Univerzita Karlova v Praze

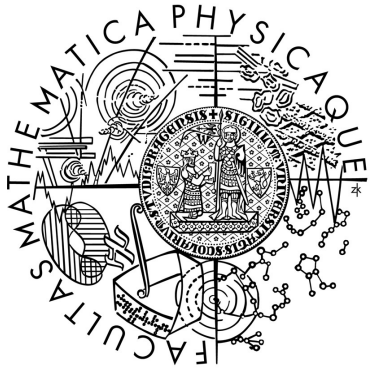
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE

2011 **Jonáš Bujok**

Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Jonáš Bujok

Nástroj pro převod PDF na text

Ústav formální a aplikované lingvistiky (207. • 32-UFAL)

Vedoucí bakalářské práce: Mgr. Jan Raab

Studijní program: Informatika
Studijní obor: obecná informatika (IOI)

Praha 2011

[Vzor: Na tomto místě mohou být napsána případná poděkování (vedoucímu práce, konzultantovi, tomu, kdo zapůjčil software, literaturu apod.)]

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona.

V Praze dne ..... podpis

Název práce: Nástroj pro převod PDF na text

Autor: Jonáš Bujok

Katedra / Ústav: Ústav formální a aplikované lingvistiky (207. • 32-UFAL)

Vedoucí bakalářské práce: Mgr. Jan Raab, Ústav formální a aplikované lingvistiky (207. • 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. 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 a univerzálnosti použití programu.

Klíčová slova: PDF, text, převaděč

Title: A Tool for Transformation of PDF to Text

Author: Jonáš Bujok

Department: Institute of Formal and Applied Linguistics (207. • 32-UFAL)

Supervisor: Mgr. Jan Raab, Institute of Formal and Applied Linguistics (207. • 32-UFAL)

Abstract: In this thesis we described an extraction procedure of text information from PDF (Portable Document Format) files. We designed, described and implemented program for this purpose. Besides the program and it's description the thesis contains information about PDF format object structure, it's syntax and logic necessary for proper understanding of text searching principles in PDF file. We also discussed filters, fonts and all other PDF Objects that the program need to process. This thesis also deals with methods and possibilities of improving program's functionality, speed, memory usage an universality of usage.

Keywords: PDF, text, convertor

Obsah

Předmluva

Vážení čtenáři této bakalářské práce, rád bych se s Vámi v tomto krátkém úvodu podělil o mé myšlenky a motivace pro vytvoření práce na tak zdánlivě jednoduché téma, jako je vytažení textu z PDF souboru. PDF soubory jsou v dnešní době velmi rozšířené a používají se hojně v počítačích, mobilních zařízeních i tiskárnách. Ovšem programů a knihoven, které s PDF pracují, je relativně málo na to, jak je tento formát populární. Sice jejich počet se zvětšuje, ale v době, kdy vznikala tato práce ještě ve formě ročníkového projektu, se jen těžko hledal program, který by dobře zvládl takový základní úkon, jako je vytáhnout z PDF souboru text. Programy, které byly k dispozici, měly různé problémy a často nevygenerovaly správně většinu speciálních znaků neanglických jazyků včetně českých. Jako hlavní problém bych viděl velikou složitost tohoto formátu. Což lze ostatně poznat už z toho, že jeho dokumentace (PDF verze 1.7 z roku 2006) má přes 1300 stran. (Podrobněji se na téma složitosti vyjadřuji v první kapitole této práce.) Dnes se situace o něco změnila a existují již celkem spolehlivé programy pro extrakci textu ze souborů tohoto formátu. Některé budou pravděpodobně i mnohem lepší, než program, který jsem napsal já, avšak tato práce může někomu dobře posloužit (a taky doufám, že poslouží) jako návod a shrnutí toho, co je k tomuto účelu potřeba vědět a naprogramovat.

[Vzor: Vlastní text bakalářské práce uspořádaný hierarchicky do kapitol a podkapitol, každá kapitola vždy na novou stranu. Je vhodné využít formátování kapitol.

Rozsah práce je udáván v normostranách. Jedna normostrana obsahuje 30 řádků po 65 znacích (řádek 30 pro číslo strany).

Písmo se doporučuje dvanáctibodové (12 pt) se standardní vzdáleností mezi řádky (řádkování 1,5). Text matematických vět se obvykle tiskne pro zdůraznění tzv. skloněným (slanted) písmem, které se podobá kurzívě. Text je zarovnán do bloku. Nový odstavec se obvykle odděluje odsazením prvního řádku.

Primárně je doporučován jednostranný tisk, přičemž oboustranný tisk není výslovně zakázán. U oboustranného tisku je třeba zohlednit úpravu správné šíře okrajů. Rub titulního listu zůstává nepotištěný.

Horní, dolní a pravý okraj 25 mm, levý okraj 40 mm. V celém textu musí být dodržena jednotná grafická úprava. Práce je tištěna na bílý papír formátu A4.

Obrázky, diagramy i tabulky se číslují tak, aby bylo možné odkázat se na ně v textu. Musí být opatřeny popisem obvyklým u vědeckých prací. Popisky tabulek, obrázků a schémat včetně jejich číslování se uvádějí pod nimi, stejnou velikostí písma jako je text práce, a pod grafickým znázorněním se uvádí pramen kurzívou a velikostí písma menší než základní text.

Zkratky použité v textu musí být vysvětleny vždy u prvního výskytu zkratky (v závorce nebo v poznámce pod čarou, jde-li o složitější vysvětlení pojmu či zkratky). Současně je připojen seznam použitých zkratek, včetně jejich vysvětlení.

Delší převzatý text jiného autora je nutné vymezit uvozovkami nebo jinak vyznačit a řádně citovat.]

1. PDF soubor

1.1. Proč vznikl

1.2. Název druhé podkapitoly v první kapitole

2. Název druhé kapitoly

2.1. Název první podkapitoly v druhé kapitole

2.2. Název druhé podkapitoly v druhé kapitole

Doslov / Závěr

[Vzor: Seznam použité literatury je zpracován podle platných standardů. Povinnou citační normou pro bakalářskou práci je ISO 690. Jména časopisů lze uvádět zkráceně, ale jen v kodifikované podobě. Všechny použité zdroje a prameny musí být řádně citovány.]

Seznam použité literatury

Seznam tabulek

Seznam použitých zkratek

PDF: Portable Document Format

[Vzor: Přílohy k bakalářské práci, existují-li (různé dodatky jako výpisy programů, diagramy apod.). Každá příloha musí být alespoň jednou odkazována z vlastního textu práce. Přílohy se číslují.]

Přílohy

[Vzor: Pevná zadní deska bakalářské práce – **není součástí elektronické verze**. Na vnitřní stranu zadní desky je vlepena obálka s CD s elektronickou verzí práce ve formátu PDF (maximální velikost 850 MB), případně přílohy, nejsou-li součástí práce.]