Rozpoznávání a indexování knižních obsahů

Jan Pokorný

NTK
50°6'14.083"N, 14°23'26.365"E
Národní technická knihovna
National Library of Technology

Východiska

- Zpřístupňování fondu knihovny pomocí centrálního vyhledávače, např. discovery systému
- Vyhledávání: podle identifikačních údajů
  podle věcných údajů
- Věcné údaje: klasifikační systémy
  tezaury
  předmětová hesla
  volně tvořená klíčová slova

Východiska

- Z údajů bibliografických záznamů budujeme vyhledávací indexy
- Indexy slouží nejen pro účely přímého vyhledávání, ale i pro generování dynamických filtrů pro zpracování vrácených výsledků (fasety a další filtry)

Kdy standardní popis nestačí

- Názvové údaje často příliš obecné nebo nevyjadřující obsah díla
- Manuální tvořený věcný popis je závislý na úsudku katalogizátora (subjektivita, nepochopení) a nemůže postihnout všechna témata obsahu díla
- Když není k dispozici fulltext nebo fulltextové indexování není vhodné/možné

Kdy standardní popis nestačí: příklad

Požadavek: Hledám knihu o *práci se souborovým systémem v jazyce Python*

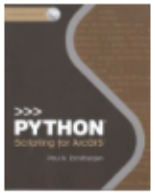
V katalogu podle názvu: nenalezeno

V katalogu podle věcného popisu: nenalezeno

Úspěšné pouze hledání obecně o jazyce Python.

Kdy standardní popis nestačí: příklad

Python : pro hackery a reverzní inženýrství
Autor: Seitz, Justin Vydáno: 2009
Umístění: regál 6D026
Vypůjčeno
Kniha



Python : scripting for ArcGIS
Autor: Zandbergen, Paul A., 1968- Vydáno: 2013
Umístění: regál 6A034
Vypůjčeno
Kniha

Python : pocket reference
Autor: Lutz, Mark Vydáno: 2010
Umístění: VŠCHT ústavy
Dostupný
Kniha

Python : referenční programátorská příručka
Autor: Beazley, David M. Vydáno: 2002
Umístění: regál 6D026
Dostupný
Kniha

Kdy standardní popis nestačí: příklad

Začínáme programovat v jazyce Python /

Hlavní autor:	Harms, Daryl D.
Další autoři:	McDonald, Kenneth, 1964-
Formát:	Kniha
Jazyk:	čeština angličtina
Vydáno:	Brno : Computer Press, 2008
Vydání:	2. opr. vyd.
ISBN:	8025121615
Předmět:	Python programovací jazyky objektově orientované programování softwarové inženýrství příručky
Tagy:	Programování (1), Python (1) Přidat

Kdy klasický popis nestačí: příklad

Kde hledat, když standardní popis nestačí?

Informační aparát knihy:

- obsah knihy (TOC)
- rejstřík knihy

Rejstřík knihy – příliš velká atomizace výrazů, příliš závislé na kontextu, ne vždy je přítomen

Obsah knihy – ideální pro získání představy o tématech, o kterých kniha pojednává

Kdy klasický popis nestačí: příklad

Obsah knihy

- Obsah knihy tvoří sám autor!
- Obsah knihy tvoří názvy kapitol, které autor odborného textu vytváří jako klíčová slova
- Název kapitoly reprezentuje obsah kapitoly

Kdy standardní popis nestačí: příklad

Využití obsahu knihy k indexování

Metoda: vytěžování klíčových slov z naskenovaných obsahů knih

(odlišné od vytěžování klíčových slov z fulltextu)

Možno určovat i váhu klíčového slova – na kolika stránkách se o tématu píše (využití počtu stran v obsahu)

Využití obsahu knihy k indexování

Postup:

1. naskenování obsahu knihy (obraz bude použit i jako náhled)
2. OCR s rozpoznáním bloků textu (tvar a kontext)
3. rozlišení textových a číselných bloků
4. eliminace cizích bloků a stopslov (označení bloků a kapitol, jiný text na stránce, apod.)
5. textová analýza s důrazem na sledování závislostí kapitol a podkapitol (kontext je velmi důležitý)
6. získání klíčových slov a převod do základních tvarů
7. uložení do bibliografického záznamu nebo jiného kontejneru

