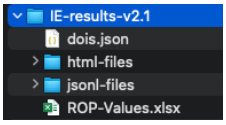
ROP ENTITY ANNOTATION GUIDELINE

Yinghao Li; September 13, 2021

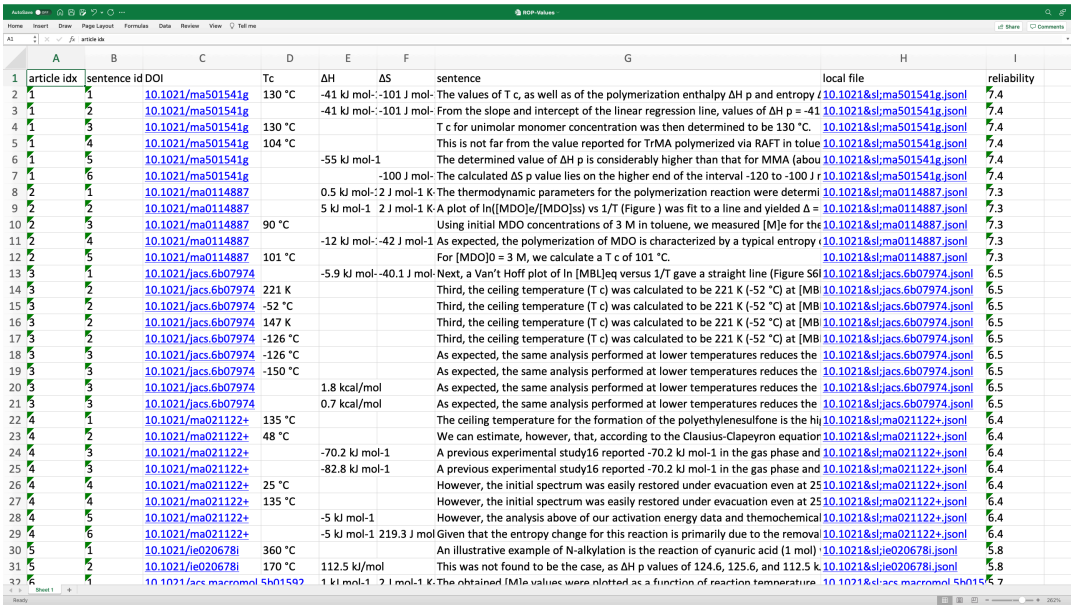
Note: check this link for the latest version of this document.

CKECKING RESULTS

The result folder should contain several folders and files.



- Html_files: the folder that contains HTML result files with polymer names and sentences containing ROP property values highlighted.
- Jsonl_files: the folder that contains jsonl result files (will be introduced later)
- Dois.json: the dois of the articles of positive results.
- ROP-Values.xlsx: A table that contains the positive sentences, articles and extracted values from them. It also has hyperlinks to the original articles as well as the local result files in “HTML_files” folder.



