Audience toolkits 1.3 release

RD2 黃彥鈞

大綱

- 產品需求
- 產品使用流程
- 各版本簡述

產品需求

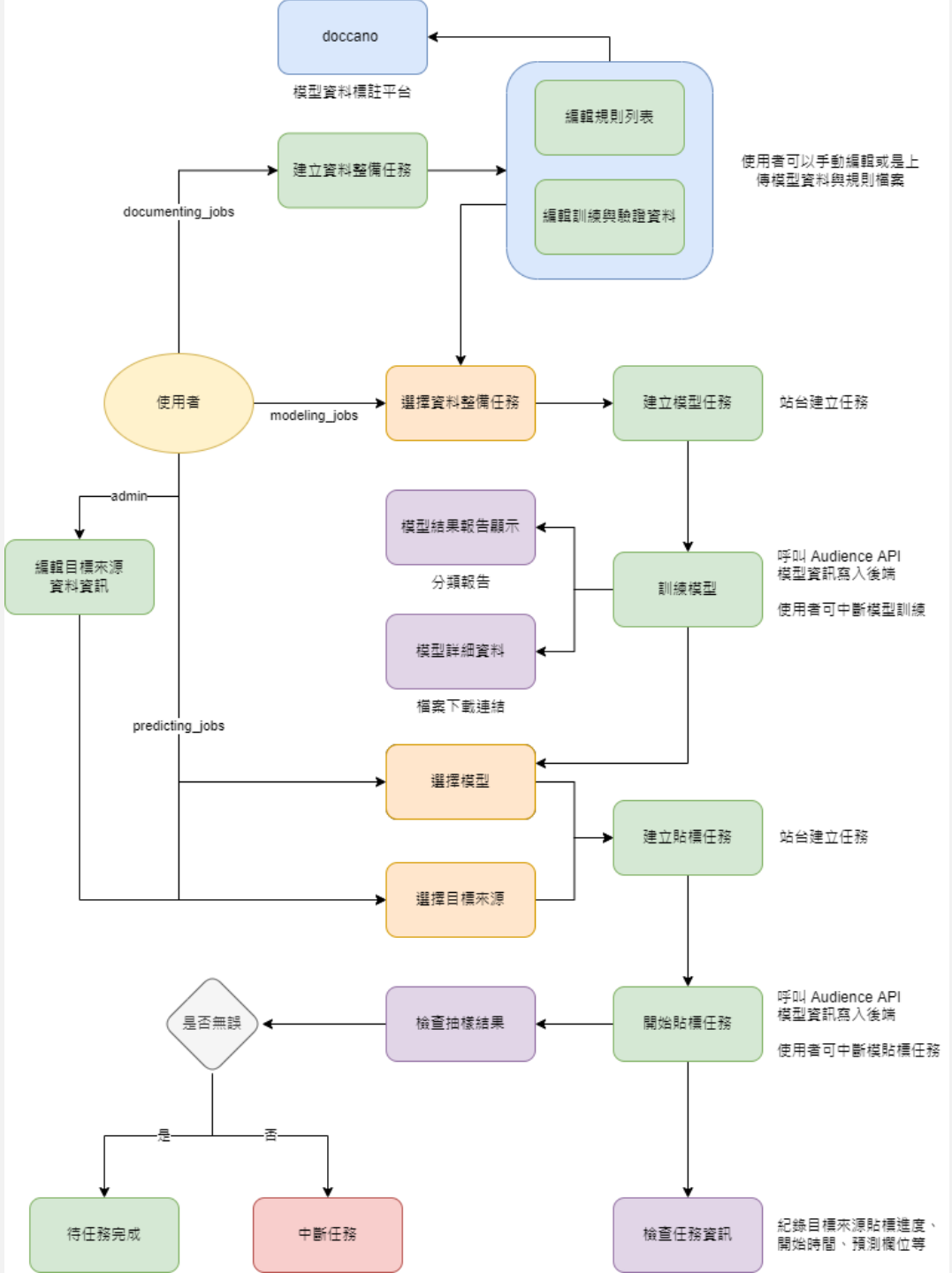
- 協助 AS1 同仁 OpView 的族群貼標所建立之自動化貼標站台
- 現行 1.3 版功能包含使用規則模型與機器學習實現文本標籤預測
- 開發產品
  - Audience Django Toolkits
  - Audience API
- 使用者可透過站台操作實現
  - 模型資料建立與維護
  - 模型資料標註 (doccano)
  - 模型訓練與維護
  - 族群貼標 ETL 任務建立與維護

