notes for open refine lesson

Einführung

Was ist open Refine?

- OpenRefine ist ein Desktop Programm, das einen Web Browser als graphisches Interface nutzt.
- Es ist nützlich, wenn man Daten in tabellarischer Form hat, also CSVs oder TSVs bei denen allerdings das Datenformat oder die Terminologie inkonsistent sind.
- Außerdem können Daten explorativ betrachtet werden. Zum Beispiel: wie oft kommt ein bestimmter Wert in einer Spalte vor

Beispiele:

https://librarycarpentry.org/lc-open-refine/01-introduction/index.html

Was noch wichtig ist wenn man mit OpenRefine arbeitet:

- Es ist keine Internetverbindung nötig. Daten und Befehle werden nicht an einen Server geschickt
- Die Original-Daten werden nicht verändert und bleiben erhalten
- Speichern passiert automatisch

Dateien einlesen

- Es können verschiedenste Datei-Formate eingelesen werden, darunter sind:
  - TSV
  - CSV
  - Excel
  - JSON
  - XML
  - Google Spreadsheets

Ein OpenRefine project erstellen:

- "Create Project"
- "Get data from This Computer"
- "Choose Files" or "Browse" and select "doaj-article-sample.csv"
- "Next"

- Bei “Character encoding” muss “UTF-8” ausgewählt werden um sicherzustellen, dass spezielle Zeichen (ß, ä, ö ü) korrekt dargestellt werden.
- Haken setzen bei “Parse next 1 line(s) as column headers” um sicherzustellen, dass die Werte der ersten Zeile als Spalten-Namen genutzt werden.
- Kein Haken bei “Parse cell text into numbers, dates, …” damit OpenRefine nicht automatische Zahlen erkennt, was zu Fehlern führen kann.
- Projekt-Name kann bei “Configure Project Name” geändert werden. Standardmäßig wird der Name der Datei als Projektname verwendet
- KEIN Haken bei ”Trim leading & trailing whitespace from strings Escape special characters with ”
- “Create Project >>” oben rechts klicken

Layout of OpenRefine, Rows vs Records

- Die Gesamtzahl der Zeilen steht oben links
- Die Anzahl der angezeigten Zeilen kann ausgewählt werden
- Über “next” “previous” usw. kann navigiert werden
- Datenmanipulation geschieht meistens über Dropdown-Menüs einzelner Spalten
- Zwei Modi um Daten anzuzeigen: “Rows” und “Records”, Jede “Row” steht für einen “Record”. Im “Record Modus” können mehrere “Rows” zu einem “Record” zusammengefügt werden.

Zellen splitten

- Momentan stehen mehrere Autoren in einer Zeile / “Row”, getrennt durch ein “|” (Pipe-Symbol). Wir können die einzelnen Autoren auf einzelne Zeilen aufteilen.
- “Authors”, “Edit Cells”, “Split multi-valued cells”, “|” als Separator auswählen
- 4009 Zeilen anstatt 1001, da jeder zusätzliche Autor einer eigenen Zeile zugeordnet wird
- Show as “Records” klicken
- Die Nummerierung ändert sich, da jetzt mehrere Zeilen/“Rows” zu einem “Record” zusammengefasst werden.

Zellen zusammenfügen

- Nachdem Zellen gesplittet wurden, können wir sie jetzt wieder zusammenfügen.
- Ein typischer Workflow mit OpenRefine wäre:
  - Zellen splitten
  - Einzelne Werte bereinigen
  - Zellen wieder zusammenführen

“Authors”, “Edit cells” “Join multi-valued cells”

- Separator bzw. Delimiter müssen bedacht ausgewählt werden.
- Sie sollen nicht in den Daten vorkommen.
- Kommas, Doppelpunkte und Semikolons sind oft schlechte Trennzeichen, da sie in den Daten vorkommen. zb. Trennung von Vorname und Nachname des Autors

Übung

- Welcher Seperator wird in der “Subjects” Spalte verwendet?
- Einzelne Werte der Subjects Spalte trennen
- Danach die Einzel-Werte wieder zusammenführen

Faceting und Filtering

Facets / Facette

- Eine **Facette** gruppiert alle Werte einer Spalte und ermöglicht es diese zu filtern und zu editieren
- “Publisher”-Spalte -> “Facet” -> “Text Facet”
- Nach Namen “name” oder Anzahl “count” sortieren
- Auswahl treffen um nur Einträge eines Wertes anzuzeigen: “Society of Pharmaceutical Technocrats”
- Bei anderem Publisher “Society of Pharmaceutical Technocrats” “Include” klicken um diese Werte der Auswahl hinzuzufügen
- “Invert” wählen um Auswahl umzukehren

Übung

Text Facet für die “license” Spalte erstellen um folgende Fragen zu beantworten:
- Wie heißt die häufigsten Lizenz? - CC BY

- Wie viele Artikel aben keine Lizenz?

