

Haystack Annotation Tool 사용법

최종 플젝 화이팅 🔥

2022.05.23 한나연

Haystack Annotation Tool ???



장점

- 하나의 passage에 최대 20개 질문 생성 가능
- 라벨링한 데이터를 SQuAD format으로 제공
- Standard question의 경우 no_answer로 태깅 가능

단점:

- 한 질문에 대해 여러개의 답변 라벨링 불가능

Haystack is an end-to-end framework that enables you to build powerful and production-ready pipelines for different search use cases.

간단한 QA pipeline부터 DPR 훈련까지 다양한 기능(?)을 제공하는 NLP, 검색관련 오픈소스 framework 입니다.

저는 annotation tool을 사용해보았습니다~

데이터 업로드 하기

The screenshot shows the Haystack annotation tool interface. The top navigation bar includes the Haystack logo, user email (anna4229@naver.com), and a menu with options: All Projects, Documents, Questions, Import, Export Labels, and User management. The main content area displays a table of projects. A 'Create project' dialog box is open, showing the following fields and buttons:

- 1**: 'Create project' button in the top right corner of the main area.
- 2**: 'Name' input field containing 'sample'.
- 3**: 'Create project' button at the bottom of the dialog box.

Nº	Title	Annotation mode	Created	Updated	Actions
1	sample	ULT	23-05-2022	23-05-2022	→ 🗑
2	완주 sample	ULT	20-05-2022	20-05-2022	→ 🗑
3	완주 train data	ULT	20-05-2022	20-05-2022	→ 🗑
4	sanple	ULT	19-05-2022	19-05-2022	→ 🗑

- 일단 회원가입 후 로그인을 진행합니다. (데이터를 업로드할 사람 한 명만 가입하면 됩니다)
- 사진의 1, 2, 3 순서에 따라 Create Project 버튼을 눌러 작업할 공간을 생성합니다.

데이터 업로드 하기

The screenshot shows the Haystack annotation tool interface. The top navigation bar includes the Haystack logo, the user email 'anna4229@naver.com', and a profile icon. Below the navigation bar, the 'Import' menu is highlighted with a red box and the number 4. A dropdown menu is open, showing 'Documents' and 'Questions', with 'Documents' highlighted by a red box and the number 5. In the 'Import documents' section, the 'CSV Batch Upload' button is highlighted with a red box and the number 6. The main area features a large dashed box with a file icon and a blue box on the right containing instructions: 'Click or drag csv file to this area to upload documents' and a schema example for 'document_text'.

- Import 탭에서 Documents에 들어가 라벨링 작업할 문서를 업로드합니다.
 - .txt, .csv 파일 모두 가능합니다.
 - 저는 csv 파일을 업로드했습니다.

데이터 업로드 하기

	document_identifier	context	document_text
1	100001	의석을 정돈하여 주시기 ...	@제207회 완주군의회(임시회) 제 1 차 본회의회의록...
2	100002	의사팀장 수고하셨습니다...	@제207회 완주군의회(임시회) 제 1 차 본회의회의록...
3	100003	다음은 의사일정 제2항 ...	@제207회 완주군의회(임시회) 제 1 차 본회의회의록...
4	100004	다음은 의사일정 제3항 ...	@제207회 완주군의회(임시회) 제 1 차 본회의회의록...
5	100101	의석을 정돈하여 주시기 ...	@제207회 완주군의회(임시회) 제 1 차 본회의회의록...

Documents

Id	Text	Status	Labeler no.	Created
59197 1	0:"의석을 정돈하여 주시기 바랍니다. 성원이 되었으므로 제207회 완주군의회 임시회 제1차 본회의 개의를 선포합니다. 먼...	NEW	2404	19-05-2022
59197 2	0:"의사팀장 수고하셨습니다. 먼저 의사일정 제1항 제207회 완주군의회 임시회 회기 결정의 건을 상정합니다. 제207회 완...	NEW	2404	19-05-2022
59200 3	{AGENDA_1': {0:"의석을 정돈하여 주시기 바랍니다. 성원이 되었으므로 제207회 완주군의회 임시회 제1차 본회의 개의...	NEW	2404	19-05-2022
59200 4	{AGENDA_2': {0:"의사팀장 수고하셨습니다. 먼저 의사일정 제1항 제207회 완주군의회 임시회 회기 결정의 건을 상정합...	NEW	2404	19-05-2022

- csv파일을 업로드하는 경우에는 텍스트가 들어있는 컬럼명을 document_text로 해야 어노테이션 툴에서 잘 인식합니다.