	A	B	C	D	E	F	G	H	I
1	article idx	sentence id	DOI	Tc	ΔH	ΔS	sentence	local file	reliability
2	1	1	10.1021/ma501541g	130 °C	-41 kJ mol-	-101 J mol-	The values of T c, as well as of the polymerization enthalpy ΔH p and entropy	10.1021&sl;ma501541g.jsonl	7.4
3	1	2	10.1021/ma501541g		-41 kJ mol-	-101 J mol-	From the slope and intercept of the linear regression line, values of ΔH p = -41	10.1021&sl;ma501541g.jsonl	7.4
4	1	3	10.1021/ma501541g	130 °C			T c for unimolar monomer concentration was then determined to be 130 °C.	10.1021&sl;ma501541g.jsonl	7.4
5	1	4	10.1021/ma501541g	104 °C			This is not far from the value reported for TrMA polymerized via RAFT in tolue	10.1021&sl;ma501541g.jsonl	7.4
6	1	5	10.1021/ma501541g		-55 kJ mol-1		The determined value of ΔH p is considerably higher than that for MMA (abou	10.1021&sl;ma501541g.jsonl	7.4
7	1	6	10.1021/ma501541g			-100 J mol-	The calculated ΔS p value lies on the higher end of the interval -120 to -100 J r	10.1021&sl;ma501541g.jsonl	7.4
8	2	1	10.1021/ma0114887		0.5 kJ mol-	2 J mol-1 K-	The thermodynamic parameters for the polymerization reaction were determi	10.1021&sl;ma0114887.jsonl	7.3
9	2	2	10.1021/ma0114887		5 kJ mol-1	2 J mol-1 K-	A plot of ln([MDO]e/[MDO]ss) vs 1/T (Figure) was fit to a line and yielded Δ =	10.1021&sl;ma0114887.jsonl	7.3
10	2	3	10.1021/ma0114887	90 °C			Using initial MDO concentrations of 3 M in toluene, we measured [M]e for the	10.1021&sl;ma0114887.jsonl	7.3
11	2	4	10.1021/ma0114887		-12 kJ mol-	-42 J mol-1	As expected, the polymerization of MDO is characterized by a typical entropy	10.1021&sl;ma0114887.jsonl	7.3
12	2	5	10.1021/ma0114887	101 °C			For [MDO]0 = 3 M, we calculate a T c of 101 °C.	10.1021&sl;ma0114887.jsonl	7.3
13	3	1	10.1021/jacs.6b07974		-5.9 kJ mol-	-40.1 J mol-	Next, a Van't Hoff plot of ln [MBL]eq versus 1/T gave a straight line (Figure S6	10.1021&sl;jacs.6b07974.jsonl	6.5
14	3	2	10.1021/jacs.6b07974	221 K			Third, the ceiling temperature (T c) was calculated to be 221 K (-52 °C) at [MB	10.1021&sl;jacs.6b07974.jsonl	6.5
15	3	2	10.1021/jacs.6b07974	-52 °C			Third, the ceiling temperature (T c) was calculated to be 221 K (-52 °C) at [MB	10.1021&sl;jacs.6b07974.jsonl	6.5
16	3	2	10.1021/jacs.6b07974	147 K			Third, the ceiling temperature (T c) was calculated to be 221 K (-52 °C) at [MB	10.1021&sl;jacs.6b07974.jsonl	6.5
17	3	2	10.1021/jacs.6b07974	-126 °C			Third, the ceiling temperature (T c) was calculated to be 221 K (-52 °C) at [MB	10.1021&sl;jacs.6b07974.jsonl	6.5
18	3	3	10.1021/jacs.6b07974	-126 °C			As expected, the same analysis performed at lower temperatures reduces the	10.1021&sl;jacs.6b07974.jsonl	6.5
19	3	3	10.1021/jacs.6b07974	-150 °C			As expected, the same analysis performed at lower temperatures reduces the	10.1021&sl;jacs.6b07974.jsonl	6.5
20	3	3	10.1021/jacs.6b07974		1.8 kcal/mol		As expected, the same analysis performed at lower temperatures reduces the	10.1021&sl;jacs.6b07974.jsonl	6.5
21	3	3	10.1021/jacs.6b07974		0.7 kcal/mol		As expected, the same analysis performed at lower temperatures reduces the	10.1021&sl;jacs.6b07974.jsonl	6.5
22	4	1	10.1021/ma021122+	135 °C			The ceiling temperature for the formation of the polyethylenesulfone is the hi	10.1021&sl;ma021122+.jsonl	6.4
23	4	2	10.1021/ma021122+	48 °C			We can estimate, however, that, according to the Clausius-Clapeyron equatior	10.1021&sl;ma021122+.jsonl	6.4
24	4	3	10.1021/ma021122+		-70.2 kJ mol-1		A previous experimental study16 reported -70.2 kJ mol-1 in the gas phase and	10.1021&sl;ma021122+.jsonl	6.4
25	4	3	10.1021/ma021122+		-82.8 kJ mol-1		A previous experimental study16 reported -70.2 kJ mol-1 in the gas phase and	10.1021&sl;ma021122+.jsonl	6.4
26	4	4	10.1021/ma021122+	25 °C			However, the initial spectrum was easily restored under evacuation even at 25	10.1021&sl;ma021122+.jsonl	6.4
27	4	4	10.1021/ma021122+	135 °C			However, the initial spectrum was easily restored under evacuation even at 25	10.1021&sl;ma021122+.jsonl	6.4
28	4	5	10.1021/ma021122+		-5 kJ mol-1		However, the analysis above of our activation energy data and themochemical	10.1021&sl;ma021122+.jsonl	6.4
29	4	6	10.1021/ma021122+		-5 kJ mol-1	219.3 J mol	Given that the entropy change for this reaction is primarily due to the removal	10.1021&sl;ma021122+.jsonl	6.4
30	5	1	10.1021/ie020678i	360 °C			An illustrative example of N-alkylation is the reaction of cyanuric acid (1 mol)	10.1021&sl;ie020678i.jsonl	5.8
31	5	2	10.1021/ie020678i	170 °C	112.5 kJ/mol		This was not found to be the case, as ΔH p values of 124.6, 125.6, and 112.5 k	10.1021&sl;ie020678i.jsonl	5.8
32	6	1	10.1021/acs.macromol.5b01592		1 kJ mol-1	2 J mol-1 K-	The obtained [M]e values were plotted as a function of reaction temperature	10.1021&sl;acs.macromol.5b015	5.7

TO VALIDATION A SENTENCE

1. Go to ROP-Values.xlsx and go through the sentences/values top-down. The results are ordered according to their reliability scores. So, results at the top of the file are more “reliable” than results below.
2. If a result is obviously correct/incorrect, you can directly add the values to the spreadsheet or discard it.

