidebar, the '模型建立任務' (Model Building Task) option is highlighted with a red box. The main area shows a '任務列表' (Task List) table with columns for 'id', '模型名稱' (Model Name), and '模型類型' (Model Type). A modal window titled '建立任務' (Create Task) is open, showing fields for '任務名稱' (Task Name) with the value 'Jab', '描述與定義' (Description and Definition) with the text '請描述模型的用途', '模型類型' (Model Type) with a dropdown menu, '特徵欄位' (Feature Field) with a dropdown menu set to '內文', and '模型資料來源' (Model Data Source) with a dropdown menu. The '模型資料來源' field is highlighted with a red box. A '送出' (Submit) button is at the bottom of the modal. In the top right corner, a '建立新任務' (Create New Task) button is also highlighted with a red box.

任務列表

id	模型名稱	模型類型
51	女性_內文_關鍵字_新聞	關鍵字規則
52	女性_內文_關鍵字_先生	關鍵字規則
53	男性_內文_關鍵字_太太	關鍵字規則
55	test5	SVM
56	重啟測試	關鍵字規則
57	已婚_內文_關鍵字	關鍵字規則
58	測試資料整備	正則表達式比
59	資料整備測試2	詞彙權重模型

Showing 31 to 38 of 38 entries

建立任務

任務名稱：
Jab

描述與定義：
請描述模型的用途

模型類型：

特徵欄位：
內文

模型資料來源：

送出

模型建立流程

每種不同的模型都有其各自的建立方式，以下會說明各種類型的建立條件與做法：

- SVM
- 隨機森林
- 關鍵字規則
- 正則表達式比對
- 詞彙權重模型

監督式學習模型

此類型的模型藉由帶有標籤的文本資料，讓機器自動學習文本中的特徵，可以使用的資料準備任務為 **監督式學習模型**。

為了訓練出此類型的模型，需要提供人工標記的資料，每種標籤建議數量為至少200篇文章。

若您發現訓練效果不佳，建議檢查各標籤的定義是否明確，標記品質是否夠好。

以下為此類型的模型列表：

- SVM
- 隨機森林

什麼時候可以使用這種模型？

當你不確定資料有什麼明確的規則或關鍵字時，或是您不希望太多人工介入，讓機器自己學習

模型任務選擇資料來源 (資料整備任務)

操作簡介-資料整備模組 I O (使用資料)

The screenshot shows a web interface for 'Test Data Preparation' (測試資料整備). On the left is a blue sidebar with navigation links: '說明書' (Manual), '資料整備任務' (Data Preparation Task), '模型建立任務' (Model Building Task), and '族群貼標任務' (Group Labeling Task). The main content area has a title '測試資料整備' and a status '完成' (Completed). Below the title, it says '由於 2022年5月5日 13:50 建立。' (Created on May 5, 2022, 13:50) and '測試資料整備' (Test Data Preparation). There are three buttons at the top right: '上傳額外測試資料' (Upload additional test data), '重新訓練' (Retrain), and '操作' (Action). The interface is divided into two main sections: '模型資訊' (Model Information) and '訓練資料' (Training Data). Under '模型資訊', it shows '模型類型：正則表達式比對' (Model type: Regular expression matching). Under '訓練資料', it shows '模型資料來源: #8 規則資料建立測試' (Model data source: #8 Rule data building test), which is highlighted with a red rectangle.

說明書

資料整備任務

模型建立任務

族群貼標任務

測試資料整備

完成

由於 2022年5月5日 13:50 建立。
測試資料整備

上傳額外測試資料 重新訓練 操作 ▾

模型資訊

模型類型：正則表達式比對

訓練資料

模型資料來源: #8 規則資料建立測試

連結至資料整備任務頁面

注意事項！

- 標註平台 **doccano** 有使用 **demo** 教學，建議先行參考後再開始貼標。
- 若需要新增 **doccano** 使用者，請洽詢 **rd2 doccano** 站台維護者。
- **資料整備任務**上傳檔案有分為 覆寫 與 新增，此功能設計理念為，若機器學習資料量過大，讓使用者可以分段上傳 (**train, dev, test** 資料)，以及規則資料集可以本地端修改好一次上傳覆寫舊資料。
- **1.3 版** 站台僅支援單標籤模型，新建立任務，是否為多標籤欄位請忽略。

注意事項2

- 機器學習資料不支援單筆新增資料與修改單筆資料標籤，原因出於防呆，確保資料內容結構一致，才能成功被模型訓練任務接受(標籤修改請於doccano 平台檢核，確認無誤再上傳至 audience 站台)。
- 若要使用機器學習資料訓練模型，請資料集至少包含 train, dev, test 資料缺一不可，詳情請參考 **任務說明**。
- 對 1.3 版站台有任何疑問請聯絡 RD2 Audience 專案負責人。