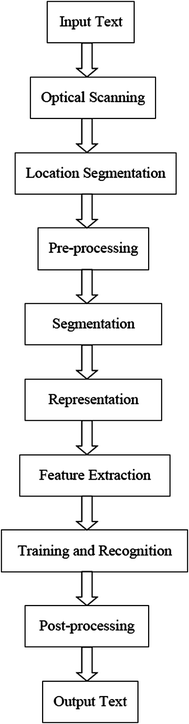
# Úvod do problematiky

Je to náročné

## Získanie textu z naskenovaných dokumentov

Na to, aby sme dokumenty mohli akokoľvek analyzovať, potrebujeme ich mať uložené v takom formáte, aby boli strojom čitateľné. Musia byť teda uložené ako text. Všetky dokumenty, ktoré v našej práci spracúvame sú uložené vo formáte PDF. Avšak iba malá časť je uložená ako text. Vo väčšine prípadov sa jedna o obrázok – naskenovaný dokument. Preto prvým krokom je získanie textu z obrázku pomocou technológie OCR. Táto technológia nie je ťažiskom práce, no je jej nevyhnutnou súčasťou.

OCR alebo optické rozoznávanie znakov sa zaoberá problémom klasifikácie optických vzorov v digitálnom obraze do príslušných alfanumerických či iných znakov [1]. Vďaka tomu sa uľahčí ich ukladanie, keďže namiesto množstva pixelov môžeme uložiť jeden alebo viac znakov. Ešte väčším benefitom je ale to, že takto uložený text môžeme prehľadávať, analyzovať a ľahko upravovať.

Tento problém bol na začiatku výskum rozpoznávania vzorov považovaný za jednoduchý [2]. Znakov je pomerne malé množstvo a ľahko sa s nimi pracuje. Problém prichádza ak sa neobmedzíme iba na latinku a jeden font – v tom prípade množstvo vzorov narastá. Ešte väčší problém ale spôsobuje ručne písaný text a kvalita spracovaného obrazu. Problematika rozpoznávania znakov sa ukázala byť komplikovanejšou. V súčasnosti je už ale dostupných množstvo komerčných aplikácii vykonávajúcich OCR na pomerne dobrej úrovni [1]. Tieto aplikácie sú ale stále veľmi závisle na kvalite vstupného obrazu, preto obzvlášť pri spracovaní menej kvalitného obrazu sa stále nemôžu porovnávať v presnosti rozpoznania znakov s ľudskými schopnosťami [1]. V našej práci používame open source OCR engine Tesseract [3].

Obrázok : Postupnosť krokov OCR systému [OCR 2017]

Najskoršie generácie OCR systémov sa spoliehali predovšetkým na techniky rozpoznávania vzorcov a spracovania obrazu [2], veľké zlepšenie ale prinieslo zapojenie metód umelej inteligencie [1]. V nedávnej minulosti prinieslo ďalšie zlepšenie, podobne ako v iných oblastiach, využitie umelých neurónových sieti (ANN). [1]

Obrázok 1 ukazuje celú postupnosť štandardných krokov OCR systémov. Za hlavné kroky OCR by sa dali označiť [4]

* analýza rozloženia dokumentu
* rozpoznanie znakov
* post-processing

### Analýza rozloženia dokumentu

Shafait považuje analýzu rozloženia dokumentu (layout analysis) za prvý krok, ktorý OCR systém vykonáva [4]. Chaudhuri tento krok nazýva *location segmentation*, no ešte pred ním uvádza krok *optical scanning* [1]. **Vysvetli čo to ten optical scanning je a čo je layout analysis.**

Pre nás je najzaujímavejší krok *pre-processing*, keďže v tomto kroku sa snažíme zlepšiť rozpoznanie textu niektorých dokumentov. Pre-processing dokumentácia Tesseract enginu, ktorý v práci používame, totiž odporúča pre zlepšenie výsledku vykonať pre-processing ešte pred spustením Tesseractu – aj keď tento OCR systém sám používa rôzne metódy spracovania obrazu [3]. ***A teraz vysvetli čo znamená ten preprocessing.***

