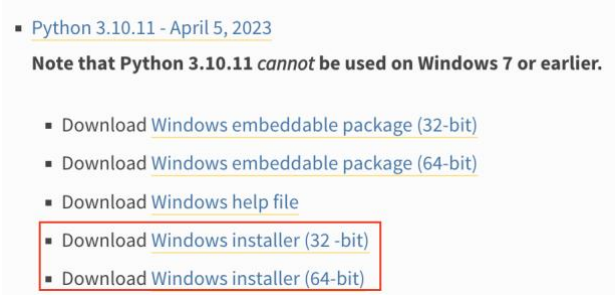
Author: Simona E. Doneva (simona.doneva@uzh.ch)
Version: 20231122

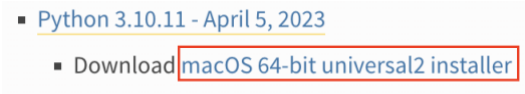
Using Prodigy for Text Categorization with custom recipe

Prerequisites

1. Download and install the latest compatible version of Python. As of 22/11/2023 this is Python 3.10.
   - For Windows: go to https://www.python.org/downloads/windows/



   - For MacOS: go to https://www.python.org/downloads/macos/



2. Setup a virtual environment
   a. virtualenv -p python3.10 prodigyenv
   b. source prodigyenv/bin/activate
3. Download and install Prodigy, see https://prodi.gy/docs/install or as follows:
   a. Open a code terminal.
   b. 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. Please do not share it with people outside of the project.

Annotation Process

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., ./prodigy/input/pilot_500_pubmed_abstracts.jsonl
   - **Configuration** file: prodigy.json
   - Terms to be **highlighted**: patterns.jsonl
   - Custom prodigy **recipe** file, e.g., recipe_textcat_patterns.py

2. Run the following command in the terminal:

python -m prodigy textcat.manual_patterns pubmed_abstracts_pilot ./input/pilot_500_pubmed_abstracts.jsonl blank:en --patterns patterns.jsonl -F ./recipe_textcat_patterns.py

Pay attention to the following parameters:

Command line parameter	Description	Comments

Author: Simona E. Doneva (simona.doneva@uzh.ch)
Version: 20231122

textcat.manual_patterns	Recipe name as defined in recipe_textcat_patterns.py	Don't change.
pubmed_abstracts_dataset_test	Reference name of the dataset that will be created by prodigy.	Can be changed by the annotator!
pubmed_abstracts.jsonl	Input data that contains the texts that will be annotated.	Make sure the .jsonl file is in the directory, from where prodigy will be run.
blank:en	Model to be used by prodigy.	Don't change.
--patterns patterns.jsonl	Reference to the keywords that should be highlighted in the interface.	Make sure the .jsonl file is in the directory, from where prodigy will be run.
-F ./recipe_textcat_patterns.py	The custom recipe to be used that will define the logic behind the interface.	Don't change.

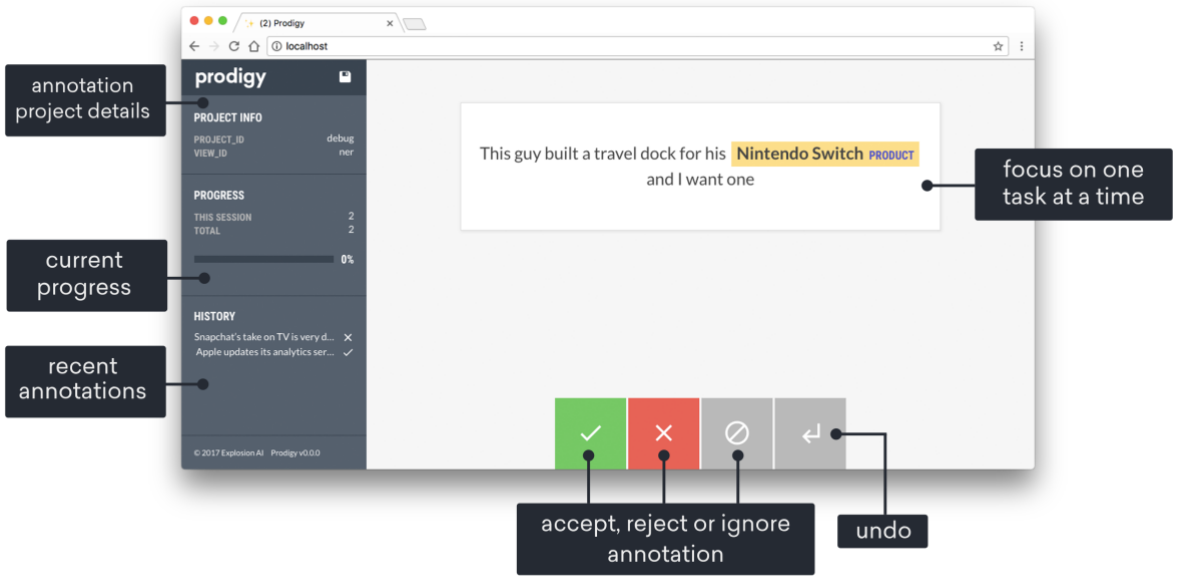
You should see the following message:

✨ Starting the web server at http://localhost:8080 ...

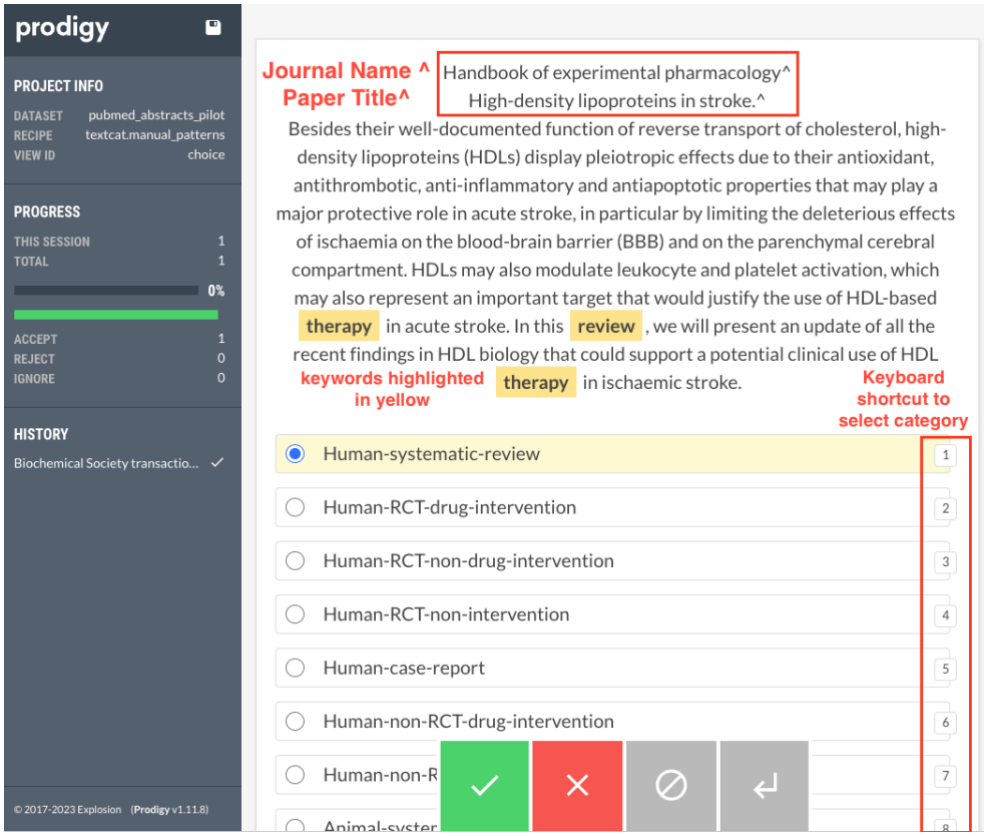
Open the app in your browser and start annotating!

3. Navigate in your web browser to http://localhost:8080
4. You will see the Prodigy interface.

The REJECT and IGNORE buttons have the same meaning for us: "this example is corrupted and not in good quality, or not relevant for my task. I want to remove this from both my training and test set." It can also just mean: "I don't know the answer and I just want to move on."



Author: Simona E. Doneva (simona.doneva@uzh.ch)
Version: 20231122



5. Use the interface to annotate the texts:

    a. Read the text.
    b. Select the fitting text category.
    c. Accept the example.
    d. Continue until the number of target annotations has been reached.
    e. Save the dataset.

6. Export the annotations with the following command:

prodigy db-out <PRODIGY_DATASET_NAME> >
./annotated_output/<ANNOTATED_DATA_FILE_NAME>.jsonl

for example

prodigy db-out pubmed_abstracts_pilot > ./annotated_output/pubmed_abstracts_pilot_500.jsonl

Author: Simona E. Doneva (simona.doneva@uzh.ch)
Version: 20231122