Název práce: Nástroj pro převod PDF na text
Autor: Jonáš Bujok
Katedra / Ústav: Ústav formální a aplikované lingvistiky (32-UFAL)
Vedoucí bakalářské práce: Mgr. Jan Raab, Ústav formální a aplikované lingvistiky (32-UFAL)

Abstrakt: V této práci je podrobně rozebrán postup extrakce textových informací z PDF (Portable Document Format) souborů a navrhnut, popsán a implementován program pro tento účel. Práce se zaměřuje hlavně na středoevropské jazyky. Kromě programu a jeho popisu jsou zde pak informace o objektové struktuře, syntaxi a logice PDF formátu nutné pro správné pochopení principu hledání textu v PDF souboru. Dále jsou zde rozebrány filtry, fonty a všechny další PDF objekty, které takový program musí umět zpracovat. Také se tato práce zabývá metodami a možnostmi vylepšení funkčnosti, rychlosti, paměťové náročnosti, spolehlivosti a univerzálnosti použití programu.