

しーず・にー ず・しーず

～東京大学・前田朗の
アプリケーション作成等・
2年間分のネタばらし～

平成24年7月26日
図書系職員のためのアプリケーション開発講習会
前田 朗

はじめに

- ▶ 2年間分の秘蔵(?)ネタを(一部隠しネタを除き)一挙公開
- ▶ 仕事とプライベートで、手広くやっています！
 - 図書受入系
 - 閲覧・サービス系
 - 施設管理・イベント系
 - 自然言語処理系(プライベートというか自己研鑽)
- ▶ 詳しく説明するのは、2つほどです(説明しだすと時間かかりすぎですから....)

コンテンツ

- お蔵出し(18件)
 - 図書受入系(10件)
 - 閲覧・サービス系(2件)
 - 施設管理・イベント系(3件)
 - 自然言語処理系(3件)
- Be-ビブリオ
- 難語レポーター

毎月ひとつのペースで何かやってたことになるようで...

必要は発明の母と申しますが...

図書館の現場には、
必要がたくさん

紹介の前に

▶ アプリ一般公開ページ

- <https://mbc.dl.itc.u-tokyo.ac.jp/products.html>

▶ 講習会資料(一般公開分)

- <https://mbc.dl.itc.u-tokyo.ac.jp/lecture/>

▶ 講習会Wiki(部内者限定)

- アクセス情報は秘密！

東京大学柏図書館の
仕事扱いのものです。

かなり適当な感想です。

MAX三つ星！



お役立ち度
一般度
使いやすさ



図書受入系 (10件)

Beービブリオ

書籍リストのブラウジングに特化したツールです。

ネットで入手できる情報源と著作権との兼ね合いが、技術よりも困難だったような。

書籍の選定にお薦め！

詳しくは公開中アプリで！

お役立ち度
一般度
使いやすさ



CiNii

Booksランチャー

CiNii Booksの書籍を
NCIDリストもしくは著者名
典拠リスト順にブラウジング

詳しくは公開中アプリで！

お役立ち度
一般度
使いやすさ



図書館の関連マップ

東京大学の図書館室の関連を2次元や3次元マップで図示

はたして、あなたの図書館の全学的な立ち位置は？

詳しくはいずれ！



お役立ち度
一般度
使いやすさ



xISBNを使った 多言語版の最新版調査

図書館の洋図書、もしかして版が古くなっていないですか？

そんなときには、このツールがおすすめです。

詳しくは講習会資料の
Webページで！



お役立ち度
一般度
使いやすさ



洋図書の納品期間は 推定できるか

グラフにしてみました....

誰か統計学的にまっとうなものを
作成してくれることを期待で
す

詳しくは講習会Wiki
(部内者用)で！

柏

お役立ち度
一般度
使いやすさ



あいまい重複調査

書籍同士でタイトル中に含まれる語が、一定の割合以上重複していたら、アラートを出すプログラム。

連想検索にある意味近い？

詳しくは講習会Wiki
(部内者用)で！

柏

お役立ち度
一般度
使いやすさ



新領域の教員著書を チェックする

CiNii Booksの「名寄せ」とWeb APIにより、特定著者の著書をまとめてリストアップできるようになりました。

問題は、著者IDが妥当かどうかの判断が人手によることと、最終処理結果のチェックがやっぱり人手によることですが...

詳しくは講習会Wiki
(部内者用)で！



お役立ち度
一般度
使いやすさ



NCIDから 和書価格を得る

NCID → ISBN → 楽天の
価格まで、すべてプログラ
ムで処理します

詳しくは講習会Wiki
(部内者用)で！



お役立ち度
一般度
使いやすさ



シリーズものの 所蔵漏れをチェック

CiNii Books(NACISIS-CAT)の親書誌と子書誌のデータを使い、図書館所蔵のシリーズものの所蔵漏れをチェックします。

もっとも、CiNii Bookの子書誌以外にも、子書誌がありえたり、複数の版をもつシリーズもあるので、作ってはみたものの使いどころがむずかしい感じです。

詳しくは講習会Wiki
(部内者用)で！



お役立ち度
一般度
使いやすさ



閲覧・ サービス系 (2件)

