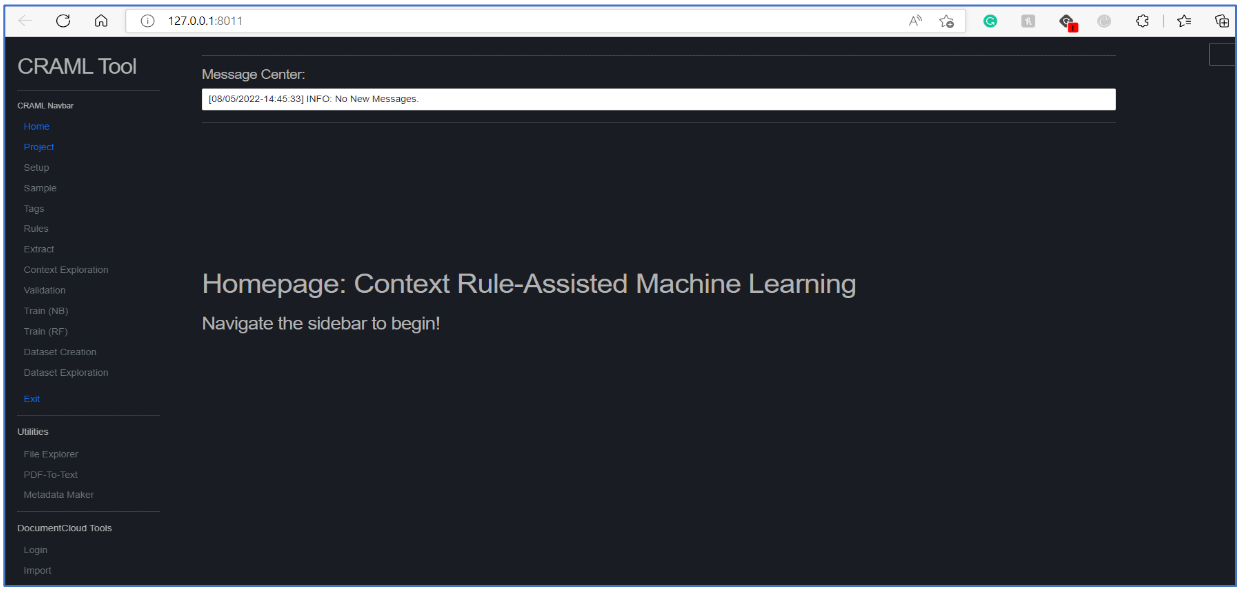
A Step-by-Step Guide to Recreate the No Poach analysis[1]

1. Start CRAML_Tool.py with Python. A browser window will appear.



Message Center:

[08/05/2022-14:45:33] INFO: No New Messages.

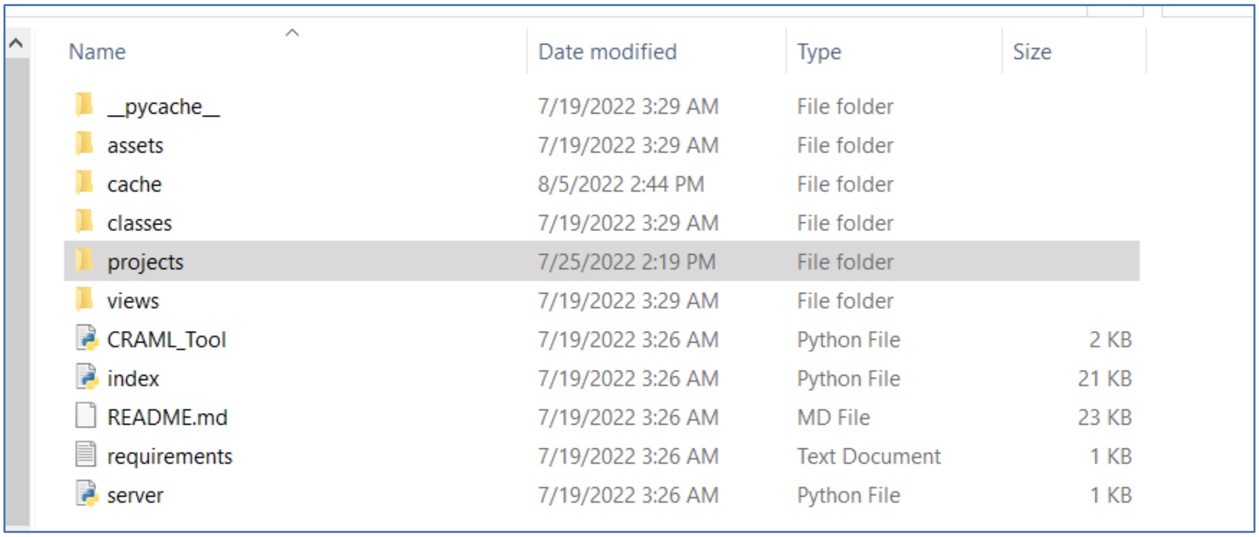
Homepage: Context Rule-Assisted Machine Learning

Navigate the sidebar to begin!

2. Download the unstructured text data and metadata to a folder.

3. Download and unzip the intermediate project files.

   a) In the folder CRAML is installed in, a “projects” folder contains all research projects.

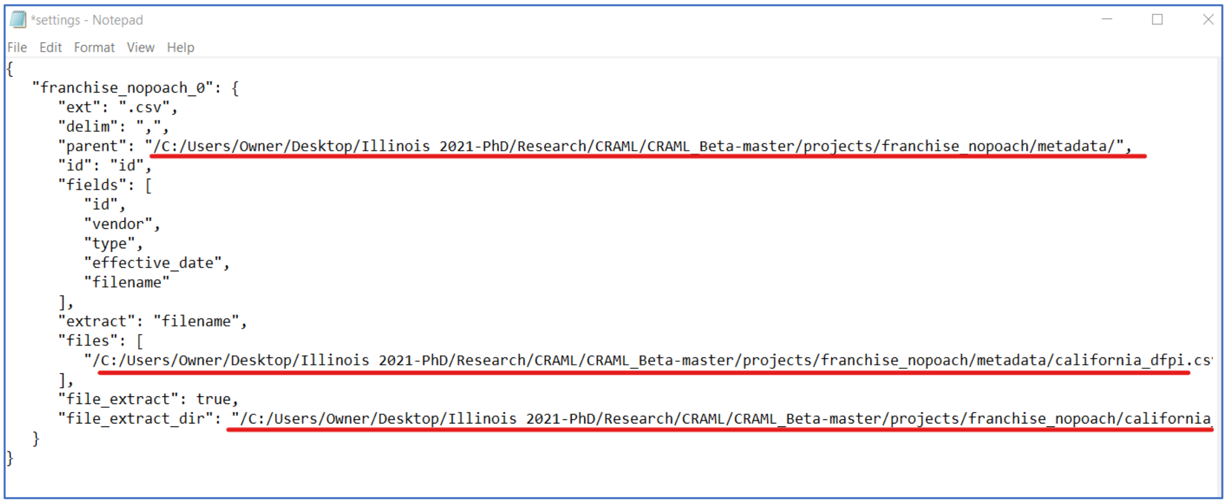


Name	Date modified	Type	Size
__pycache__	7/19/2022 3:29 AM	File folder	
assets	7/19/2022 3:29 AM	File folder	
cache	8/5/2022 2:44 PM	File folder	
classes	7/19/2022 3:29 AM	File folder	
projects	7/25/2022 2:19 PM	File folder	
views	7/19/2022 3:29 AM	File folder	
CRAML_Tool	7/19/2022 3:26 AM	Python File	2 KB
index	7/19/2022 3:26 AM	Python File	21 KB
README.md	7/19/2022 3:26 AM	MD File	23 KB
requirements	7/19/2022 3:26 AM	Text Document	1 KB
server	7/19/2022 3:26 AM	Python File	1 KB

