



národní
úložiště
šedé
literatury

A Machine for Automatic Subject Indexing Using ToC

Pokorný, Jan
2018

Dostupný z <http://www.nusl.cz/ntk/nusl-380921>

Dílo je chráněno podle autorského zákona č. 121/2000 Sb.

Licence Creative Commons Uveďte původ-Neužívejte komerčně-Nezpracovávejte 4.0

Tento dokument byl stažen z Národního úložiště šedé literatury (NUŠL).

Datum stažení: 09.05.2024

Další dokumenty můžete najít prostřednictvím vyhledávacího rozhraní nusl.cz .



A Machine for Automatic Subject Indexing Using ToC

National Library of Technology, Prague

Reasons for development of the machine

- Subject description performed by a **librarian-cataloger** is inefficient:
 - expensive human labour
 - time delay
 - limited ability to understand and describe a subject
- When document **titles** are too general or does not express the document meaning and contents
- When document fulltext is not available or its using is not allowed

Example

TASK: find a book on **advanced using the file system in Python**


OPAC by title: 0 results

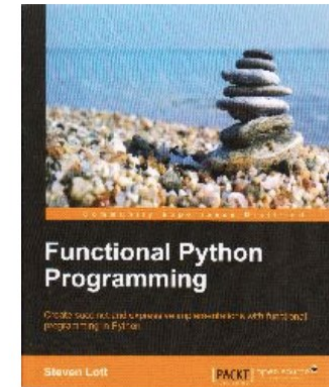
OPAC by subject: 0 results

We will find books on Python but have no tools to specify embedded subtopics.

Example

Functional Python programming : create succinct and expressive implementations with functional programming in Python

Hlavní autor:	<u>Lott, Steven F.</u>
Médium:	 <u>Kniha</u>
Jazyk:	<u>English</u>
Vydáno:	<u>Birmingham ; Mumbai : Packt Publishing, 2015</u>
Edice:	<u>Community experience distilled (Packt Publishing)</u>
Žánr/forma:	<u>příručky</u>
ISBN:	<u>978-1-78439-699-2</u>
Témata:	<u>programovací jazyk Python</u> <u>funkcionální programování</u>



Using nested information instruments

- Book index
- Table of contents

Book index – items too atomized without expressed context (a list of separated words and short terms in alphabetical order playing a role of keys to locating information contained in the book)

Table of contents – brings overview of all important topics contained

Table of contents - advantages

- ToC is created by the author himself
- ToC is set from chapter and subchapter titles which are mostly created by the author as keywords in basic grammar forms (adv over fulltext processing)
- Chapter titles indicate chapter contents
- Precise image of contained important topics and subtopics keeping their relations and document structure (title of the work -> section name -> chapter name -> subchapter name and so on)
- ToC contain links to page numbers (we can measure importance of topics)

Table of Contents

1. Introduction	3
2. Game market overview	5
2.1 General market overview	5
2.2 Market penetration effectiveness	6
2.3 Blockchain and mobile games	8
2.4 Game market competitive environment	11
2.5 Mobile gaming apps profitability	11
2.6 Other blockchain-based games	13
3. Business model	14

Common problems

- High number of ToC forms and variations (layout, typography, design, structure, embedding)
- Pages with ToC often contain other texts (to be removed)
- Dependency between chapter-subchapters and book title-chapters

Table of Contents

Getting Started.....	3
Benefits.....	3
Word versions.....	3
Get the template.....	4
Using the Template.....	5
Sample text.....	5
Arrangement of your ETDR.....	5
Basic formatting requirements.....	5
Fonts.....	5
Line spacing.....	6
Margins.....	6
Footnotes/Endnotes.....	6
Page Numbers.....	6
Styles.....	7
Configure Word for working with styles.....	8
Styles Used in the ETDR Template.....	10
Apply a different style.....	10
Modify a style.....	10
Copying/Pasting.....	11
Section and page breaks.....	11
Table of Contents.....	12
Add a new chapter.....	13
Add a new subdivision heading within a chapter.....	13
Figures and Tables.....	13
Images.....	13
PowerPoint slides.....	13

- 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

OBSAH

Úvod	5	Kresba štětcem	92
[I.] TECHNIKA 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

PRINT
38 East 29th Street
New York, NY 10016
Phone: 212-447-1400
Fax: 212-447-5231
Email: info@printmag.com