当日貸出

- 図書館システムのオフライン貸出を流用します。
- 目的外利用なので、当日貸出用の仕掛けだけPerlで作成です
- 東京大学柏図書館の自動化書庫内雑誌の館内閲覧や、図書館端末の使用手続きで活躍中！

詳しくは講習会Wiki
(部内者用)で！

お役立ち度
一般度
使いやすさ



柏図書館の 学内利用者用端末

柏図書館の学内利用者用端末
をPCからセットアップ

セキュリティをいかにかけたか
ドキュメントとして残しています

詳しくは講習会Wiki
(部内者用)で！

お役立ち度
一般度
使いやすさ



施設管理・ イベント系 (4件)

ブックトーク用 メールフォーム

柏図書館のブックトークの会場定員
オーバーを防ぐため事前登録制に。

そのためのメールフォームです。携帯
ブラウザからも使えるのがポイント！

詳しくは講習会Wiki
(部内者用)で！



お役立ち度
一般度
使いやすさ



「東京大学柏図書館 ブックトーク」 Facebookページ

Facebookページをいかに、図書館イベントで活用できるかの事例になればと。

東京大学柏図書館長をはじめ、担当職員の尽力によるところ大です。

詳しくは公開中ページで！



お役立ち度
一般度
使いやすさ



Googleカレンダーを使った 会議室予約状況可視化

▶ 現在、作成中！

詳しくはいずれ！

柏

お役立ち度
一般度
使いやすさ



クイックタイムムービーの 字幕タイミングを 指定秒遅らせる

映画上映会の際に、業者が送ってきたクイックタイム画像と字幕のずれを、Perlで無理やり直しました。

たぶん、図書館的に次に使う機会は
まずないかも

詳しくは講習会Wikiで！

柏

お役立ち度
一般度
使いやすさ



自然言語処理系 (3件)

難語レポーター

- ▶ 蔵書の評価用に作成です。が、図書館関係外のほうが需要あるかも
- ▶ 独自（勝手）理論で作成していますので、一般公開してから性能を評価するはめに....
- ▶ でも、面白いのであとで解説します。

詳しくは公開中アプリで！

お役立ち度
一般度
使いやすさ



Wikipediaタガー

- ▶ 図書の選定用に作成です。が、もう図書館を越えた一般的な感じが...
- ▶ というか、誰かが作ってそうで見当たらないので、勢い余って自作してしまったという、オチがついています。
- ▶ あとで具体的に、図書選定での応用を考えないと。