[1] Acknowledgements. Thank you to Patricia Tabarani who helped prepare this step-by-step replication guide.

b) Download the intermediate files and unzip folders that contain the metadata, keywords, and the settings files of the project “franchise_nopoach.” Once downloaded, extract the files, and save them in the “projects” folder created previously.

c) In the settings.json file in each project directory, update the path to point it to the directory that has the cleaned data and the file that contains the metadata. For example, see below:



```
{
   "franchise_nopoach_0": {
      "ext": ".csv",
      "delim": ",",
      "parent": "/C:/Users/Owner/Desktop/Illinois 2021-PhD/Research/CRAML/CRAML Beta-master/projects/franchise nopoach/metadata/",
      "id": "id",
      "fields": [
         "id",
         "vendor",
         "type",
         "effective_date",
         "filename"
      ],
      "extract": "filename",
      "files": [
         "/C:/Users/Owner/Desktop/Illinois 2021-PhD/Research/CRAML/CRAML_Beta-master/projects/franchise_nopoach/metadata/california_dfpi.cs
      ],
      "file_extract": true,
      "file_extract_dir": "/C:/Users/Owner/Desktop/Illinois 2021-PhD/Research/CRAML/CRAML_Beta-master/projects/franchise_nopoach/california
   }
}
```

Additional Guidance on Working with CRAML

Users can work with their own data inside CRAML.

Interacting with CRAML

a) New directories created in the “projects” folder outside of CRAML will appear as projects in CRAML. Project directories deleted outside of CRAML will be deleted in CRAML.
b) Likewise, creating a new project inside CRAML creates a directory in the projects folder. As users perform analysis in CRAML, new folders and files are created in each project folder to save the output and enable users to interactively engage with the output.

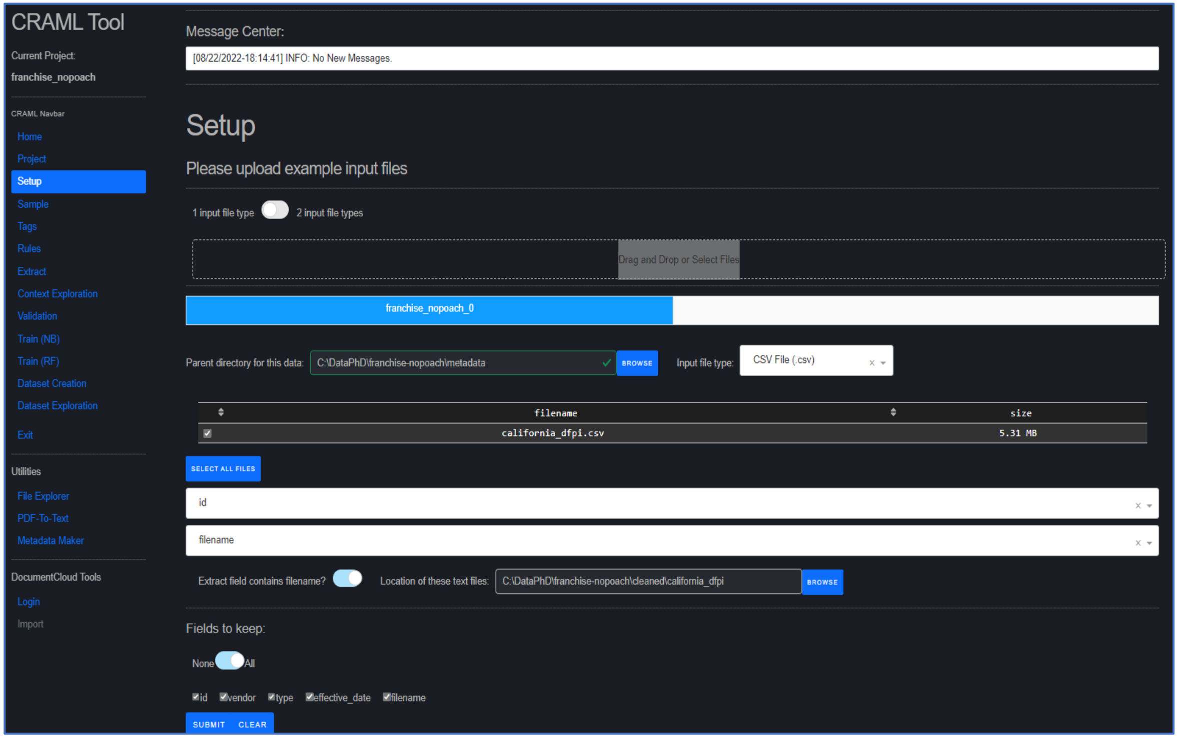
Setting Up a project on CRAML:

Setup is a challenging part of working in CRAML because text comes in many varieties and there's no one standard way of having a massive text corpus.

Inputs must be in a form that CRAML will accept. CRAML is designed to be text format agnostic, and accepts CSV, JSON, XML and other formats, but for unusual cases, users need some programming experience to input data into CRAML.

“Extract fields contains filename?” refers to whether the metadata refers to specific text files.

"Fields to keep:" refers to whether output datasets should include metadata fields.



Current Project: franchise_nopoach

Message Center:

[08/22/2022-18:14:41] INFO: No New Messages.

Setup

Please upload example input files

1 input file type ○ 2 input file types

franchise_nopoach_0

Parent directory for this data: C:\DataPhD\franchise-nopoach\metadata

Input file type: CSV File (.csv)

	filename	size
☑	california_dfpi.csv	5.31 MB

id

filename

Extract field contains filename? ●

Location of these text files: C:\DataPhD\franchise-nopoach\cleaned\california_dfpi

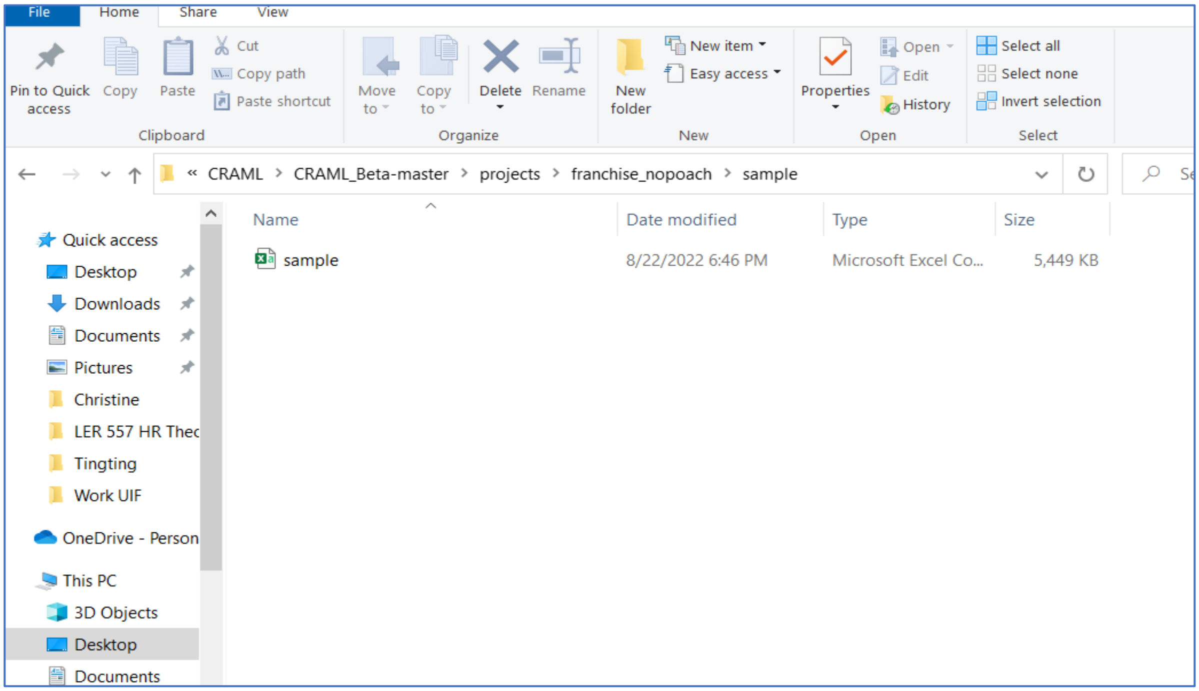
Fields to keep:

None ● All

☑id ☑vendor ☑type ☑effective_date ☑filename

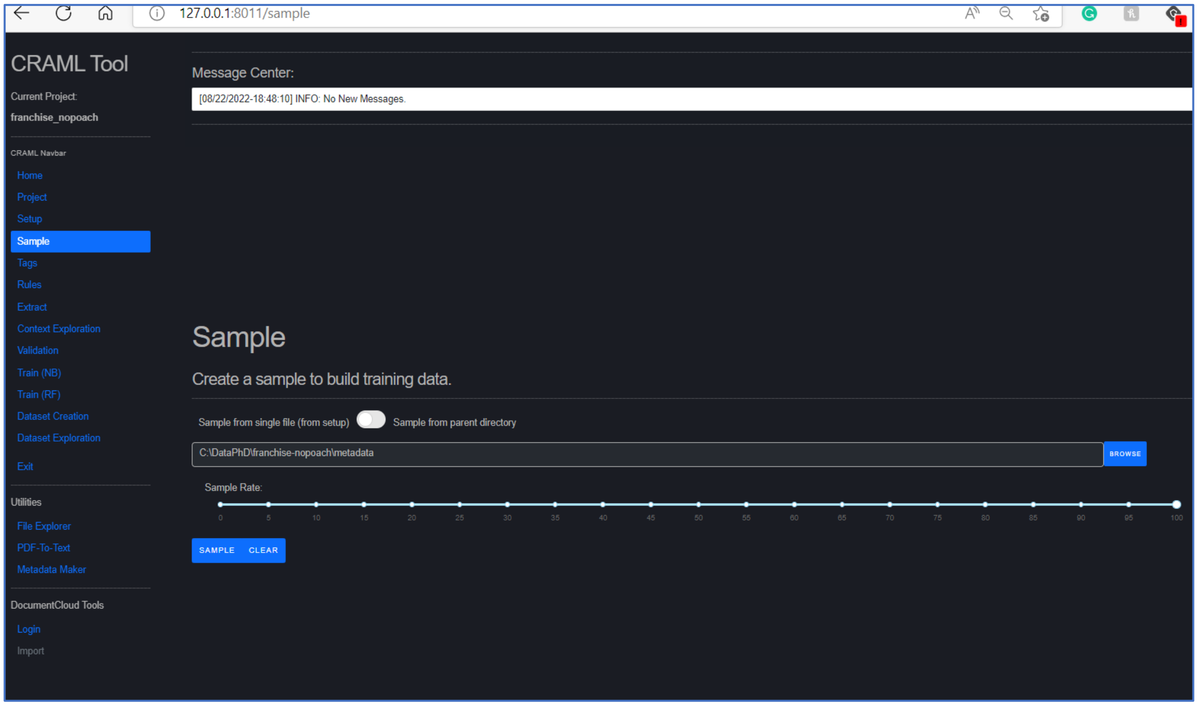
Sampling Data

Sampling selects a percentage of documents and stores the results in sample.csv for further analysis.



« CRAML › CRAML_Beta-master › projects › franchise_nopoach › sample

Name	Date modified	Type	Size
sample	8/22/2022 6:46 PM	Microsoft Excel Co...	5,449 KB



Current Project:
franchise_nopoach

Message Center:

[08/22/2022-18:48:10] INFO: No New Messages.

Sample

Create a sample to build training data.

Sample from single file (from setup) Sample from parent directory

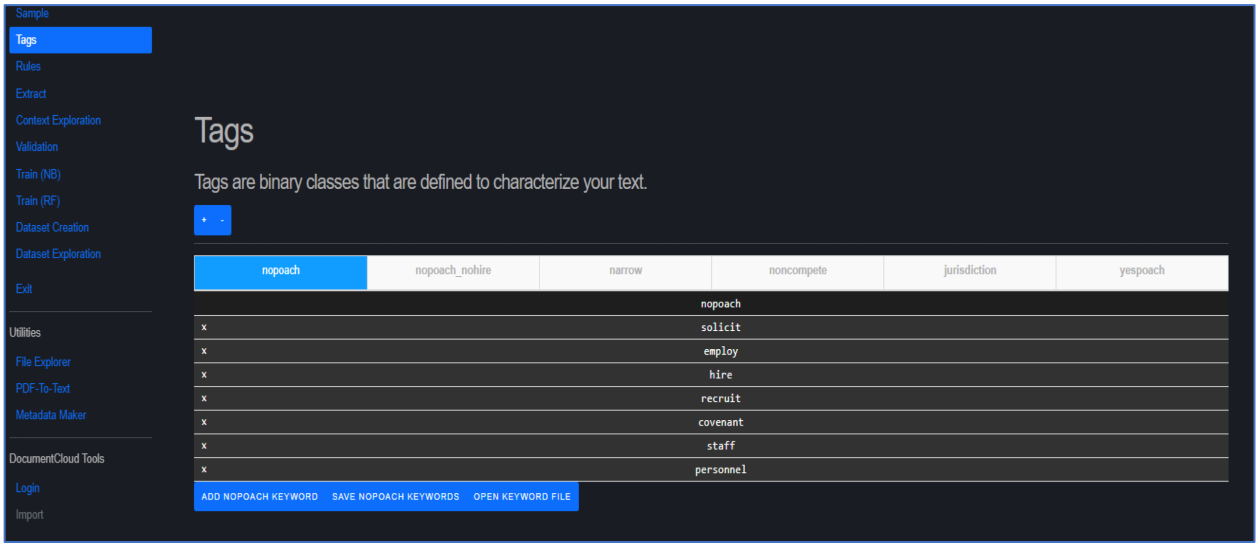
C:\DataPhD\franchise-nopoach\metadata

Sample Rate: 100

Tags and Keywords

Keywords and tags are stored in keywords.json.

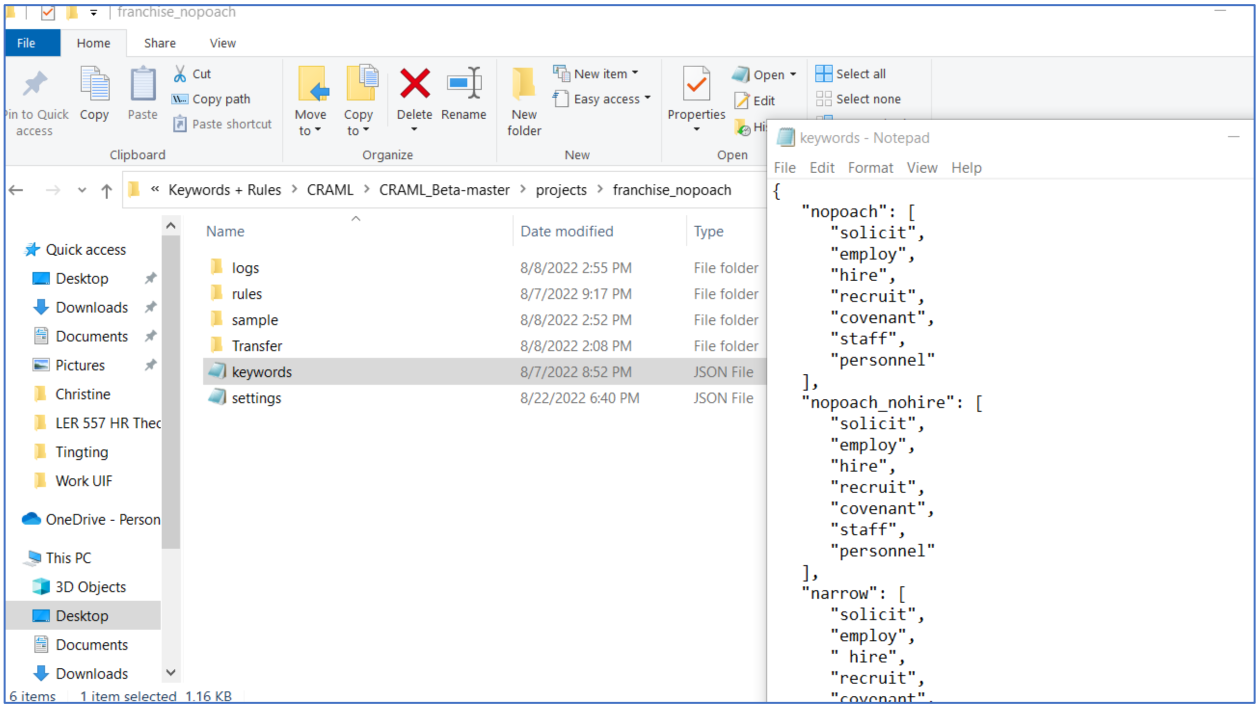
In CRAML, "Tags" appear on the left side of the screen. Keywords can be added in CRAML (click "Add Keyword" then "Save Keyword") or outside of the browser.



