onepress

ŁUKASZ ŻYŁA

DZIENNIKARSTWO DANYCH I DATA STORYTELLING

Redaktor prowadzący: Barbara Gancarz-Wójcicka

Helion S.A.
ul. Kościuszki 1c, 44-100 Gliwice
tel. 32 231 22 19, 32 230 98 63
e-mail: *onepress@onepress.pl*
WWW: *http://onepress.pl* (księgarnia internetowa, katalog książek)

Drogi Czytelniku!
Jeżeli chcesz ocenić tę książkę, zajrzyj pod adres
http://onepress.pl/user/opinie?dzidan
Możesz tam wpisać swoje uwagi, spostrzeżenia, recenzję.

ISBN: 978-83-283-8312-8

Printed in Poland.

Spis treści

Rozdział 5.

Czyszczenie i agregacja danych

Kiedy przymierzamy się do stworzenia materiału dziennikarskiego, najpierw, rzecz jasna, szukamy informacji. Czasami poszukujemy jej w ciemno, czasami kierując się intuicją, a innym razem po prostu logicznie wyciągając wnioski, potrafimy je z pewnym prawdopodobieństwem przewidzieć.

W przypadku dziennikarstwa danych zamiast informacji poszukujemy danych. Nie pozostajemy przy samej informacji; zamiast tego sprawdzamy, na podstawie jakich danych została utworzona (o tym, że informacje składają się z danych, pisałem w rozdziale 4.). Natomiast poszukiwanie źródeł informacji, czyli danych, nie różni się niczym innym od poszukiwania informacji, tzn. tak samo często kierujemy się intuicją w poszukiwaniach, co również zostało opisane. W przypadku tradycyjnego dziennikarstwa to poszukiwanie zajmuje najwięcej czasu, potem czeka nas już tylko przygotowywanie materiału (tekstu, wideo, reportażu radiowego) i oddanie do redakcji.

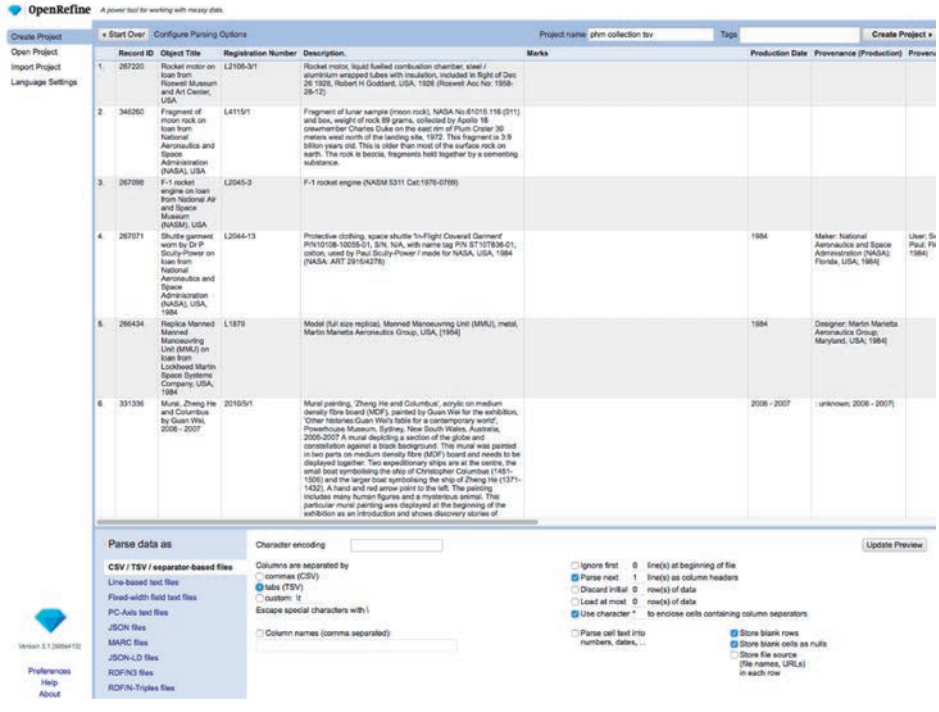
Przygotowanie i oddanie do składu zajmuje zdecydowanie mniej czasu aniżeli wyszukiwanie informacji. Odmiennie jest z materiałem dziennikarstwa danych. Po wyszukaniu danych konieczna jest żmudna praca przy ich przygotowywaniu i analizie. Zdarza się często, że jakość danych, które pozyskaliśmy (zwłaszcza danych publicznych), pozostawia wiele do życzenia. Trzeba sobie zdawać sprawę, że instytucje (zwłaszcza publiczne), które gromadzą dane, nie robią tego, aby można było na ich podstawie prowadzić jakieś analizy. Dane gromadzone są często do celów administracyjnych czy biurokratycznych. Dlatego urzędnicy niespecjalnie przywiązują wagę do jakiejkolwiek standaryzacji prezentacji danych. Mając jednak dane w postaci tabelarycznej w takich formatach jak *.csv* czy *.xls*, możemy je odpowiednio wyczyścić, sformatować, a następnie posortować i przefiltrować.

