大学図書館員のための データクレンジングのはじめ方

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

「蔵書目録、機関リポジトリ、デジタルアーカイブズ… 見てくれないか、どこもメタデータだらけだ。僕たちはこれらテキストデータを整理整頓しつつ生活しなくてはならない。」

どこかの図書系職員のひとりごと

自己紹介

- 東京大学情報システム部情報基盤課学術情報チーム上席係長
- キャリアの半分が図書館システム担当部署
- テキスト処理のできる図書系職員
  - 「キーワード自動抽出システム」の開発担当
  http://gensen.dl.itc.u-tokyo.ac.jp/
- ちょっとしたツールの作成と公開
  - 図書系職員のためのアプリケーション開発講習会
  https://mbc.dl.itc.u-tokyo.ac.jp/products.html
- ここ数年はNIIの「大学図書館員のためのIT総合研修」講師も

データクレンジングとは？

講師の理解（よい定義が見つからなかったので）

- データをきれいにすること
  - **表記ゆれ**の修正（正規化）
  - **欠損値**の値をセット
  - **異常値**の検出
  - **外れ値**の除去（統計や機械学習の場合）

  など
- 機械学習や統計の**前処理として重要視**される
- **一般にデータクリーニング**も同じ意味で使われる
  - そもそもクレンジング(cleansing)自体が和製英語？
    - https://ja.wikipedia.org/wiki/%E3%82%AF%E3%83%AC%E3%83%B3%E3%82%B8%E3%83%B3%E3%82%B0
- **テキストクレンジング**という対象をテキストに限定した語も

大学図書館員ならメタデータクレンジングでは

- **メタデータに特化**したスキルを身に着けよう
  - より特化した中で業界としてのノウハウを得よう
- メタデータの**品質**の向上に寄与できる
  - 作成するメタデータは作成者に依存するが、複数人だと…
- 対象の**学問分野の知識がなくてもとりくめる**
  - 欠損値や表記ゆれ等をチェックすることはできる

図書館業務における
データクレンジング経験談

ここからは自身の経験で話します

- この講演の前に、いくつかのデータマイニング・機械学習・統計の本でデータクレンジングの記述を読んでみました
- しかし数値データ向けの話が多く、テキストデータについては表記ゆれを直す、といった簡単な記述くらいしかなさそうです
- 開き直って自身の経験をベースに話します

とっさに気づきにくいものも一度気づけば明確

この絵にある**植物以外**のものとは？

「草虫図扇面」
出典：国立文化財機構所蔵品統合検索システム
(https://colbase.nich.go.jp/collection_items/tnm/TA-708?locale=ja)

目ではみえないノイズ

- 半角スペースは見えない
  - 文字列の先頭や末尾に半角スペースが入ることがある
    - コピー＆ペースト時の手違いでよくあるのでは
  - 2つ連続の半角スペースは、単一の半角スペースと見分けづらい
- 全角スペースと半角スペースは見分けづらい
- コントロールコードはまったく見えない

誤字・脱字

- 英語に限らず**スペルミスは普通にある**
- ファットフィンガー問題
  - 指が太いと**キーボードの隣の文字**を打ちやすい？
- **担々麺と坦々麺**の区別はつきますか？

表記ゆれ

- 表記ゆれを統一（正規化）する
- 全角・半角の違い、略語、異体字、算用数字と漢数字、ハイフンの有無など
- しかし、IDで同定できればよい、との考えもありえる
  - 例えば著者IDで本人が同定できれば、都度本人が何と名乗ってもよいはず

ルールのあるデータ

- 統制語彙やコード表の利用
- ISBN, ISSN, NCIDといったコード
- メタデータスキーマが定める表記ルール
  - 文字・数値の別
  - 雑誌の巻号表記