產品使用流程

- 主要流程
- 資料整備任務
- 模型建立任務
- 族群貼標任務

主要流程1

- 資料整備任務
- 模型建立任務
- 族群貼標任務

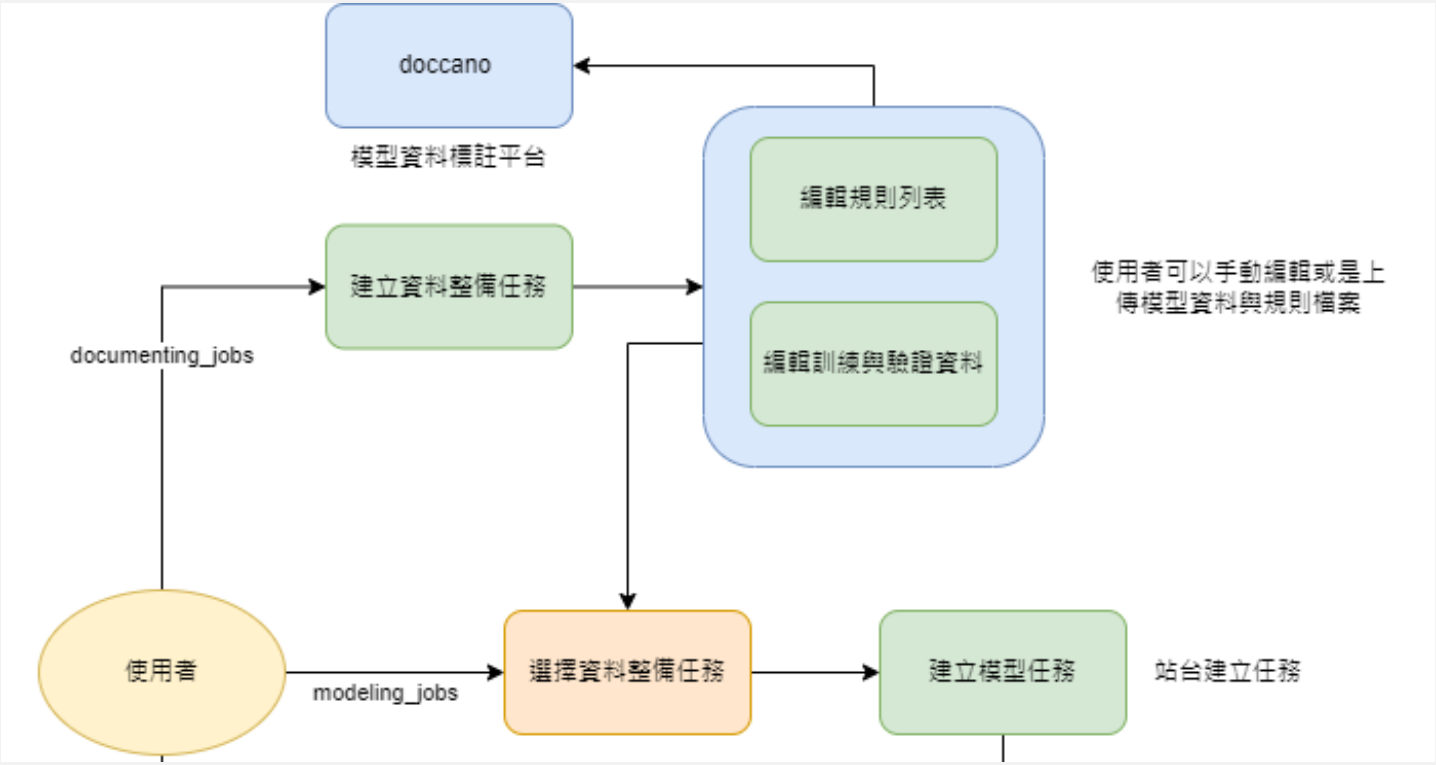


主要流程2

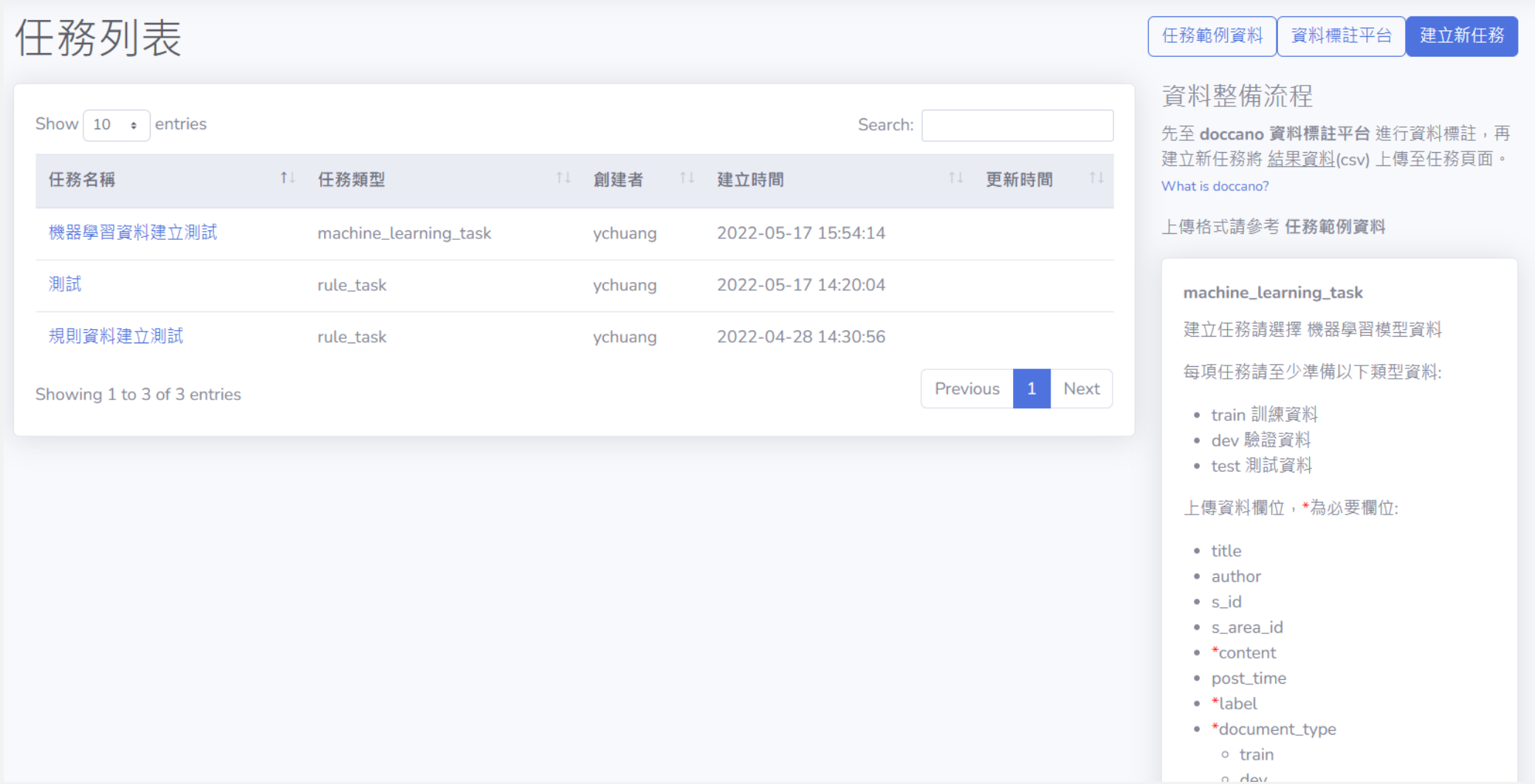

AUDIENCE
TOOLKITS
說明書
資料整備任務
模型建立任務
族群貼標任務
說明書
ychuang
Audience Toolkits 族群訓練工具包
藉由此工具，您可以從0開始訓練一個屬於你的機器學習模型，並且進一步應用至OpView的族群標籤。
什麼是Audience族群標籤
Audience，aka族群，藉由透過網路上內容作者發布的內容，判斷其背後可能代表的族群類型。這邊假設網路內容可用的資訊如發文頻道、 發文作者名稱、發文內容等資訊，會透露出一些某些族群的特有資訊。目前支援可用的欄位名稱有 標題、內容、來源、來源網站、作者。
舉個例子
在dcard討論區中，作者名稱有個固定的格式，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。