5.1. Pozbądź się śmieci

Na początek zajmiesz się czyszczeniem danych. Aby odnaleźć jakieś wzorce w danych, wywnioskować jasne i klarowne informacje, musisz te dane wyczyścić. Co to oznacza? Musisz zrobić wszystko, żeby dane stały się zrozumiałe i przejrzyste w formie tabeli. Tabele nie mogą zawierać pustych miejsc (np. spacji czy pustych komórek), wszystkie litery powinny być ustandaryzowane, w kolumnach mają się znajdować dane tylko jednego typu. Nie da się przecież pracować na tabeli, jeśli w kolumnie *X* umieszczone są numery i nazwy miast — takiej kolumny nie da się np. zsumować.

Na tym etapie pracy zachęcam Cię do korzystania z różnych narzędzi i do raczej intuicyjnego podejścia. Jeżeli dysponujesz małą tabelą w formacie *.xls*, w której znajdują się połączone kolumny, w dodatkowych wierszach jakieś opisy itp., to może wystarczy po prostu wyczyścić ją ręcznie w Google Sheets? Lecz jeśli wierszy masz ponad 100, lepiej skorzystać z automatycznych rozwiązań. Jednym z najbardziej popularnych narzędzi do sporych zbiorów jest Open Refine. Ta aplikacja (do 2012 r. funkcjonowała jako Google Refine) to narzędzie ułatwiające porządkowanie i przetwarzanie danych, umożliwiające import danych z wielu formatów: *.tsv*, *.csv*, *.*sv*, *.xls* i *.xlsx*, *.json*, *.xml*, *.rdf*, *.xml*. Open Refine ułatwia sortowanie i dzielenie danych, a także ich przetwarzanie z wykorzystaniem tzw. wyrażeń regularnych. Swoją pracę można w łatwy sposób eksportować do popularnych formatów (*.csv*, tabela *.html* czy pliki Excela i *.odt*). Z Open Refine można korzystać w systemach linuksowych, macOS oraz w Windowsach. Narzędzie dystrybuowane jest na licencji *open source*.

Aplikację pobierzesz ze strony openrefine.org. Żeby ją zainstalować, postępuj zgodnie z instrukcjami. Aplikacji używa się w przeglądarce. Aby zacząć pracę z plikiem z danymi, kliknij *Create project*, a następnie skorzystaj z możliwości narzędzia i wybierz sposób pobrania danych. Oprócz tradycyjnego pobrania pliku z komputera możesz także umieścić link do adresu URL czy bezpośrednio połączyć program z arkuszem Google'a. W aplikacji pojawi się tabela, a pod nią dodatkowe możliwości jej przetworzenia (rysunek 5.1).



Rysunek 5.1. Open Refine to rozbudowany program do czyszczenia i agregacji danych. Na początku z pewnością wymaga sporej uwagi i skupienia.