## Klasifikácia dokumentu

### Konečný užívateľ výhod

Dokumenty, ktoré sú v našej práce analyzované vznikajú na základe zákona o registri partnerov verejného sektora a o zmene a doplnení niektorých zákonov z 25. októbra 2016[5]. Pre našu prácu je zaujímavý §11, ktorý hovorí o identifikácii konečného užívateľa výhod. Ten je ale definovaný zákonom o ochrane pred legalizáciou príjmov z trestnej činnosti a o ochrane pred financovaním terorizmu a o zmene a doplnení niektorých zákonov. Podľa tohto zákona sa ako konečný užívateľ výhod (ďalej len KUV) označí

1. „fyzická osoba ktorá skutočne ovláda alebo kontroluje právnickú osobu, fyzickú osobu – podnikateľa alebo združenie majetku, a každá fyzická osoba, v prospech ktorej tieto subjekty vykonávajú svoju činnosť alebo obchod“ [6]
2. členovia jej vrcholového manažmentu ak žiadna fyzická osoba nespĺňa kritéria aby mohla byť zapísaná podľa a) [6]
3. „Konečným užívateľom výhod je aj fyzická osoba, ktorá sama nespĺňa kritériá podľa odseku 1 písm. a), b) alebo písm. c) druhého a štvrtého bodu, avšak spoločne s inou osobou konajúcou s ňou v zhode alebo spoločným postupom spĺňa aspoň niektoré z týchto kritérií.“[6]

Kritéria na zapísanie KUV podľa a) sa líšia v závislosti od typu spoločnosti. Ak ide o právnickú osobu, ktorá nie je združením majetku ani emitentom cenných papierov, fyzická osoba je označená ako KUV podľa a) ak

1. má priamy alebo nepriamy podiel alebo ich súčet najmenej 25 % na hlasovacích právach v právnickej osobe alebo na jej základnom imaní vrátane akcií na doručiteľa, [6]
2. má právo vymenovať, inak ustanoviť alebo odvolať štatutárny orgán, riadiaci orgán, dozorný orgán alebo kontrolný orgán v právnickej osobe alebo akéhokoľvek ich člena, [6]
3. ovláda právnickú osobu iným spôsobom, ako je uvedené v prvom a druhom bode, [6]
4. má právo na hospodársky prospech najmenej 25 % z podnikania právnickej osoby alebo z inej jej činnosti. [6]

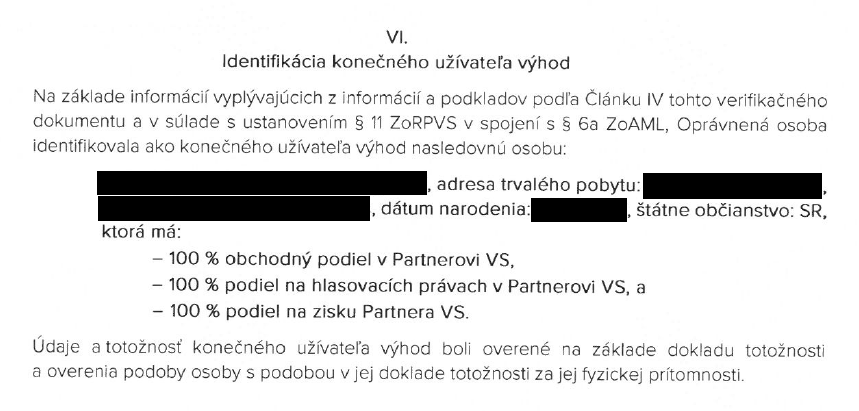
„Ak ide o fyzickú osobu – podnikateľa, fyzická osoba, ktorá má právo na hospodársky prospech najmenej 25 % z podnikania fyzickej osoby – podnikateľa alebo z inej jej činnosti“[6] bude označená za KUV.

Pokiaľ sa jedná o združenie majetku, fyzická osoba je označená ako KUV podľa a) ak