- 즉, 다른 컬럼이 csv파일에 있어도 document_text 컬럼 데이터만 인식합니다.

- document_identifier 컬럼은 라벨링 후 export할 때 사용될 수 있다는데, 일단 문서를 업로드하면 맘대로 id가 부여되어서 언제 어떻게 사용되는지 저도 잘 모르겠습니다;; (다행인건 id값이 1씩 커져서 그냥 사용중입니다)

- 문서가 크면 시간이 조금 걸리는데, 잘 업로드 되었다면 오른쪽 사진과 같습니다.

Standard question 만들기

1









sample Documents Questions Import Export Labels User management

sample / Questions

Standard questions

Show document level questions

2 Add question

Question text	Category	Created	Type	Updated	Actions
<이벤트>의 기간은?	A	21-05-2022	globally available	21-05-2022	 
이번 회의의 결정 사항은 무엇인가?	B	19-05-2022	globally available	19-05-2022	 
다음 회의는 언제인가?	A	19-05-2022	globally available	19-05-2022	 
회의의 주제는 무엇인가?	A	19-05-2022	globally available	19-05-2022	 

- Questions 탭에서 Add question 버튼을 눌러 질문과 질문의 유형을 추가합니다.
- standard question을 생성하면 라벨링 시 기본 질문으로 뜨기 때문에, 질문이 거의 고정되어 있는 경우 미리 정의하면 편리합니다.
 - But, 라벨링 시 standard question을 변경할 수 없습니다.


라벨링 작업자 추가하기

The screenshot displays the 'User management' section of a software interface. At the top, a dark blue navigation bar contains links for 'All Projects', 'Documents', 'Questions', 'Import', 'Export Labels', and 'User management' (highlighted with a red box). Below this, a breadcrumb trail shows 'User management'. The main content area features a table with columns 'No', 'First name', 'Last name', and 'Actions'. The table lists four users: 1. 성진 하, 2. 재학 이, 3. 태일 김, 4. 찬국 문. Each user has a gear icon in the 'Actions' column, with the first one highlighted by a blue box. A 'Create new user' button is located in the top right corner of the main area (highlighted with a red box). A 'Create user' modal form is open in the center, containing fields for 'First name' (peppa), 'Last name' (pig), and 'Email' (peppapig@gmail.com, highlighted with a red box). A 'Create user' button is at the bottom of the modal.

No	First name	Last name	Actions
1	성진	하	
2	재학	이	
3	태일	김	
4	찬국	문	

- User management >> Create new user >> 이름, 이메일 입력
 - 이메일만 제대로 입력하면 됩니다.
- 깃허브 프로젝트 초대하듯이 invitation같은거 없고, 그냥 방금 설정한 이메일&비번으로 로그인합니다.
- 파란색 네모 안의 설정을 눌러 각 user의 비밀번호를 변경할 수 있습니다.
 - 추가된 user는 모든 프로젝트에 접근할 수 있습니다.

라벨링 시작하기

 **Haystack**
annotation tool

anna4229@naver.com

All Projects

Documents

Questions







Import

Export Labels

User management

Projects

Create project

Nº	Title	Annotation mode	Created	Updated	Actions
1	완주 sample	DEFAULT	20-05-2022	20-05-2022	 
2	완주 train data	DEFAULT	20-05-2022	20-05-2022	 
3	sanple	DEFAULT	19-05-2022	19-05-2022	 

문서가 들어있는 프로젝트로 들어갑니다.

라벨링 시작하기

🏠 / 완주 train data

Documents

Id	Text	Status	Labeler no.	Created	Updated	Actions
60347 1	@제207회 완주군의회(임시회) 제 1 차 본회의회의록@ 의석을 정돈하여 주시기 바랍니다. 성원이 되었으므로 제207회 ...	NEW	2404	20-05-2022	20-05-2022	 
60347 2	@제207회 완주군의회(임시회) 제 1 차 본회의회의록@ 의사팀장 수고하셨습니다. 먼저 의사일정 제1항 제207회 완주군...	NEW	2404	20-05-2022	20-05-2022	 

라벨링할 문서를 선택합니다.