Źródło: materiały własne

Program domyślnie ustawi pierwszy wiersz jako nagłówek. Możesz także ustawić sposób parsowania Twojego pliku, np. w przypadku *.tsv* separacja komórek powinna być zastosowana za pomocą tabulatorów, a w przypadku *.csv* za pomocą przecinków. Program zazwyczaj wszystkie te informacje wykrywa i automatycznie Ci w tym pomaga, poprzez dolny panel możesz jednak dokonać manualnie kilku zmian w sposobie pobierania Twojej tabeli.

Gdy uda Ci się już zaimportować całą tabelę, możesz sobie poklikać różne miejsca, aby bliżej zapoznać się z interfejsem. Do manipulowania danymi służą funkcje dostępne poprzez kliknięcie ikonki trójkąta przy nagłówkach kolumn. Daną kolumnę możesz np. zwinąć, klikając *view* i dalej *Collapse this column*; jeżeli chcesz, aby kolumna z powrotem się pojawiła, wystarczy kliknąć puste miejsce po niej. Kolumny możesz też przesuwać w prawo lub w lewo, aby wygodniej Ci się na nich pracowało. Możesz także zmieniać nazwy kolumn przy użyciu funkcji *rename*. Na bocznym panelu z lewej strony możesz prześledzić historię swoich zmian i powrócić do wcześniejszych ustawień.

Po zakończeniu pracy możesz wyeksportować projekt do różnych formatów: *.csv*, *.tsv*, *.xls*, *.xlsx* czy tabeli w dokumencie *.html*. Jeżeli pracujesz z bardzo dużymi zbiorami danych, możesz zwiększyć alokację zużywania pamięci dla swojej aplikacji. Możesz to zrobić zgodnie z instrukcją znajdującą się na stronie kodu źródłowego: *https://github.com/OpenRefine/OpenRefine/wiki/FAQ%3A-Allocate-More-Memory*.

W celu pokazania możliwości programu wykorzystam zbiór danych ze strony dane.gov.pl, zawierający listę beneficjentów (projektów) Funduszy Europejskich w latach 2014 – 2020. Plik bez problemu wyszukasz na stronie (ponieważ instytucje publiczne często zmieniają adresy URL swych zasobów, nie będę podawał aktualnego linku).

5.1.1. Sortowanie

Dane z tabeli możesz posortować, co przyda Ci się do późniejszej analizy. Do tego wykorzystasz funkcję `sort`, która pojawi się po kliknięciu ikony trójkąta przy nazwie kolumny. Daną kolumnę możesz posortować ze względu na tekst (od A do Z lub na odwrót), liczby (od najmniejszych do największych lub na odwrót), daty (od najwcześniejszych lub najpóźniejszych) i format *boolean*, czyli dane, które mają wartość prawdy lub fałszu. Metodą „przeciągnij i upuść" (ang. *drag & drop*) możesz ustalić, w jakiej kolejności mają się pojawić np. puste komórki lub te z błędem.

5.1.2. Wyszukiwanie fasetowe

Open Refine daje Ci możliwość analizowania każdej kolumny np. pod względem powtarzanych cech w wierszach. Cechy informacji, które powtarzają się w komórkach, to fasety i na ich podstawie w Open Refine tworzy się filtry. Proces ten nazywa się wyszukiwaniem fasetowym. Możliwości wyszukiwania fasetowego to przede wszystkim sposobność spojrzenia na dane z różnych perspektyw. Metoda ta nie zmienia wartości komórek. Przetestuj ją na kolumnie *Nazwa beneficjenta*.

Wybierz *facet*, a następnie *Text facet*. W lewym panelu pojawia się efekt funkcji. Zmień *Sort by* na *count*. Jak widzisz, na pierwszym miejscu jest Łódź, która otrzymała fundusze na 121 projektów. Widzisz także *21497 choices*, co oznacza, że łącznie 21 497 podmiotów uzyskało dofinansowanie na swoje projekty.