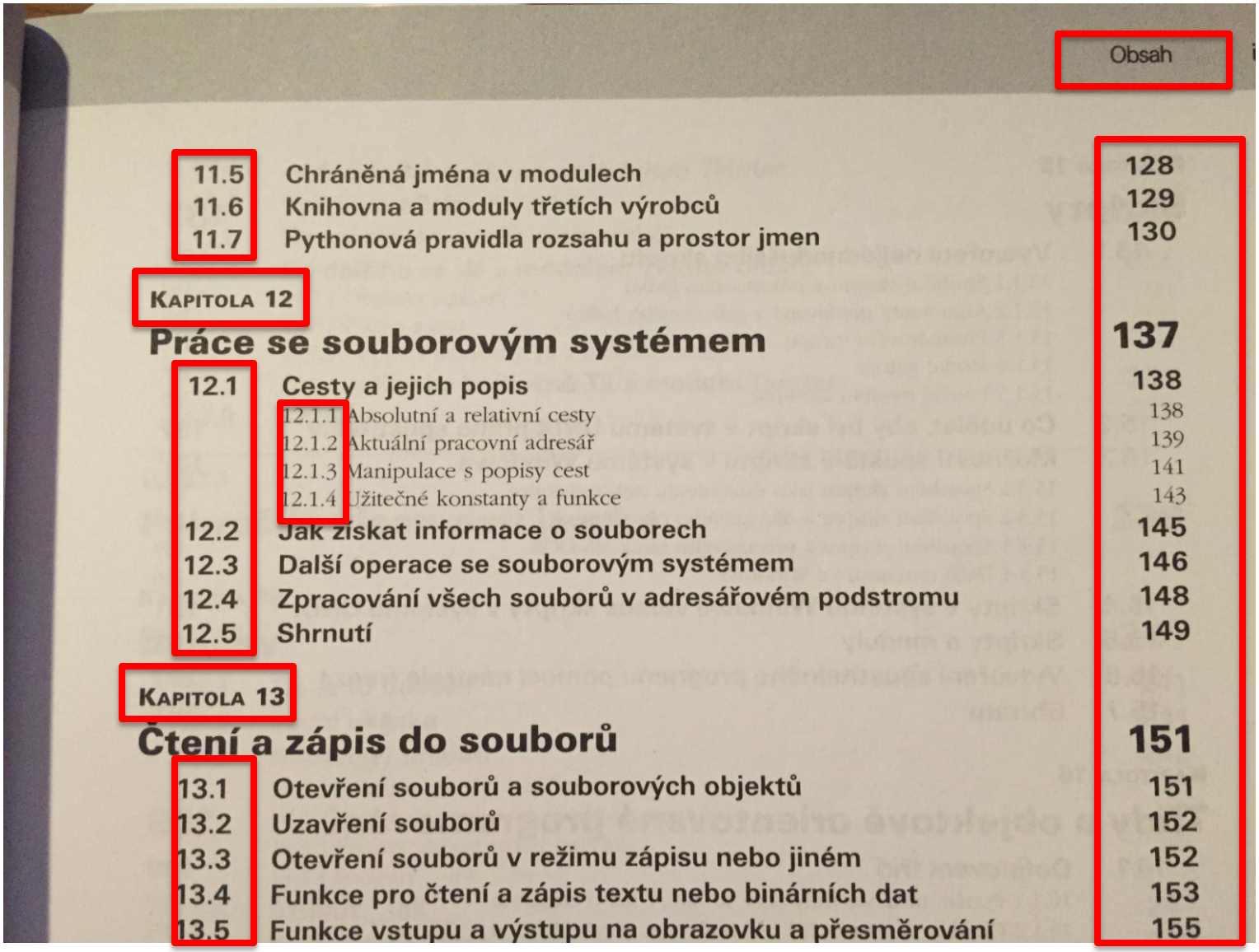
Problémy

Problém rozložení textu, typografie a designu

Knihy mají obsah řešen různě z hlediska jeho struktury, členění do kapitol a podkapitol, layoutu na stránce, typografického a grafického provedení.

Na stránce s obsahem je často i jiný text a mnoho textových prvků, které je třeba odfiltrovat.

Problémy

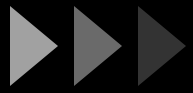


Obsah

Problémy

OBSAH

Problémy



Problémy

Problém rozpoznání závislostí

Typicky vztah *kapitola-podkapitola*, ale často i *název knihy-kapitola-podkapitola*.

Mnohdy značně ztíženo typograficky či rozložením na stránce.

Problémy

Moduly – jaké? V kontextu názvu *Python* → moduly v jazyce Python

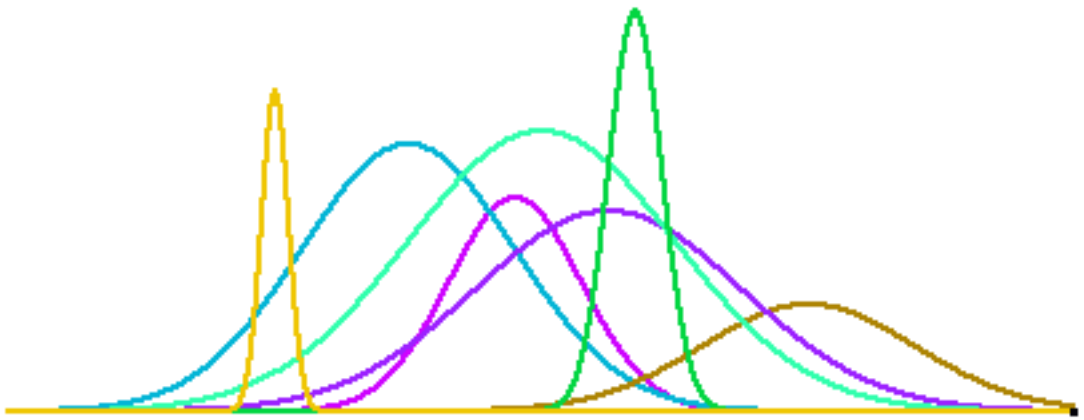
Základy – jaké? V kontextu kapitoly Moduly → Základy modulů v jazyce Python

Úložiště

- do bibliografického záznamu (do pole 505 v MARC nebo do vybraného prvku v XML) → výrazné zvětšení záznamu
- mimo bibliografický záznam do pomocného úložiště s prolinkováním v rámci lokálního systému
- do externího systému pro využití více knihovnami (podobně jako se pracuje s obálkami knih)

Další využití – detekce trendů

- Detekce trendů – seznam témat, o kterých se za určité období nejvíce píše
- Jaká klíčová slova se v daném období nejčastěji vyskytují v daných oborech → časová osa s kulminujícími klíčovými slovy, žebříčky nejpopulárnějších témat



Realizace

- projekt NTK v rámci programu VISK 2015
- na podzim 2015 k dispozici beta verze
- propojení s AKS pro ukládání do záznamu
- indexování v AKS nebo v discovery systému, který AKS sklízí
- využitelné jako pomocný index při vyhledávání či podle požadavků konkrétních katalogů a vyhledávačů