Editor-in-chief
Joyce Rutter Kaye

Guest art director
Abbott Miller

Art director
Stephanie Skirvin

Managing editor
Todd Pruzan

Senior editor
Jeremy Lehrer

Editor-at-large
Martin Fox

Associate art director
Elizabeth Chen

Assistant editor
Andrew Yang

Editorial intern
Lisa Case

Copy editors
Ariana Donalds
Caitlin Dover
Mark Higa

Contributing editors
Roy R. Behrens
Colin Berry
John Canemaker
Michael Dooley
Cathy Fishel
Steven Heller
Rich Hoxsey
Harold Martin
Rick Poynor
Ellen Shapiro
Lisa Trollback
Anthony Vagnoni
Tom Vanderbilt

Subscribers:
Send subscription orders
and inquiries to:
PRINT

print

Comment 13
Notes about this issue, and PRINT's redesign.
by Joyce Rutter Kaye

Contributor 14
Where we're calling from.

Letters 16
"If my company bans this issue of PRINT, might other periodicals
be banned as well?"

F.O.B. 18
Government comics from Japan, clean graffiti from England, a note
on our new type, and more.

Shelf Life 26
New album covers, book jackets and packaging: the best and worst.

Monologue 28
Going Public
A Napster-like revolution promises to spread design language and
its tools to the masses.
by Ellen Lupton

Observer 33
Being there
The prophet of the digital revolution omitted plenty when they
made their '90s proclamations.
by Rick Poynor

Newsstand 37
No Surrender
Double Take was once America's most celebrated new arts
journal. Why didn't it stay afloat?
by Jason Zengerle

Dialogue 40
Glenn Horowitz
On the sale of the Herbert Matter Archive to Stanford
University,
by Steven Heller

45 In Print
Vol.1 / No.1
Reflection on PRINT's June 1940 debut. by Martin Fo

106 Collective Soul
Designers' resource site, a FlightCheck Studio extension, Optimo Didot type, and more.
by Rich Hoxsey

112 Type
FF Celeste Sans
A new "retrospective traditional" by Christopher Burke
by Paul Shaw

114 Books
The Pushpin Graphic, by Seymour Chwast
Review by Allen Shapiro
The Business of Holidays, edited by Maud Lavin
Review by Collin Berry
Freedom Fries, by Steve Brodner
Review by Edward Sorel
In the Shadow of No Towers, by Art Spiegelman
Review by Peter Kuper

118 Event
Hard News
Visa Pour L'Image in Perpignan, France
Review by Rhonda Rubinstein

120 Frames
Out of Film
Digital technology has made the end of traditional filmmaking a
virtual certainty.
by Joseph Kennedy

128 End Product
Book Paptism
Melcher Media's waterproof books are both
submersible and eco-friendly.
by Caitlin Dover

Design
Type Culture

PRINT (ISSN 0032-9510) is published 6 times a year in February, April, June, August, October, and December, by F&W Publications, 4700 E. Galbraith Road, Cincinnati, OH 45236. Subscription rates: US \$57 for one year; \$107 for two years; outside the US \$98. Phone 877-860-9145 (US and Canada) 386-246-3361 (toll-free). Email: brain@fwpub.com. Single copies of PRINT are distributed exclusively by Populision SA, Sheridan House, 11016A Westwood Hill, Rose, CA 94663, USA. Phone 415-273-8911. Email: populision@populision.com. Periodical postage paid at Cincinnati, OH, and at additional mailing offices. Postmaster: Send address changes to Print, PO Box 402035, Palm Coast, FL 32142. Printed in the USA. Publisher: Steve Pign.

00

CONTENTS

01

INTRODUCTION

- 01. Purpose of Guidelines
- 01. Stockholders

03

TRAVEL

- 04. Air Travel
- 06. Car Rental
- 08. Lodging
- 09. Meals
- 11. Miscellaneous

12

EXPENSES

- 13. Entertainment
- 14. Flowers & Gifts
- 15. Office Supplies
- 16. Telecom
- 17. Misc Expenses

18

GENERAL INFO

- 19. General Info
- 20. Approvals
- 21. Personal Costs

24

INDEX

- 23. Index

Содержание:



04 Вступление

- 05 Обращение Председателя правления
- 06 Обращение



08 Макроэкономические условия
деятельности Банка в 2006 году



12 Финансовые итоги 2006 года.
Ресурсная политика Банка

- Корпоративный бизнес Банка 16
- Обслуживание частных клиентов 22
- Деятельность Банка на рынке ценных бумаг 24
- Международное сотрудничество 27
- Регулярный бизнес Банка 28