문서에 대답이 있는 경우

MARK DOCUMENT AS DONE ☐

SHOW LABELS OF ALL USERS ☐

Questions

U

No questions prepared yet. Please use the Question tab to add questions for your project.

1

ADD CUSTOM QUESTION

Annotation Document

Search

@제207회 완주군의회(임시회) 제 1 차 본회의회의록@

의석을 정돈하여 주시기 바랍니다. 성원이 되었으므로 제207회 완주군의회 임시회 제1차 본회의 개의를 선포합니다. 먼저 의사팀장으로부터 의회 관련 사항에 대한 보고가 있겠습니다. 의사팀장은 보고하여 주시기 바랍니다.

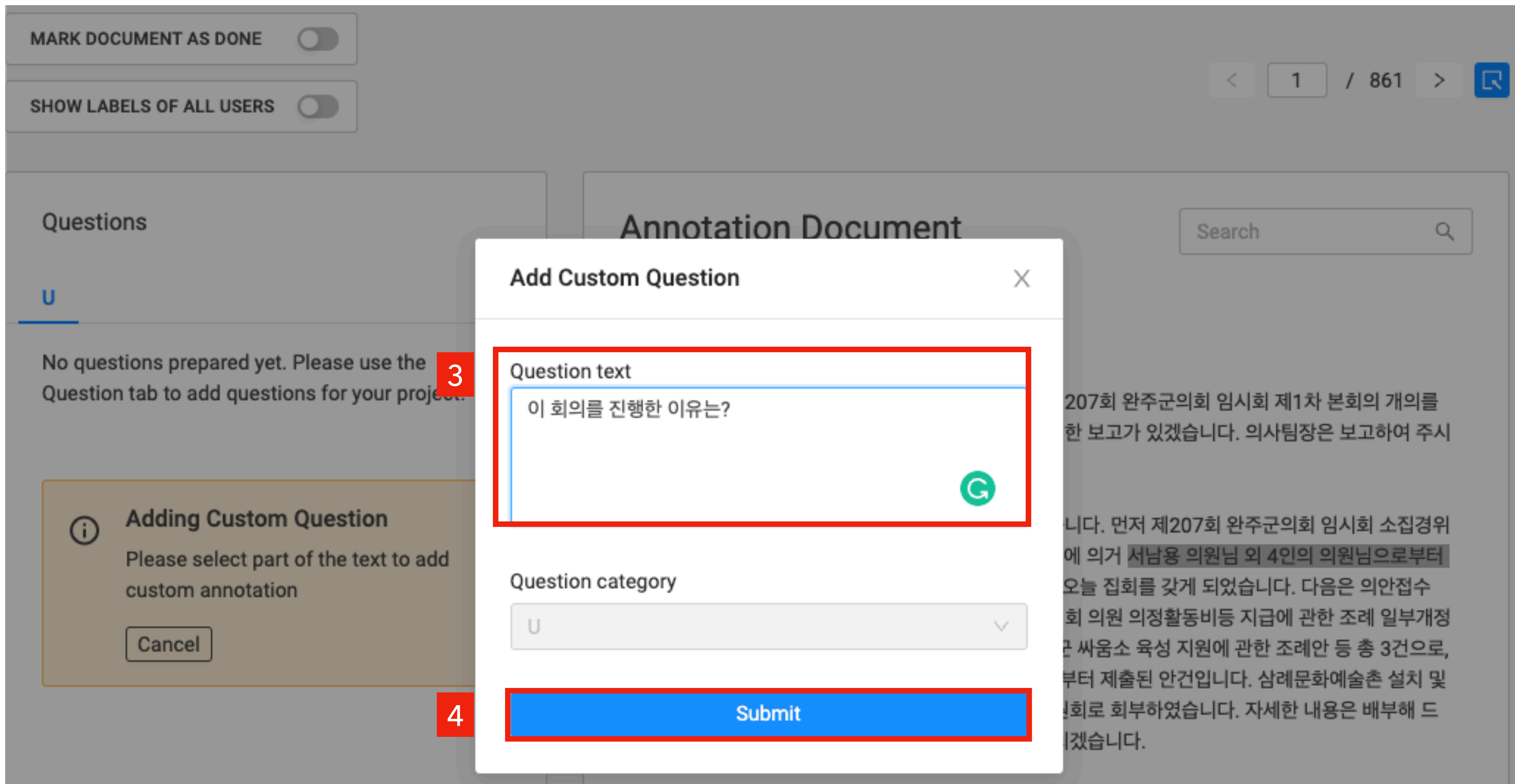
2

의사팀장 이은미 입니다. 의회관련 사항을 보고 드리겠습니다. 먼저 제207회 완주군의회 임시회 소집경위를 보고 드리겠습니다. 지방자치법 제45조 제2항의 규정에 의거 서남용 의원님 외 4인의 의원님으로부터 임시회 집회 요구가 있어 지난 8월6일 집회 공고를 하고 오늘 집회를 갖게 되었습니다. 다음은 의안접수 및 회부사항입니다. 먼저 의원발의 안건입니다. 완주군의회 의원 의정활동비등 지급에 관한 조례 일부개정조례안, 완주군 향토문화유산 보호 및 관리 조례안, 완주군 싸움소 육성 지원에 관한 조례안 등 총 3건으로, 해당 상임위원회로 회부 하였습니다. 다음은 완주군수로부터 제출된 안건입니다. 삼례문화예술촌 설치 및 운영조례 일부개정 조례안 등 총 40건으로 해당 상임위원회로 회부하였습니다. 자세한 내용은 배부해 드린 유인물을 참조하여 주시기 바랍니다. 이상 보고를 마치겠습니다.

기본적으로 설정되어있는 질문은 없습니다. 가이드라인의 질문 유형과 예시를 참고해 질문을 생성합니다.

1) ADD CUSTOM QUESTION 버튼을 누릅니다.

2) 마우스로 드래그해 정답을 먼저 선택합니다.



3) 질문을 작성합니다.

