ツール

項番	ツール名	配置場所 as Role	機能概要	機能説明	入力	出力	実装言語	使用するOSS/Pythonモジュール
❶	全文検索システム	Server A	着目する話題を記述したインターネットリソースの巡回と、その全文の収集	①インターネットリソースを巡回するクローラーと、②全文検索エンジンからなる。	巡回するURLの指定	Ⓔ Elasticsearch Repository	Java	FESS elasticsearch
❷	collect	Server B	全文検索システムの新着記事の抽出	①全文検索システム から Get Request over HTTP により前回処理した以降の巡回記事を取り出し、②.list.txt ファイルに登録済の記事を除外し、③結果を HTML フォーマットで書き出す。	前回処理以降の日数	Ⓝ 新着記事の抽出結果 yyyymmdd.html	Ruby	—
❸	Webブラウザ	Server B	新着記事抽出結果の確認	①新着記事抽出ツールの抽出結果を確認し、②人間が有為と判断した記事をショートカットファイル(.url)に書き出す。	新着記事抽出結果のブラウズ	bookmark(ショートカット形式)	不要	Firefox / Chrome
❹	dir2list	Server B	ショートカット群とショートカットリストの間の相互変換	ファイルシステムに階層的に展開されたショートカット群(dir)と、1ファイルからなるショートカットリスト(.list.txt)の間の相互変換を行う。	Ⓐ ショートカット群(dir)のルートディレクトリパス名	Ⓛ ショートカットリスト(.list.txt)ファイル名	Ruby	—
	list2dir				Ⓛ ショートカットリスト(.list.txt)ファイル名	Ⓐ ショートカット群(dir)のルートディレクトリパス名		
❺	dir2keyword	Server B	ショートカット群付属のキーワードファイルとキーワードリストの間の相互変換	ファイルシステムに階層的に展開されたショートカット群ディレクトリ(dir)にあるキーワードファイル(keywords.txt)と、1ファイルからなるキーワードリスト(.keywords.txt)の間の相互変換を行う。	Ⓐ ショートカット群(dir)のルートディレクトリパス名	Ⓚ キーワードリスト(.keywords.txt)ファイル名	Ruby	—
	keyword2dir				Ⓚ キーワードリスト(.keywords.txt)ファイル名	Ⓐ ショートカット群(dir)のルートディレクトリパス名		
❻	crawl	Server B	ショートカット群をクロウルし、その指し示すコンテンツをローカルファイルにダウンロード	①ファイルシステムに階層的に展開されたショートカット群(dir)をクロウルし、②その指し示すコンテンツを(ショートカットのディレクトリ名に.crawledを付加したディレクトリ上の)ローカルファイルにダウンロードする。	Ⓐ ショートカット群(dir)のルートディレクトリパス名	Ⓑ ショートカットのディレクトリ名に.crawledを付加したディレクトリ上のローカルファイル(HTML または PDF)	Ruby	—
❼	merge keywords	Server B	キーワードリストの同義語リスト化	1ファイルからなるキーワードリスト(.keywords.txt)から同義語リストを生成する。	Ⓚ キーワードリスト(.keywords.txt)ファイル名	Ⓢ 同義語リスト(synonyms.txt)	Ruby	—
❽	html2plaintext	Server B	ダウンロードコンテンツのプレインテキスト化	ダウンロードコンテンツからHTMLのタグ情報などを取り去りプレインテキスト化する。テキストは形態素解析ツールによって形態素に分解し、その区切りは“/”などにより明示しておく。 (現時点ではまだⓈ同義語リストは参照していないが、将来形態素のクレンジングに使用予定)	Ⓑ ショートカットのディレクトリ名に.crawledを付加したディレクトリ上のローカルファイル(HTML または PDF)	ショートカットのディレクトリ名に.plaintextを付加したディレクトリ上のローカルファイル(形態素分解済みテキスト)	Python	janome / mecab / juman++
❾	digest	Server B	プレインテキスト化したコンテンツを適当な長さに要約する	プレインテキスト化した形態素分解済みテキストを、オンプレミスまたはクラウドの分類処理にかけられる分量になるように要約する。	ショートカットのディレクトリ名に.plaintextを付加したディレクトリ上のローカルファイル(形態素分解済みテキスト)	Ⓒ ショートカットのディレクトリ名に.digestを付加したディレクトリ上のローカルファイル(要約済みテキスト)	Python	gensim (Doc2Vec)
❿	bookmarks	Server C	ファイルシステムに階層的に展開されたショートカット群を Tree View 形式で表示する	ファイルシステムに階層的に展開されたショートカット群を treeview.js によって扱い、Webブラウザから TreeView形式で見えるようにする。	Ⓛ ショートカットリスト(.list.txt)ファイル名	Webブラウザでブラウズ	Ruby (Ruby on Rails)	treeview.js
⓫	classify	Server B	収集した記事を分類・タグ付けする	オンプレミスまたはクラウドの分類処理により、収集した要約済み記事を分類・タグ付けする。	Ⓒ ショートカットのディレクトリ名に.digestを付加したディレクトリ上のローカルファイル(要約済みテキスト)	Ⓐ ショートカット群(dir)のルートディレクトリパス名	Python	要検討

Role の異なるサーバを物理的に同じサーバ上に配置してもよい

実装なし
インストールと設定のみ