3. If it is hard to tell, you can go the local HTML file by clicking the corresponding link. The sentences are listed at the top of the file as well as highlighted in the body. You can also go to the original website by clicking the doi link for more information.

TO GIVE FEEDBACK

The **preferred** feedback method is manually annotated sentences/articles with NER labels and positive/negative sentence-level labels. You can also provide an unformatted feedback file in natural language.

The following section describes how to provide NER labels and positive/negative sentence-level labels.

DOCCANO

Doccano is a powerful open-source web-based NLP annotation tool. It has modern, stable and user-friendly annotation interface.

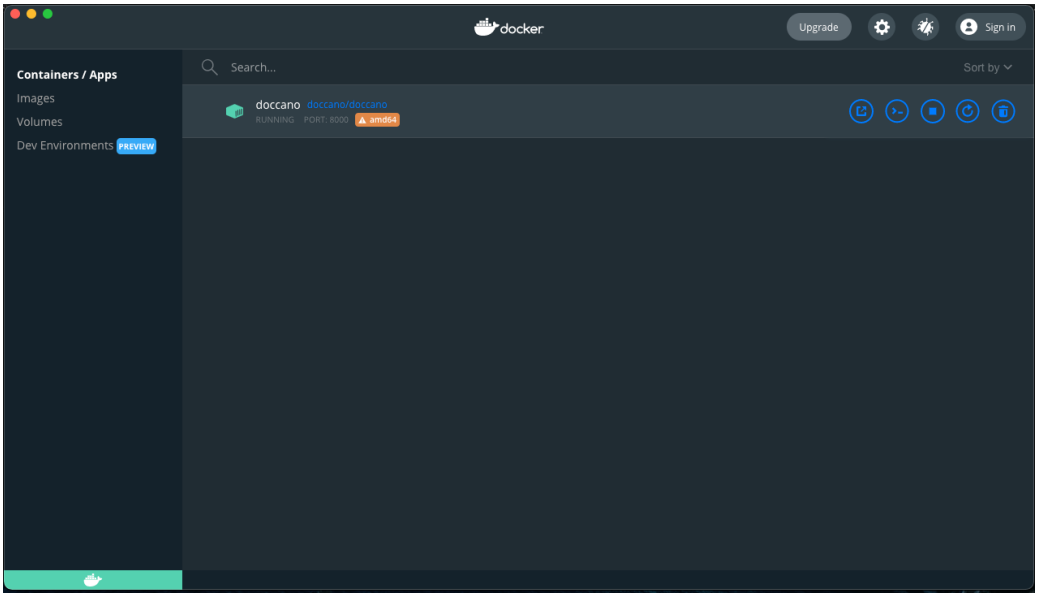
INSTALLATION

Recommended method: using Docker (https://www.docker.com/)

1. Download and install Docker: https://www.docker.com/get-started
2. Start Docker process.
3. Follow the instruction on https://github.com/doccano/doccano in "Docker" section to install Doccano in your Docker. Notice that you need to specify your username and password.
4. Start the docker and go to http://127.0.0.1:8000/ to annotate data.

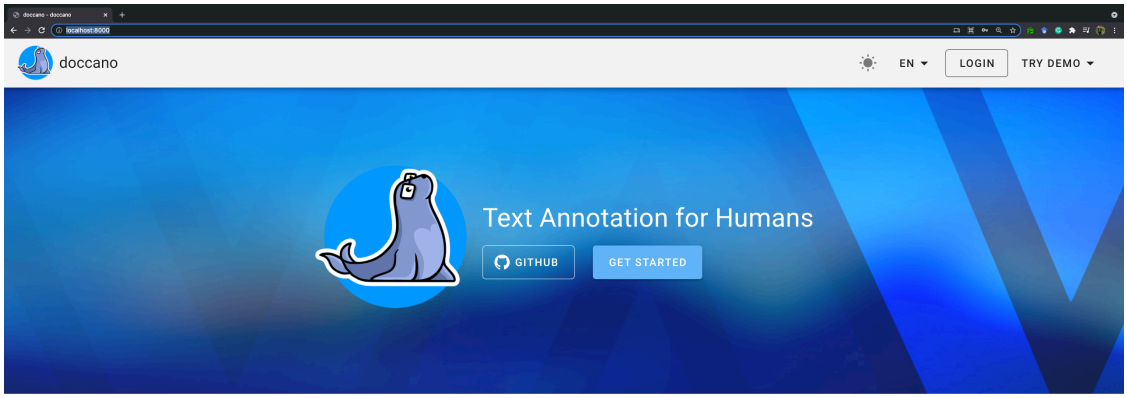
Other methods: Follow the instruction on https://github.com/doccano/doccano. Notice that other installation methods are not tested.