4) Submit 버튼을 눌러 완료합니다.

MARK DOCUMENT AS DONE

SHOW LABELS OF ALL USERS

< 1 / 861 >

Questions

U

5

● 이 회의를 진행한 이유는?

1

ADD CUSTOM QUESTION

Annotation Document

Search

@제207회 완주군의회(임시회) 제 1 차 본회의회의록@

5

의석을 정돈하여 주시기 바랍니다. 성원이 되었으므로 제207회 완주군의회 임시회 제1차 본회의 개의를 선포합니다. 먼저 의사팀장으로부터 의회 관련 사항에 대한 보고가 있겠습니다. 의사팀장은 보고하여 주시기 바랍니다.

의사팀장 이은미 입니다. 의회관련 사항을 보고 드리겠습니다. 먼저 제207회 완주군의회 임시회 소집경위를 보고 드리겠습니다. 지방자치법 제45조 제2항의 규정에 의거 서남용 의원님 외 4인의 의원님으로부터 임시회 집회 요구가 있어 지난 8월6일 집회 공고를 하고 오늘 집회를 갖게 되었습니다. 다음은 의안접수 및 회부사항입니다. 먼저 의원발의 안건입니다. 완주군의회 의원 의정활동비등 지급에 관한 조례 일부개정조례안, 완주군 향토문화유산 보호 및 관리 조례안, 완주군 싸움소 육성 지원에 관한 조례안 등 총 3건으로, 해당 상임위원회로 회부 하였습니다. 다음은 완주군수로부터 제출된 안건입니다. 삼례문화예술촌 설치 및 운영조례 일부개정 조례안 등 총 40건으로 해당 상임위원회로 회부하였습니다. 자세한 내용은 배부해 드린 유인물을 참조하여 주시기 바랍니다. 이상 보고를 마치겠습니다.

5) 생성한 질문과 답변을 확인합니다.

6) 동일한 문서에 질문을 추가하고 싶은 경우, ADD CUSTOM QUESTION 버튼을 눌러 1~5과정을 반복합니다.

MARK DOCUMENT AS DONE

SHOW LABELS OF ALL USERS

< 1 / 861 >

Questions

U

이 회의를 진행한 이유는?

1

완주군 싸움소 육성 지원에 관한 조례안을 다룬 회의는?