資料整備任務1

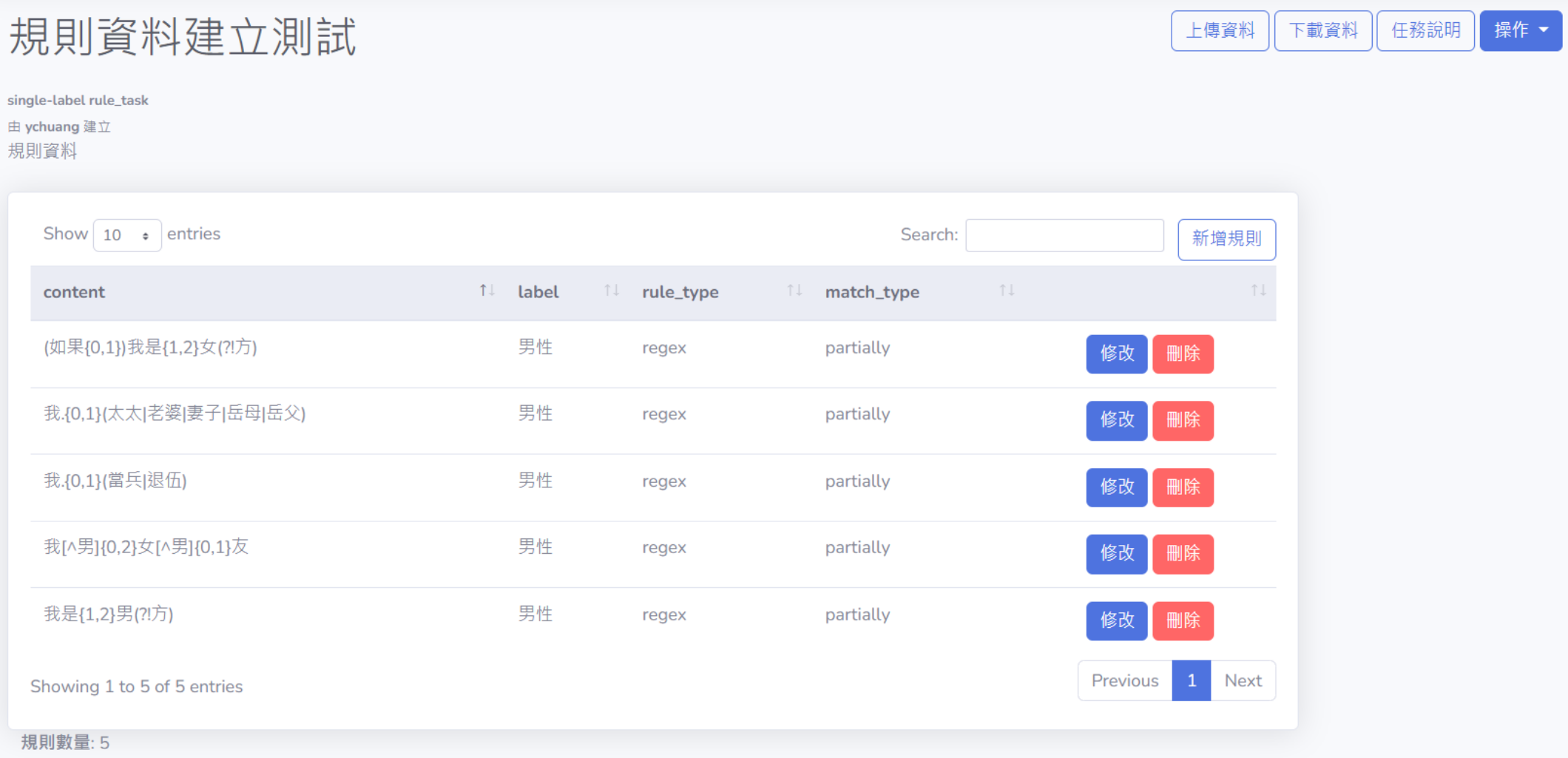
- 使用 **doccano** 進行資料快速標註 demo
- **資料整備任務**上傳模型資料與規則資料
- 使用者可在任務中新增、修改與刪除資料
- 建立好的資料整備任務可以供模型訓練任務所使用



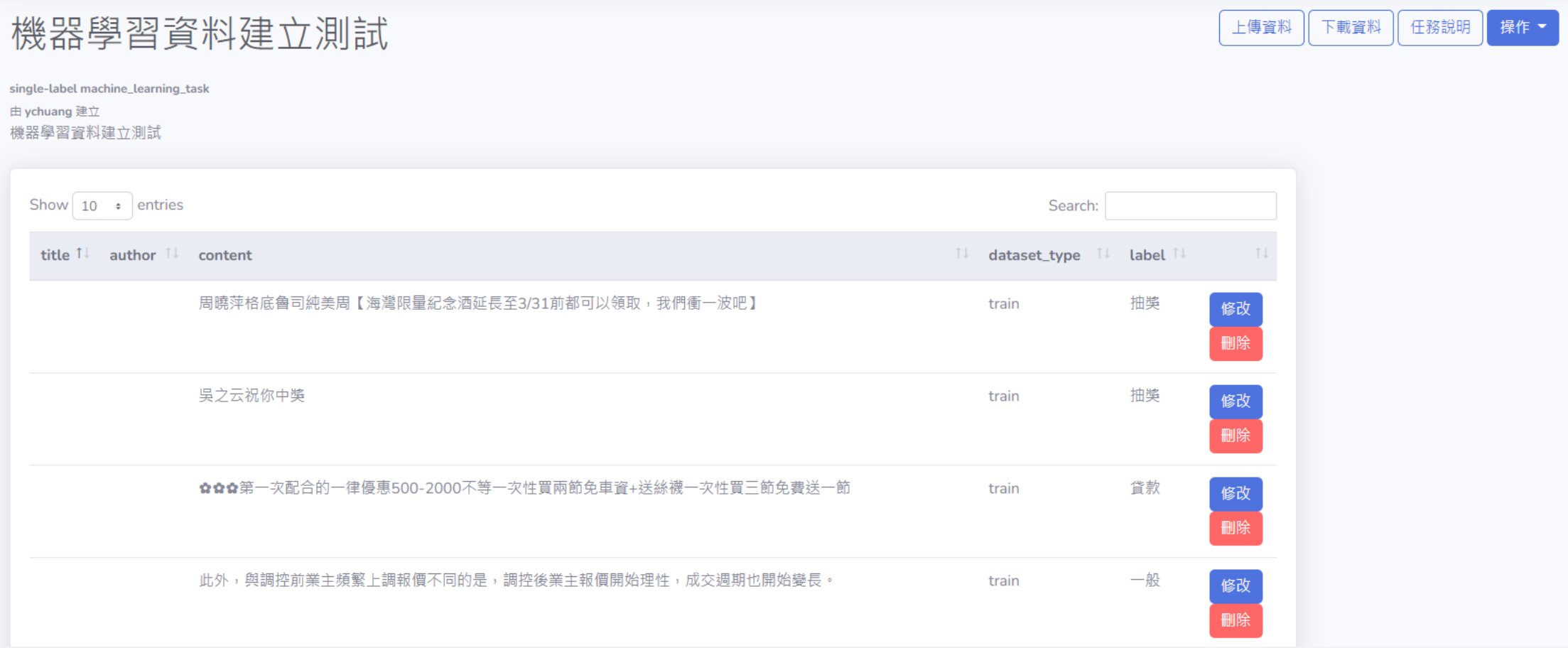
資料整備任務2


任務列表
任務範例資料
資料標註平台
建立新任務
Show 10 entries
Search:
任務名稱
任務類型
創建者
建立時間
更新時間
機器學習資料建立測試
machine_learning_task
ychuang
2022-05-17 15:54:14
測試
rule_task
ychuang
2022-05-17 14:20:04
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
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
s_id
s_area_id
*content
post_time
*label
*document_type
train