Your Docker UI (if the version of Docker you installed is with UI) should look like this if the doccano container is running:

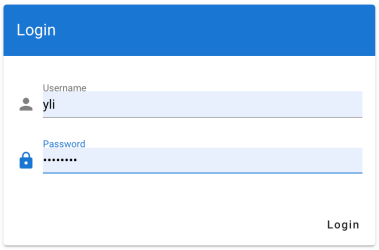


CREATING PROJECT

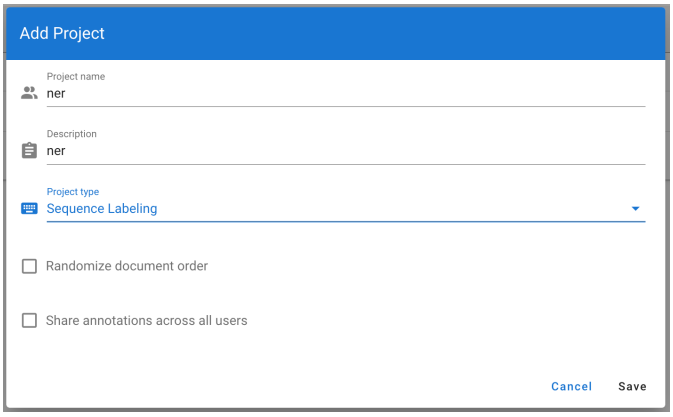
Once you have successfully installed and ran image, you should be able to see the web page below with URL http://localhost:8000/ or http://127.0.0.1:8000/



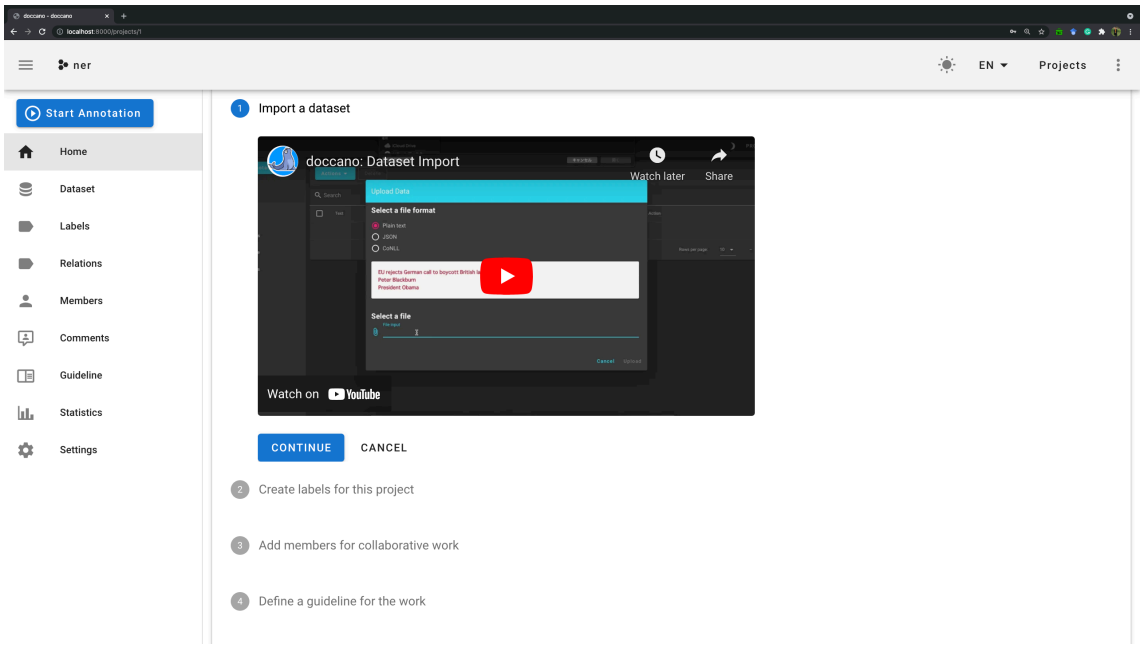
The first thing to do is logging into the system with the username and password you just specified.



Then, you need to create a project to hold the dataset. You can choose whatever project name and description you like. For project type, you need to choose sequence labeling. Save the project



Once you have created the project, you should be able to see the page below:



Doccano has (not that) detailed video instruction on how to annotate data on this platform. It should cover most of the cases. You can give it a look if you are interested.

