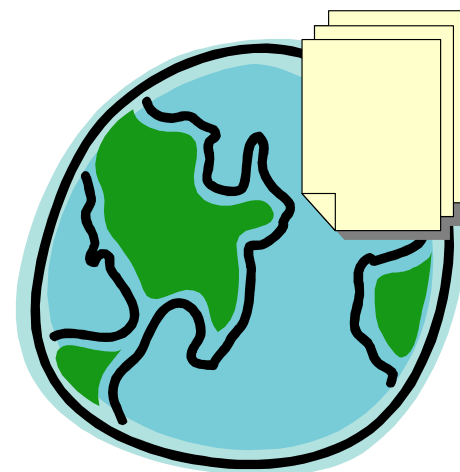
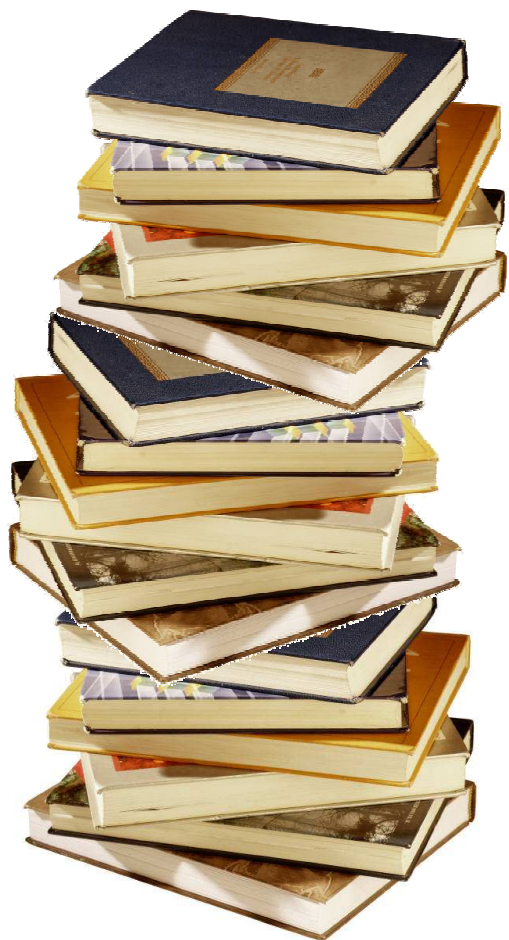