  など

キー(ID)と値が合っていない

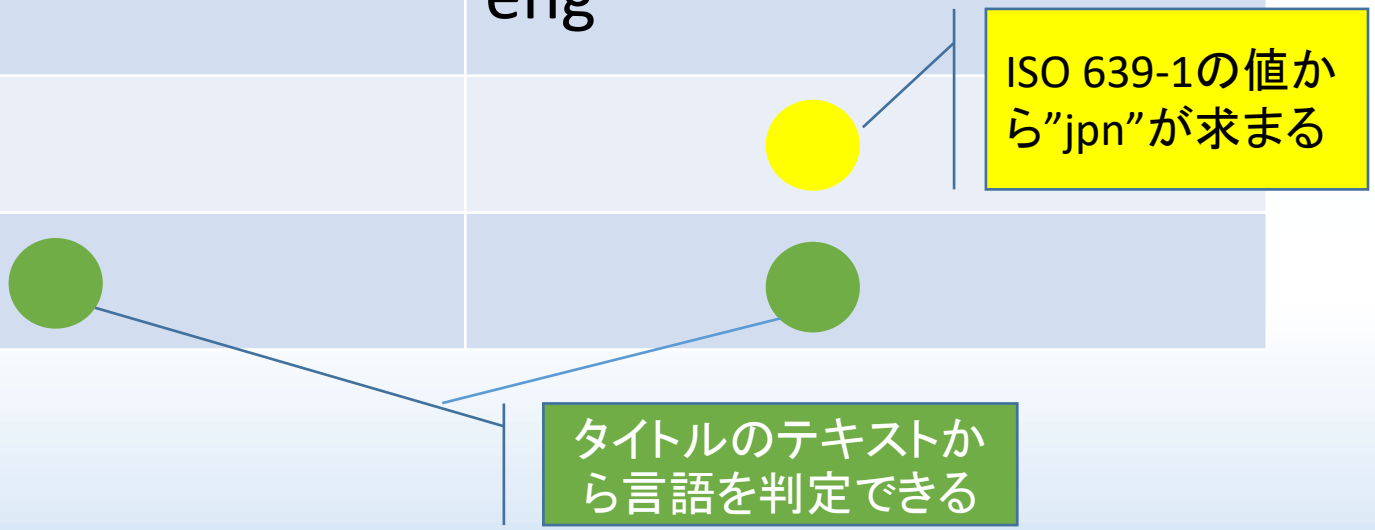
- IDが決まれば値が一意に決まるはずが…
- Excelで行コピーのつもりが連続値にしてしまうなど

NCID	タイトル
AN00081826	大学の図書館
AN00081827	大学の図書館
AN00081828	大学の図書館
AN00081829	大学の図書館

