Lehrstuhl Angewandte Informatik IV
Datenbanken und Informationssysteme
Prof. Dr.-Ing. Stefan Jablonski

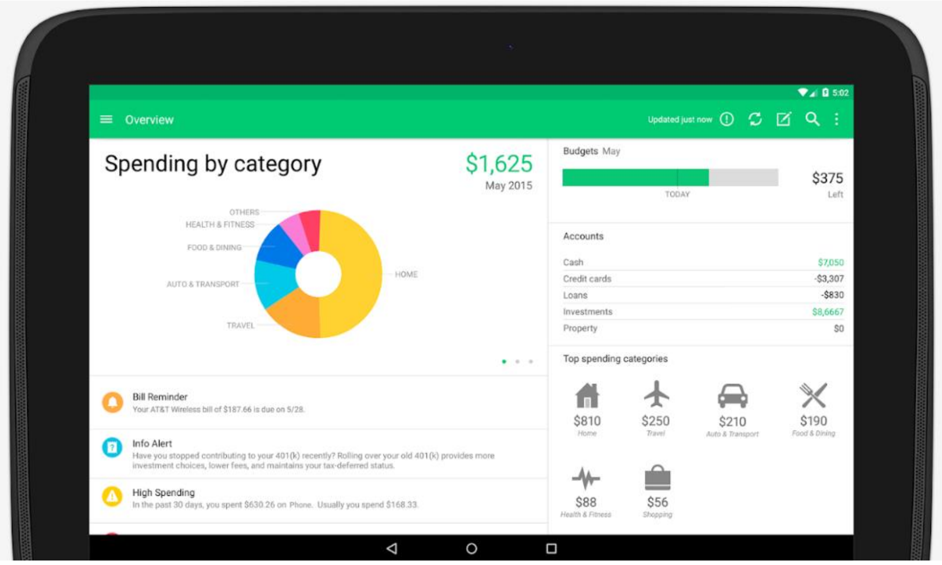
Institut für Angewandte Informatik
Fakultät für Mathematik, Physik und Informatik
Universität Bayreuth

UNIVERSITÄT
BAYREUTH

Master Thesis: Receipt analysis for enhancing financial transactions using NLP technology

Context:

Businesses track their cash flow to monitor and analyse their expenditures and revenues. This can be used for automatically generating reports for strategical realignment. Online banking and electronic payment methods like credit card, debit card, PayPal, Google Pay etc. help to gather data which is used in categorizing the bank postings. For instance, a transaction containing the keyword 'cinema' may be categorized as 'entertainment'. Figure 1 shows the app "Mint" which addresses the private usage. But you can see how different transactions are categorized and then presented in a report.



Accounts

Cash	$7,050
Credit cards	-$3,307
Loans	-$830
Investments	$8,6667
Property	$0

Bill Reminder
Your AT&T Wireless bill of $187.66 is due on 5/28.

Info Alert
Have you stopped contributing to your 401(k) recently? Rolling over your old 401(k) provides more investment choices, lower fees, and maintains your tax-deferred status.

High Spending
In the past 30 days, you spent $630.26 on Phone. Usually you spend $168.33.

Figure 1: Mint (https://play.google.com/store/apps/details?id=com.mint&hl=en_GB)

Problem Statement:

Online banking and electronic payment methods are very useful as input data for suchlike systems. However, offline payments or manual credit transfers are still in use and prevent a full-blown cash flow analysis as no digital registration is generated, but printed receipts are available. NLP (Natural Language Processing) provides techniques for analysing those receipts, to extract the single items, and to load these transactions into the financial analysis system.

An additional issue is the inaccuracy of some transactions payed with debit cards for instance. Imagine buying several items from your local supermarket. Your digital recordings will contain the single debit from the supermarket only, but not the items (2x apple, water, toilet paper, toothpaste) in particular, which could also be further categorized (food, sanitary products).