Zaawansowane przetwarzanie tekstu dla potrzeb bibliotek cyfrowych



Stanisław Osiński
stanislaw.osinski@man.poznan.pl

Ilość dostępnych informacji
rośnie bardzo szybko



[Problemy dostępu do informacji
pojawiają się na wielu różnych polach



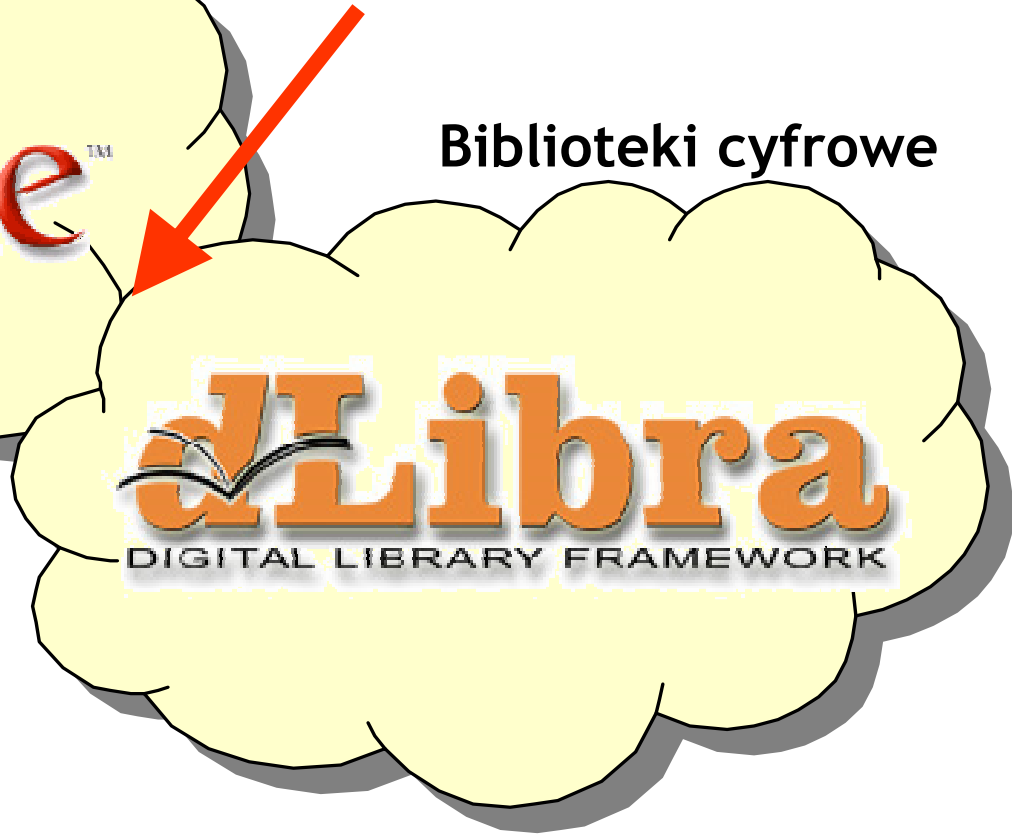
Google™

dLibra
DIGITAL LIBRARY FRAMEWORK

Czy wyszukiwanie informacji i biblioteki cyfrowe mają wspólne obszary?



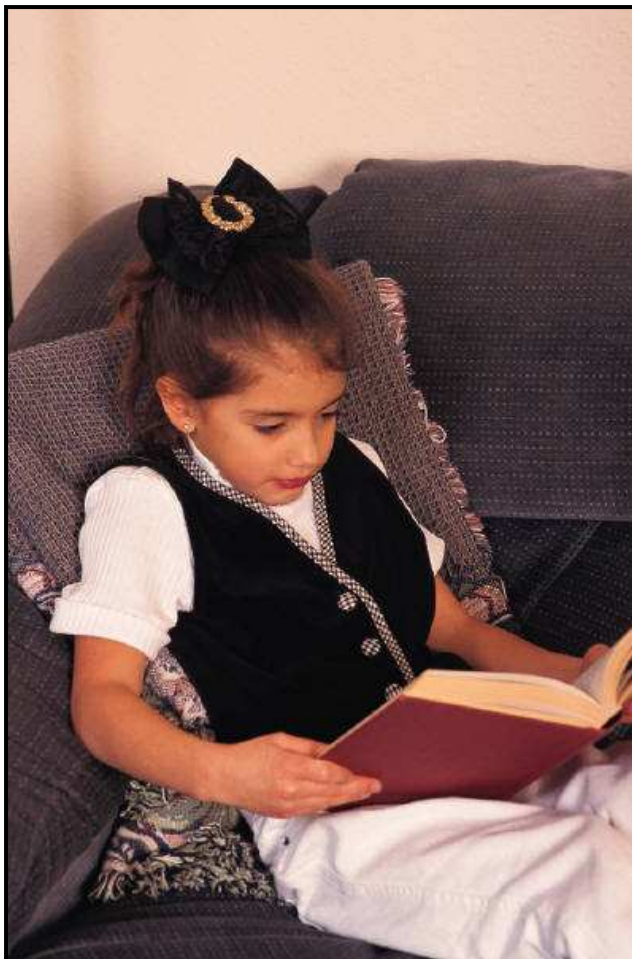
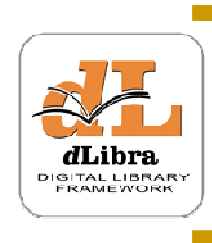
Wyszukiwanie informacji
(Information Retrieval)