2

ADD CUSTOM QUESTION

Annotation Document

Search

@제207회 완주군의회(임시회) 제 1 차 본회의회의록@

의석을 정돈하여 주시기 바랍니다. 성원이 되었으므로 제207회 완주군의회 임시회 제1차 본회의 개의를 선포합니다. 먼저 의사팀장으로부터 의회 관련 사항에 대한 보고가 있겠습니다. 의사팀장은 보고하여 주시기 바랍니다.

의사팀장 이은미 입니다. 의회관련 사항을 보고 드리겠습니다. 먼저 제207회 완주군의회 임시회 소집경위를 보고 드리겠습니다. 지방자치법 제45조 제2항의 규정에 의거 서남용 의원님 외 4인의 의원님으로부터 임시회 집회 요구가 있어 지난 8월6일 집회 공고를 하고 오늘 집회를 갖게 되었습니다. 다음은 의안접수 및 회부사항입니다. 먼저 의원발의 안건입니다. 완주군의회 의원 의정활동비등 지급에 관한 조례 일부개정조례안, 완주군 향토문화유산 보호 및 관리 조례안, 완주군 싸움소 육성 지원에 관한 조례안 등 총 3건으로, 해당 상임위원회로 회부 하였습니다. 다음은 완주군수로부터 제출된 안건입니다. 삼례문화예술촌 설치 및 운영조례 일부개정 조례안 등 총 40건으로 해당 상임위원회로 회부하였습니다. 자세한 내용은 배부해 드린 유인물을 참조하여 주시기 바랍니다. 이상 보고를 마치겠습니다.

서로 대응되는 질문-대답은 동일한 색상으로 표시됩니다.

문서에 대답이 없는 경우

MARK DOCUMENT AS DONE ☐

SHOW LABELS OF ALL USERS ☐

Questions

U

No questions prepared yet. Please use the Question tab to add questions for your project.

1

ADD CUSTOM QUESTION

Annotation Document

Search

2

@제207회 완주군의회(임시회) 제 1 차 본회의회의록@

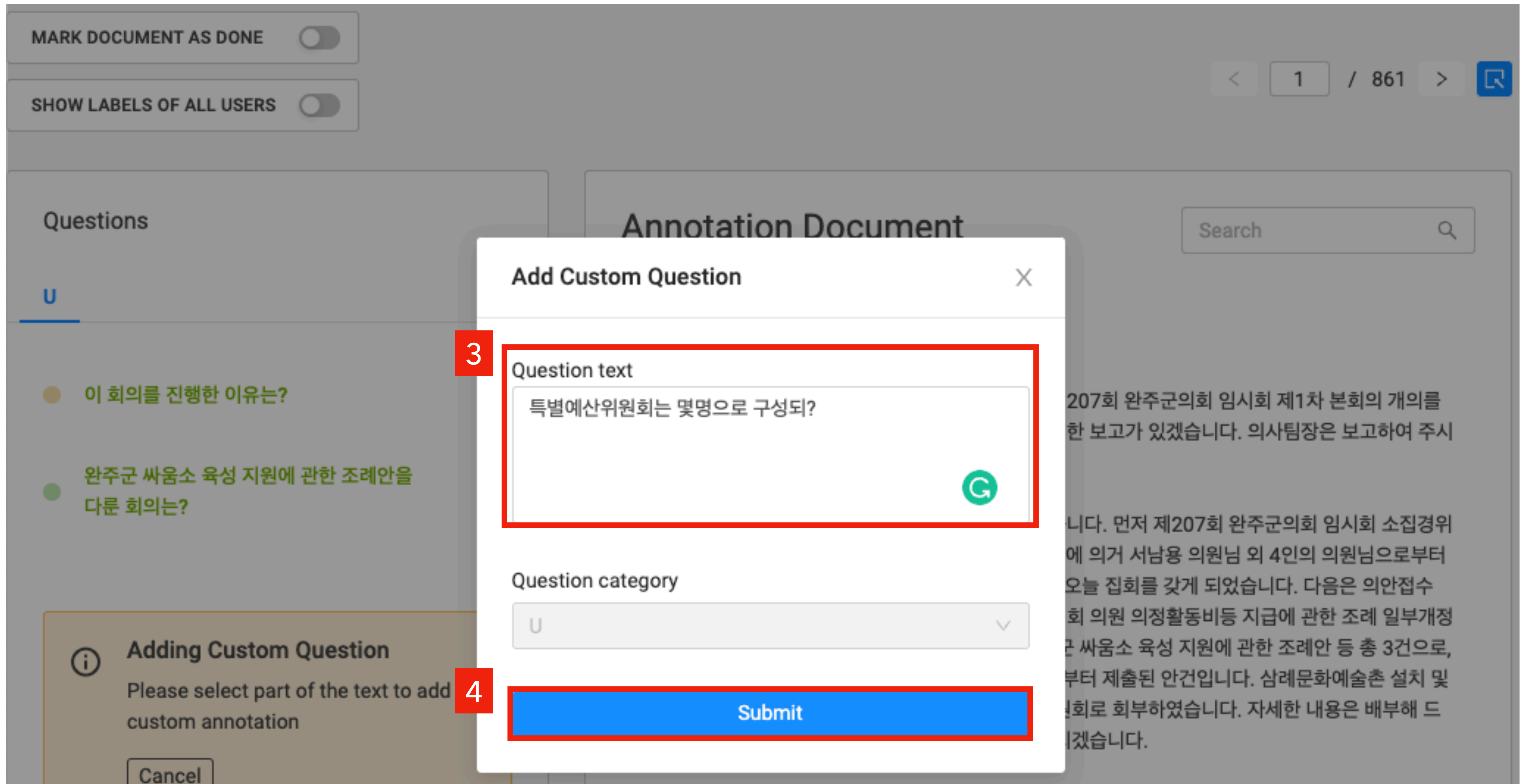
의석을 정돈하여 주시기 바랍니다. 성원이 되었으므로 제207회 완주군의회 임시회 제1차 본회의 개의를 선포합니다. 먼저 의사팀장으로부터 의회 관련 사항에 대한 보고가 있겠습니다. 의사팀장은 보고하여 주시기 바랍니다.