資料整備任務3

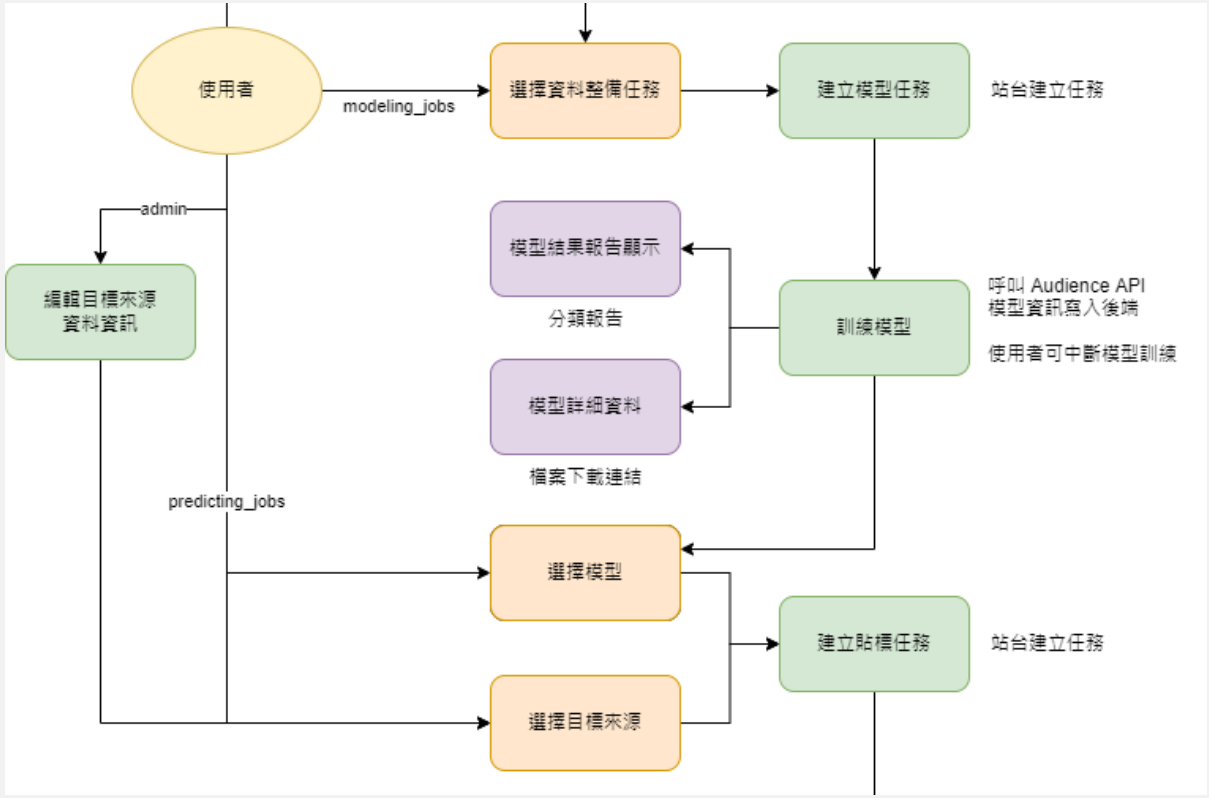

規則資料建立測試
上傳資料
下載資料
任務說明
操作
single-label rule_task
由 ychuang 建立
規則資料
Show 10 entries
Search:
新增規則
content
label
rule_type
match_type
(如果{0,1})我是{1,2}女(?!方)
男性
regex
partially
修改
刪除
我.{0,1}(太太|老婆|妻子|岳母|岳父)
男性
regex
partially
修改
刪除
我.{0,1}(當兵|退伍)
男性
regex
partially
修改
刪除
我[^男]{0,2}女[^男]{0,1}友
男性
regex
partially
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
Previous
1
Next
規則數量: 5

資料整備任務4


機器學習資料建立測試
上傳資料
下載資料
任務說明
操作
single-label machine_learning_task
由 ychuang 建立
機器學習資料建立測試
Show 10 entries
Search:
title
author
content
dataset_type
label
周曉萍格底魯司純美周【海灣限量紀念酒延長至3/31前都可以領取，我們衝一波吧】
train
抽獎
修改
刪除
吳之云祝你中獎
train
抽獎
修改
刪除
✿✿✿第一次配合的一律優惠500-2000不等一次性買兩節免車資+送絲襪一次性買三節免費送一節
train
貸款
修改
刪除
此外，與調控前業主頻繁上調報價不同的是，調控後業主報價開始理性，成交週期也開始變長。
train
一般
修改
刪除