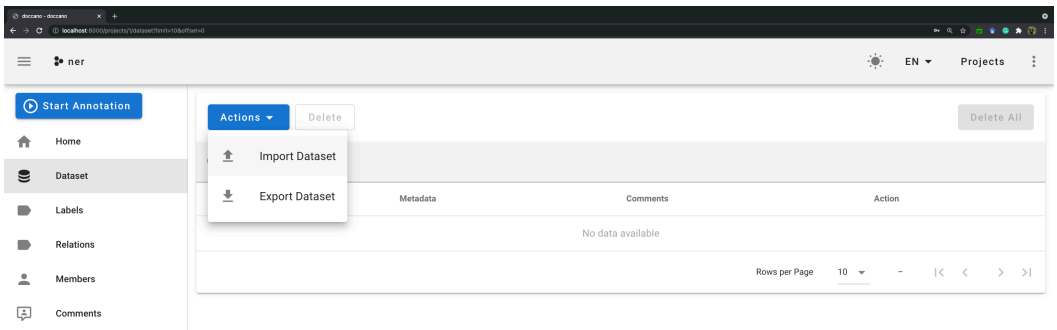
ANNOTATING DATA

The automatically generated ROP property sentences and polymer names are currently not perfect. To improve the performance of the models, we need larger training dataset. Doccano allows us to directly modify the predicted labels, which can be much faster than annotating a document from scratch.

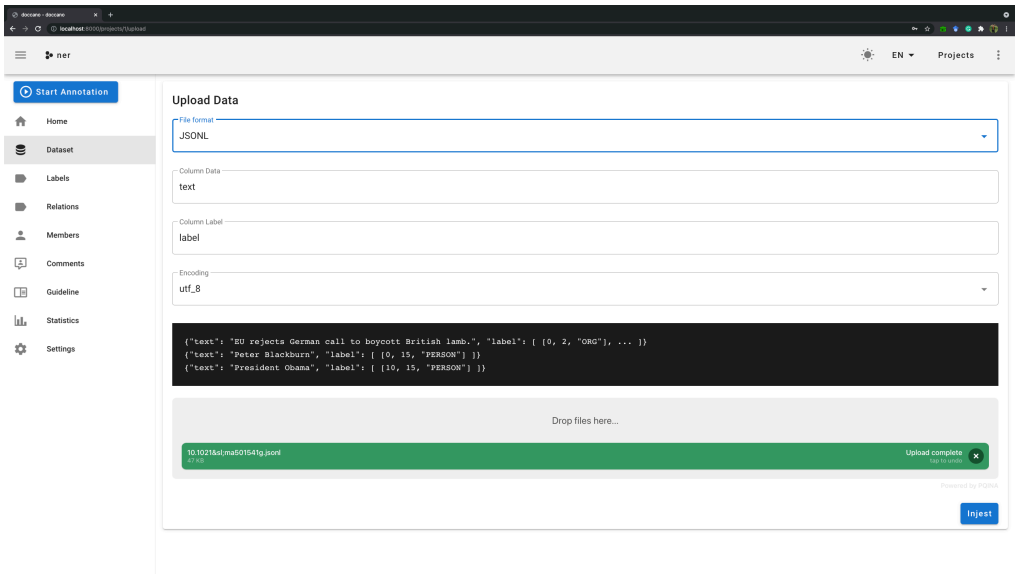
The input of Doccano is in JSONL format, which are provided in the “jsonl_files” folder. The files have the same names apart from the suffix as the result files in the “html_files” folder, so you can seach for them by the file name with your file manager. I do not know how to implement a “reveal in folder/finder” function in the Excel tables thus we have to stick to this less convenient way of finding files.

For the documents you think that are worth labeling, to correct the labels generated by the automated NLP system or introducing new annotations, you can follow the procedures below:

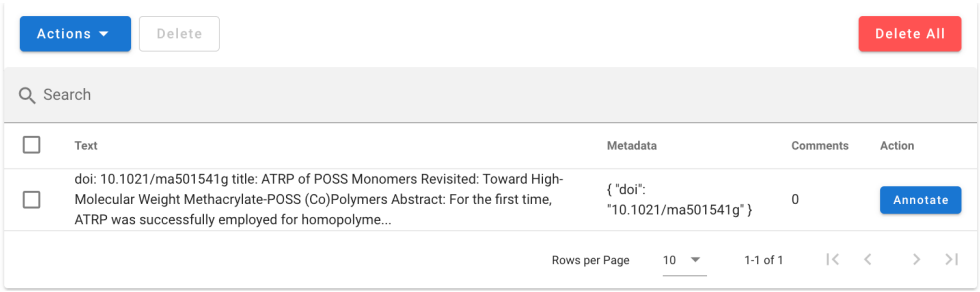
1) Import datafile: In Doccano webpage, click “Dataset”, “Actions”, “Import dataset”