Tags

Tags are binary classes that are defined to characterize your text.

nopoach
solicit
employ
hire
recruit
covenant
staff
personnel



Name	Date modified	Type
logs	8/8/2022 2:55 PM	File folder
rules	8/7/2022 9:17 PM	File folder
sample	8/8/2022 2:52 PM	File folder
Transfer	8/8/2022 2:08 PM	File folder
keywords	8/7/2022 8:52 PM	JSON File
settings	8/22/2022 6:40 PM	JSON File

```
{
   "nopoach": [
      "solicit",
      "employ",
      "hire",
      "recruit",
      "covenant",
      "staff",
      "personnel"
   ],
   "nopoach_nohire": [
      "solicit",
      "employ",
      "hire",
      "recruit",
      "covenant",
      "staff",
      "personnel"
   ],
   "narrow": [
      "solicit",
      "employ",
      " hire",
      "recruit",
      "covenant",
```

Extracting chunks

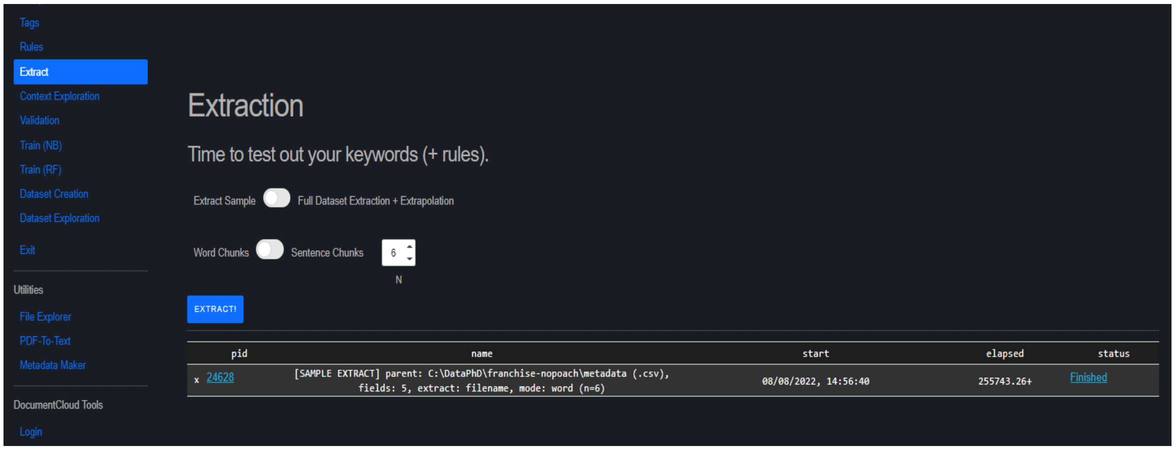
a) Extract chunks of text containing keywords from the sample.

b) Choose "Word" or "Sentence" length chunks and adjust the N (number of words or sentences to the left and right of the keyword that will be extracted). As small an N as necessary to understand the context is advised.

c) Click "EXTRACT!" A folder with the process identification number for the extraction (pid) will appear in the projects folder under "csv".

d) The csv file in the pid folder will contain all extracted text from the sample.

e) Users who do not wish to create ML models can extrapolate the rules onto the full dataset by selecting "Full Dataset Extraction + Extrapolation."

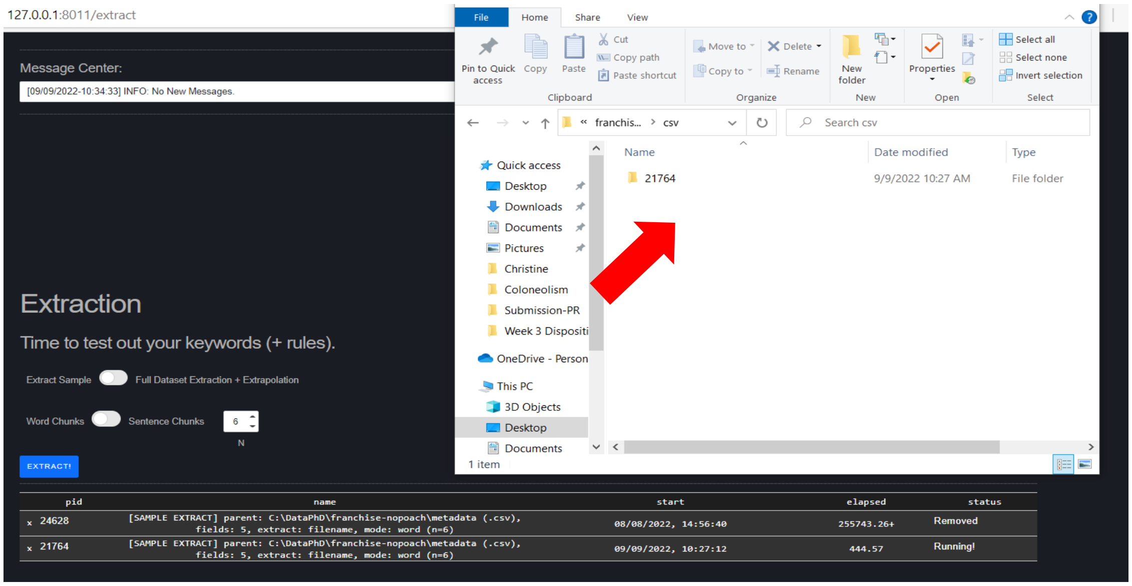


Extraction

Time to test out your keywords (+ rules).

N: 6

pid	name	start	elapsed	status
x 24628	[SAMPLE EXTRACT] parent: C:\DataPhD\franchise-nopoach\metadata (.csv), fields: 5, extract: filename, mode: word (n=6)	08/08/2022, 14:56:40	255743.26+	Finished


127.0.0.1:8011/extract

Message Center:

[09/09/2022-10:34:33] INFO: No New Messages.

Extraction

Time to test out your keywords (+ rules).