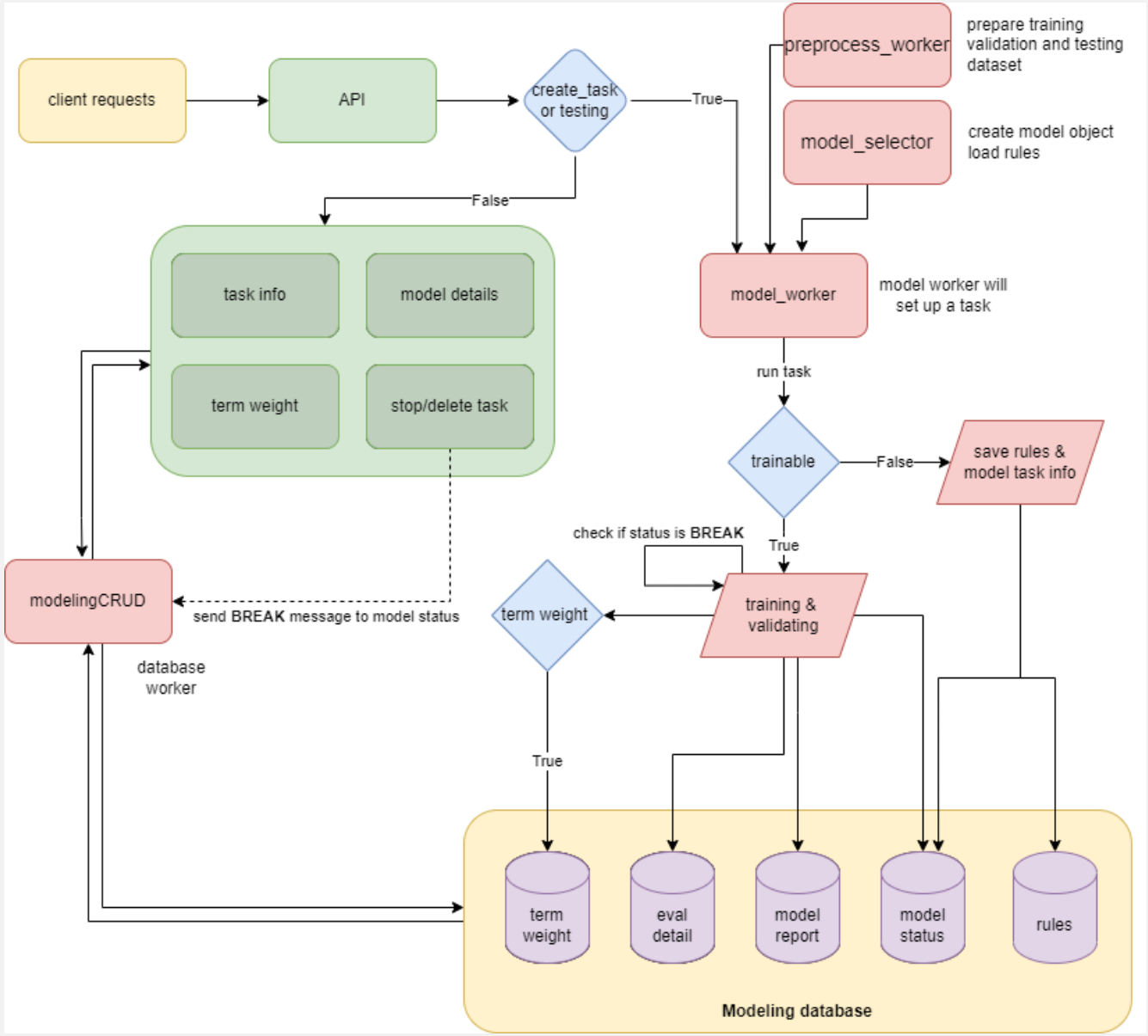
模型建立任務1

- 透過資料整備任務的模型或規則資料來訓練模型
- 現行模型種類
  - 關鍵字模型
  - 正則表達式模型
  - 隨機森林模型
  - 詞彙權重模型

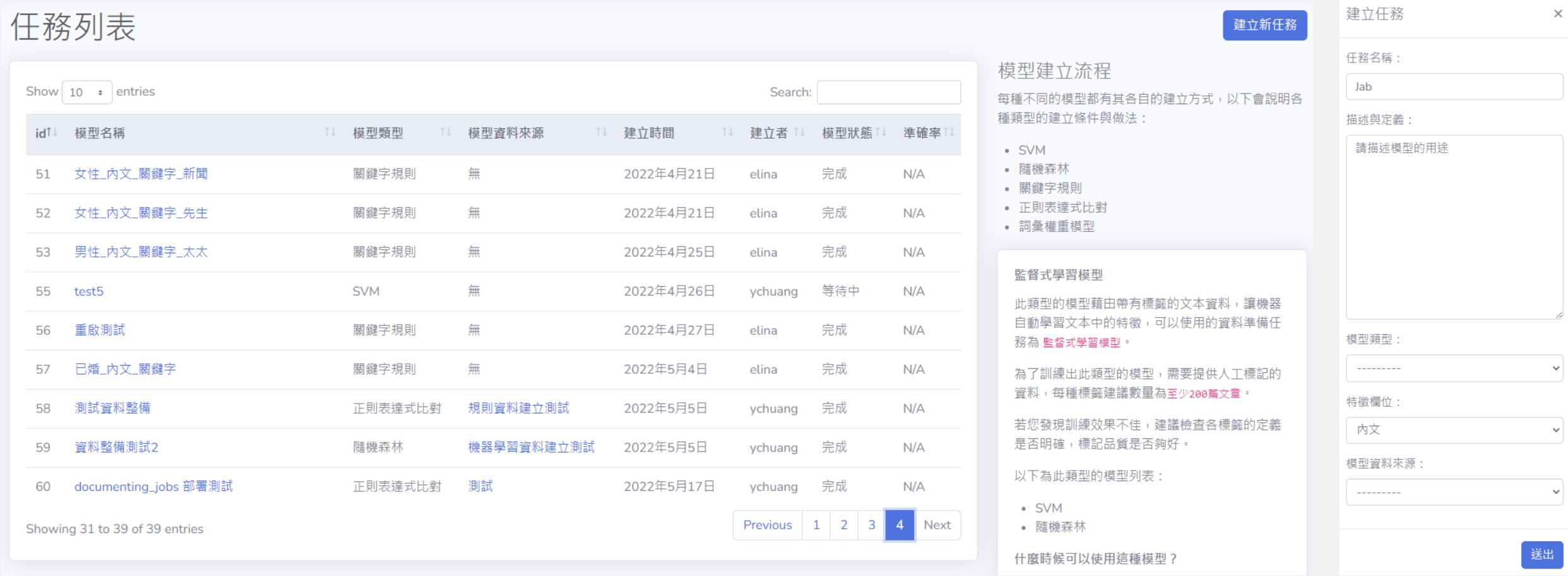


模型建立任務2

- Audience modeling api 流程



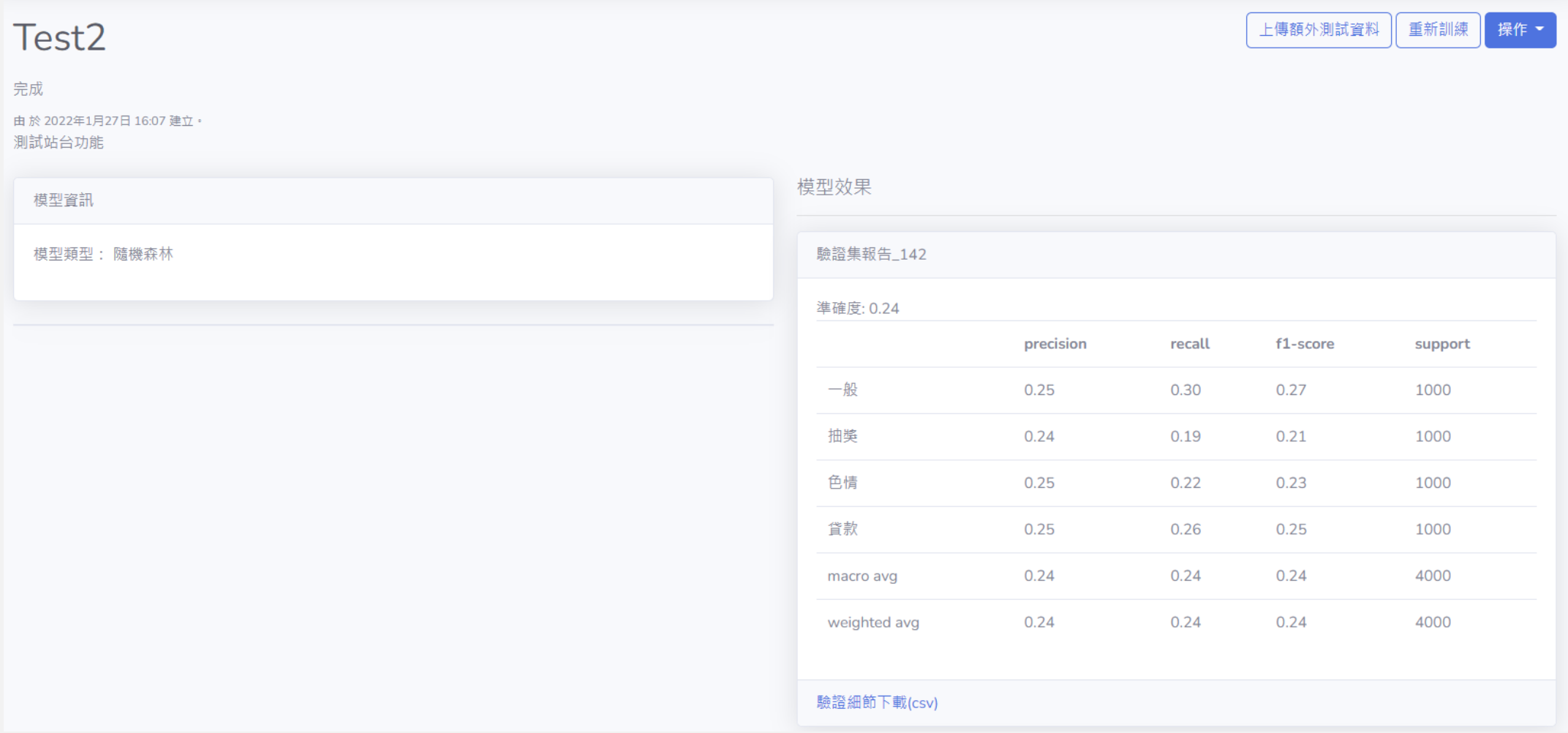
模型建立任務3


任務列表
建立新任務
Show 10 entries
Search:
id
模型名稱
模型類型
模型資料來源
建立時間
建立者
模型狀態
準確率
51 女性_內文_關鍵字_新聞 關鍵字規則 無 2022年4月21日 elina 完成 N/A
52 女性_內文_關鍵字_先生 關鍵字規則 無 2022年4月21日 elina 完成 N/A
53 男性_內文_關鍵字_太太 關鍵字規則 無 2022年4月25日 elina 完成 N/A
55 test5 SVM 無 2022年4月26日 ychuang 等待中 N/A
56 重啟測試 關鍵字規則 無 2022年4月27日 elina 完成 N/A
57 已婚_內文_關鍵字 關鍵字規則 無 2022年5月4日 elina 完成 N/A
58 測試資料整備 正則表達式比對 規則資料建立測試 2022年5月5日 ychuang 完成 N/A
59 資料整備測試2 隨機森林 機器學習資料建立測試 2022年5月5日 ychuang 完成 N/A
60 documenting_jobs 部署測試 正則表達式比對 測試 2022年5月17日 ychuang 完成 N/A
Showing 31 to 39 of 39 entries
Previous 1 2 3 4 Next
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？
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：