欠損値

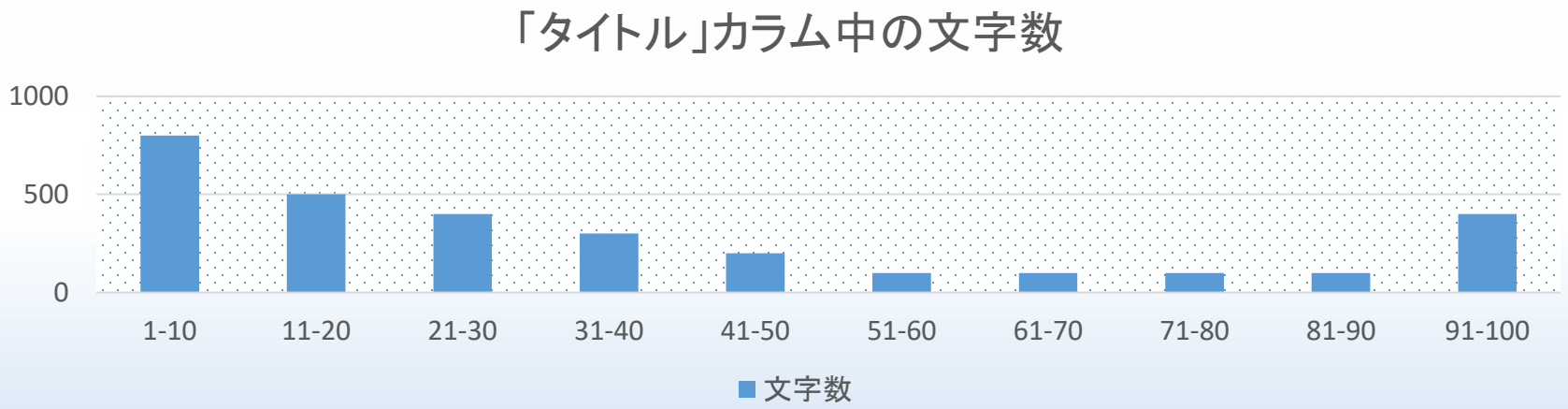
他のカラムの値から欠損値を埋められることがある

タイトル	タイトルの言語コード（ISO 639-1）	タイトルの言語コード (ISO 639-2)
Title A	en	eng
タイトルB	ja	
Title C		



異常値

- データ中の異常な値（数値）を確認する
- いっけん、メタデータでは関係ないようにみえるが
- データ項目のテキストを文字数（数値）に変換し、それをグラフ化して確認してみる。下図の例から見えてくることとは？



OpenRefineはなぜすごいのか

見ているだけでは足りない…

「見てはいるが、観察していない。差は歴然だ。例えば、君も玄関からこの部屋までの階段は何度も見ているね。」

アーサー・コナン・ドイル著、大久保ゆう訳

「ボヘミアの醜聞」よりシャーロック・ホームズのセリフ

https://www.aozora.gr.jp/cards/000009/files/226_31222.html

※青空文庫 CC-BY

ChatGPTがデータクレンジングに推奨

OpenRefine (formerly Google Refine): OpenRefineは、Windowsデスクトップ上で使用できるオープンソースのデータクレンジングツールです。データの整形、クリーニング、変換を行うのに役立ちます。CSVファイルやExcelファイルなどのデータ形式を扱うことができ、数多くのデータクリーニング操作がサポートされています。

公式ウェブサイト: OpenRefine

（以下、略）

いくつか試した限りでは、OpenRefineをまっさきに推奨してきます

半構造データ(XMLやJSON)を表形式に

- これだけでデータの一覧性が大幅によくなる


```
"paper_title": {
        "ja": "図書館員のための個人プロジェクトによる学術情報システムスキルアップ",
        "en": "Improving the skills of information professionals. Learning through personal projects : ideas for librarians to improve the skills for academic information systems"
    },
    "authors": {
        "ja": [
            {
                "name": "前田朗"
            }
        ],
        "en": [
            {
                "name": "Akira Maeda"
            }
        ]
    },
    "published_paper_owner_roles": [
        "lead"
    ],
```