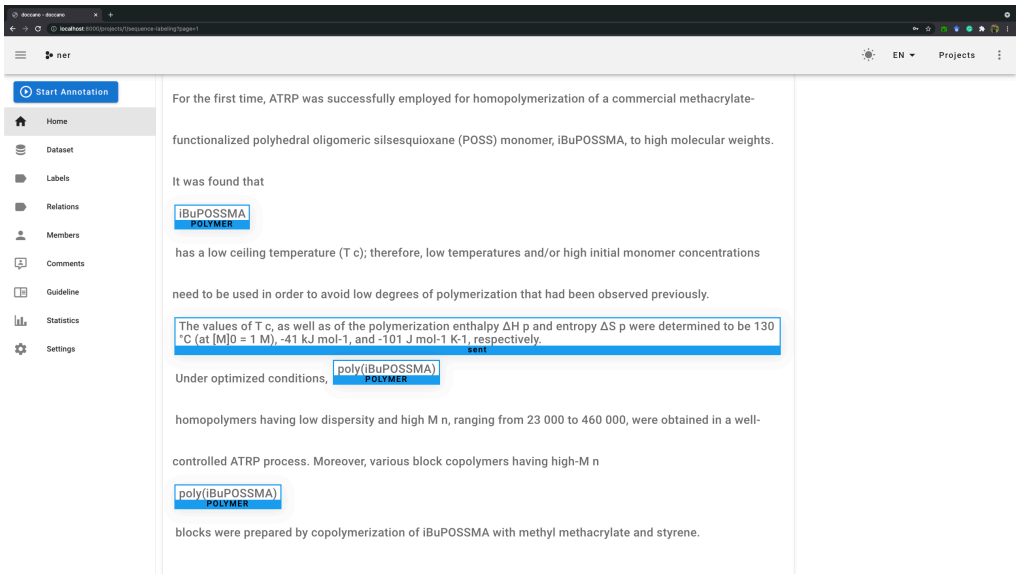
Choose “JSONL” data format and drop the candidate file into the box, click “inject”



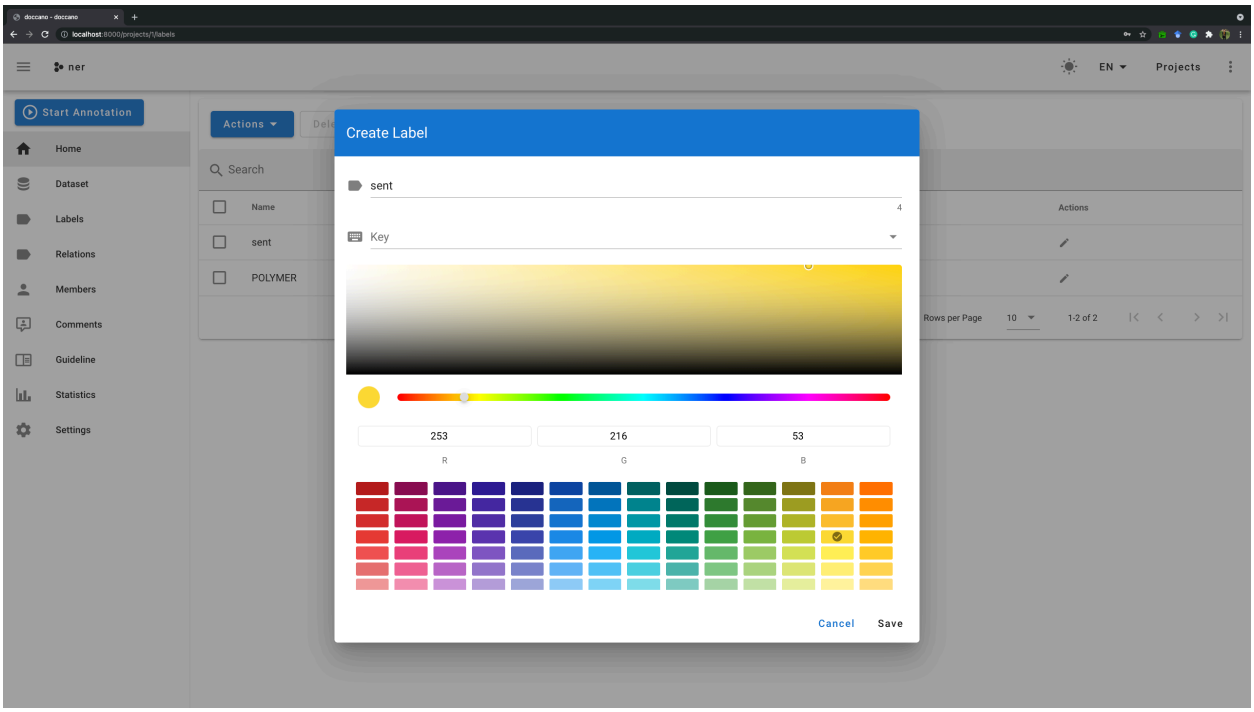
Goto “Dataset” or use “Backward” button, you will be able to see the file you just imported