의사팀장 이은미 입니다. 의회관련 사항을 보고 드리겠습니다. 먼저 제207회 완주군의회 임시회 소집경위를 보고 드리겠습니다. 지방자치법 제45조 제2항의 규정에 의거 서남용 의원님 외 4인의 의원님으로부터 임시회 집회 요구가 있어 지난 8월6일 집회 공고를 하고 오늘 집회를 갖게 되었습니다. 다음은 의안접수 및 회부사항입니다. 먼저 의원발의 안건입니다. 완주군의회 의원 의정활동비등 지급에 관한 조례 일부개정조례안, 완주군 향토문화유산 보호 및 관리 조례안, 완주군 싸움소 육성 지원에 관한 조례안 등 총 3건으로, 해당 상임위원회로 회부 하였습니다. 다음은 완주군수로부터 제출된 안건입니다. 삼례문화예술촌 설치 및 운영조례 일부개정 조례안 등 총 40건으로 해당 상임위원회로 회부하였습니다. 자세한 내용은 배부해 드린 유인물을 참조하여 주시기 바랍니다. 이상 보고를 마치겠습니다.

Custom question은 no_answer 태깅이 안되서 대안책으로 @문자를 정답으로 태깅했습니다.

1) ADD CUSTOM QUESTION 버튼을 누릅니다.

2) no_answer 데이터를 라벨링하고 싶은 경우에는 대답으로 title의 첫 번째 문자인 @를 선택합니다.



- 3) 동일하게 질문을 작성합니다.
- 4) Submit 버튼을 눌러 완료합니다.

5

MARK DOCUMENT AS DONE ☐

SHOW LABELS OF ALL USERS ☐

Questions

U

이 회의를 진행한 이유는?

1

완주군 싸움소 육성 지원에 관한 조례안을 다룬 회의는?

2

특별예산위원회는 몇명으로 구성되?

3

ADD CUSTOM QUESTION

Annotation Document

Search

@제207회 완주군의회(임시회) 제 1 차 본회의회의록@

의석을 정돈하여 주시기 바랍니다. 성원이 되었으므로 제207회 완주군의회 임시회 제1차 본회의 개의를 선포합니다. 먼저 의사팀장으로부터 의회 관련 사항에 대한 보고가 있겠습니다. 의사팀장은 보고하여 주시기 바랍니다.

