

R001_Fall_2023_AI_Reproducibility – Context Labelling

Instructions for setting up the development environment

1. You may use Jupyter Notebook (recommended)/Google Colab with any Python 3.8+ stable environment.
2. Clone the below repository.
<https://github.com/lamps-lab/ai-reproducibility>
3. required folder structure.

```
Citing_Paper_contexts
|----- RS_001_MLRC_2022_01.json
|----- ...
|----- RS_149_ICDAR_2018_16.json
citation_context_counts_for_cited_papers.json
R001_Citation_Context_Labelling_Shared.ipynb
```

Instructions for using jupyter notebook

1. Locate the directory where folder structure is set up
2. Run the jupyter notebook "R001_Citation_Context_Labelling_Shared.ipynb"
3. when interruptions occur - look for the "temp_labelling.json" file within the Citing_Paper_contexts_labels" folder for backup of the current working file

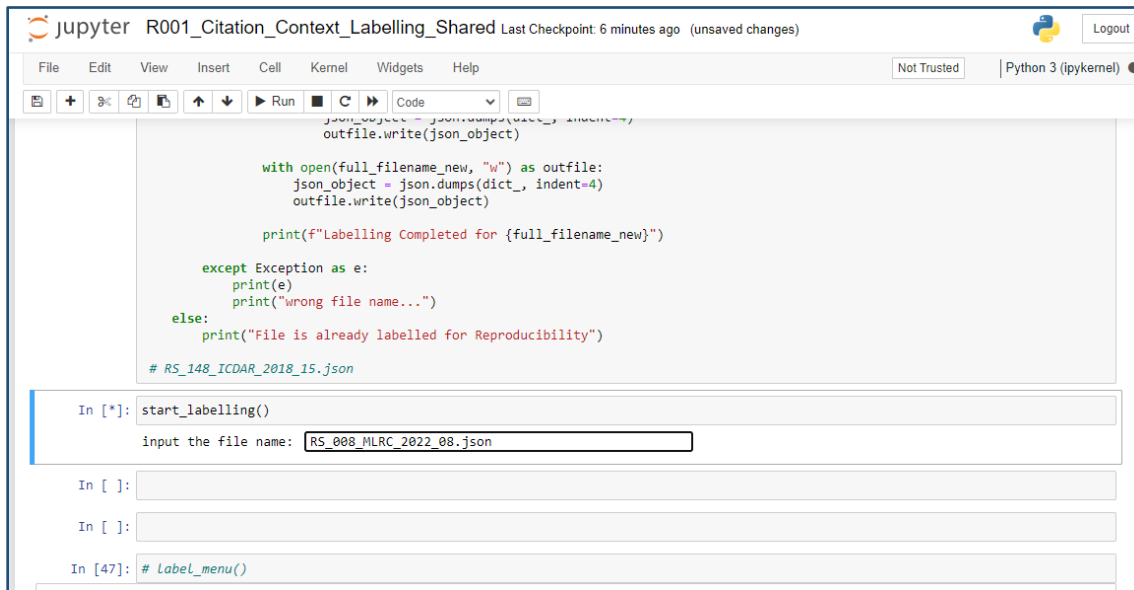
Instructions for labelling

1. Run all the cells up to the `start_labelling()` cell
2. Enter a .json filename from the list `available_files_for_labelling` (ex: `RS_148_ICDAR_2018_15.json`)
3. use one of the five values (-2, -1, 0, 0.5, 1) as the input label score
4. use the below guideline when choosing the label score.

Score	Label	Definition	Example
1	Strong	Containing words or phrases that indicate an effort to reproduce, replicate, or repeat the experiments or obtain consistent, quantitative, or qualitative results	We present a replication study of BERT pretraining (Devlin et al., 2019), which includes a careful evaluation of the effects of hyperparameter tuning and training set size.
0.5	Weak	The software was used for preprocessing or comparison, but it is unclear whether an attempt to reproduce results was conducted	We investigated open information extraction methods such as REVERB (Fader et al., 2011) and OLLIE (Mausam et al., 2012).
0	Neutral	Simply mentioning the cited paper without any attempts to run the implementation or verify the results	Pre-training methods that learn directly from raw text have revolutionized NLP over the last few years (Devlin et al., 2018).
-1	P-NR	An unsuccessful attempt to redo the experiments due to the unavailability of resources - Process not reproducible	Dataset source or location was not provided in [38].
-2	O-NR	An unsuccessful attempt to reproduce the reported results - Outcome not reproducible	Because we could not obtain the same F1-score using the code provided, [10]...