詳しくは公開中アプリで！

お役立ち度
一般度
使いやすさ



専門用語コレ・モカモ

▶ 東京大学の図書館室の蔵書バリエーションを利用した、関連語提示ツール

詳しくは公開中アプリで！

お役立ち度
一般度
使いやすさ



Beービブリオ

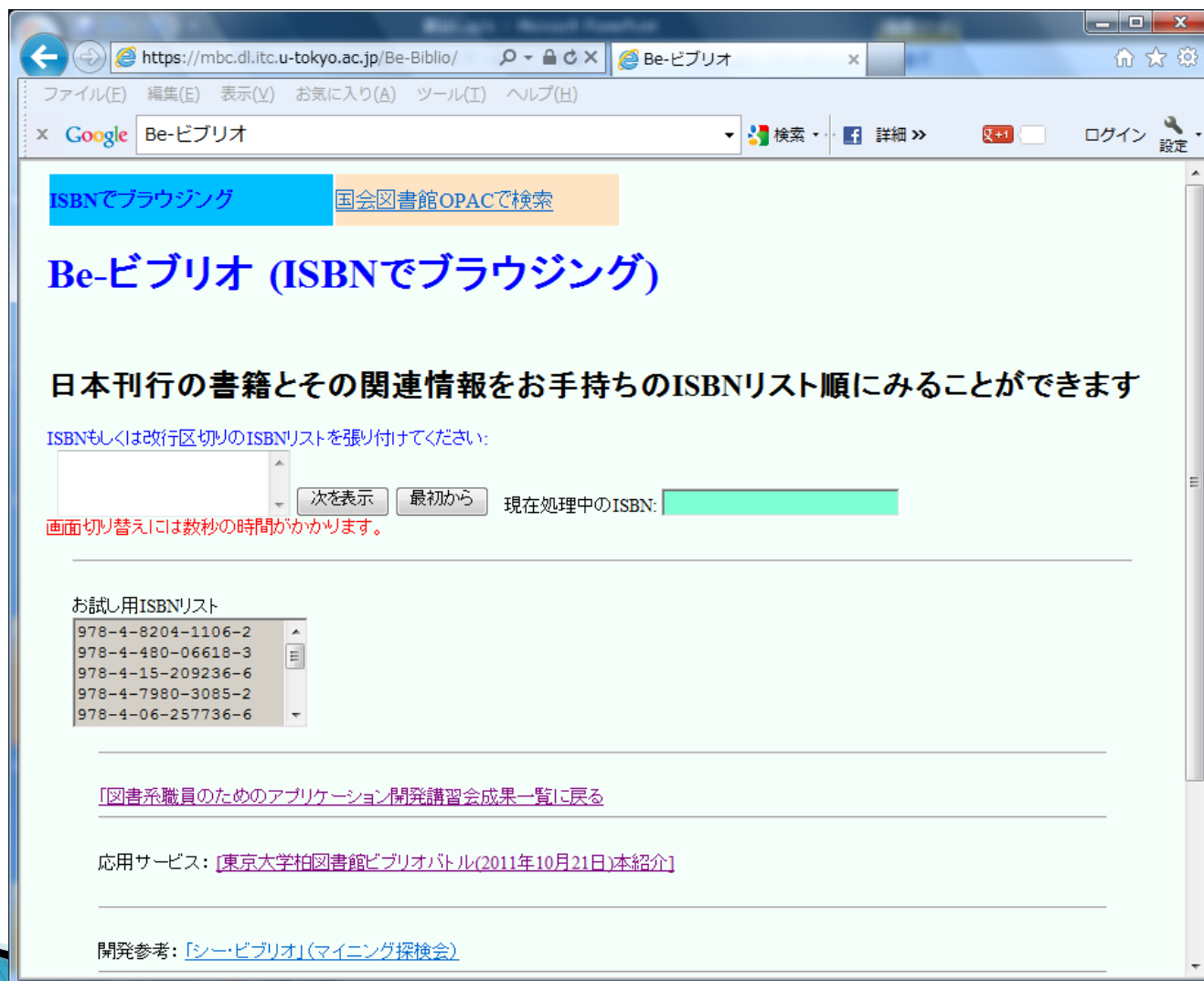
～書籍関連情報との 奮闘記と業務利用～

<https://mbc.dl.itc.u-tokyo.ac.jp/Be-Biblio/>

Webには書籍関連情報がたくさん？

- ▶ じつは、著作権を考えるとそう甘くないのです。
- ▶ くわえて、Web API等で入手できる情報も限られていて...
- ▶ 最後は、楽天ブックスに直球で「つかって問題ないですか？」と聞きました。ここまでやって、わかったこともあります。

ISBNリストで図書をブラウジング



図書リストからの書籍内容チェックに使えます！

書籍情報 表示画面例

書誌情報
(国会図書館 or
楽天ブックス)