Extract Sample ○ Full Dataset Extraction + Extrapolation

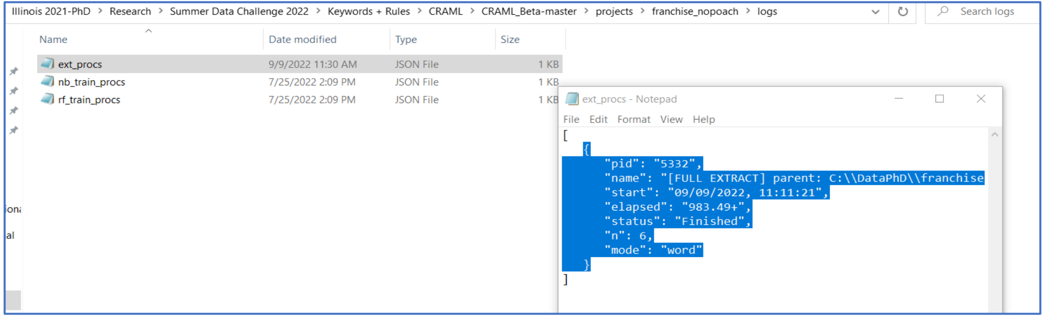
Word Chunks ○ Sentence Chunks 6 N

EXTRACT!

	pid	name	start	elapsed	status
x	24628	[SAMPLE EXTRACT] parent: C:\DataPhD\franchise-nopoach\metadata (.csv), fields: 5, extract: filename, mode: word (n=6)	08/08/2022, 14:56:40	255743.26+	Removed
x	21764	[SAMPLE EXTRACT] parent: C:\DataPhD\franchise-nopoach\metadata (.csv), fields: 5, extract: filename, mode: word (n=6)	09/09/2022, 10:27:12	444.57	Running!

Name	Date modified	Type
21764	9/9/2022 10:27 AM	File folder

f) The "logs" folder contains an "ext_procs.json" file that contains the information on each extraction. Manually erasing the pid as shown in the screenshot below will remove the entry in CRAML.


Illinois 2021-PhD › Research › Summer Data Challenge 2022 › Keywords + Rules › CRAML › CRAML_Beta-master › projects › franchise_nopoach › logs

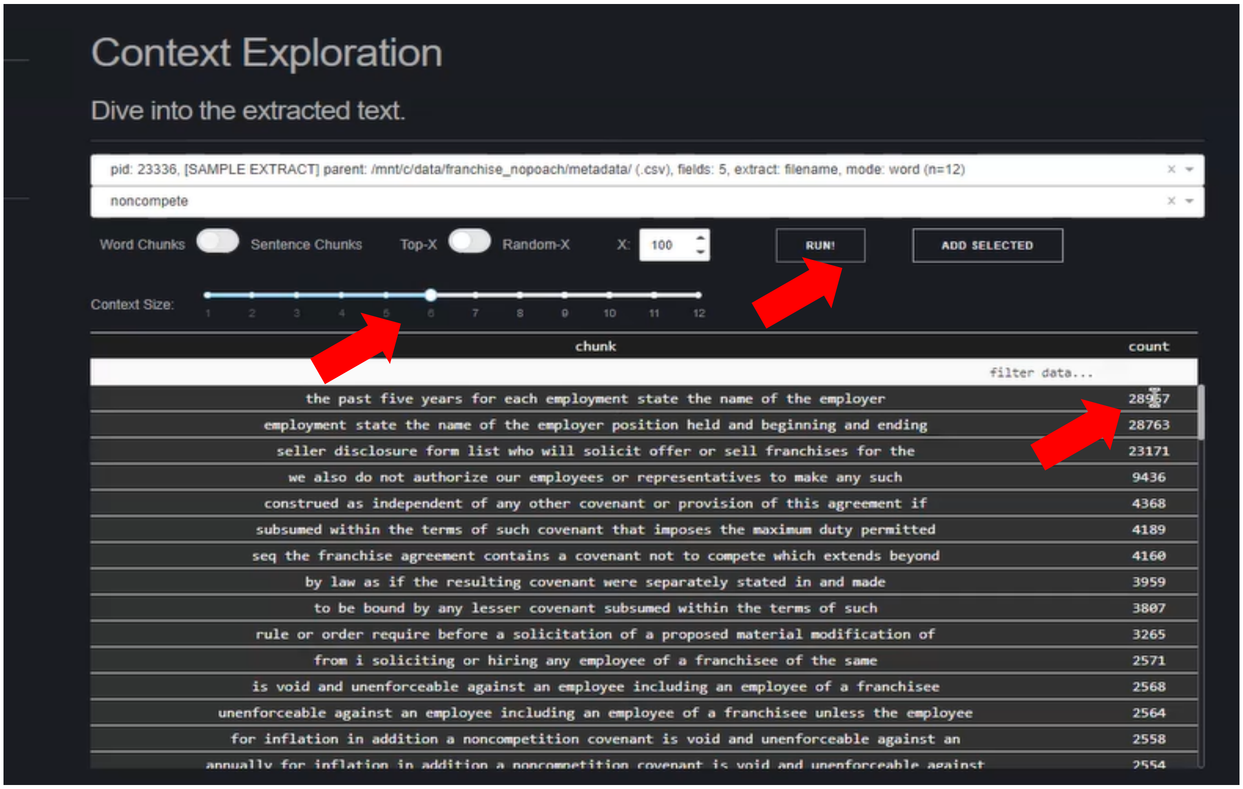
Name	Date modified	Type	Size
ext_procs	9/9/2022 11:30 AM	JSON File	1 KB
nb_train_procs	7/25/2022 2:09 PM	JSON File	1 KB
rf_train_procs	7/25/2022 2:09 PM	JSON File	1 KB

ext_procs - Notepad

```
[
    {
        "pid": "5332",
        "name": "[FULL EXTRACT] parent: C:\\DataPhD\\franchise
        "start": "09/09/2022, 11:11:21",
        "elapsed": "983.49+",
        "status": "Finished",
        "n": 6,
        "mode": "word"
    }
]
```

Exploring context of extracted text data

a) "Context Exploration" allows users to identify the most common NGrams in the extracted text that contain keywords within a specific keyword.



Context Exploration

Dive into the extracted text.

pid: 23336, [SAMPLE EXTRACT] parent: /mnt/c/data/franchise_nopoach/metadata/ (.csv), fields: 5, extract: filename, mode: word (n=12)