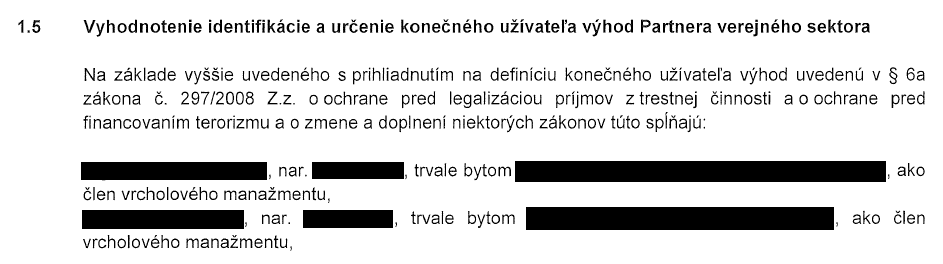
1. je zakladateľom alebo zriaďovateľom združenia majetku
2. má právo vymenovať, inak ustanoviť alebo odvolať štatutárny orgán, riadiaci orgán, dozorný orgán alebo kontrolný orgán združenia majetku alebo ich člena alebo je členom orgánu, ktorý má právo vymenovať, inak ustanoviť alebo odvolať tieto orgány alebo ich člena,
3. je štatutárnym orgánom, riadiacim orgánom, dozorným orgánom, kontrolným orgánom alebo členom týchto orgánov,
4. je príjemcom najmenej 25 % prostriedkov, ktoré poskytuje združenie majetku, ak boli určení budúci príjemcovia týchto prostriedkov; ak neboli určení budúci príjemcovia prostriedkov združenia majetku, za konečného užívateľa výhod sa považuje okruh osôb, ktoré majú významný prospech zo založenia alebo pôsobenia združenia majetku.[[1]](#footnote-1)[6]

To, kto je zapísaný ako KUV pre danú firmu je dostupné na rpvs.sk spolu s dokumentom, ktorý popisuje ako bol KUV identifikovaný a zdôvodňuje, prečo bola daná osoba zapísaná ako KUV. Zjednodušene, v našej práci sa snažíme rozdeliť spoločnosti na také

* ktorých KUV je skutočným vlastníkom
* ktorých skutočný vlastník nie je zo štruktúrovaných dát známy, keďže ako KUV je zapísaný štatutár



Obrázok : Príklad ako môže vyzerať časť dokumentu v prípade, že ako KUV bol zapísaný majiteľ



Obrázok : Príklad ako môže vyzerať časť dokumentu v prípade, že ako KUV bol zapísaný štatutár

Mám vstup zo štrukturovaných dát. Potrebujem funkciu jeKonečnýUžívateľ(Zoznam ľudí), ktorá zistí, či tí ľudia sú skutočnými majiteľmi, prípadne či ovládajú spoločnosť (teda splňajú znaky KUV podľa nejakého zákona).

Potrebujem Name Entity Recognociosion. Potrebujem nahradiť zapísaného KUV ako KUV -> nech sa model učí potvrdiť/poprieť výrok „KUV je skutočným vlastníkom“. Pre jeKonečnýUžívateľ(jeKonečnýUžívateľ)

[1] A. Chaudhuri, K. Mandaviya, P. Badelia, and S. K. Ghosh, “Optical Character Recognition Systems,” *Optical Character Recognition Systems for Different Languages with Soft Computing*, pp. 9–41, 2017, doi: 10.1007/978-3-319-50252-6\_2.

[2] S. Mori, C. Y. Suen, and K. Yamamoto, “Historical review of OCR research and development,” *Proceedings of the IEEE*, vol. 80, no. 7, pp. 1029–1058, Jul. 1992, doi: 10.1109/5.156468.

[3] “Tesseract documentation,” *Tesseract OCR*. http://tesseract-ocr.github.io/ (accessed Jan. 21, 2021).

[4] F. Shafait and R. Smith, “Table Detection in Heterogeneous Documents,” *Detecting tables in document images is important since not only do tables contain important information, but also most of the layout analysis methods fail in the presence of tables in the document image. Existing approaches for table detection mainly focus on detecting tables in single columns of text and do not work reliably on documents with varying layouts. This paper presents a practical algorithm for table detection that works with a high accuracy on documents with varying layouts (company reports, newspaper articles, magazine pages, . . . ). An open source implementation of the algorithm is provided as part of the Tesseract OCR engine. Evaluation of the algorithm on document images from publicly available UNLV dataset shows competitive performance in comparison to the table detection module of a commercial OCR system.*, p. 9.

[5] Slov-lex, “315/2016 Z.z. - Zákon o registri partnerov verejnéh...,” *Slov-lex*. https://www.slov-lex.sk/pravne-predpisy/SK/ZZ/2016/315/20170201 (accessed Jan. 20, 2021).

[6] Slov-lex, “297/2008 Z.z. - Zákon o ochrane pred legalizáciou p...,” *Slov-lex*. https://www.slov-lex.sk/pravne-predpisy/SK/ZZ/2008/297/ (accessed Jan. 20, 2021).

1. Pre viac informácii o KUV a spôsobe identifikácie pozri 297/2008 Z.z § 6a [↑](#footnote-ref-1)