書籍タイトルに
ついての
Web上のキー
ワードを集計

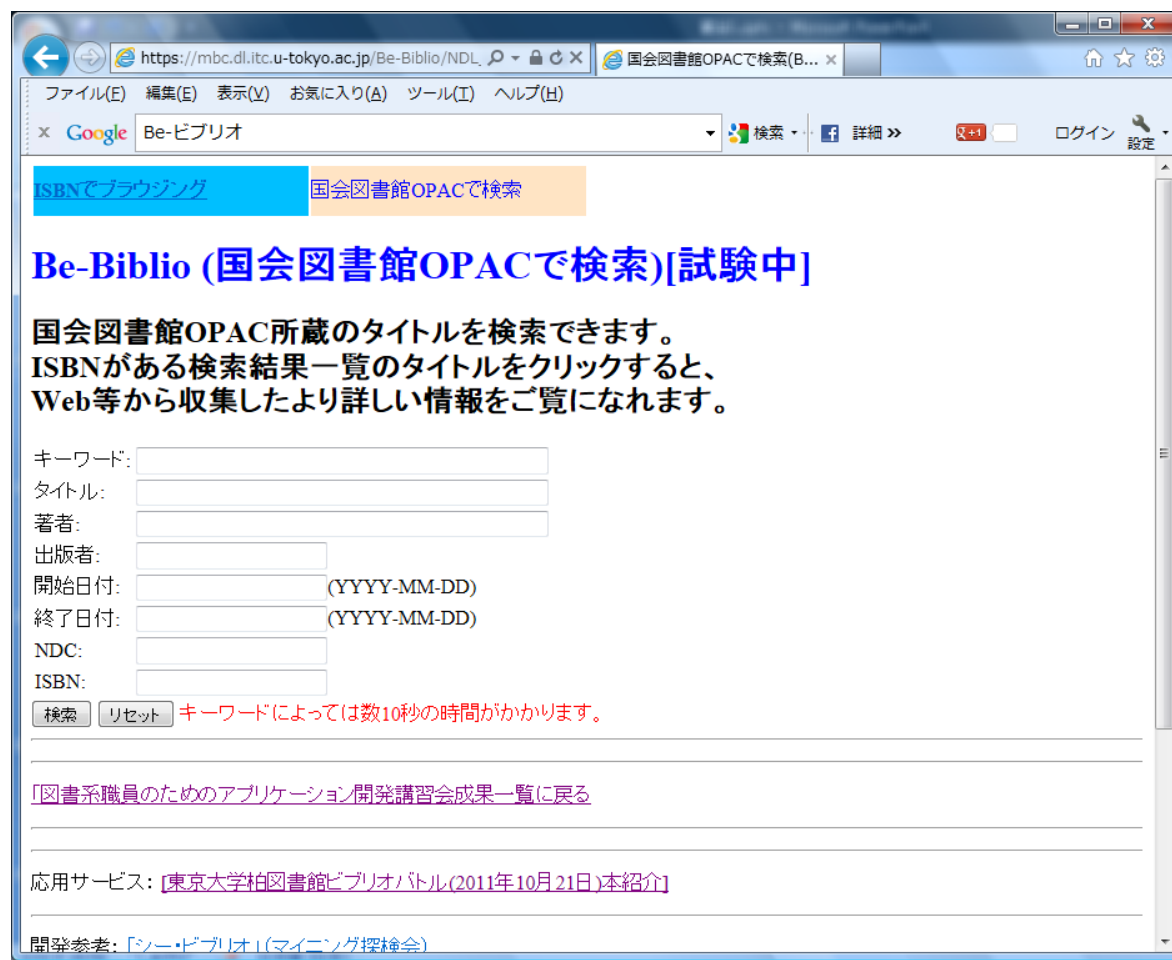
書籍タイトルに
ついての
Googleブログ
検索結果

内容紹介
(楽天ブックス)

書影
(楽天ブックス)

The screenshot shows the Be-Publio website interface. At the top, there's a navigation bar with 'Be-Publio (ISBNでブラウジング)' and a search bar. Below the header, a section titled '日本刊行の書籍とその関連情報をお手持ちのISBNリスト順に' is visible. A large blue box on the left side of the page is labeled '国会図書館サーチ' (National Diet Library Search). To the right of this box, a callout points to a section labeled '内容紹介 (楽天ブックス)' (Content Introduction (Rakuten Books)). Below the blue box, there's a list of keywords under the heading 'Yahoo! plus 関連Web' (Yahoo! plus Related Web). A callout points to this list, stating '書籍タイトルについてのWeb上のキーワードを集計' (Collecting keywords from the Web about the book title). Further down, another callout points to a section labeled 'Googleブログ検索' (Google Blog Search), stating '書籍タイトルについてのGoogleブログ検索結果' (Google Blog search results about the book title). On the right side of the page, there's a section labeled '楽天ブックス 内容紹介' (Rakuten Books Content Introduction) and a section labeled '楽天ブックス 書影' (Rakuten Books Book Cover). A callout points to the book cover image, stating '書影 (楽天ブックス)' (Book Cover (Rakuten Books)).

