

# Receipt digitization with python-tesseract

Sabyasachi Chakrabarty, Íñigo Vicente Hernández,  
Marie Jamroszczyk, Faranak Foroughian, Nishta Agarwal, Lion Wolf

## Objective

This project aims to establish an image processing pipeline to digitize shopping receipts consisting of these steps:

- Preprocessing of receipt image data (possibly with opencv2)
- Optical Character Recognition with python-tesseract
- extract item:price pairs from receipts
- Storing, sorting and processing of information in database
- extract statistics (such as monthly non-food/food expenses; anual sweets consumption; etc.)

## Requirements

- 1) The system must handle shopping receipts from different stores:
  - different stores produce vastly different receipts which might require different processing of the images.
  - Our main goal is to include a pipeline for grocery stores such as Aldi, Edeka, Penny etc. (inclusion of other store types is an optional goal).
- 2) The system must be able to deal with wrinkles and folds or bad image quality in receipts
  - Not all images are suitable for OCR, we want to be able to process reasonable examples of image data
- 3) The extraction of items and their corresponding prices is the main goal of our project.
- 4) Sorting and possibly classifying items into groups (e.g. non-food/food/sweets etc.)
- 5) statistics about item categories and the corresponding buying behaviour over time
  - these statistics will be given as graphs (if appropriate)

## Help

The system will include a man page.

## Coding Practices

- 1) The teaching team will be able to test a shopping receipt on a working prototype.
- 2) The source code will adhere to OOP practices.
- 3) The source code will adhere to the PEP 8 styleguide.
- 4) Errors such as bad OCR will be handled appropriately.
- 5) Unit tests for all methods that read in data from external sources or user input and all methods that manipulate data will be included.