Task:

Your task is to evaluate on helpful NLP techniques initially, especially OCR (Optical Character Recognition), (context-based)[1] Spelling Correction and Word Classification (might be extended by Abbreviation Detection and Expansion if necessary). Further on, these techniques are part of a monolithic software architecture which you will design and implement in a next step. The software system is able to capture receipts (e.g. in form of a *.pdf file) and automatically extract the items registered on that receipt. Rudimentary classification algorithms will associate the items to categories. It is your choice if you build the "financial-insight-software" from scratch, or if you extend existing open-source applications. For the latter, https://github.com/awesome-selfhosted/awesome-selfhosted#money-budgeting-and-management might be a good starting point.

Goal:

The goal of the thesis is a software system for financial analysis including an NLP module for integrating paper printed receipts. Amongst others, the thesis is evaluated in terms of the architecture, usability and functional scope of the software. Also, the scientific methods in which the thesis is positioned in the state-of-the-art frontier of NLP research as well as an overview of related tools and their functional scope form the basis of the grading.

Material:

The subsequent suggested material serves as a starting point and should be extended during the course of your work.

Related Work

- https://www.researchgate.net/profile/Rafi_Ullah13/publication/324059015_Optical_Character_Recognition_Engine_to_extract_Food-items_and_Prices_from_Grocery_Receipt_Images_via_Templating_and_Dictionary-Traversal_Technique/links/5abb8355a6fdcc6c46797984/Optical-Character-Recognition-Engine-to-extract-Food-items-and-Prices-from-Grocery-Receipt-Images-via-Templating-and-Dictionary-Traversal-Technique.pdf
- https://link.springer.com/chapter/10.1007/978-3-030-20890-5_35
- http://www.diva-portal.org/smash/get/diva2:1215460/FULLTEXT01.pdf
- "Text Extraction from Bills and Invoices", H. Sidhwa et al., 2018
- "Automatic reading and interpretation of paper invoices", C. Bostrom, 2016
- "Recognition of Invoices from Scanned Documents", H. T. Ha, 2017
- "Extracting structured data from invoices", X. Holt & A, Chisholm, 2018
- "Evaluation of pre-processing techniques for the analysis and recognition of invoice documents", P. v. Zyl, 2015

[1] Context-based spelling correction uses domain knowledge like, for instance, domain-specific dictionaries.

Background Knowledge

- OCR: "Character recognition systems – A guide for Students and Practioners", M. Cheriet, 2007 (UBT library, available online)
- Spelling Correction: https://www.youtube.com/watch?v=oWsMIW-5xUc&list=PLLssT5z_DsK8HbD2sPcUIDfQ7zmBarMYv [Episodes 20 - 22]
- Ontologies:
  - https://protegewiki.stanford.edu/wiki/Main_Page
  - https://protegewiki.stanford.edu/wiki/Protege4Pizzas10Minutes
  - http://mowl-power.cs.man.ac.uk/protegeowltutorial/resources/ProtegeOWLTutorialP4_v1_3.pdf
- Abbreviation Detection and Expansion (can maybe be solved by Spelling Correction, too): http://www.cs.cmu.edu/~./wammar/pubs/icetea.pdf

Tools and Frameworks

- OCR: Tesseract > Tutorials:
  - Java: https://www.geeksforgeeks.org/tesseract-ocr-with-java-with-examples/
  - Python: https://www.pyimagesearch.com/2017/07/10/using-tesseract-ocr-python/
- Spelling Correction:
  - Easy-to-use framework: https://github.com/pragnakalp/spellcheck-using-dictionary-in-python
- Word Classification: reuse available ontologies, e.g.
  - http://www.productontology.org/doc/Supermarket and/or
  - http://www.productontology.org/doc/Grocery_store

Training and testing data

- manual search
- Maybe a tool can be used to generate receipt images that "look like scanned receipts" (cf. also the contained project report for further information; attention: further tools might be needed to add noise to the image): https://github.com/billstark/receipt-scanner