**Instructions for invoiceReader project**

1. create new folder

2. navigate to created folder

3. open cmd by typing in the Windows Explorer search field and navigate to this folder

4. create a virtual environment typing (python -m venv <name of venv like InvoiceReader>

5. activate your virtual environment while you#re in your root folder typing

( .\ InvoiceReader\Scripts\activate )

6. install required packages creating requirements.txt and typing the required

packages like this:

numpy

pandas

scipy

matplotlib

pillow

opencv-python

jupyter

7. now install the packages typing (pip install -r requirements.txt)

8. install tesseract-OCR, set environment variable and

9. see these pages and read for more info

* (https://tesseract-ocr.github.io/tessdoc/Data-Files.html)
* (https://tesseract-ocr.github.io/tessdoc/Installation.html)

only in case, if you want to do OCR opertions in other languages than English.

10. consider to download the tessdata, it's needed for language specific trained data in our case deu.traineddata from

* (https://github.com/tesseract-ocr/tessdata\_best)

11. now install pytesseract typing (pip install pytesseract)

12. visit (<https://spacy.io/usage>) and set the following configuartions in ordert to download German specific spacy.

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

13. Pay attention to activating your virtual environment in your root directory or folder, means navigate to your root directory and type cmd instead oft the path in the Windows explorer and after that acivate your venv as described above.

14.now type

(pip install -U spacy ) to install spacy and then

(python -m spacy download de\_core\_news\_sm ) 🡪 to install German pre-trained model

15. convert all pdf invoices to images using (PdfToImg\_Converter1.py)

16. **Now create the 02\_Data\_Preparation.ipynb file in jupyter notebook and run the file as it is to prepare data for labeling and transforming into specific format. For more details see code comments!**

17. now transform manually your .csv file from step 16 to .txt file by saving the .csv file in (tab separated txt) format in Excel. Now put the saved .txt file to the root directory of your project where your .ipynb files are. **See data-74To95\_text.txt** for more details on that in the root directory.

18. now run the 03\_Data\_Preprocessing.ipynb as the file name explains what to do. See code comments for more info

after processing your data you need to save your TrainData and TestData in pickle format, so take at step 18

19. now visit (<https://spacy.io/usage/training>) and then go to 🡪 Quickstart and set configs like this:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

20. and then go to the right bottom corner and download this config file (base\_config.cfg) and put to your target folder

21. now open the config file in jupyter notebook and at line 3 you see this command:

( python -m spacy init fill-config ./base\_config.cfg ./config.cfg )

22. copy that command without parenthesis then click on 🡪 New and then 🡪 Terminal from jupyter notebook

23. now navigate to the folder in our case (1\_InvoiceReaderProject) where the config file is and then

24. execute the command from step 21 and done !

Ein Bild, das Text, Screenshot, drinnen enthält.

Automatisch generierte Beschreibung

25. now visit this link again (<https://spacy.io/usage/training>) and scroll down to 🡪 Preparing Training Data and copy the code preprocess.py 🡪 make new file in jupyter note and rename to preprocess.py 🡪 paste the copied code from above there and edit the code as bellow:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

26. create the folders output and data in your project folder, if you haven’t yet

27. now go to your pre-opened Terminal in jupyter notebook and run 🡪 python .\preprocess.py

This file is from step 25 to remember And done, now we have saved our data in spacy format

28. if you haven’t yet, create now the output directory and type the following command:

Python -m spacy train .\config.cfg –output .\output\ --paths.train .\data\train.spacy --paths.dev .\data\test.spacy

Make sure you see something like this output, this ist the training process for spacy NER-Model

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

29. Now run the file 04\_Predictions step by step in ordert o understand how the NER-Mode works. See code comments for more.

This file is already cleaned and save das **predictions.py** for you. No need to save as .py file for predictions.

30. Last step 🡪 just run the file 05\_Final\_predictions.ipynb by opening and running step by step in jupyter notebook