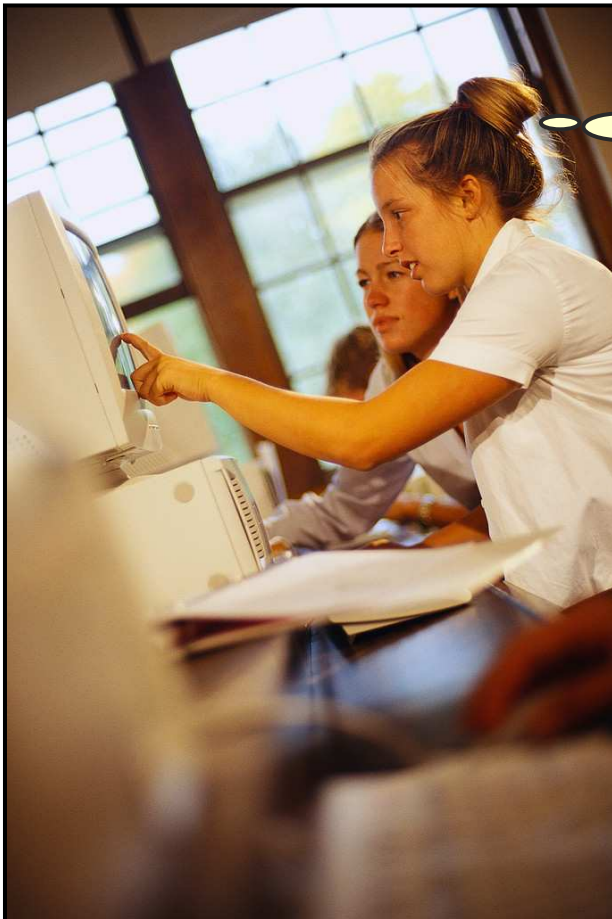
Biblioteki cyfrowe



Skorzystać mogą zarówno czytelnicy jak i bibliotekarze



Czytelnicy chcą szybko
dotrzeć do właściwej informacji



Biografia
Fryderyka Chopina

chopin biografia

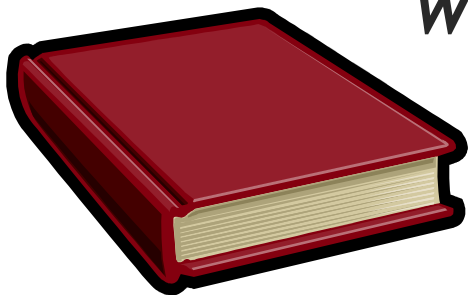
szopen biografia

„w *chołdzie* chopinowi”

biografie kompozytorów



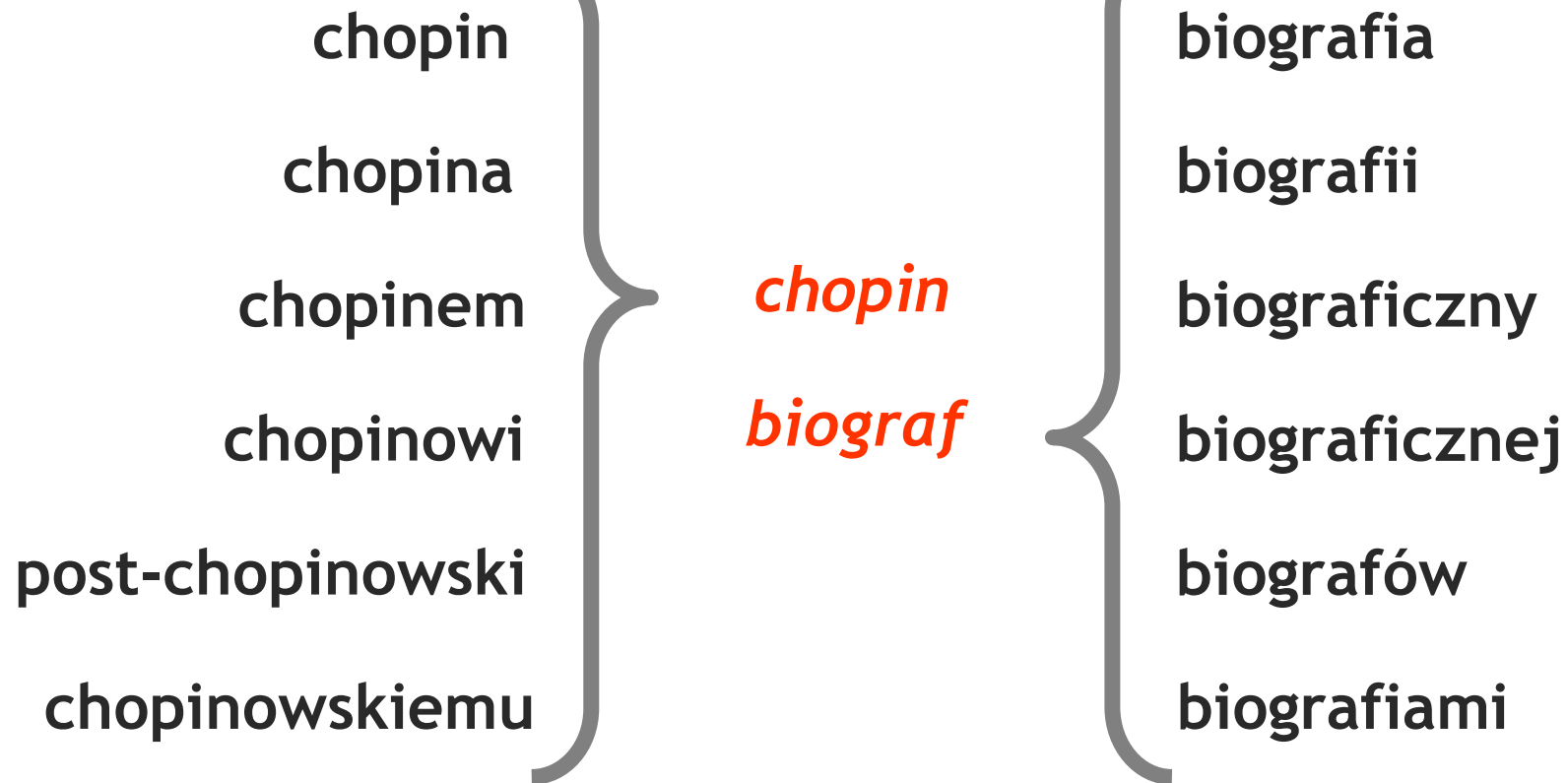
Bardzo prosta wyszukiwarka
nie zawsze się sprawdza