検索インターフェイスも用意



The screenshot shows a web browser window with the URL <https://mbc.dl.itc.u-tokyo.ac.jp/Be-Biblio/NDL>. The page title is "Be-Biblio (国会図書館OPACで検索)[試験中]". The main content area contains a search form with the following fields: "キーワード:", "タイトル:", "著者:", "出版者:", "開始日付:" (with a "(YYYY-MM-DD)" hint), "終了日付:" (with a "(YYYY-MM-DD)" hint), "NDC:", and "ISBN:". Below the form are "検索" (Search) and "リセット" (Reset) buttons. A red text note states: "キーワードによっては数10秒の時間がかかります。" (It may take several tens of seconds depending on the keyword). At the bottom, there are links: "「図書系職員のためのアプリケーション」開発講習会成果一覧に戻る" and "応用サービス: [東京大学柏図書館ビブリオバトル(2011年10月21日)本紹介]". The footer text is "開発参考: 「シー・ビブリオ」(マイニング探検会)".

Be-Biblio (国会図書館OPACで検索)[試験中]

国会図書館OPAC所蔵のタイトルを検索できます。
ISBNがある検索結果一覧のタイトルをクリックすると、
Web等から収集したより詳しい情報をご覧になれます。

キーワード:
タイトル:
著者:
出版者:
開始日付: (YYYY-MM-DD)
終了日付: (YYYY-MM-DD)
NDC:
ISBN:

キーワードによっては数10秒の時間がかかります。

[「図書系職員のためのアプリケーション」開発講習会成果一覧に戻る](#)

応用サービス: [\[東京大学柏図書館ビブリオバトル\(2011年10月21日\)本紹介\]](#)

開発参考: [「シー・ビブリオ」\(マイニング探検会\)](#)

国会図書館アクセス用のPerlモジュールも開発。公開中

情報源について

Web上ではさまざまな書籍関連情報があるが...



著作権と網羅性を考えると、使える情報は限られる！

楽天の「内容紹介」と「書影」 → 楽天に直に問い合わせて問題ないことを確認

Yahoo! 検索結果(抜粋) からのキーワード抽出

Web上の書評 → Googleブログ検索がよさそう

書籍サイトの「レビュー」について

- ▶ Amazon, Google, 楽天とも、APIでのレビュー情報の提供はなし
 - Google, 楽天は、スクレイピングをすれば技術的には取得可能
 - Google (10/500)
 - 楽天 (23/500)
 - AmazonはHTMLのタグ付の関係で、レビュー部分そのものを取り出すことは難しい(加えて、AmazonはHTMLのタグを頻繁に変更します[経験談])
- ▶ スクレイピングの場合、少なくとも相手のサーバーに負荷をかけることは許されない(業務妨害)
- ▶ 「公開された著作物」ではおそらくある
- ▶ 引用といった場合
 - フレーズ程度？
 - 1文程度？
- ▶ 著作者
 - Amazonのレビューは、Amazonが著作権を持つ
 - Googleや、楽天は？ → 楽天は問い合わせたらダメでした

結論だけ言うと、レビュー情報の2次利用は現状難しいです

書籍サイトの「内容紹介」は使えるか

▶ [Bookデータベース]

- 「内容紹介」でよく見かける[Book]データベースは、日外アソシエーツの売り物
- よって、[Book]データベースの著作権は、Amazonや楽天ではなく日外アソシエーツにあると考えられる
- ただし、[Book]データベース以外にも「内容紹介」はある。書籍サイトから日外アソシエーツの許諾だけでは不十分
- 日外アソシエーツから購入すれば済む話だが、個人で購入できる金額ではない

▶ 楽天ブックスAPI

- APIで「内容紹介」をかなり取得できる(139/500)

