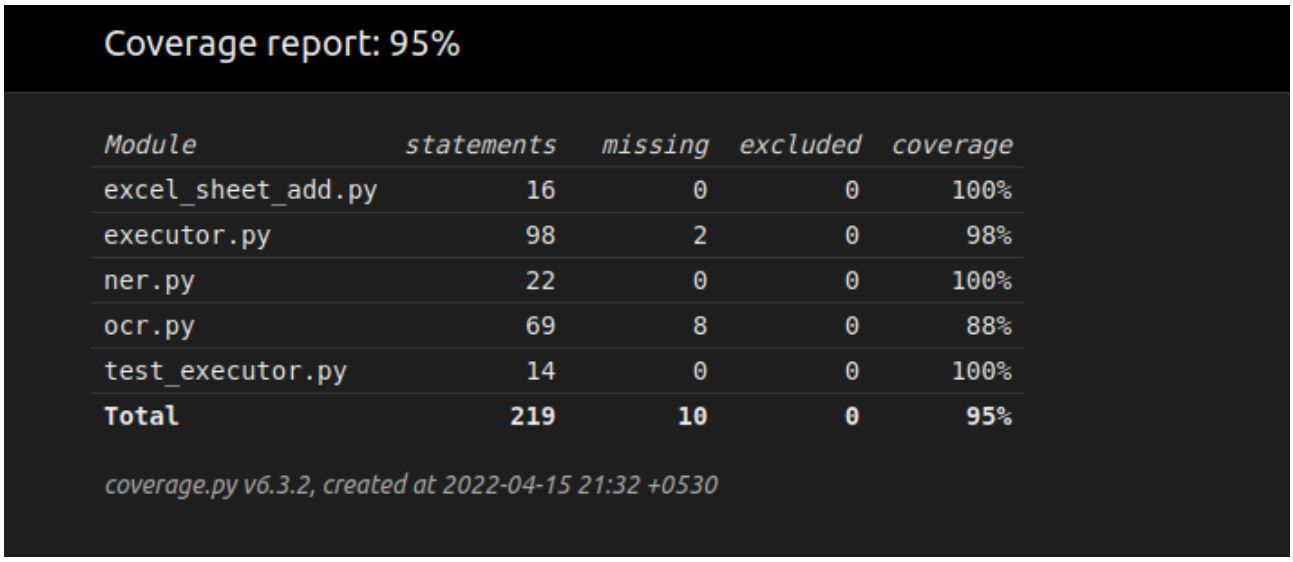
DESIGN DOC

Coverage



Coverage report: 95%

Module	statements	missing	excluded	coverage
excel_sheet_add.py	16	0	0	100%
executor.py	98	2	0	98%
ner.py	22	0	0	100%
ocr.py	69	8	0	88%
test_executor.py	14	0	0	100%
Total	**219**	**10**	**0**	**95%**

coverage.py v6.3.2, created at 2022-04-15 21:32 +0530

Design Decisions

- To find title of the book, used pytesseract to get the text corresponding to the maximum height of the surrounding box for the text detected.
- To find author and publisher, used spacy with "en_core_web_lg", where used “PERSON” label for author and “ORG” label for publishers
- Used xlrd to write values to excel from within python

Description of Files

- excel_sheet_add.py : has functions to add new rows to excel shhet
- executor.py: is the main file that calls all other files to perform the task
- ner.py: Does nlp to obtain author and publishers
- ocr.py: Uses pytesseract to obtain text from image
- runner.py: wrapper around executor.py, it calls the main method of executor.py
- test_executor.py: has unit tests to check the code