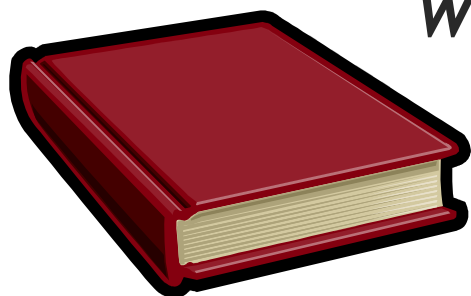
*W hołdzie Chopinowi:
rys biograficzny kompozytora*

chopin biografia	0 publikacji
szopen biografia	0 publikacji
w <i>chołdzie</i> chopinowi	0 publikacji
biografie kompozytorów	196 publikacji

Lematyzacja usuwa końcówki fleksyjne



Lematyzacja usprawnia wyszukiwanie



*W hołdzie Chopinowi:
rys biograficzny kompozytora*

chopin biografia

1 publikacja



szopen biografia

0 publikacji

w *chołdzie* chopinowi

0 publikacji

biografie kompozytorów

196 publikacji

Słownik synonimów znajdzie
podobne słowa i pojęcia



szopen

chopin

szopen

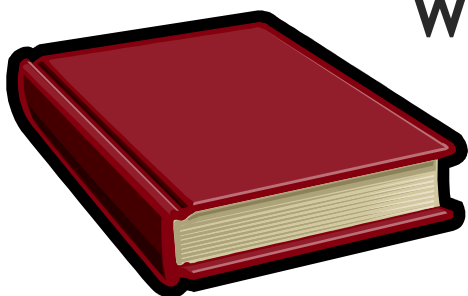
fryderyk
szopen

Fryderyk Chopin

Fryderyk Szopen

Frédéric Chopin

Słownik synonimów
również **usprawnia wyszukiwanie**



W hołdzie Chopinowi:
rys biograficzny kompozytora

chopin biografia

1 publikacja



szopen biografia

1 publikacja



w *chołdzie* chopinowi

0 publikacji

biografie kompozytorów

196 publikacji

Słownik ortograficzny poprawi błędy



chłodzie (87%)
hałodzie (72%)
hołodzie (63%)

~~chołodzie~~
chłodzie

~~chołodzie~~
hołodzie

hołodzie (99%)
chłodzie (12%)
hałodzie (2%)





Słownik ortograficzny poprawi błędy

Google [WWW](#) [Grafika](#) [Grupy dyskusyjne](#) [Nowości](#) [Katalog](#)
kwśmewki [Szukanie zaawansowane](#)
[Ustawienia](#)
© Szukaj w Internecie © Szukaj na stronach kategorii: Polski

WWW

Czy chodziło Ci o: [kwaśniewski](#)

Nie znaleziono stron zawierających wszystkie podane słowa.

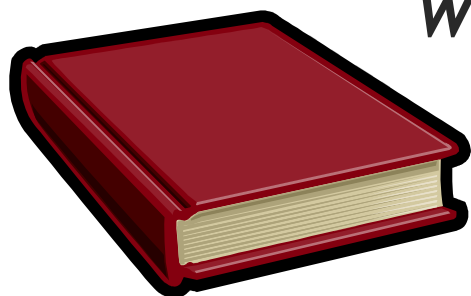
Podana fraza - **kwśmewki** - nie została odnaleziona.

Podpowiedzi:

- Sprawdź, czy wszystkie słowa zostały poprawnie napisane.
- Spróbuj użyć innych słów kluczowych.
- Spróbuj użyć bardziej ogólnych słów kluczowych.

