実習
OpenRefineの使い方

2023年9月24日
大学図書館研究会第54回全国大会
東京大学情報システム部・前田朗

実習内容

1. インストール
2. 講師操作の繰り返し
3. 持参データ
4. 自由練習・個別相談

インストール

ダウンロードと解凍

1. OpenRefine入手
   https://openrefine.org/
   （「openrefine」で検索）
2. 「Download」ボタン
3. Zipファイル解凍
4. デスクトップに配置

設定・起動

1. 使用可能メモリを増やす
   ① openrefine.l4j.ini
   ② パラメータを3000M (3BG)に
      -Xms256M
      -Xmx1024M
2. 起動する
   ① openrefine.exe
   ② コマンドプロンプト起動確認
   ③ Webブラウザ上で画面表示確認
3. セキュリティソフトが警告を出すことがあるが許可

Microsoft Defenderの例

Windows によって PC が保護されました

Microsoft Defender SmartScreen は認識されないアプリの起動を停止しました。このアプリを実行すると、PC が危険にさらされる可能性があります

「詳細情報」
→　「実行」で許可

画面表示言語設定

1. 「Language settings」
2. プルダウンで「日本語」
3. 「Change language」

その他

- OpenRefineの終了
  - コマンドプロンプト停止
- うっかりWebブラウザを閉じてしまったら
  http://127.0.0.1:3333/
- データリセット
  1. 「既存プロジェクト」
  2. 「作業ディレクトリ閲覧」
  3. OpenRefineの終了
  4. 「作業ディレクトリ」内のファイルを全削除
  5. OpenRefineの起動

講師操作の繰返し
(repeat after me)

JAPAN SEARCH

- 国立国会図書館が提供する日本のデジタルアーカイブのポータルサイト
  https://jpsearch.go.jp/
- Web APIによりメタデータを取得できる
  https://jpsearch.go.jp/static/developer/webapi/

トップ2000件取得

Webブラウザでアクセス後、
ctrl+sで名前をつけて保存

https://jpsearch.go.jp/api/item/search/jps-cross?keyword=葛飾北斎&from=0&size=500

https://jpsearch.go.jp/api/item/search/jps-cross?keyword=葛飾北斎&from=500&size=500

https://jpsearch.go.jp/api/item/search/jps-cross?keyword=葛飾北斎&from=1000&size=500

https://jpsearch.go.jp/api/item/search/jps-cross?keyword=葛飾北斎&from=1500&size=500

全件取得するAPIも用意されているが、
手作業で行うのは難しい

データの読み込み

1. 「ファイルの選択」
2. ファイル（複数可）を指定
3. 「次へ」
4. 複数ファイル選択の場合は、選択ファイルを確認の上、「パースオプションの指定」
   （単一ファイルのときは不要）

パースオプション指定→プロジェクト作成

1. アイテム１件の範囲を指定
2. クリックで選択確定
3. プレビューを表形式データを確認し、「プロジェクトの作成」
4. 「行」「レコード」件数確認

JAPAN SEARCHでのアイテム範囲選択例
実習ではCOMMONのみ取得ください
（アーカイブズごとの独自項目は取り込まない）

表の表示

- 「行」モードと「レコード」モードを使い分ける
  - レコードの判定は先頭カラムによる
  - 文字列フィルタなどに影響
- 「前へ」「次へ」で表示ページ切り替え。１ページの表示件数を切り替え可。

カラム（列）の操作

- 各カラム（列）の操作はヘッダ行の▼プルダウンから行う
- 全カラムの操作や一部特殊処理は一番左の「全て」から行う

カラムの並び替え・削除
（レコードを正しく認識させる）

1. 「全て」の▼
2. 「カラムの編集」
3. 「カラムの並び替え・削除」
4. ユニークキーとなるカラムを、一番先頭に
5. 「OK」
6. 「レコード」モード件数確認

この研修のJAPAN SEARCHの例では次のいずれか

- 「common-id」を先頭に
- 先頭の「File」を削除

ファセット

1. 対象カラムの▼
2. 「ファセット」>「文字列ファセット」
3. 左カラムのファセット表示確認

タイトル、著者、出版社などで試してみる

クラスタリング

1. ファセット欄「クラスタリング」
   - 「全ては表示できません」メッセージが出た場合は、「カウントを制限してください」で上限緩和（有効性はマシンスペック依存）
2. 類似した文字列がグループ化表示されていることを確認
3. 方法を「最近傍法」に変更し別の結果が出ることを確認する
   - 「半径」を大きくしヒット増
   - 「文字ブロック」を小さくしヒット増

文字列フィルタ

1. 対象カラムの▼
2. 「文字列フィルタ」
3. 文字列を入力
4. 「反転」を試す
5. 正規表現マッチを試す

出版年で「明治」+「数値（半角）」のパターンをマッチさせる

^明治¥d+$

ソート

1. 対象カラムの▼
2. 「ソート」

星と旗でマーキング

1. マーキング作業
    1. 「全て」の星と旗マーク確認
    2. 星でマーキング
    3. 旗でマーキング
2. マーキング後の処理
    1. 「全て」の▼
    2. 「ファセット」＞「星ファセット」
    3. 「ファセット」＞「旗ファセット」

自動セーブとやり直し

- セーブ後から実行
  1. 青ダイヤのアイコン
  2. 既存のプロジェクト
  3. 「プロジェクト名」の値
- やり直し
  「取り消す/やり直す」

外部ファイル出力

- 「出力」＞「Excel2007+(.xlsx)
- Excelで一行１レコードにしたいときは、複数行の原因のカラムを「セル編集」＞「多値のセルを結合」で１行にまとめておく

この研修のJAPAN SEARCHの例では（あれば）「File」を削除し「category」と「subcategory」ともに「多値のセルを結合」すれば１行１レコードになる

外部データとの照合

1. 対象カラムの▼
2. 「照合（名寄せ）」＞「照合開始(reconcile)」
3. Wikidata(ja)を設定に追加
   https://wdreconcile.toolforge.org/ja/api
4. 「Wikipedia(ja)」
5. 「照合開始」
6. 対象カラムの▼
7. 「照合（名寄せ）」＞「アクション」＞「最優候補とセルをマッチさせる」
8. 書き代わりとリンク生成を確認

JAPAN SEARCHの「contributor」（寄与者）に、Wikidata(ja)で照合をかけてみる

持参データ

OpenRefines向きの持参データでない場合はこちらを

- ERDB-JP
  - https://erdb-jp.nii.ac.jp/
  - 「ドキュメント」→「検索/エクスポート」
- NACSIS-CAT図書100分の1サイズ
  - https://maedaak.github.io/biginDatacleansing4Librarian/
  - 自機関の新JAIRO Cloud
  - 「機関リポジトリURL」 + /oai?verb=ListRecords&metadataPrefix=jpcoar&**from=2023-04-01&until=2023-09-23**
  - （2ページめ以降）

    「機関リポジトリURL」 + /oai?verb=ListRecords&metadataPrefix=jpcoar**&resumptionToken=xxxx**
- researchmap文献
  - https://api.researchmap.jp/**xxxx**/published_papers?limit=1000

自由練習・個別相談