# Variables and parameters in *configs.py*

This document describes the variables and parameters that are set in *configs.py.*

Essential settingsThese variables and parameters are *required* to be set in order for the pipeline to function correctly.

* **run\_name:** the run\_name is the identifier by which you will find the data and results corresponding to the current run. To illustrate, the run\_name *test\_pipeline* will lead to the raw text being stored under *raw\_papers\_test\_pipeline,* and the processed text under *processed\_papers\_test\_pipeline.*
* **path\_data\_folder:** the path where you want the raw text and the processed text to be stored. Note that it is recommended to specify a different path than the *path\_raw\_pdf’s,* which is the path to where you stored the pdf’s that you want to extract triplets from.
* **path\_texts\_entropy\_calculation:** the path to the texts that are used for the calculation of cross-categorical term entropy. If you use the standard entropy values provided in the repository, or you already calculated entropy values, this may be left empty.
* **path\_raw\_pdf:** the path to the folder with pdf’s that you want to extract triplets from. It is recommended for this to be a different folder than *path\_data\_folder.*
* **path\_fewshot\_examples:** the path to the fewshot\_examples that are used for extracting the triplets. Note that the fewshot\_examples should be adjusted based on the target domain for a good performance of the LLM.
* **path\_metadata:** the path to the arXiv metadata.
* **path\_memory\_estimate:** the path to the memory estimates of the LLM. Note that the memory estimates are specific to the model type, transformers version, torch version, flash-attention version and GPU. For new memory estimations, one can use the Textwiz library: ***INSERT LINK***
* **num\_cpus:** the number of CPU’s to use for multiprocessing.

## Optional settings

### General

* **seed:** the seed used for reproducibility
* **clear\_before\_run:** if you set this value to True, during the *load\_data.py* phase, all files in the results and data folder will be erased, to have a fresh run. Note that it is recommended to leave this option to False, and simply have a fresh run by changing the *run\_name.*
* **logging\_level\_cmd:** the level of logs that are displayed in the command window. When using slurm this is the level of logs displayed in the .err file.
* **logging\_level\_file:** the level of logs that are displayed in the log file, which can be found in the results folder, in a sub-folder named *logs.*

### Data loading