©2005 Google

Korekta błędów ortograficznych usprawnia wyszukiwanie



*W hołdzie Chopinowi:
rys biograficzny kompozytora*

chopin biografia

1 publikacja



szopen biografia

1 publikacja



w hołdzie chopinowi

1 publikacja



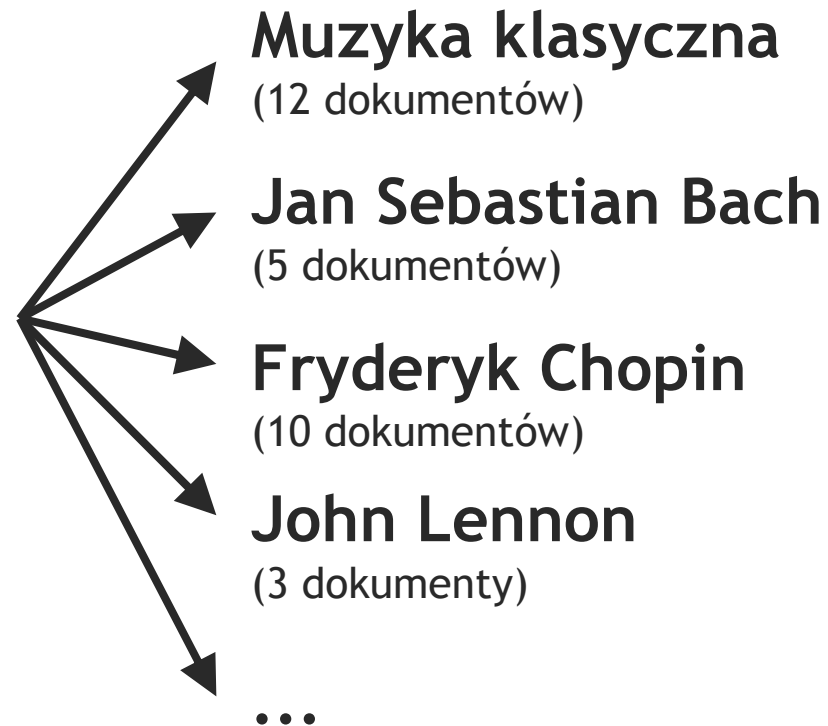
biografie kompozytorów

196 publikacji

Duże ilości wyników można automatycznie grupować



**biografie
kompozytorów**
(196 dokumentów)



Duże ilości wyników można automatycznie grupować



The screenshot displays the Carrot2 search engine interface. At the top, there is a search bar containing the text 'text mining' and a 'Search' button. Below the search bar, there are navigation links: 'components', 'admin', 'large query', 'demonstration', and 'what is carrot?'. A dropdown menu shows 'Process: YahooAPI, LINGO, Dynamic Tree' and another dropdown shows 'Download results: 200'. On the left side, there is a sidebar with a tree view of search results, categorized under 'sub topics'. The main content area shows a list of search results, each with a number, a title, a brief description, and a URL.

Sort: [flat] [group] [score]

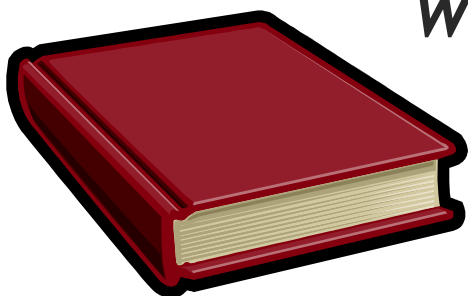
sub topics

- All groups (168)
- Information Extraction (13)
- Knowledge Mining System (12)
- Text Retrieval (8)
- Workshop on Text Mining (9)
- Text Analysis (11)
- Text Document (12)
- Search Text (9)
- Unstructured Data (8)
- Research (7)
- Text Mining Tools (10)
- Mining Knowledge from Text Collections (8)
- Nuggets (4)
- Analytics Technology (5)
- SAS Text Miner (6)
- Maximizing Text-mining Performance (4)
- Develop for Text Mining (5)
- Visual Association Rules for Text Mining (4)
- SAS Uncovers Text-mining Deal (6)
- Haym Hirsh's Publications Text (2)
- Wikipedia the Free Encyclopedia (2)
- Seminar in Linguistics (4)
- Text Mining Operations (3)
- (Other) (16)

- Open Directory - Reference: Knowledge Management: Knowledge Discovery: Text Mining**
the entire directory only in Knowledge_Discovery/Text_Mining. See also: About Scatter/Gather - Using text clustering as a way to group document according to the overall similarities in their content. ...
Eidetica - Netherlands firm offers search and text mining solutions on a hosting basis ... and a resources section. Text Mining, Web Mining, Information Retrieval and Extraction from ...
http://dmoz.org/Reference/Knowledge_Management/Knowledge_Discovery/Text_Mining
- Wikipedia: Text mining**
Wikipedia Free Encyclopedia's article on 'Text mining'
http://en.wikipedia.org/wiki/Text_mining
- Text Mining**
... The general idea of text mining - getting small "nuggets" of desired information out of "mountains" of ... retrieval (IR) itself. Currently text mining is enjoying a surge of interest ...
http://www.scils.rutgers.edu/~msharp/text_mining.htm
- Marti Hearst: What Is Text Mining?**
... What Is Text Mining? Marti Hearst. SIMS,UC Berkeley ... Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information ...
<http://www.sims.berkeley.edu/~hearst/text-mining.html>
- SAS | Data and Text Mining**
... latest enhancements include the addition of text mining capabilities that enable you to quickly ... predict new business opportunities. Text mining capabilities enable you to apply such ...
<http://www.sas.com/technologies/analytics/datamining>
- text mining and web-based information retrieval reference**
... Text Mining, Web Mining, Information Retrieval and Extraction from the WWW References. Text mining overview article ... Their works include text mining, medical text understanding ...
http://filebox.vt.edu/users/wfan/text_mining.html
- Text Mining Workshop, April 13, 2002 (Arlington, VA)**
(Image designed by Justin T. Giles) April 13, 2002 Workshop. Hyatt Regency, Crystal City, Arlington, Virginia. Extracting content from text continues to be an important research problem for information processing and management. ... the enormous volume of online textual material, effective yet scalable approaches to text mining will be needed ... A one-day workshop on Text Mining is being held in conjunction ...
<http://www.cs.utk.edu/tmw02>