noncompete

X: 100

Context Size: 6

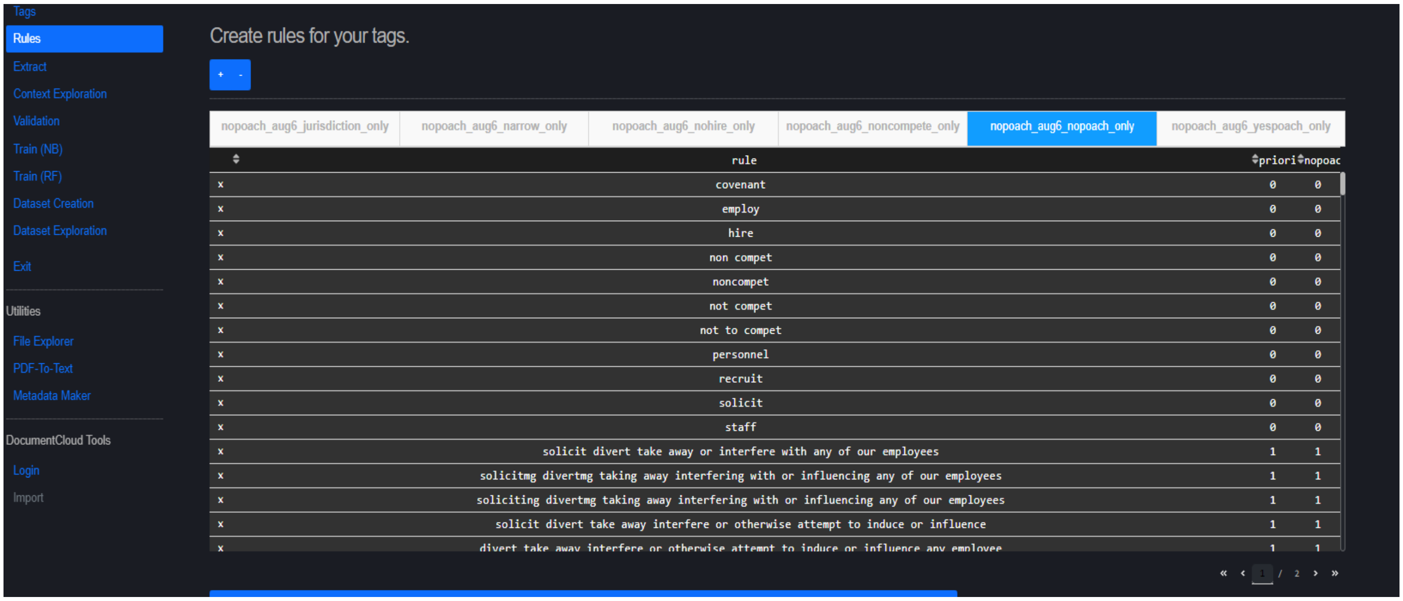
chunk	count
the past five years for each employment state the name of the employer	28957
employment state the name of the employer position held and beginning and ending	28763
seller disclosure form list who will solicit offer or sell franchises for the	23171
we also do not authorize our employees or representatives to make any such	9436
construed as independent of any other covenant or provision of this agreement if	4368
subsumed within the terms of such covenant that imposes the maximum duty permitted	4189
seq the franchise agreement contains a covenant not to compete which extends beyond	4160
by law as if the resulting covenant were separately stated in and made	3959
to be bound by any lesser covenant subsumed within the terms of such	3807
rule or order require before a solicitation of a proposed material modification of	3265
from i soliciting or hiring any employee of a franchisee of the same	2571
is void and unenforceable against an employee including an employee of a franchisee	2568
unenforceable against an employee including an employee of a franchisee unless the employee	2564
for inflation in addition a noncompetition covenant is void and unenforceable against an	2558
annually for inflation in addition a noncompetition covenant is void and unenforceable against	2554

Rules

a) Rules are stored in csv files in the project/rules folder.

b) Each rule has a priority level and a 0/1 tag.

c) Keywords are given a priority level zero – meaning the first thing that CRAML is going to do is get all of the chunks of text that has the keyword and code it 0. Higher priority rules overwrite lower priority rules.



Create rules for your tags.

+ -
nopoach_aug6_jurisdiction_only
nopoach_aug6_narrow_only
nopoach_aug6_nohire_only
nopoach_aug6_noncompete_only
nopoach_aug6_nopoach_only
nopoach_aug6_yespoach_only

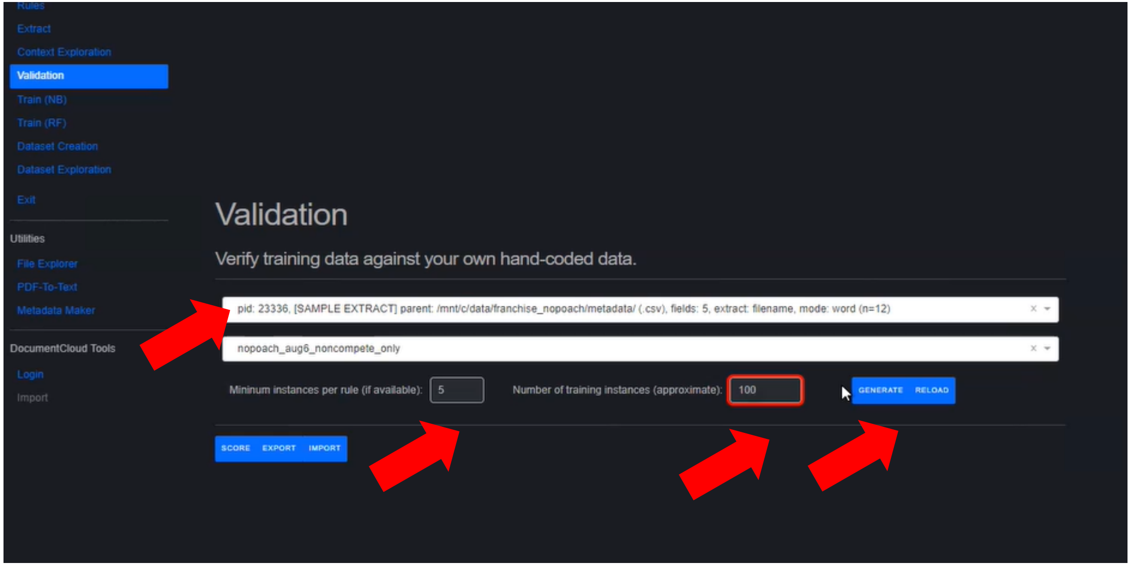
	rule	priori	nopoac
x	covenant	0	0
x	employ	0	0
x	hire	0	0
x	non compet	0	0
x	noncompet	0	0
x	not compet	0	0
x	not to compet	0	0
x	personnel	0	0
x	recruit	0	0
x	solicit	0	0
x	staff	0	0
x	solicit divert take away or interfere with any of our employees	1	1
x	solicitmg divertmg taking away interfering with or influencing any of our employees	1	1
x	soliciting divertmg taking away interfering with or influencing any of our employees	1	1
x	solicit divert take away interfere or otherwise attempt to induce or influence	1	1
x	divert take away interfere or otherwise attempt to induce or influence any employee	1	1

« ‹ 1 / 2 › »

Validation

a) To understand the interaction between the rules and the extracted text, Validation allows uers to set a minimum parameter for how much observations per rule.

b) Users can validate within CRAML and/or export validation files to third party RAs to independently validate.



Validation

Verify training data against your own hand-coded data.

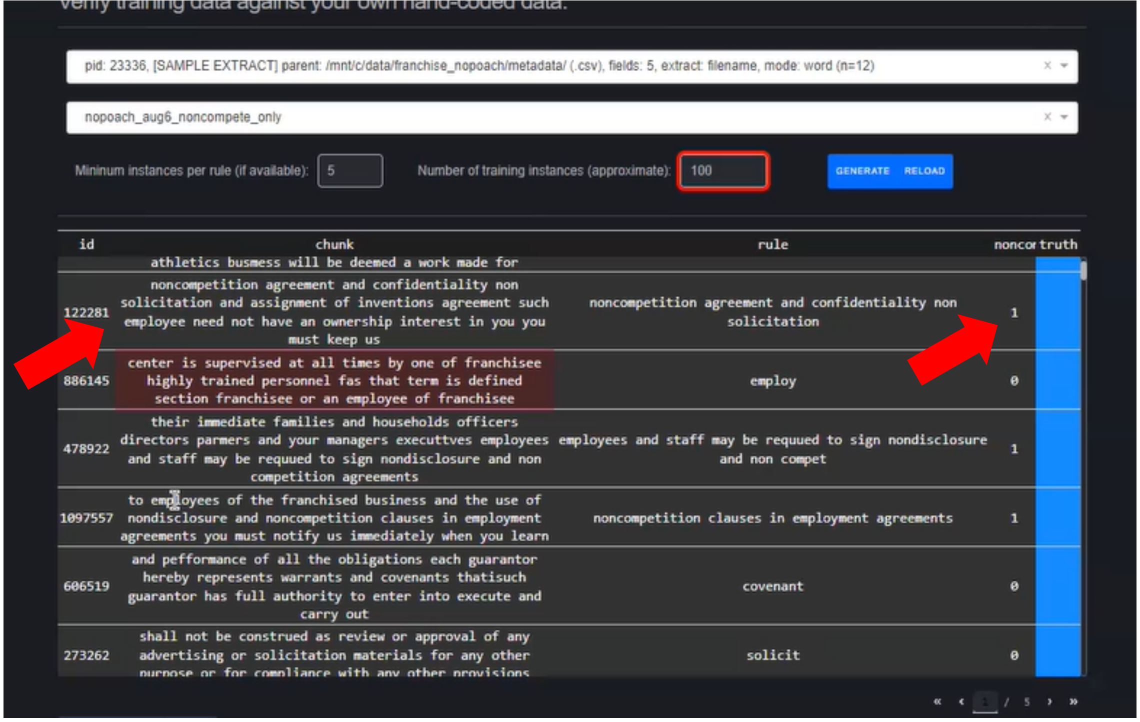
pid: 23336, [SAMPLE EXTRACT] parent: /mnt/c/data/franchise_nopoach/metadata/ (.csv), fields: 5, extract: filename, mode: word (n=12)

nopoach_aug6_noncompete_only

Mininum instances per rule (if available): 5

Number of training instances (approximate): 100

c) In CRAML, users can go over the chunks and see if they align with the rule



Verify training data against your own hand-coded data.