Example Screenshots

Screenshot 01: Input a file name for labelling



The screenshot shows a Jupyter Notebook titled "R001_Citation_Context_Labelling_Shared". The code in the cell is as follows:

```
json_object = json.dumps(dict_, indent=4)
outfile.write(json_object)

with open(full_filename_new, "w") as outfile:
    json_object = json.dumps(dict_, indent=4)
    outfile.write(json_object)

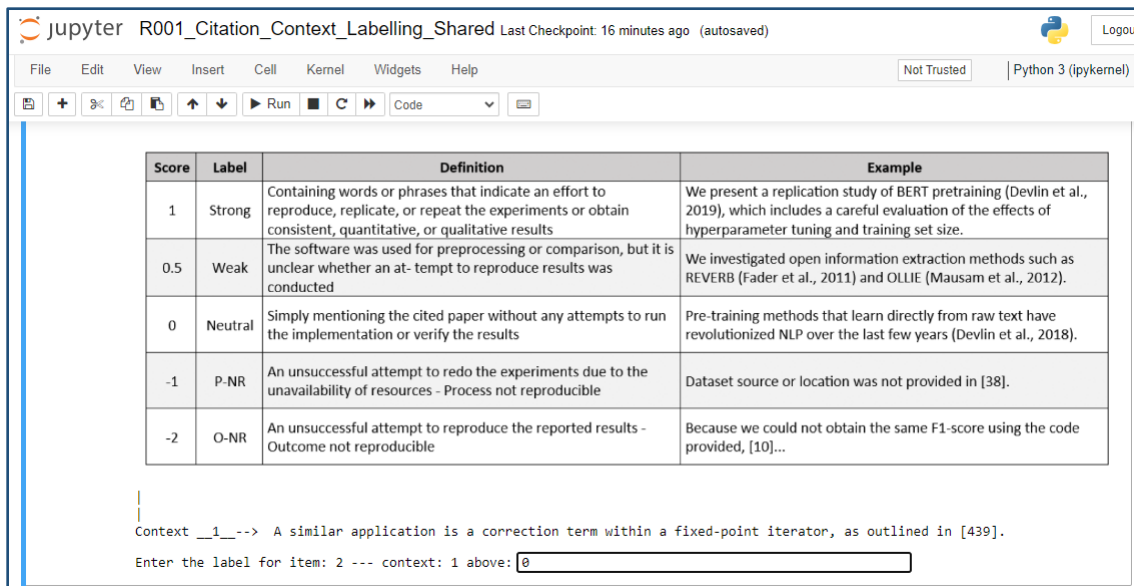
print(f"Labelling Completed for {full_filename_new}")

except Exception as e:
    print(e)
    print("wrong file name...")
else:
    print("File is already labelled for Reproducibility")

# RS_148_ICDAR_2018_15.json
```

Below the code, the input prompt "In [*]: start_labelling()" is followed by a text input field containing "RS_008_MLRC_2022_08.json". Below this, there are two empty input fields for "In []:" and "In []:". At the bottom, the prompt "In [47]: # Label_menu()" is shown.

Screenshot 02: Input a label score for a single citation context



The screenshot shows a Jupyter Notebook titled "R001_Citation_Context_Labelling_Shared". The code in the cell is as follows:

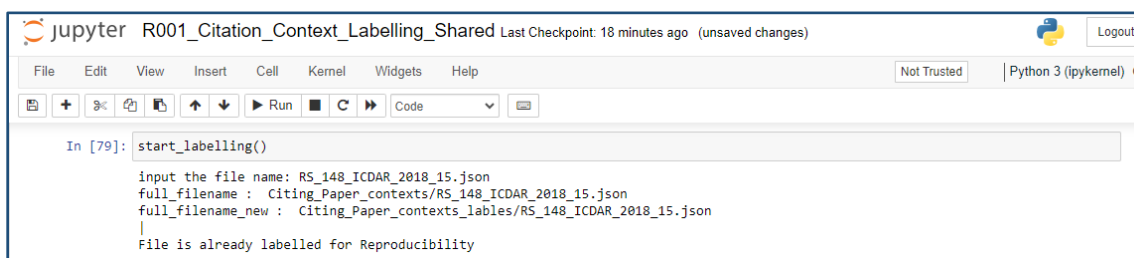
```
Context __1__--> A similar application is a correction term within a fixed-point iterator, as outlined in [439].
Enter the label for item: 2 --- context: 1 above: 0
```

Below the code, there is a table with 4 columns: Score, Label, Definition, and Example.

Score	Label	Definition	Example
1	Strong	Containing words or phrases that indicate an effort to reproduce, replicate, or repeat the experiments or obtain consistent, quantitative, or qualitative results	We present a replication study of BERT pretraining (Devlin et al., 2019), which includes a careful evaluation of the effects of hyperparameter tuning and training set size.
0.5	Weak	The software was used for preprocessing or comparison, but it is unclear whether an at-tempt to reproduce results was conducted	We investigated open information extraction methods such as REVERB (Fader et al., 2011) and OLLIE (Mausam et al., 2012).
0	Neutral	Simply mentioning the cited paper without any attempts to run the implementation or verify the results	Pre-training methods that learn directly from raw text have revolutionized NLP over the last few years (Devlin et al., 2018).
-1	P-NR	An unsuccessful attempt to redo the experiments due to the unavailability of resources - Process not reproducible	Dataset source or location was not provided in [38].
-2	O-NR	An unsuccessful attempt to reproduce the reported results - Outcome not reproducible	Because we could not obtain the same F1-score using the code provided, [10]...

Below the table, the input prompt "Enter the label for item: 2 --- context: 1 above:" is followed by a text input field containing "0".

Screenshot 03: Ignoring already labelled files



The screenshot shows a Jupyter Notebook titled "R001_Citation_Context_Labelling_Shared". The code in the cell is as follows:

```
In [79]: start_labelling()

input the file name: RS_148_ICDAR_2018_15.json
full_filename : Citing_Paper_contexts/RS_148_ICDAR_2018_15.json
full_filename_new : Citing_Paper_contexts_lables/RS_148_ICDAR_2018_15.json
File is already labelled for Reproducibility
```