# Using Prodigy for Named Entity Annotation

1. Download and install the latest version of Python, see <https://www.python.org/downloads/> for details.
2. Download and install Prodigy, see <https://prodi.gy/docs/install> or as follows:
   1. Open a code terminal.
   2. Run the following pip command:

pip install prodigy -f https://XXXX-XXXX-XXXX-XXXX@download.prodi.gy

The code XXXX-XXXX-XXXX-XXXX is our academic Prodigy license. It will expire in 3 months (~02.08.2023). Please do not share it with people outside of the project.

1. Make sure you have the necessary data files in the directory from where you are going to start prodigy.

* **Dataset** for annotation, e.g.: clinical\_trials\_random\_official\_500\_from\_1497.jsonl
* Colour **configuration file** for the labels:

prodigy.json

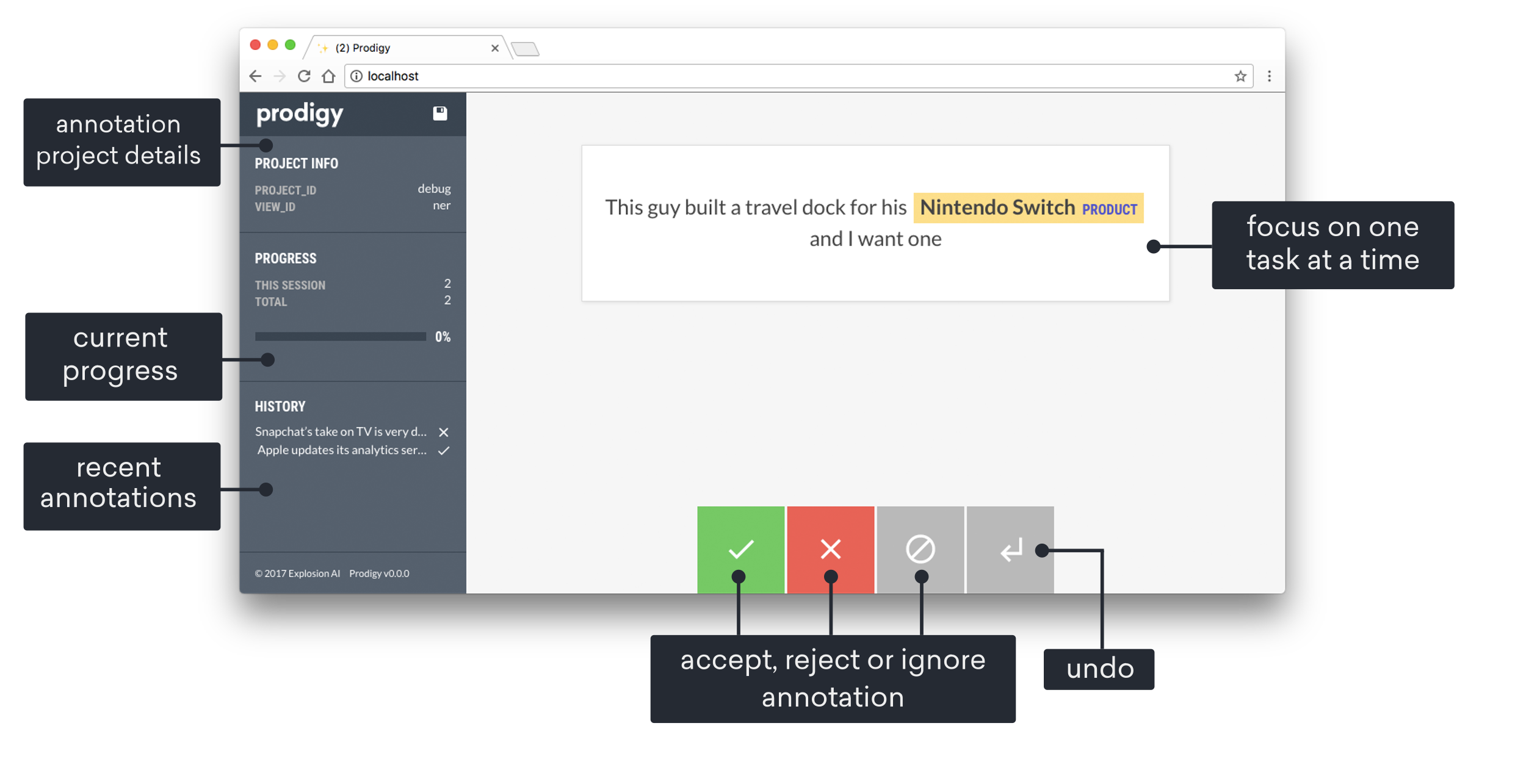
1. Run the following command in the terminal, where:

* PRODIGY\_DATASET\_NAME = the name of the local prodigy model used to identify the project, e.g. clinical\_trials\_ds
* ANNOTATION\_DATA\_FILE = .jsonl file with the texts for annotation

prodigy ner.manual <PRODIGY\_DATASET\_NAME> blank:en ./<ANNOTATION\_DATA\_FILE\_NAME>.jsonl --label DRUG,BEHAVIOURAL,SURGERY,DIET,RADIOTHERAPY,REHABILITATION,I-OTHER,DISEASE

e.g.

prodigy ner.manual simona\_ct\_ds\_100\_neuro blank:en ./prodigy\_data/ct\_random\_neurological\_100\_20230615.jsonl --label DRUG,BEHAVIOURAL,SURGICAL,RADIOTHERAPY,PHYSICAL,OTHER,CONDITION,CONTROL

1. Navigate to <http://localhost:8080>
2. You will see the Prodigy interface:
3. Use the interface to annotate the texts:
   1. Select the label you want to add.
   2. Mark the word(s) that should have the label.
   3. You can remove a label if you think you’ve made a mistake.

A screenshot of a computer

Description automatically generated with medium confidence

* 1. Accept the example.
  2. Continue until the number of target annotations has been reached.
  3. Save the dataset.

A picture containing text, screenshot, font, logo

Description automatically generated

1. Export the annotations with the following command:

prodigy db-out <PRODIGY\_DATASET\_NAME> > ./<ANNOTATED\_DATA\_FILE\_NAME>.jsonl

# Other tasks

**Create a dataset and load annotations**

prodigy db-in simona\_ct\_ds\_review ./annotated\_data/simona\_annotated\_100.jsonl

prodigy db-in amelia\_400\_review ./annotated\_data/amelia\_annotated\_400\_neuro.jsonl

**Compare two datasets for the relevant annotation sets**

prodigy review amelia\_to\_review ben\_to\_review

prodigy review neuro\_merged\_all simona\_to\_review,ben\_to\_review,amelia\_to\_review --label DRUG,BEHAVIOURAL,SURGICAL,RADIOTHERAPY,PHYSICAL,OTHER,CONDITION,CONTROL

prodigy review matching\_samples\_reviewed matched\_ncd\_samples,matched\_ncd\_samples --label DRUG,BEHAVIOURAL,SURGICAL,RADIOTHERAPY,PHYSICAL,OTHER,CONDITION,CONTROL

**See performance with increasing training data**

prodigy train-curve --ner neuro\_merged\_all --show-plot

**Create spacy datasets for training**

data-to-spacy ./spacy\_datasets --ner matched\_ncd\_samples,neuro\_merged\_all --eval-split 0.2