Call for Bachelor Thesis

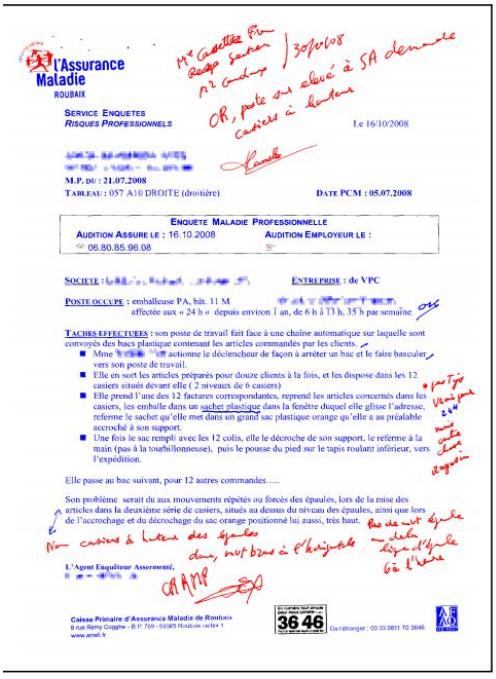
FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

Handwritten and Printed Text Separation in Historical Documents

Objective of this work:

With the increase of digitized documents, automatic document analysis has become extremely important. The presentation of administrative documents for public introduces varieties of document types, content, quality and structure. Fundamentally speaking, documents can be skewed, noisy, overlapped with graphics, i.e., lines, unconstrained annotations, stamps.

In this thesis, existing technologies for visual semantic analysis of unstructured data will be investigated. The dataset of the work consists of 5595 images of tax forms from 1988. In order to provide a textual representation of the documents automatic text and pattern recognition processes have to be applied. An optical character recognition (OCR) system recognizes either printed or handwritten text. Hence, the task of the thesis is to separate machine printed text from handwritten text in scanned documents before feeding it to an OCR system. As a first step, the images will be preprocessed (e.g. cropping, noise filtering). Then the images will be segmented (e.g. line/word segmentation). As the last contribution the segments will be fed into a deep learning binary classifier to be separated into handwritten and printed text segments.


L'Assurance Maladie
ROUBAIX

SERVICE ENQUETES
RISQUES PROFESSIONNELS

Le 16/10/2008

M.P. DU : 21.07.2008
TABLEAU : 057 A10 DROITE (droitière)
DATE PCM : 05.07.2008

ENQUETE MALADIE PROFESSIONNELLE
AUDITION ASSURE LE : 16.10.2008
AUDITION EMPLOYEUR LE :
06.80.85.96.08

SOCIETE :
ENTREPRISE : de VPC

POSTE OCCUPE : emballeuse PA, bât. 11 M
affectée aux « 24 h » depuis environ 1 an, de 6 h à 13 h, 35 h par semaine

TACHES EFFECTUEES : son poste de travail fait face à une chaîne automatique sur laquelle sont convoyés des bacs plastique contenant les articles commandés par les clients.
- Mme actionne le déclencheur de façon à arrêter un bac et le faire basculer vers son poste de travail.
- Elle en sort les articles préparés pour douze clients à la fois, et les dispose dans les 12 casiers situés devant elle (2 niveaux de 6 casiers)
- Elle prend l'une des 12 factures correspondantes, reprend les articles concernés dans les casiers, les emballe dans un sachet plastique dans la fenêtre duquel elle glisse l'adresse, referme le sachet qu'elle met dans un grand sac plastique orange qu'elle a au préalable accroché à son support.
- Une fois le sac rempli avec les 12 colis, elle le décroche de son support, le referme à la main (pas à la tourbillonneuse), puis le pousse du pied sur le tapis roulant inférieur, vers l'expédition.

Elle passe au bac suivant, pour 12 autres commandes.....

Son problème serait du aux mouvements répétés ou forcés des épaules, lors de la mise des articles dans la deuxième série de casiers, situés au dessus du niveau des épaules, ainsi que lors de l'accrochage et du décrochage du sac orange positionné lui aussi, très haut.

L'Agent Enquêteur Assermenté,

Caisse Primaire d'Assurance Maladie de Roubaix
3646

The project work will be supervised by **Prof. Dr. Harald Sack, Tabea Tietz and Oleksandra Vsesviatska**, **Information Service Engineering at Institute AIFB, KIT, in collaboration with FIZ Karlsruhe.**

Keywords:
Knowledge Graphs, Cultural Heritage, NLP

Pre-requisites:
Knowledge of Programming with Python.

Contact persons:
Tabea Tietz
tabea.tietz@fiz-karlsruhe.de
Oleksandra Vsesviatska
oleksandra.vsesviatska@fiz-karlsruhe.de

aifb
Institute of
Applied Informatics and
Formal Description Methods
http://www.aifb.kit.edu/

KIT
Karlsruhe Institute of Technology