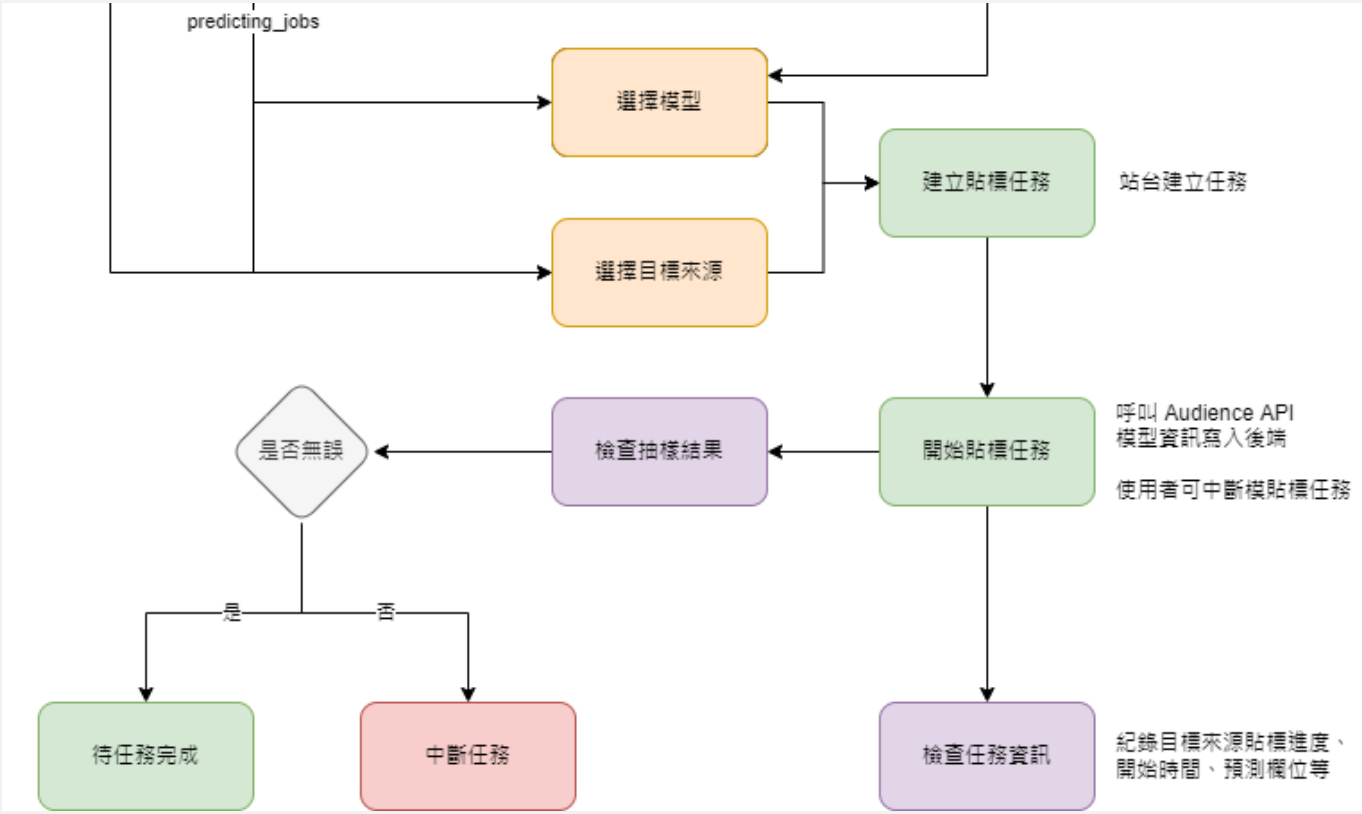
送出

模型建立任務4


Test2
上傳額外測試資料
重新訓練
操作
完成
由 於 2022年1月27日 16:07 建立。
測試站台功能
模型資訊
模型類型： 隨機森林
模型效果
驗證集報告_142
準確度: 0.24
precision recall f1-score support
一般 0.25 0.30 0.27 1000
抽獎 0.24 0.19 0.21 1000
色情 0.25 0.22 0.23 1000
貸款 0.25 0.26 0.25 1000
macro avg 0.24 0.24 0.24 4000
weighted avg 0.24 0.24 0.24 4000
驗證細節下載(csv)

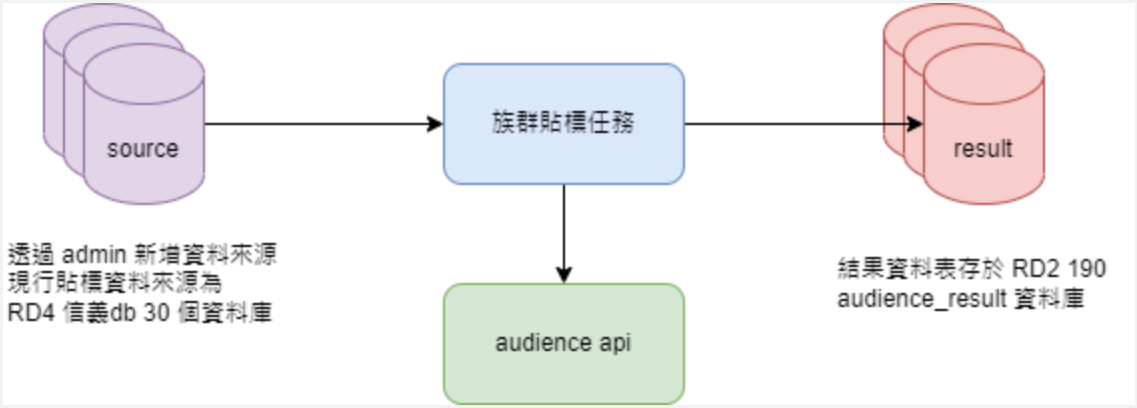
族群貼標任務1

- 選擇欲應用之模型，和要預測的資料來源與時間範圍
- 開始貼標任務後，使用者可任務進行中查看已貼標之抽樣結果資料
- 透過檢查抽樣資料，可以決定是否中斷任務或待其繼續完成貼標
- 查看任務資訊以獲得任務狀態與結果