– 6

Filter

- Text Filter können verwendet werden um nur Zeilen / “Rows” anzuzeigen, die einen bestimmten “Text” enthalten:
- “Title”, “Text filter”, “synthesis”

Andere Facetten

- Numeric oder Numerische
- Timeline “Zeitstrahl”
- Scatterplots “Streudiagramm”

weitere Facetten unter “Customized facets”

Aufgabe

- Alle Einträge anzeigen, die keine DOI haben
  - “DOI” “Facets” > “Customized facets” > “Facet by blank” > “True”

Aufgabe

- Text Facette auf der “language” Spalte bilden und die Variationen von “EN” und “English” korrigieren
  - “language” > “Facet” > “Text facet” > Maus über “Englisch” > “Edit” > “EN” eingeben > “Apply”

Clustering

Die Cluster-Funktion gruppiert ähnliche aber inkonsistente Werte und ermöglicht es diese zu verändern bzw. anzuglechen

- Die “Cluster” werden automatisch auf Grund verschiedener Algorithmen gewählt.
- Eventuell muss mit den verschiedenen Algorithmen experimentiert werden um das gewünschte Ergebnis zu erhalten
- Standardmäßig wird der häufigste Wert genutzt um die anderen Werte anzupassen.
- Aber es kann auch ein anderer Wert ausgwählt werden

  - "Author"-Spalte -> "Edit cells" -> "Split multi-valued cells" -> "|" als Separator auswählen
  - "Edit cells" -> "Cluster and edit"
  - "key collision"-Methode und "fingerprint"- Keying Function auswählen
  - Algorithmen wechseln. Welcher funktioniert am Besten?
  - Werte zusammenführen / mergen

Spalten verändern und sortieren

"All" erste Spalte "Edit columns" -> "Re-order / remove columns" ....

- drag and drop Spalten Namen um sie neu anzuordnen oder zu entfernen
- Spalten umbenennen "Edit column" > "Rename this column"
- Spalten sortieren: "Title" > "sort" > "a-z", "sort" > "permanently"

Einführung Transformations

- Manche Änderungen können mit den bisher erlernten Methoden nicht realisiert werden. Zum Beispiel:
  - Daten aus einer einzelnen Spalte in mehrere Spalten splitten
  - Standardisierung eines Datenformats in einer Spalte ohne die tatsächlichen Werte zu ändern (z.B. Standardisieren eins Datums formats oder Satzzeichen entfernen)
  - Teile eines Strings aus einem längeren String extrahieren (z.B. ISBNs in einer Zitation)
- Transformationen sind in einer extra Sprache geschrieben, genannt: GREL (General Refine Expression Language)

Aktivität

- Text Facet von "Publisher" erstellen:
  - "Publisher"- Spalte > "Facet" "Text facet"
  - "Edit cells" -> "Common transforms" -> "Trim leading and trailing whitespace"
  - Hat sich die Facette geändert? Eventuell Refresh klicken

Transformations schreiben

“Edit cells” -> “Transform”

GREL supports two types of syntax:

```
value.function(options)
function(value, options)
```

Use Facets and the GREL expression value.toTitlecase() to put the titles in Title Case

`Facet by publisher`

- “Publisher”- Spalte > “Facet” “Text facet”
- Auswählen von “Akshantala Enterprises” und “Society of Pharmaceutical Technocrats” (kann mit “include” gemacht werden)
- Alle Title sind großgeschrieben
- Dropdown Menü in der “Title”-Spalte anklicken
- “Edit cells” -> “Transform”
- Mit “OK” bestätigen

Undo und Redo

Auf Undo und Redo klicken:

- Einzelne Schritte anklicken um sie Rückgängig zu machen.
- Wieder darauf klicken führt die Schritte wieder aus
- Der gesamte Workflow kann in eine JSON Datei exportiert werden
- JSON Dateien mit Open-Refine Befehlen können importiert und ausgeführt werden

Transformation von Kalenderdaten und Zahlen

- Alle Facetten und Filter entfernen
- Bei der “Date”- Spalte, “Edit cells” -> “Transform”
- GREL expression “value.toDate(”dd/MM/yyyy“)” und mit “Ok” bestätigen
- Werte werden jetzt in Grün dargestellt und folgen einem Standard-format (ISO 8601)
- Nun können Funktionen ausgeführt werden die speziell für Datumsangaben geschrieben wurden:

- “Date”- Spalte > “Edit column” > “Add column based on this column”.
- Bei “New column name” “Formatted-Date” reinschreiben
- In GREL expression Box: “value.toString(”dd MMMM yyyy“)”

Transformation advanced

- “Author”-Spalte > “Edit cells” > “Split multi-valued cells” > Separator “|”
- “Facet” > “Custom text facet”
- Expression box: “value.contains(”,“).toString()”
- “Authors”-Spalte > “Edit cells” > “Transform”
- Expression box: “value.match(/(.),(.)/)”
- / bedeutet es werden Regular Expression verwendet, .* steht für eine beliebige anzahl an beliebigen Zeichen
- wir erhalten ein Array bzw. Liste
- In der Expression Box muss der Befehl erweitert werden : “value.match(/(.),(.)/).reverse().join(” “)”

Daten exportieren

- Auf Exportieren klicken und verschiedene Dateien auswählen