<http://carrot.cs.put.poznan.pl/carrot2-remote-controller/>

Grupowanie wyników
może przynieść dodatkowe usprawnienia



*W hołdzie Chopinowi:
rys biograficzny kompozytora*

chopin biografia

1 publikacja



szopen biografia

1 publikacja



w hołdzie chopinowi

1 publikacja



biografie kompozytorów

12 kategorii



Bibliotekarze chcą łatwo udostępniać nowe zasoby



Słowa
kluczowe

Opisy,
streszczenia

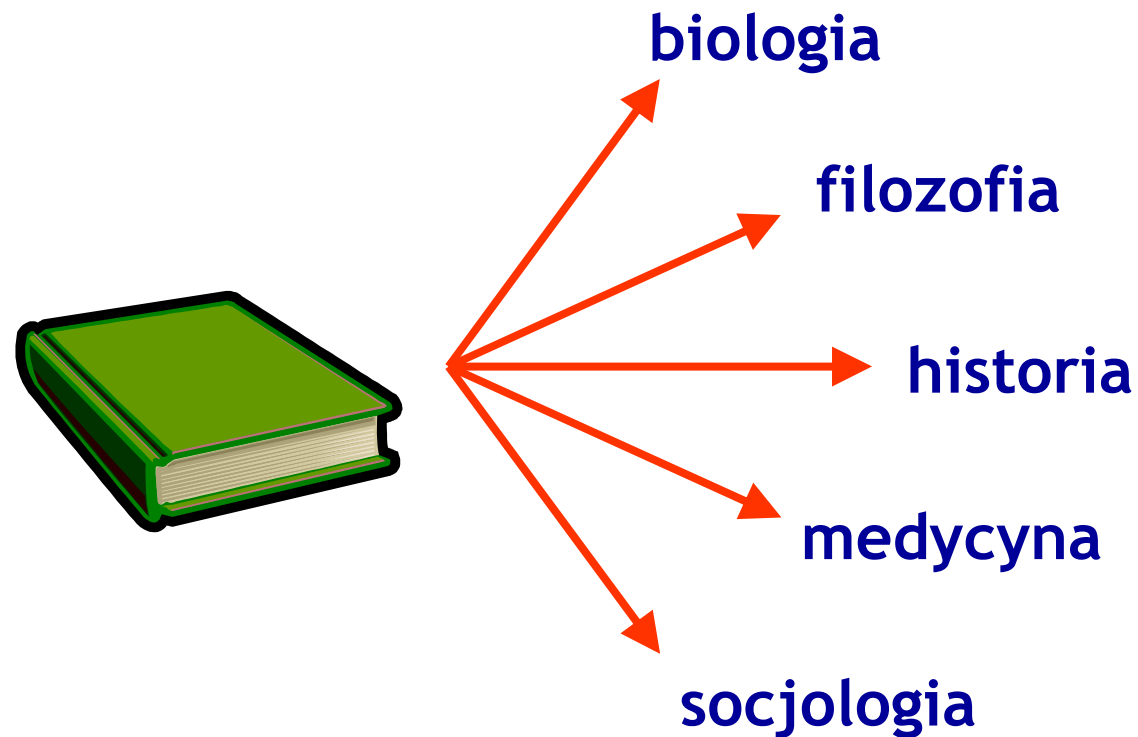
Definicje,
wiedza



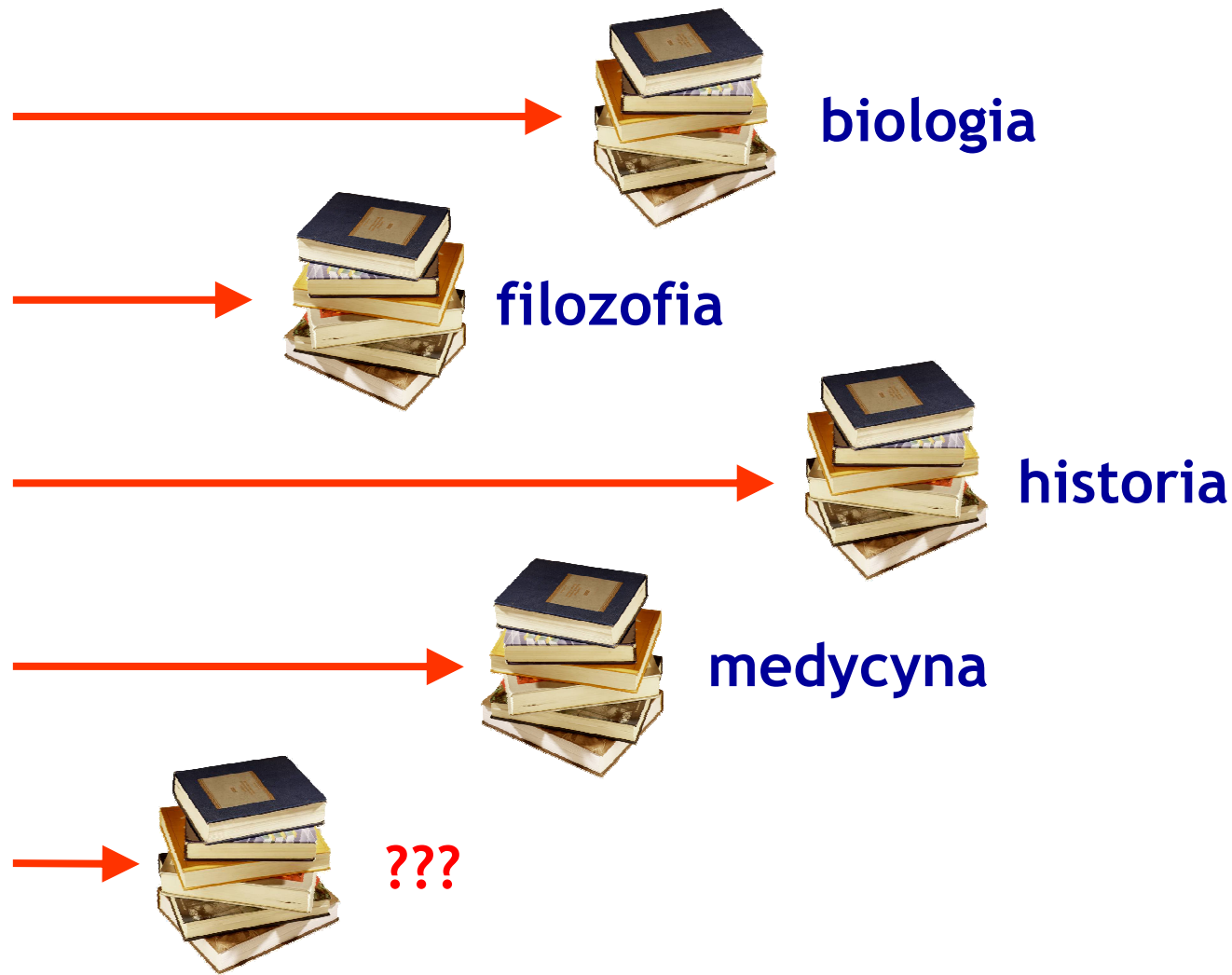
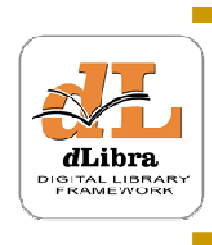
Bardzo duża liczba nowych publikacji
może wymagać częściowej automatyzacji



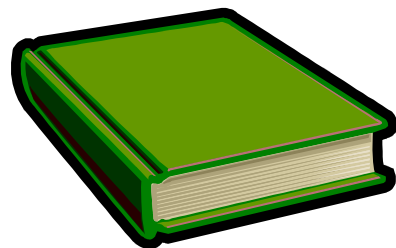