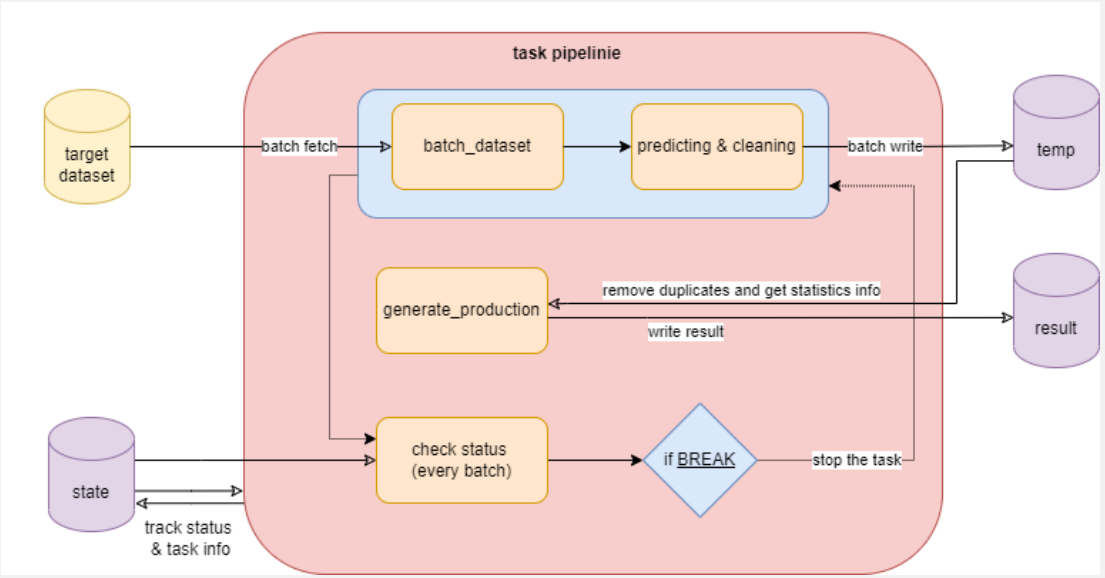
族群貼標任務2

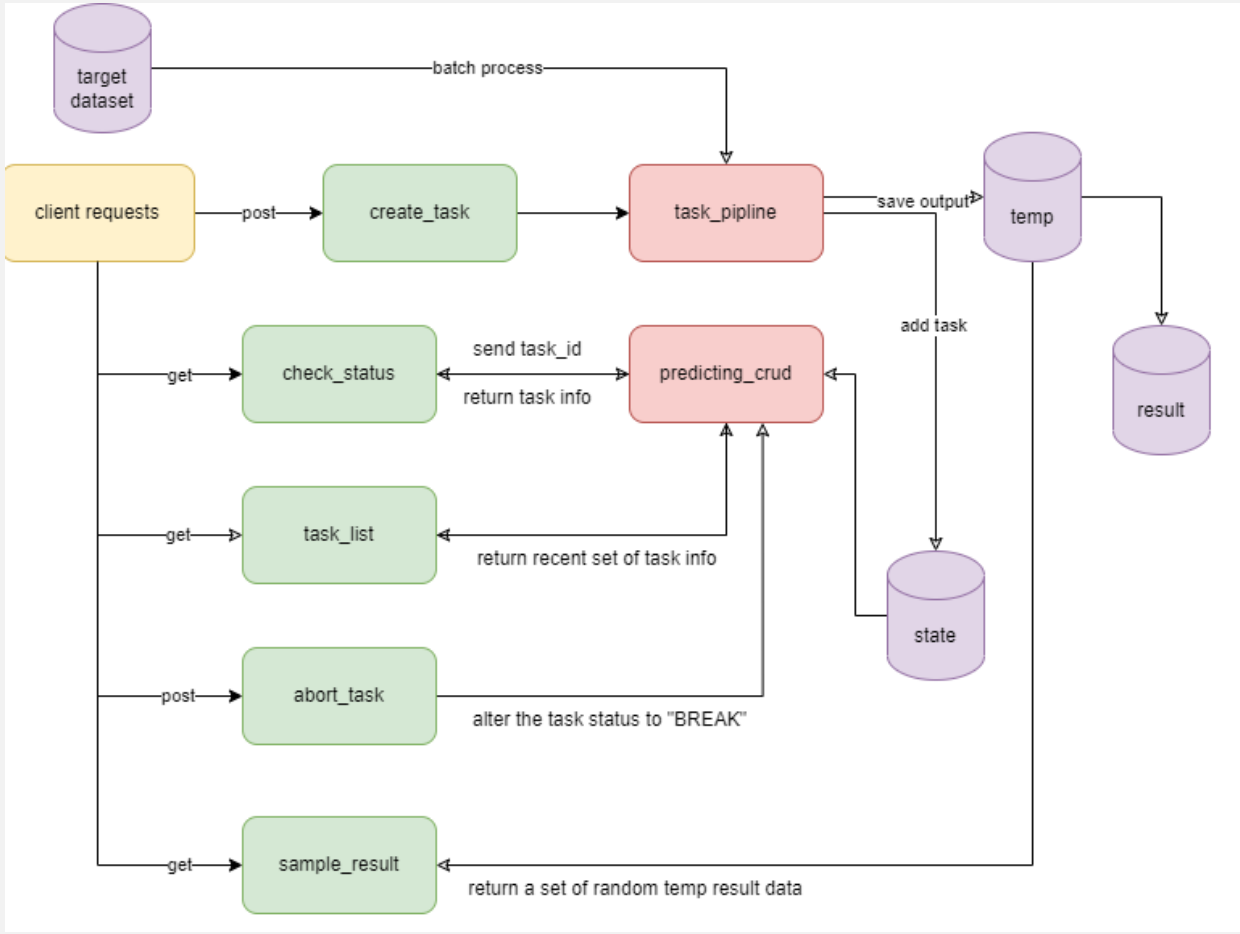
- 預測資料來源目前設定為 RD4 信義資料庫中 30 個來源，涵蓋臉書、滴卡、新聞社群與評論等資料。
- 使用者可以選擇欲貼標之來源與時間範圍(2019-2022) 來執行貼標。
- 貼標結果儲存於 RD2 190 資料庫 audience_result