▶ Google

- APIで「内容紹介」を入手できるケースは少ない(0/500)

書籍サイトの「書影」は使えるか

- ▶ 「書影」には著作権が存在する
- ▶ AmazonやGoogle、楽天が提供する「書影」は著作権についての責任を各社に負ってもらえるとも考えられるが...
- ▶ もし、上記の業者が著作者への許諾をとっていないとすると、フルサイズの「書影」の利用は見合わせたほうが無難では。
- ▶ そもそも画面上で小さい書影しか使えないのであれば、ここは書影を一切使わないという割り切りも考えられる

出版社の「RSS」ニュースは使えるか

出版社によっては、RSSで自社のニュース
(新刊情報が多い)を提供しているが...

- ▶ RSSは公表された著作物といえるか？
- ▶ 新着情報の提供を目的としたRSSについて、データの蓄積利用は問題ないか
- ▶ リンクも張ることができないRSSを引用扱いにできるか

著作権もさることながら、RSSで出版ニュースを提供している出版社が少ないこと、ほとんどが新刊のお知らせであることを考えると、リスクやコストを負うメリットは少ないのでは。

出版社の「ツイッター」は使えるか

出版社によっては、ツイッターで自社の出版物の宣伝を行っているところがあるが...

- ▶ ツイッターは公表された著作物といえるか？
- ▶ 短期間での利用を目的としたツイッターにおいてデータの蓄積利用は問題ないか
- ▶ 「引用」にするには全体の一部である必要があるが、もともと情報の少ないツイッターではフレーズくらいしか残らない？
- ▶ ツイッターのつぶやきの一部＋ツイッターアカウントで引用を明記できるか？
- ▶ リツイートも著作権上問題ではないかとの意見もあるようだが...

著作権もさることながら、ツイッターで情報を提供している出版社が少ないこと、ほとんどが新刊のお知らせであることを考えると、リスクやコストを負うメリットは少ないのでは。

Amazonは使えるか？

- ▶ APIでは、「内容紹介」を入手できない
- ▶ レビューはAmazonが著作権をもつ。スクレイピングすれば(技術的には)入手できそうだが、タグ付けの関係できれいには取り出せない(加えてAmazonはよくHPのタグを変えます → 根性がないとシステムを維持できない)
- ▶ 書影はAPIで大きいサイズのもの(リンク)を入手できるが、書影はデザインに著作権があるので、あまりサイズが大きいのも考えもの
- ▶ キャッシュは24時間制限
- ▶ 1秒間に1回のアクセス制限
- ▶ 放送での使用は不可能

調べれば調べるほど、ネガティブな材料だらけ...

Web上の書評は使えるか

- ▶ さすがにこれは「公表された著作物」では。
- ▶ 引用については、「文章の一部抜粋」+「Webページへのリンク」で
- ▶ スクレイピング不要。ブログ(Web)検索エンジンのAPIが結果として返す「スニペット」(抜粋)を使える
- ▶ APIの利用
 - 試した限りでは、Googleの「ブログ検索」がもっともよい情報が得られる。ただし、Web APIについてはよくわからなかった
 - Yahoo! ブログ検索は比較的簡単に使えるが...
 - BINGはブログ検索がなかった
 - Yahoo! Web検索だと、だいたいAmazonなど書店サイトの書籍情報ページがヒットしてしまう...

情報の「引用」から離れる

- ▶ 書籍タイトルの関連語
 - 東京大学OPAC + “言選Web”
の手法を使用

難語レポーター

～大学図書館員、ここまで
できてもいいんじゃない？～

https://mbc.dl.itc.u-tokyo.ac.jp/nango_report/

「言選Web」と教員連携のススメ

- ▶ 東京大学情報基盤センター・中川教授、経済学部資料室・小島講師、柏図書館・前田の合同プロジェクト
- ▶ ちなみに、前田は開発を担当
- ▶ 情報系（情報基盤センター）の先生に師事（論文の共著を含む共同プロジェクト）をすることで、図書館員が通常もちえないだけの知識や技術を得ることができたようです（経験談）。
- ▶ この「難語レポーター」も「言選Web」のひとつの応用です。