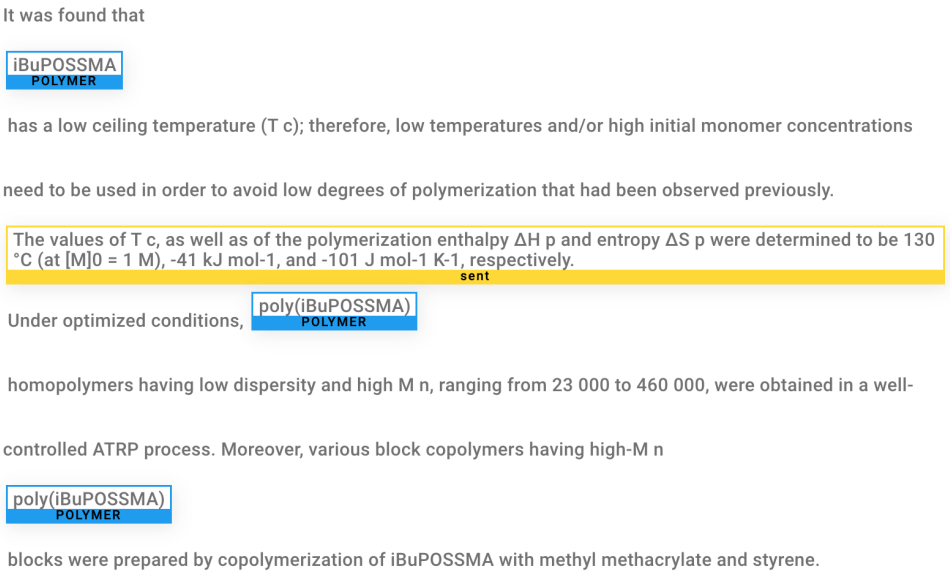
2) Annotating data: click the “Annotate” button, you will see the annotation interface with some words/sentences already labeled by the BERT models



The above figure shows 2 entity types: “sent” and “POLYMER”. “sent” entity represents the sentences that are predicted to contain ROP property values. “POLYMER” entity represents the polymer names. You can change their colors in the “Labels” panel to make it easier to separate them apart (“Labels” -> “Action”)

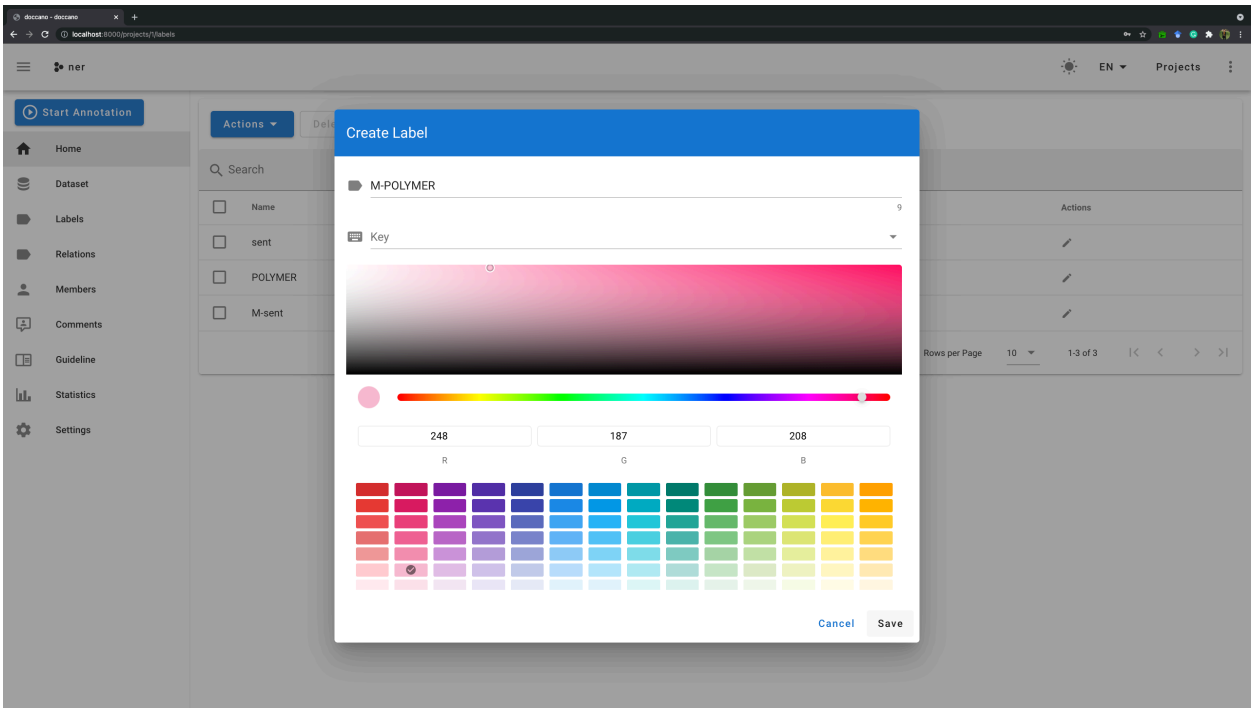


Result:

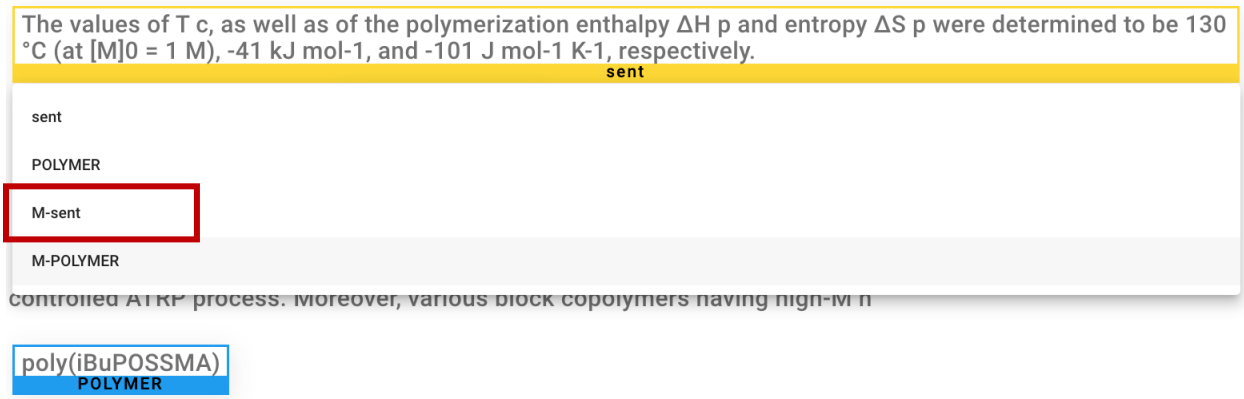


NOTICE: You should check the results in the Excel table and HTML result files *BEFORE* going through this labeling step. This system is not suitable for direct result presentation.

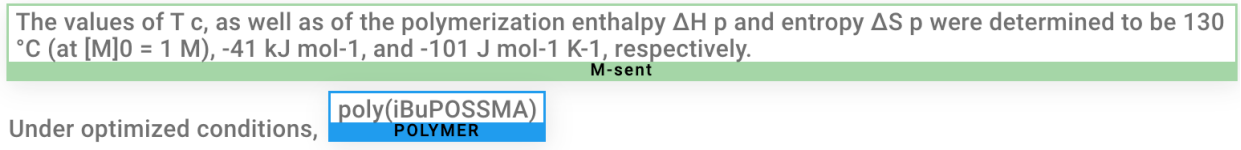
To insert your labels or confirm model output, you should create entity labels starting with "M-", in this case, "M-sent" and "M-POLYMER" (the labels are case-sensitive). Notice that entity labels not starting with "M-" will be discarded during my future post processing. You can also assign a hotkey to each entity label:



To confirm an annotation, click that entity and, in the drop-down list, select the corresponding entity label started with "M-":

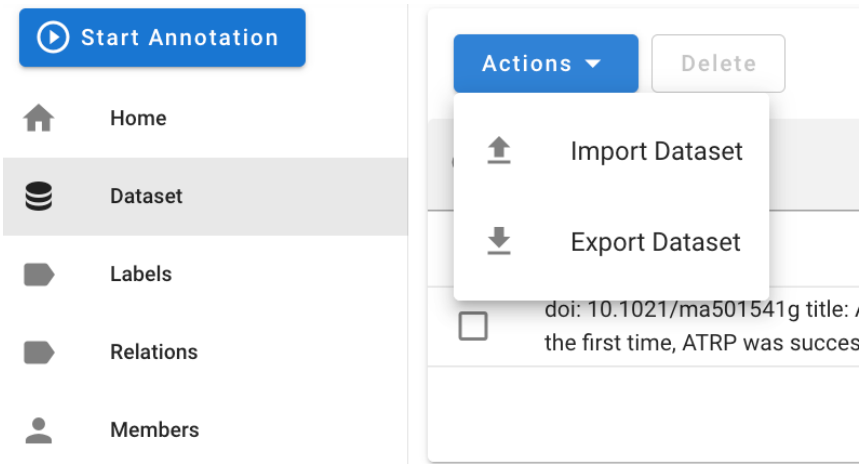


The entity label is successfully changed:



To insert an annotation, simply select the text that belongs to the entity and choose the corresponding entity label started with “M-” in the drop-down list. To delete an annotation, hover your curser above the entity and click the “X” button at the top-left corner.

3) Saving results: You can export your dataset with the “Export dataset” button in the “Action” list in “Dataset” section.



Notice that the platform only allows you to export the entire dataset in one JSONL file. It is suggested to label some document -> export results -> delete all document in the dataset -> import and label other documents. This procedure keeps your workspace clean.

Important: according to the GitHub page, you need to use `` `docker container stop doccano -t 5` `` to stop your doccano container while keeping your work. Your progress may be lost otherwise.