pid: 23336, [SAMPLE EXTRACT] parent: /mnt/c/data/franchise_nopoach/metadata/ (.csv), fields: 5, extract: filename, mode: word (n=12)

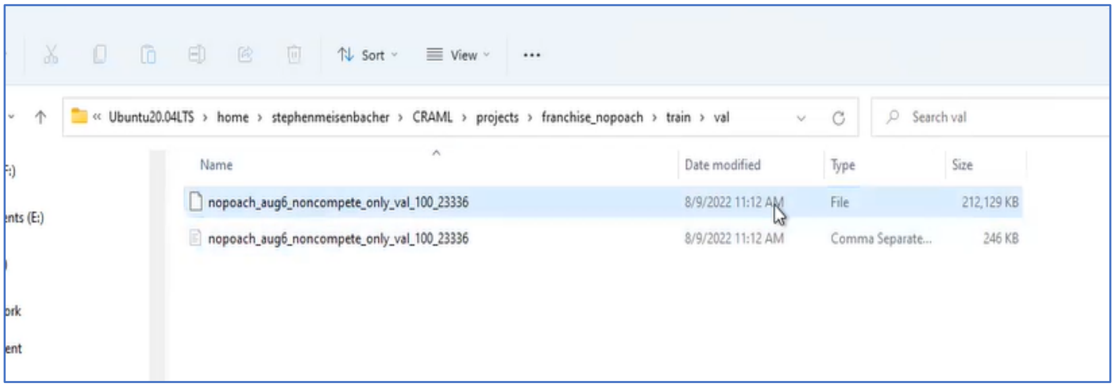
nopoach_aug6_noncompete_only

Mininum instances per rule (if available): 5

Number of training instances (approximate): 100

id	chunk	rule	noncor	truth
	athletics busmess will be deemed a work made for			
122281	noncompetition agreement and confidentiality non solicitation and assignment of inventions agreement such employee need not have an ownership interest in you you must keep us	noncompetition agreement and confidentiality non solicitation	1	
886145	center is supervised at all times by one of franchisee highly trained personnel fas that term is defined section franchisee or an employee of franchisee	employ	0	
478922	their immediate families and households officers directors parmers and your managers executtves employees and staff may be requued to sign nondisclosure and non competition agreements	employees and staff may be requued to sign nondisclosure and non compet	1	
1097557	to employees of the franchised business and the use of nondisclosure and noncompetition clauses in employment agreements you must notify us immediately when you learn	noncompetition clauses in employment agreements	1	
606519	and pefformance of all the obligations each guarantor hereby represents warrants and covenants thatisuch guarantor has full authority to enter into execute and carry out	covenant	0	
273262	shall not be construed as review or approval of any advertising or solicitation materials for any other purpose or for compliance with any other provisions	solicit	0	

d) Validation files are saved in the projects under train/val

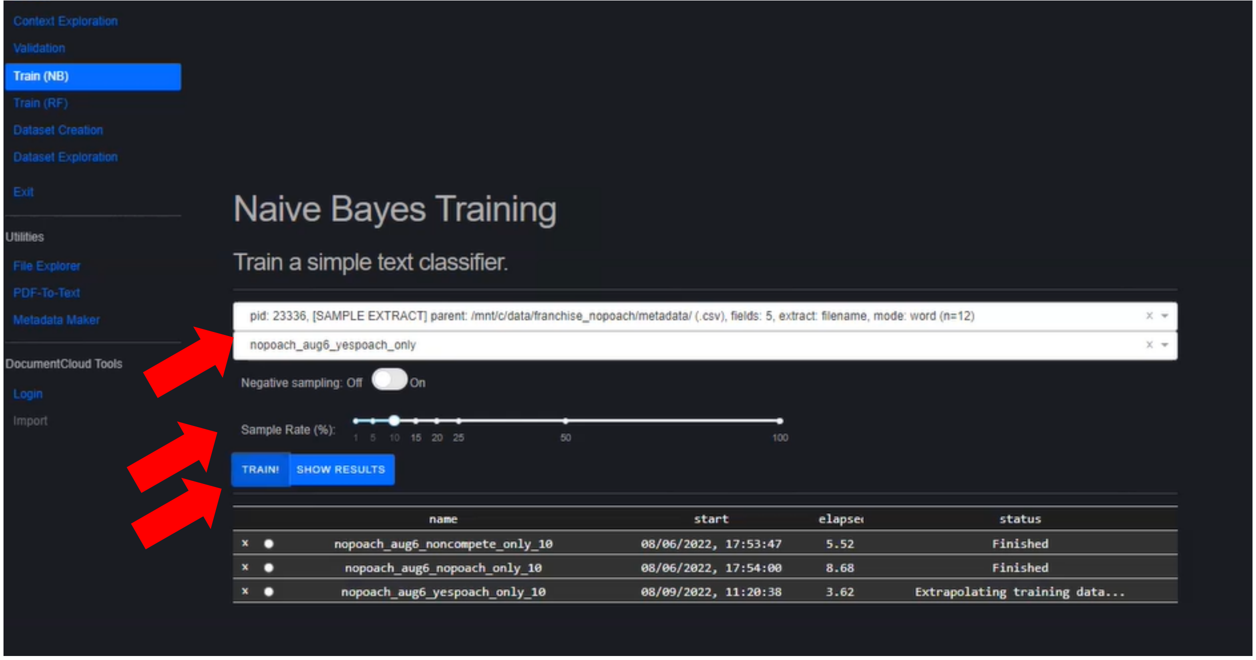


« Ubuntu20.04LTS › home › stephenmeisenbacher › CRAML › projects › franchise_nopoach › train › val

Name	Date modified	Type	Size
nopoach_aug6_noncompete_only_val_100_23336	8/9/2022 11:12 AM	File	212,129 KB
nopoach_aug6_noncompete_only_val_100_23336	8/9/2022 11:12 AM	Comma Separate...	246 KB

Building Training Data and Training ML models in CRAML

a) Under Train, select the sample, the validated rule file, and then click "train."



Naive Bayes Training

Train a simple text classifier.

pid: 23336, [SAMPLE EXTRACT] parent: /mnt/c/data/franchise_nopoach/metadata/ (.csv), fields: 5, extract: filename, mode: word (n=12)

nopoach_aug6_yespoach_only

Negative sampling: Off On

Sample Rate (%):

		name	start	elapse	status
×	●	nopoach_aug6_noncompete_only_10	08/06/2022, 17:53:47	5.52	Finished
×	●	nopoach_aug6_nopoach_only_10	08/06/2022, 17:54:00	8.68	Finished
×	●	nopoach_aug6_yespoach_only_10	08/09/2022, 11:20:38	3.62	Extrapolating training data...

b) Users may need to augment resulting training data, either with several options in CRAML or outside. If the sample has a lot of "1s" but few "0s", of vice versa, it won't perform very well.

c) Clicking on "SHOW RESULTS" will report the F-Score and relevant information on the classifier produced.