Algorytm klasyfikujący przypisuje dokument do kategorii



Algorytm klasyfikujący wybierze odpowiednie miejsce w katalogu



Algorytm streszczający wybiera ważne fragmenty tekstu

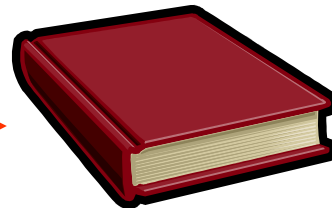


*„Najwybitniejszy polski
kompozytor, a także
wybitny pianista”*

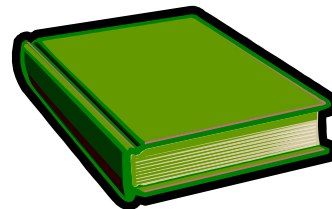
*„W 1838 roku Chopin
poznaje starszą od niego
o 6 lat i dominującą nad
nim George Sand”*

*„Umarł w 1849 roku, w
otoczeniu najbliższych osób”*

Algorytm streszczający zasugeruje opis publikacji



*„Najwybitniejszy polski
kompozytor, a także
wybitny pianista”*



*„Istnieją kontrowersje, kto był
głównym autorem zwycięstwa” (???)*

Istnieją algorytmy wyszukujące
nazwiska, daty, miejsca etc.



