Issue	Assignee(s)	Estimated time	Actual time	Completion status	Notes
#10: NER framework	Calin Georgescu	2 hours	1 hour	Done	Still requires some bugfix, as some of the tokens are not recognized properly
#11: Export to pdf	Rebecca Andrei	2 hours	4 hours	Done for Scenario 1, will need to be modified for Scenario 2	Required a lot of research into how to write and format text in a pdf using PDFBox, which then needed to be uploaded to the frontend to allow the user to download it. For Scenario 1, we allow the user to export the text with annotations in a .txt file (for extracting data to train an LLM), and a .pdf. The PDF does not contain paragraphs or newlines, due to some preprocessing bugs, but it does display all the text correctly within a page and creates pages automatically if it needs more than what is available.
#12: Pre processing	Paul Stan	3h	6h	Partially done	Required way more research than anticipated. Figured out out to split text into different sections, but still need to understand how to differentiate between content and captions
#13: Table parsing	Stefan Bud, Radu Vasile	8 hours	22 hours	Done	A lot of research was necessary for this task, into how tables are represented in PDF files, as well as how to identify lines in a document

- What worked well:
  - From the previous sprint, we became more accustomed to the PDFBox library, thus it was easier to come up with ideas.
  - Previous research is also very helpful in developing further issues
- What difficulties we encountered:
  - There were very few sources that we found regarding table parsing, thus it required a lot of time to brainstorm ideas, and find something that works.
  - The client was unavailable for a long period of time, thus we were unsure whether we are on the right track.
  - Creating PDF files proved to be harder than expected, due to strict formatting requirements.
- What should we adjust for the next sprint:
  - We should not wait too long for the client to confirm every development decision, and take some of them ourselves.
  - We notice that time estimates are pretty inaccurate, therefore we should try to provide better estimates.