Data Creation

a) Creating a dataset from the machine learning model runs the entire pipeline of tools, and requires users:

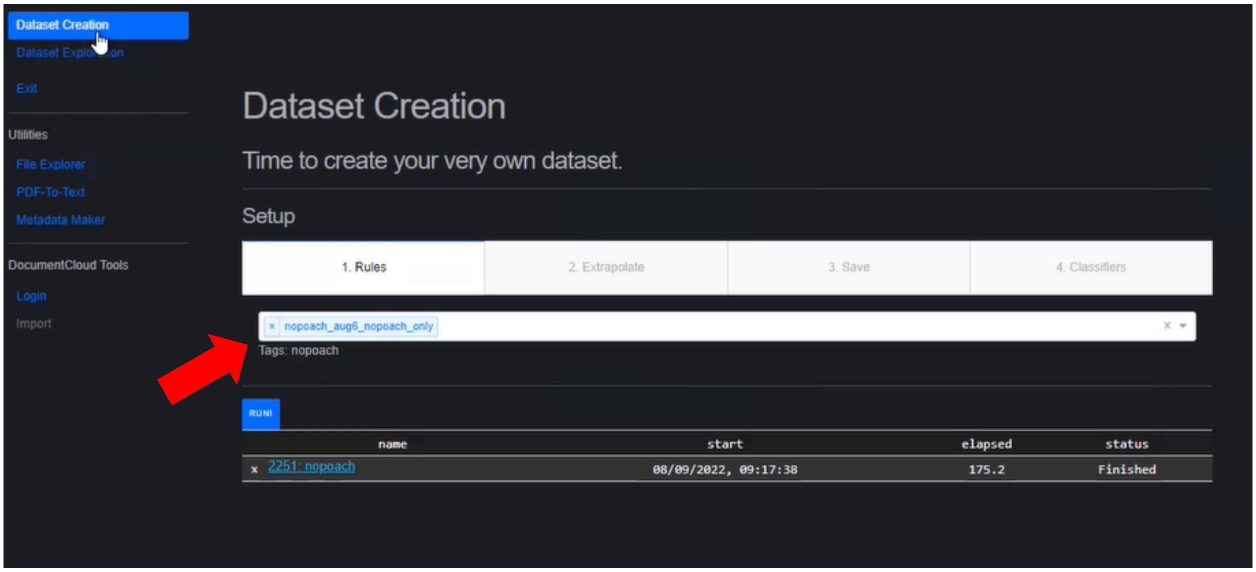
a. select a rules file,

b. choose the context window,

c. whether to keep the text or put it into a database,

d. select a trained ML classifier,

e. click “Run.”



Dataset Creation

Time to create your very own dataset.

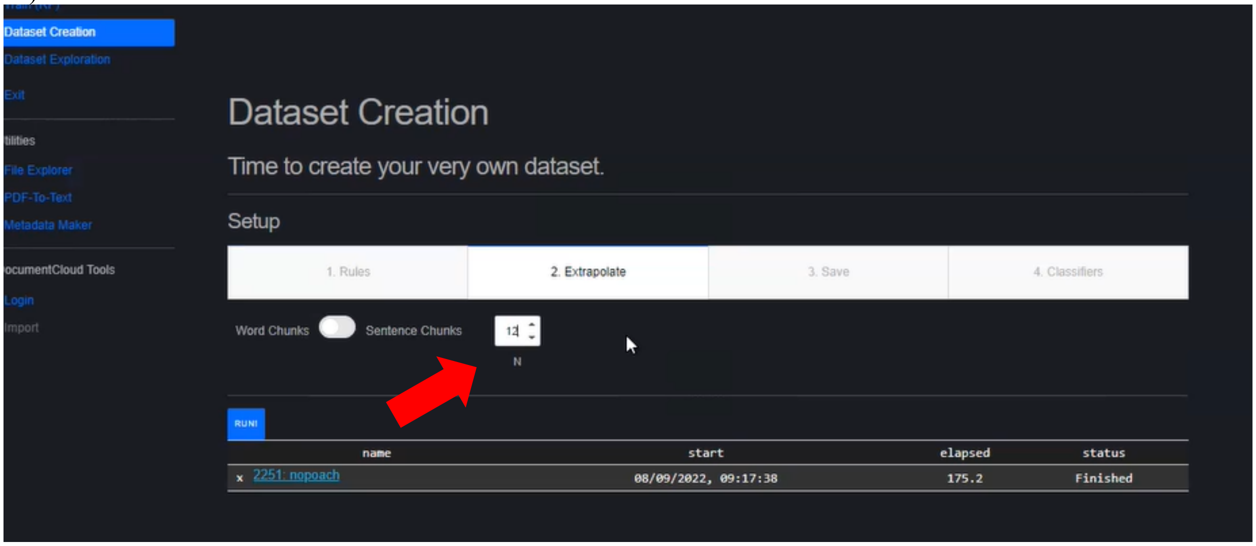
Setup

× nopoach_aug6_nopoach_only

Tags: nopoach

	name	start	elapsed	status
x	2251: nopoach	08/09/2022, 09:17:38	175.2	Finished

2)



Dataset Creation

Time to create your very own dataset.

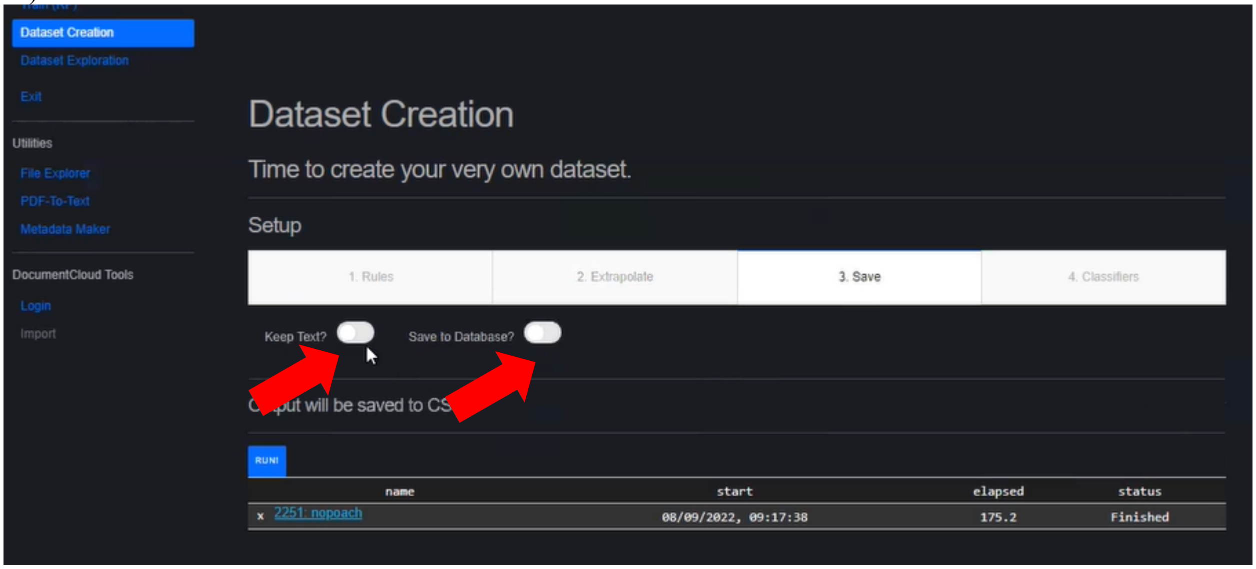
Setup

Word Chunks Sentence Chunks 12

N

	name	start	elapsed	status
x	2251: nopoach	08/09/2022, 09:17:38	175.2	Finished

3)



Dataset Creation

Time to create your very own dataset.

Setup

Keep Text? ○ Save to Database? ○

C put will be saved to CS

	name	start	elapsed	status
x	2251: nopoach	08/09/2022, 09:17:38	175.2	Finished