<http://gensen.dl.itc.u-tokyo.ac.jp/>

難語レポーター実行例

The screenshot shows a web browser window with the URL `https://mbc.dl.itc.u-tokyo.ac.jp/n`. The page title is '難語レポーター'. The browser's menu bar includes 'ファイル(E)', '編集(E)', '表示(V)', 'お気に入り(A)', 'ツール(I)', and 'ヘルプ(H)'. Below the menu bar is a Google search bar with the text '検索' and a '詳細 >>' link. The main content area has the title '難語レポーター' and the instruction '日本語テキスト中の難しい語をレポートします。' (Report difficult words in Japanese text). Below this is a text input field containing the sample text: '1920年代から1960年代初頭までの物理学史、化学史、人類学史、薬理学史、そしてより哲学に近い一種の生命論等をカバーする。'. At the bottom of the input field are two buttons: '実行' (Execute) and 'クリア' (Clear). To the right of these buttons is a red warning message: '画面切り替えには数秒の時間がかかります。' (It takes a few seconds to switch screens).

難語レポーター

日本語テキスト中の難しい語をレポートします。

日本語テキストを張り付けて、「実行」ボタンをクリックください。

1920年代から1960年代初頭までの物理学史、化学史、人類学史、薬理学史、そしてより哲学に近い一種の生命論等をカバーする。

実行 クリア 画面切り替えには数秒の時間がかかります。

画面例は「近刊検索β」にあった
「昭和前期の科学思想史」(勁草書房)の内容紹介

レポート1

学術度判定(独自手法)

1. 学術度判定

最大値の用語により判定します(ノイズの場合は、下表の妥当な用語より判断ください)

学術度が高いです

用語	学術(ac.jp)サイトヒット件数	全サイトヒット件数	学術度値	学術度判定
化学史	843	4210	0.20024	高い
人類学史	164	984	0.16667	高い
生命論	1030	11000	0.09364	学術度あり
薬理学史	1	13	0.07692	学術度あり
一種	91200	2760000	0.03304	判定不可
哲学	208000	6300000	0.03302	判定不可

学術(ac.jpドメイン)限定のWeb検索と、
限定なしのWeb検索の
ヒット件数の比率で学術度を判定

レポート2

レア語をチェック(独自手法)

2. レア語チェック

用語	用語生起コスト	Yahooヒット件数
薬理学史	10.50799	13
人類学史	10.25990	984
化学史	10.09175	4210
生命論	11.12021	11000
一種	11.05643	2760000
哲学	10.91562	6300000

IPADICの[形態素生起コスト]をもとにした
独自の用語生起コストと、
Yahoo!検索件数をもとにレポート

レポート3

テキスト中の使用文字

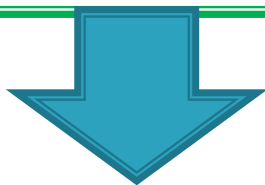
3. テキスト中の使用文字

項目	件数	%
トータル	62	-
ラテン	0	0
ひらがな	16	25
カタカナ	2	3
漢字	30	48
非常用漢字	0	0
中学以上	1	1
小学6年生	1	1
小学5年生	0	0
小学4年生	7	11
小学3年生	7	11
小学2年生	4	6
小学1年生	10	16

テキスト中の常用漢字(学年別)、
ひらがなの比率などをレポート

開発の目的と既存研究

- ▶ 文の難易度を判定できると(シー・ビブリオも参照)
 - 本の推薦・選定 or 蔵書の評価
 - その他、一般向けにも使えるかも
- ▶ 「難読度」の既存研究
 - 帯2 (教科書コーパスによる機械学習)
 - リーダビリティ・リサーチ・ラボ(ひらがなや文章長さ述語数の組合せ)



一定量のテキストが必要(書名のような短いフレーズでは厳しい)
図書の内容紹介の難読度と本体の難読度が必ずしも一致しない

近刊検索βの「書名」「内容」「目次」情報を 帯2にかけた結果1(小学生)