의사팀장 이은미 입니다. 의회관련 사항을 보고 드리겠습니다. 먼저 제207회 완주군의회 임시회 소집경위를 보고 드리겠습니다. 지방자치법 제45조 제2항의 규정에 의거 서남용 의원님 외 4인의 의원님으로부터 임시회 집회 요구가 있어 지난 8월6일 집회 공고를 하고 오늘 집회를 갖게 되었습니다. 다음은 의안접수 및 회부사항입니다. 먼저 의원발의 안건입니다. 완주군의회 의원 의정활동비등 지급에 관한 조례 일부개정조례안, 완주군 향토문화유산 보호 및 관리 조례안, 완주군 싸움소 육성 지원에 관한 조례안 등 총 3건으로, 해당 상임위원회로 회부 하였습니다. 다음은 완주군수로부터 제출된 안건입니다. 삼례문화예술촌 설치 및 운영조례 일부개정 조례안 등 총 40건으로 해당 상임위원회로 회부하였습니다. 자세한 내용은 배부해 드린 유인물을 참조하여 주시기 바랍니다. 이상 보고를 마치겠습니다.

- 5) no_answer 유형의 질문-대답도 동일하게 생성된 것을 확인할 수 있습니다.
- 6) 현재 문서에 대한 라벨링을 완료했다면, > 버튼을 눌러 다음 문서로 이동합니다.

라벨링 완료 후

완주 train data Documents Questions Import Export Labels User management

/ 완주 train data / Labels

All project labels My labels

All labels

Id	Question text	Answer	
361864	이 회의를 진행한 이유는?	서남용 의원님 외 4인의 의원님으로부터 임시회 집회 요구	
361865	완주군 싸움소 육성 지원에 관한 조례안을 다룬 회의는?	제207회 완주군의회 임시회 제1차 본회의	→ 🗑
361870	특별예산위원회는 몇명으로 구성되?	@	→ 🗑

Export answers
Export table in excel
Export table in CSV
Export in squad format

Export Labels >> Export answers >> Export in squad format 버튼을 순서대로 눌러 .json파일을 다운받습니다.
다른 확장자도 잘 다운되지만, 어차피 .json으로 읽을꺼니까 .json으로 다운받습니다.

라벨링 완료 후

```
1 {
2   "data": [
3     {
4       "paragraphs": [
5         {
6           "qas": [
7             {
8               "question": "1_제178회 완주군의회 임시회 회기는 언제로 결정됐어?",
9               "id": 341139,
10              "answers": [
11                {
12                  "answer_id": 362212,
13                  "document_id": 603489,
14                  "question_id": 341139,
15                  "text": "4월 20일부터 4월 27일까지 8일간",
16                  "answer_start": 144,
17                  "answer_end": 165,
18                  "answer_category": null
19                }
20              ],
21              "is_impossible": false
22            },
23            {
24              "question": "3_제178회 완주군의회 임시회 회기를 발의한 사람은 누구인가?",
25              "id": 341140,
26              "answers": [
27                {
28                  "answer_id": 362213,
29                  "document_id": 603489,
30                  "question_id": 341140,
31                  "text": "김상식 위원장 외 3인의 의원",
32                  "answer_start": 119,
33                  "answer_end": 135,
34                  "answer_category": null
35                }
36              ],
37              "is_impossible": false
38            }
39          ],
40          "context": "@제178회 완주군의회(임시회) 제 1 차 본회의회의록@\\n\\n의사일정",
41          "document_id": 603489
42        }
43      ]
44    },
45  ],
46}
```

왼쪽은 하나의 context에서 2쌍의 질문-대답을 생성한 예시입니다.

도움이 될만한 자료

- haystack 웹사이트:

<https://annotate.deepset.ai/projects/3564>

- 유튜브 튜토리얼(영어 발음을 알아듣기는 힘들지만 따라하면 수월함):

https://www.youtube.com/watch?v=4pPHSmd_FM0

- haystack 깃허브 주소:

https://github.com/deepset-ai/haystack/tree/master/annotation_tool

- haystack docs:

<https://haystack.deepset.ai/docs/intromd>