講師のresearchmapの業績リスト(JSONフォーマット)の一部抜粋
人がみても読みやすくは作られているが、一覧性がよいわけではない

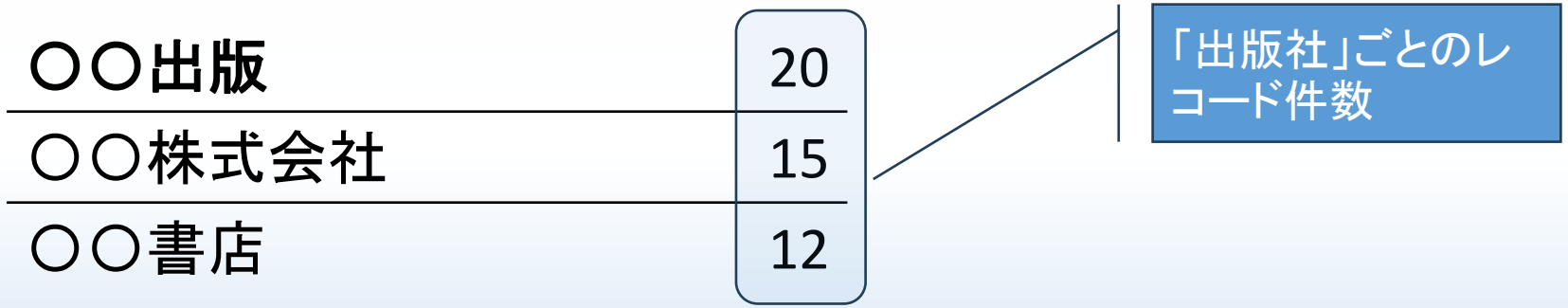
正規表現によるパターンマッチ

- 正規表現はテキスト処理では、定番とさえいえる
  - 単純な前方一致や中間任意よりも複雑な文字列パターンを指定できる
  - 大体のテキストエディタでサポート
  - Excelに対するOpenRefineの利点のひとつ
- パターンに合致するレコードを取り出せる
  - 逆にパターンにマッチしない（問題データ）も取り出せる
  - パターンの例: メールアドレス, URL, ISBN
- しかし、チェックデジットの確認はできない
  - 関数やプログラミングの領域となる
  - OpenRefineでもGRELという言語があり、ISBNのチェックデジットチェックも可

ファセットによるデータの通覧

- ファセット
  - カラム内の一致した値でレコードをグルーピング
  - データを圧縮して見せるので情報を確認しやすい

例：「出版社」カラムをテキストファセット化



クラスタリングによる名寄せ

- クラスタリング
  - ファセットをさらに似たもの同士でグループ化
    - つまり「似て非なる」情報が集まる
  - 「表記ゆれ」や「誤記」の可能性があるものを確認できる
  - パラメータ指定で挙動を調整できる

例:「著者」カラムをクラスタリング

Maeda Akira
Akira maeda
Maeda, Akira

レコードのマーキング機能

- 気になったレコードにワンクリックでマーキング
- マーキングは次の２種類
  - スター　→　Good Dataマーキング用
  - 旗　→　Bad Dataマーキング用

外部データとの照合

- 基本機能で使用可能
  - Wikidata record link (en)（Wikipedia(en)語彙) ※設定不要で使用可能
  - Wikidata (ja) (Wikipedia(ja)語彙）
  - VIAF (バーチャル国際典拠ファイル）

  など
- 拡張機能(RDF Extension)で各種RDFとの照合が可能
  - NDC （(RDF形式で配布されている）

  など
- 外部プログラム(Fuseki)を使えば任意のTSVとの照合も可能

完全一致ではなく、**あいまいマッチであることに注意**

OpenRenfineは目録業務にも？

MARC21レコードをOpenRefineに取り込むことができる
（しかし、同じMARCでもNACSIS-CATのCATPレコードは取り込めない）

OpenRefine以外で使えそうなツール

- Microsoft Excel　→　事務用アプリケーションの定番
- pandas (Python)　→　表データ処理、プログラミングができることが前提
- ChatGPT　→　試したところ表記ゆれを調べてくれた。他にもできそう。
- 講師の自作ツール類
  - Perlモジュール Lingua::LanguageGuesser （テキストの言語判定）
    http://gensen.dl.itc.u-tokyo.ac.jp/LanguageGuesser/LanguageGuesser_demo_ja.html
  - NACSIS-CAT雑誌所蔵の形式チェック
    https://mbc.dl.itc.u-tokyo.ac.jp/shs/
  - Fuurin Checker　(Sudachi辞書を使った表記ゆれチェック）
    https://github.com/maedaak/FuurinChecker
  - junii2データ診断
    https://mbc.dl.itc.u-tokyo.ac.jp/junii2checker/

参考情報

- 大学図書館員のためのIT総合研修 2023年度（テーマ：WebAPIを使ったデータの入手とその整備）
  - WebAPIとデータクレンジングについての３日コースの研修
  - https://contents.nii.ac.jp/hrd/it/2023/result

参照文献

- 「応用基礎としてのデータサイエンス」（講談社） ISBN:987-4-530789-2
  - 「欠損値」、「外れ値」、「異常値」についての記述を参照
- 「テキストマネジメント」（岩波書店） ISBN:978-4-00-029899-5
  - データクリーニングとして、「テキストの正規化」、「不要テキスト削除」、「欠損値の補完」、「不要語削除」の記述あり
- 「実践自然言語処理」（オライリー） ISBN:978-4-87311-972-4
  - 「ファットフィンガー問題」についての記述を参照
- OpenRefine https://openrefine.org/

ここからはいよいよ実習！