Cコード販売対象	小学生					
	1年生	2年生	3年生	4年生	5年生	6年生
児童	2		6	32	60	180
学参I(中・小)						6
学参Ⅱ(高校)						
検定教科書						
専門				8		
教養						
実用				4	5	54
一般	4	6	15	16	65	192
雑誌扱	6	2	9	36	55	126
婦人						12

近刊検索βの「書名」「内容」「目次」情報を 帯2にかけた結果2(中学生以上)

Cコード販売対象	中学生			高校生			大学生
	1年生	2年生	3年生	1年生	2年生	3年生	
児童	168	160	144	50			13
学参I(中・小)	14	40	9	20			
学参Ⅱ(高校)	7	16	9		11	12	
検定教科書			9				
専門	21	160	261	340	275	312	455
教養	7	24	90	130	110	84	208
実用	126	72	180	30	44		104
一般	567	1,080	1,971	1,540	858	408	1,209
雑誌扱	287	440	504	150	55	36	156
婦人	21	8	9				

児童・学生向きの本の内容紹介は、
大人(親)が読むことを前提に書かれているのが多いのかも？

「リーダビリティ・リサーチ・ラボ」による リーダビリティの公式

$$Y = -0.148X_1 + 1.585X_2 - 0.117X_3 - 0.126X_4 + 15.581$$

Y = 学年

X1 = 文章中の平仮名の割合

X2 = 1文の平均述語数

X3 = 1文の平均文字数

X4 = 1文の平均文節数

[http://readability.nagaokaut.ac.jp/research/html/modules/tin
yd0/](http://readability.nagaokaut.ac.jp/research/html/modules/tin
yd0/)

文章ではなく用語重視のアプローチ

- ▶ 短いテキストでも有効に働くかもしれない
 - 本の内容紹介のような短文向き
- ▶ 当初は、日本語WordNetや国会図書館件名標目における用語の上位・下位に着目
 - 下位語(専門度が高いと想定)が必ずしも難しい概念とはいえない
- ▶ そこで、まずは用語の**レア度**に着目
 - めったに出現しない語 ⇒ 難しい語とみなす

レア語抽出

2-Yahoo! ヒット件数

Webでヒット件数が少ない語 =
(わかりやすい)レア語

しかし

テキスト中の全用語についてWeb(Yahoo!)検索を行うのでは、処理時間がかかってしまう

そこで

用語生起コストをもとに、レア度が高いと思われる上位4件についてのみ、Yahoo!ヒット件数を調べた

レア語抽出

3-処理フロー

形態素解析(和布蕪)

形態素

用語生成(言選Web)

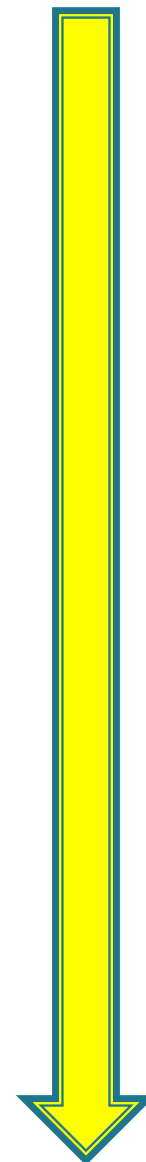
全用語

各用語の「拡張単語コスト」を計算

上位4件の用語

Yahoo! ヒット件数調査

上位4件の用語



学術度判定

- ▶ 学術度 = 国内学術サイトのヒット件数 ÷ 日本語Webサイトのヒット件数
- ▶ レア語と同様に用語生起コストの上位4件のみ判定（レスポンスタイムを稼ぐ）
- ▶ 判定基準

学術度	判定
0.5より上	学術度が非常に高い
0.1～0.5	学術度が高い
0.05～0.1	学術度あり
0.01～0.05	学術度判定不可
0～0.01	学術度が低い

[課題]テキスト全体としての評価

- ▶ テキスト全体としての評価をどのように計算で出すかは、これからの課題
- ▶ 現状では、テキスト中のもっとも学術度の高い語で学術度を判定させているが、テキスト量(対象となる用語数)が多いと、学術度を高く判定しかねない