Fryderyk Chopin
George Sand
Jarosław Iwaszkiewicz

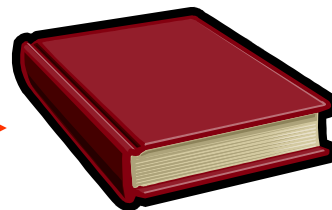
1 marca 1810 roku
17.10.1749
Po roku 1723

(w) Żelazowej Woli
(do) Paryża
Szkoła Główna Muzyki (???)

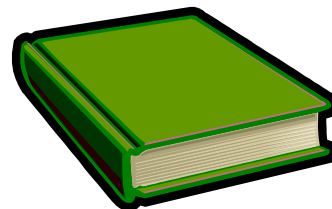
ul. Wiejska 1, 00-000 Warszawa

134 zł 12 gr
12,00 zł
6 EUR

Prowadzone są prace nad automatyczną ekstrakcją metadanych



Autor: **Jarosław Iwaszkiewicz**
Słowa kluczowe: **romantyzm, fyderyk chopin, biografia**



Autor: **Józef Piłsudski (???)**
Słowa kluczowe: **marszałek, biografia, uczył do gimnazjum (???)**

Możliwe jest automatyczne wyszukiwanie definicji pojęć



„Tlen jest niemetalem z szóstej grupy głównej”

„Poznań to jedno z najstarszych i największych polskich miast, położone nad rzeką Wartą”

„Poznań jest ważnym ośrodkiem przemysłu spożywczego” (???)

Możliwe jest automatyczne wyszukiwanie definicji pojęć



Google

Web Images Groups News Froogle Local more »

define: digital library

Search

Advanced Search
Preferences

Sign in

Web

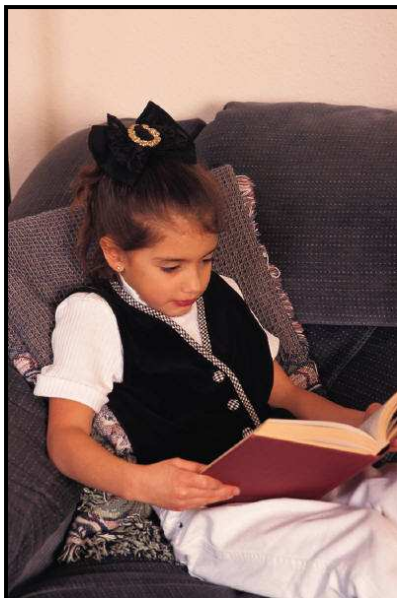
Related phrases: [perseus digital library](#) [digital library federation](#) [list of digital library projects](#) [alexandria digital library](#) [digital library services](#) [digital library systems](#) [acm digital library](#) [greenstone digital library](#) [california digital library](#) [informedia digital library](#)

Definitions of **digital library** on the Web:

- Collection of texts, images, etc., encoded so as to be stored, retrieved, and read by computer. Digital printing
www.sir.arizona.edu/resources/glossary.html
- The meaning of the phrase digital library varies tremendously, but one simple definition is the use of computers to store library materials appearing in electronic (digital) format.
www.law.harvard.edu/library/collections/digital/guidelines_glossary.php
- Digital libraries can include reference material or resources accessible through the World Wide Web. Digitized portions of a library's collection or original material produced for the web can also be included in a digital library.
www.collectionscanada.ca/vrc-rvc/s34-151-e.html
- an integrated set of services for capturing, cataloging, storing, searching, protecting, and retrieving information
www.wtec.org/loyola/digilibs/d_01.htm

<http://www.google.com>

Skorzystać mogą zarówno czytelnicy jak i bibliotekarze



lematyzacja

synonimy

ortografia

grupowanie wyników

klasyfikacja

streszczenia

ekstrakcja informacji

ekstrakcja metadanych

ekstrakcja definicji



Dziękuję za uwagę



Zaawansowane przetwarzanie tekstu
dla potrzeb bibliotek cyfrowych

Stanisław Osiński
stanislaw.osinski@man.poznan.pl