族群貼標任務3

- Audience predicting api 流程

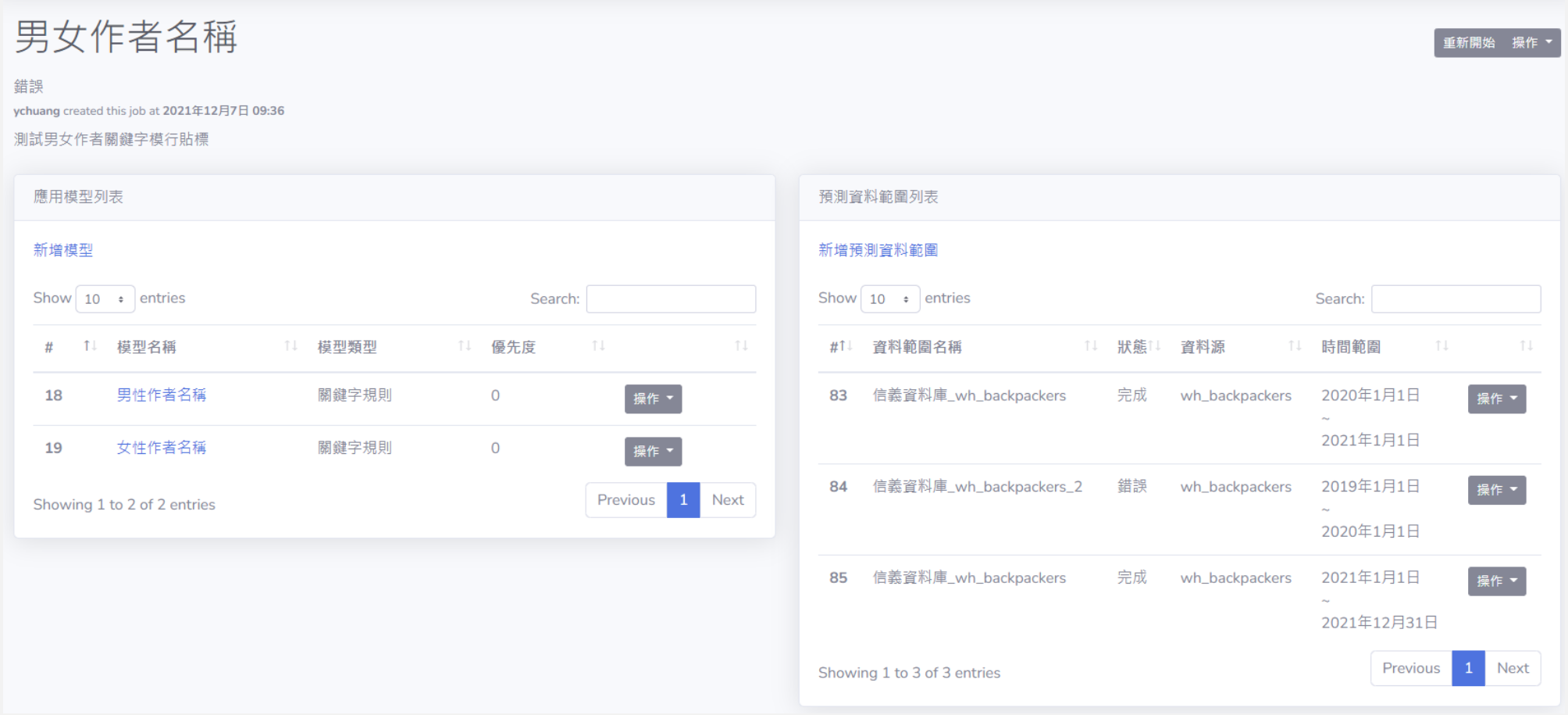




族群貼標任務4


任務列表
建立新任務
Show 10 entries
Search:
#
任務名稱
任務描述
應用模型數
預測範圍數
建立者
建立時間
狀態
34 男性_內文_正規表達式_特定資料庫 特定資料庫 3 35 elina 2022年4月20日 11:57 完成
35 男性_內文_關鍵字_新聞 男性_內文_關鍵字_新聞 1 8 elina 2022年4月20日 14:11 完成
36 女性_內文_特定資料庫 女性_內文_特定資料庫 3 27 elina 2022年4月21日 14:12 完成
37 女性_內文_關鍵字_新聞 女性_內文_關鍵字_新聞 1 8 elina 2022年4月22日 09:57 完成
38 已婚_內文_關鍵字 已婚_內文_關鍵字 1 14 elina 2022年5月5日 13:44 完成
族群貼標流程
設定好欲此用的模型列表，與目標資料範圍，系統就會排定貼標任務。

族群貼標任務5


男女作者名稱
重新開始
操作
錯誤
ychuang created this job at 2021年12月7日 09:36
測試男女作者關鍵字模行貼標
應用模型列表
新增模型
Show 10 entries
Search:
模型名稱 模型類型 優先度
18 男性作者名稱 關鍵字規則 0 操作
19 女性作者名稱 關鍵字規則 0 操作
Showing 1 to 2 of 2 entries
Previous 1 Next
預測資料範圍列表
新增預測資料範圍
Show 10 entries
Search:
資料範圍名稱 狀態 資料源 時間範圍
83 信義資料庫_wh_backpackers 完成 wh_backpackers 2020年1月1日 ~ 2021年1月1日 操作
84 信義資料庫_wh_backpackers_2 錯誤 wh_backpackers 2019年1月1日 ~ 2020年1月1日 操作
85 信義資料庫_wh_backpackers 完成 wh_backpackers 2021年1月1日 ~ 2021年12月31日 操作
Showing 1 to 3 of 3 entries
Previous 1 Next

族群貼標任務6

任務ID	來源	貼標進度	任務狀態	開始時間	應用模型	預測對象	貼標率	檢查點	錯誤訊息
0823d822afd211ecb688d45d6456a14d	wh_backpackers	SUCCESS	finish	2022-03-30 10:35:16	KEYWORD_MODEL	author	28.75	None	None
bbaf84eaafe111ecb688d45d6456a14d	wh_backpackers	FAILURE	None	2022-03-30 12:27:39	KEYWORD_MODEL	author	None	2019-07-18 12:00:00	missing), unterminated subpattern at position 3
d22c54fa57e711ec880f04ea56825bad	wh_backpackers	SUCCESS	no_data	2021-12-08 13:29:32	keyword_model	author	None	None	None

族群貼標任務7

任務ID	來源作者	標籤	爬取時間	來源ID	命中內容	命中規則
0823d822afd211ecb688d45d6456a14d	WH_F0003_jansony	/male	2020-01-01T02:07:00	WH_F0003	jansony	[[('男性', [[('janson', 0)]])]]
0823d822afd211ecb688d45d6456a14d	WH_F0003_sarah5365	/female	2020-01-01T03:22:00	WH_F0003	sarah5365	[[('女性', [[('sara', 0), ('sarah', 0)]])]]
0823d822afd211ecb688d45d6456a14d	WH_F0003_l4783grace	/female	2020-01-01T09:17:00	WH_F0003	l4783grace	[[('女性', [[('grace', 5)]])]]
0823d822afd211ecb688d45d6456a14d	WH_F0003_karlliy	/male	2020-01-01T10:12:00	WH_F0003	karlliy	[[('男性', [[('karl', 0)]])]]
0823d822afd211ecb688d45d6456a14d	WH_F0003_annajy	/female	2020-01-01T13:17:00	WH_F0003	annajy	[[('女性', [[('anna', 0)]])]]
0823d822afd211ecb688d45d6456a14d	WH_F0003_hansw	/male	2020-01-01T14:22:00	WH_F0003	hansw	[[('男性', [[('hans', 0)]])]]
0823d822afd211ecb688d45d6456a14d	WH_F0003_violet yang	/female	2020-01-01T15:18:00	WH_F0003	violet yang	[[('女性', [[('violet', 0)]])]]
0823d822afd211ecb688d45d6456a14d	WH_F0003_liuashley	/female	2020-01-01T15:26:00	WH_F0003	liuashley	[[('女性', [[('ashley', 3)]])]]
0823d822afd211ecb688d45d6456a14d	WH_F0003_samchow0226	/male	2020-01-01T15:35:00	WH_F0003	samchow0226	[[('男性', [[('sam', 0)]])]]
0823d822afd211ecb688d45d6456a14d	WH_F0003_Pan Cindy	/female	2020-01-01T17:29:00	WH_F0003	Pan Cindy	[[('女性', [[('cindy', 4)]])]]
0823d822afd211ecb688d45d6456a14d	WH_F0003_Rogerc k	/male	2020-01-01T18:23:00	WH_F0003	Rogerc k	[[('男性', [[('roger', 0)]])]]
0823d822afd211ecb688d45d6456a14d	WH_F0003_sharonmak1209	/female	2020-01-01T19:20:00	WH_F0003	sharonmak1209	[[('女性', [[('sharon', 0)]])]]
0823d822afd211ecb688d45d6456a14d	WH_F0003_Chris_edward	/male	2020-01-01T20:00:00	WH_F0003	Chris_edward	[[('男性', [[('chris', 0), ('edward', 6), ('ward', 8)]])]]
0823d822afd211ecb688d45d6456a14d	WH_F0003_totopeggy	/female	2020-01-01T22:00:00	WH_F0003	totopeggy	[[('女性', [[('peggy', 4)]])]]

各版本簡述

Audience API

版號	說明
2.0	開發貼標任務 API
2.1	優化貼標 pipline 修改 API 格式
2.2	開發模型任務 API
2.3	開發前處理模組
2.4	開發資料整備任務 API

Audience Django Toolkits

版號	說明
1.1	重構 predicting_jobs
1.2	重構 modeling_jobs
1.3	開發 documenting_jobs

Audience Django Toolkits 1.4
Audience API 2.5
預計新增多標籤模型訓練，提供深度學習版本模型，以及設計模型訓練與效能優化的 client 界面