Funkcję tę można zastosować także do wartości liczbowych. Aby jednak tego dokonać, musisz się zapoznać z inną możliwością Open Refine. Zwróć uwagę na kolumnę *Poziom unijnego dofinansowania w procentach*. Wartości zapisane są z przecinkiem, co oznacza, że nie są one liczbami. Cyfry dziesiętne w programie Open Refine oddzielane są kropką. W tym przypadku są to przecinki, zatem wartości w tej kolumnie to tekst. Aby wykorzystać funkcję do wyszukiwania fasetowego na liczbach, trzeba wartości z kolumny przekonwertować na wartość liczbową.

W programie masz do czynienia z czterema typami danych: tekst (*string*), liczba (*integer*), *boolean* (wartość prawda lub fałsz), *data*. Każda komórka zawierająca jakąś wartość musi być jednym z tych czterech typów danych. W celu analizy danych możesz dokonać transformacji jednego typu na inny.

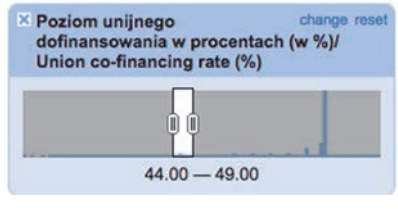
Aby przekonwertować wartość tekstową na liczbową, musisz najpierw zamienić przecinek na kropkę, a następnie dokonać transformacji. Do tego możesz użyć zaawansowanej możliwości tworzenia tzw. wyrażeń regularnych za pomocą Google Refine Expression Language, czyli GREL. GREL działa tak samo jak tworzenie formuł w Excelu. Kliknij więc ikonę trójkąta kolumny, potem *Edit cells* i wybierz *Transform*. W ramce, gdzie zapisana jest nazwa `value`, musisz wpisać następującą regułę: `value.replace(,,',','.').toNumber()`.

`value.replace(,,',','.')` oznacza, że w wartości komórki (`value`) odnajdujesz przecinek (pierwszy znak w apostrofie — `','`), który zastępujesz kropką (drugi znak w apostrofie `'.'`). Przecinek i kropka w apostrofach oddzielone są przecinkiem.

`toNumber()` — program dokonuje transformacji wartości (`value`) na typ liczbowy (`number`).

Kliknij *OK*. Jak widać, wartości zmieniły kolor na zielony. Oznacza to, że są teraz typem liczbowym.

Możesz teraz skorzystać z funkcji fasetowej. Kliknij *facet*, a następnie *numeric facet*. W lewym panelu pojawiła się ramka z grafiką (rysunek 5.2).



Rysunek 5.2.
Suwak regulujący zakres liczbowy.
Źródło: materiały własne

Suwakiem możesz np. ustawić obszar 40% dotacji. Możesz teraz dodać kolejny filtr, czyli *facet*, z nazwami beneficjentów. Kliknij przy kolumnie z nazwami beneficjentów *facet* i *text facet*. W ten sposób w lewym panelu pojawi się lista podmiotów, które otrzymały dotacje na poziomie 40% wartości projektu.

5.1.3. Wykrywanie duplikatów

W zbiorach danych często będziesz miał do czynienia z powtarzającymi się rekordami. Duplikaty to rekordy, które pojawiają się minimum podwójnie. W tym fragmencie posłużysz się przykładem z zagranicznej bazy danych, a konkretnie z The Museum of Applied Arts and Sciences z Sydney, którą można znaleźć w serwisie data.nsw.gov.au (wszelkiego rodzaju zbiory muzealne to dobre miejsca, na których można szlifować umiejętności czyszczenia baz). Obecnie muzeum umożliwia dostęp do swych zbiorów poprzez API. Ty jednak skorzystasz z pliku *.tsv*, który wcześniej został pobrany z API jako plik *.json*.