30 Факторы успешного развития

- 32 Информационные технологии
- 33 Кадровая политика
- 34 Общественная деятельность



36 Ответность
и аудиторское заключение

- 38 Финансовая отчетность по РСБУ
- 43 Финансовая отчетность по МСФО

Embedded items and relations

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

Moduls – which? In context of book title *Python* → Moduls in Python language

Basics – which? In context of chapter *Moduls* and book title *Python* → Basics of modules in Python language

Workflow

1. Scanning book pages containing ToC
2. OCR with text block detection (learning)
3. Eliminating irrelevant text blocks
4. Resolving word and numeric blocks
5. Text analysis with focus on the context and position of ToC items
6. Keyword extraction and removal of stop words
7. Subject classification assignment
8. Validation (learning)
9. Usage

Scanning book pages contained ToC

- An operator have to scan all pages where a ToC is present in a book.
- Typically it is spread over more pages.
- Careful work reduces error rates in the subsequent steps (missing page, a bent corner, too solid bending, crooked text lines or a book scanned askew).
- The result of scanning is a bitmap object in the form of a graphic file such as TIFF or JPEG.

OCR with text block detection

- OCR software must be able to detect the blocks that contain text on a page.
- OCR software may identify some blocks incorrectly (layout, typography, design, structure) --> need for human intervention
- Learning from new layout situations to work better and better

Irrelevant text block elimination

- Irrelevant texts and graphic objects must be distinguished and removed from processing.
- For the next steps, we need to work only with text blocks that contain ToC items.

Resolution of word and numeric blocks

- ToCs usually list chapters with the name of each chapter and the page number where each chapter begins
- In addition, the page number for the next chapter determines the range of pages that can be used to calculate a relevance scale or to display the importance of a chapter in thematic clouds.
- Therefore, the software must retrieve the name of each chapter and its location (pages) for each ToC.
- Formal markers removal ("Chapter 2", "Book first", "Index" etc.)

Text analysis with focus on context of ToC items

- Scientific books typically use a multilevel structure for chapters in which subchapters are embedded into parent chapters keeping their context.
- We have to handle with the exact order of items and their position in their ToC tree
- Based on layout (embedded items are often indented) or on typography (embedded items often have smaller or thinner fonts)
- The meaning of child items needs to be extended by utilizing context from all parent items in order to create definite and significant keywords.

Keyword extraction and removal of stop words

- Harvested keywords can be optimized and transformed according to needs, deduplicated etc.
- The machine can also remove keywords from chapters which are considered to be marginal (e.g., if a chapter is under 2 pages, it can be assumed that such a topic is presented marginally).

Subject classification assignment

- Finally, the system suggests terms for selected subject classification systems that best match each book.
- This functionality uses a learning mechanism that tracks what keyword combinations have been applied to which classification terms.
- The system currently supports Universal Decimal Classification, Conspectus, and PSH. Others can be added.

UPDATE FOLDERS

000002325

000002505

000002579

000003580

000004104 000005010 000005014 000005018 

000006074

000006804

000006806 

000006873

000006929

000006931

000007721

Sysno: 000006804

ADD

SAVE

RE-SCORE

VIEW TOC

Fyzikální chemie II / Josef P. Novák a kolektiv

score	psh	keywords	konspekt	nerizene	bez slovníku	
60.5	<input type="checkbox"/>	Faradayovy zákony elektrolýzy				
60.4	<input type="checkbox"/>	Schrödingerova rovnice				
60.4	<input type="checkbox"/>	Ohmův zákon				
47.0	<input type="checkbox"/>	fyzikální chemie	<input type="checkbox"/>	fyzikální chemie	<input type="checkbox"/>	Fyzikální chemie
15.2	<input type="checkbox"/>	galvanické články	<input type="checkbox"/>	galvanické články		
15.2	<input type="checkbox"/>	kinetická teorie plynů	<input type="checkbox"/>	teorie plynů		
15.2	<input type="checkbox"/>	fázová rozhraní	<input type="checkbox"/>	fázové rozhraní		
15.2	<input type="checkbox"/>	molekulárně kinetická teorie	<input type="checkbox"/>	kapaliny - kinetická teorie		

Technologies used

- Open source
- Web-based Java application
- Solr dicts (classification schemas)
- Tools for text analysis (Morphodita, OpenNLP etc.)
- Angular for GUI

What next

- Recognition/assessment of related words groups (lists)
(trees = oak, beech, maple, pine, aspen, ...)
→ the group name is deductive from the specific list items
- Intellectual proces which needs a knowledge