

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](#)  Přidat
(1)



▶▶▶ 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

		Obsah
11.5	Chráněná jména v modulech	128
11.6	Knihovna a moduly třetích výrobců	129
11.7	Pythonová pravidla rozsahu a prostor jmen	130
KAPITOLA 12		
Práce se souborovým systémem		137
12.1	Cesty a jejich popis	138
	12.1.1 Absolutní a relativní cesty	138
	12.1.2 Aktuální pracovní adresář	139
	12.1.3 Manipulace s popisy cest	141
	12.1.4 Užitečné konstanty a funkce	143
12.2	Jak získat informace o souborech	145
12.3	Další operace se souborovým systémem	146
12.4	Zpracování všech souborů v adresářovém podstromu	148
12.5	Shrnutí	149
KAPITOLA 13		
Čtení a zápis do souborů		151
13.1	Otevření souborů a souborových objektů	151
13.2	Uzavření souborů	152
13.3	Otevření souborů v režimu zápisu nebo jiném	152
13.4	Funkce pro čtení a zápis textu nebo binárních dat	153
13.5	Funkce vstupu a výstupu na obrazovku a přesměrování	155



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
11.5	Chráněná jména v modulech	128
11.6	Knihovna a moduly třetích výrobců	129
11.7	Pythonová pravidla rozsahu a prostor jmen	130
KAPITOLA 12		
Práce se souborovým systémem		137
12.1	Cesty a jejich popis	138
	12.1.1 Absolutní a relativní cesty	138
	12.1.2 Aktuální pracovní adresář	139
	12.1.3 Manipulace s popisy cest	141
	12.1.4 Užitečné konstanty a funkce	143
12.2	Jak získat informace o souborech	145
12.3	Další operace se souborovým systémem	146
12.4	Zpracování všech souborů v adresářovém podstromu	148
12.5	Shrnutí	149
KAPITOLA 13		
Čtení a zápis do souborů		151
13.1	Otevření souborů a souborových objektů	151
13.2	Uzavření souborů	152
13.3	Otevření souborů v režimu zápisu nebo jiném	152
13.4	Funkce pro čtení a zápis textu nebo binárních dat	153
13.5	Funkce vstupu a výstupu na obrazovku a přesměrování	155



Problémy

OBSAH

Úvod	5	Kresba štětcem	92
[I.] TECHNKA KRESBY		Frotáž	94
Jak začít	11	Monotyp	95
Portrét	15	Rytina do nitrolaku a škrábací papír	95
Zátiší	32	Konečná úprava kreseb, adjustace, uložení	98
Krajina	34	Slova závěrem, ale i do začátku	101
Kresba lidského těla	41	SLOVNÍČEK	
Typy kreslířů	52	UMĚLECKÝCH SLOHŮ A SMĚRŮ	109
[II.] MATERIÁL A TECHNICKÉ POSTUPY		SLOVNÍČEK	
(I.) Kreslicí prostředky se širokou stopou	59	ODBORNÝCH VÝRAZŮ A NÁZVŮ	121
(II.) Prostředky s užší stopou (hrotové)	72	Doporučená literatura	133
(III.) Materiály	76	Minimum znalosti o devadesáti mistrech	
Lavírovaná kresba	87	světové a naši kresby	137



Problémy

- 130 □ Stavitelé věží
- 136 □ Metafora a filosofie jazyka

- 148 □ Výtahy z Vesmíru
- 149 □ Svět a světlo
- 151 □ Vytrhávání z kontextu
- 153 □ Nebezpečnost umění
- 155 □ Kámen a strom
- 156 □ Vidění vidění
- 158 □ Přírodní, přirozené a umělé
- 160 □ Jaký příběh, takový svět



Problémy

Problém rozpoznání závislostí

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

Mnohdy značně ztíženo typograficky či rozložením na stránce.



Problémy

5. <i>Moduly</i>	118
Proč používat moduly?	118
Základy	119
Moduly jsou jmenné prostory	120
import	123
Opětovné načítání modulů	124
Drobnosti	127
Oblíbené problémy	133
Shrnutí	137
Cvičení	138

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

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

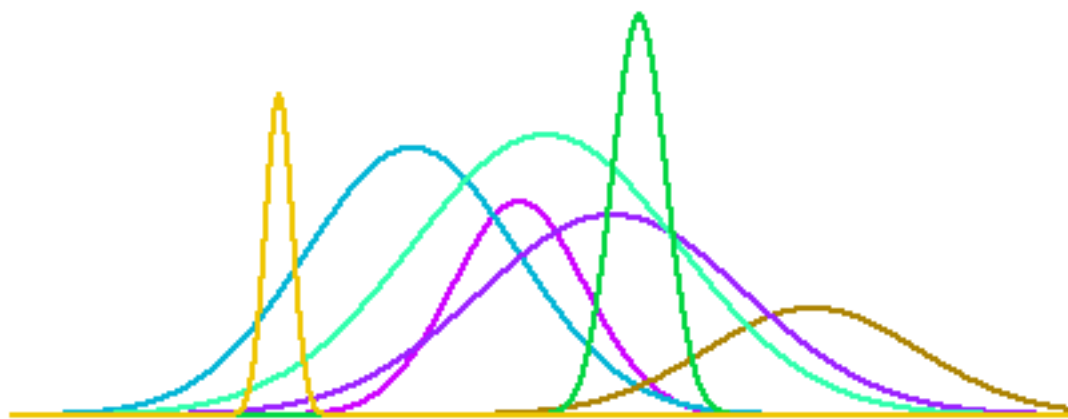


Úložiště

- ❖ 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