Plik umieść w programie w podany wcześniej sposób. Aby sprawdzić liczbę zduplikowanych rekordów, posłuż się w tej bazie polem *Registration Number*. Następnie wybierz *Facet*/*Customized facets*/*Duplicates facet*. Jak widzisz, w bocznym panelu z wartością *true* znajduje się 667 rekordów. Oznacza to, że 667 rekordów bazy danych jest zduplikowanych. Po kliknięciu *true* w kolumnie *Registration Number* zauważysz na górze puste pola. Oznacza to, że program traktuje puste pola jako zduplikowane. Musisz je wyodrębnić. Kliknij więc *Registration Number*/*Facet*/*Customized facets*/*Facet by blank*. Jeśli zaznaczysz *false*, przekonasz się, że teraz rekordów zduplikowanych jest 37. Aby przyjrzeć się tym rekordom, wybierz dodatkowo *text facet*. Posortuj teraz te rekordy liczbą (*count*). Zauważ, że na pierwszym miejscu znalazł się rekord, który ma nawet 4 duplikaty. Następnie spróbuj nieco inną metodą sprawdzić duplikaty w kolumnie *Record ID*. Jako że fasetowe wyszukiwanie duplikatów nie przeszukuje wartości liczbowych (*integers*), zacznij od posortowania tej kolumny. Wybierz *Sort by* na pasku nad nagłówkami kolumn i dalej *By Rekord ID*, zaznacz *numbers* i wybierz *smallest first*. Podczas sortowania zaznacz też *Reorder rows permanently*, aby zmiany w sortowaniu zostały zapisane na trwałe. Niebawem dowiesz się, jak usunąć duplikaty.

5.1.4. Zastosowanie filtru tekstowego

Spróbuj teraz znaleźć w zbiorach muzeum jakieś rekordy, które mogą dotyczyć naszego kraju. Sprawdź tytuły prac. Wybierz *Object Title*, a następnie *Text filter*. W bocznym panelu wpisz `Poland`. Brawo, znalazłeś 7 rekordów! Filtrem tekstowym możesz teraz sprawdzić kolumnę kategorii zbioru dzieł. Po wybraniu kolumn *Categories* i *Text filter* wprowadź `|`. W ten sposób sprawdzisz rekordy, które dotyczą więcej niż jednej kategorii. Jak widzisz, większość rekordów dotyczy więcej niż jednej kategorii. Wpisz teraz podwójny podzielnik `||`. Okazuje się, że jest 8 rekordów z błędnie wpisanymi podziałami kategorii. Dzięki filtrowaniu tekstowemu nie tylko możesz wyszukać różnego rodzaju dane, ale także wykryć nieprawidłowości.

5.1.5. Transformacje komórek

Oprócz filtrowania i analizy zbioru danych Open Refine pozwala Ci także na transformacje i edycję komórek. W programie umożliwia to zakładka *Edit cells*. Czasami musisz np. „oczyścić" wartości w komórkach, połączyć je czy usunąć niepotrzebne spacje.

Transformacje możesz zacząć od usuwania białych znaków (np. niepotrzebnych spacji znajdujących się na początku lub na końcu wartości w komórkach). Sprawdź więc tym razem w bazie dotacji unijnych kolumnę z tytułami projektów. Aby wykryć białe znaki, wybierz *Tytuł projektu*/*Edit cells*/*Common transforms*/*Trim leading and trailing whitespace*. Jak widzisz na rysunku 5.3, program usunął białe znaki w 1217 komórkach.

Rysunek 5.3. Program umożliwia szybkie usuwanie zbędnych komórek.

Źródło: materiały własne

Należy dodać, że tego typu operacje przeprowadza się tylko na typach tekstowych, czyli *string*, a nie numerycznych, czyli *integer*. Możesz także skorzystać z metody *Collapse consecutive whitespace*, która usuwa białe znaki ze środka stringów.

Jeżeli Twoja baza oparta jest na danych pochodzących ze strony internetowej, może się zdarzyć, że komórki będą zawierać pozostałości kodu HTML-a. Na przykład litera „ś” w kodzie HTML-a zapisywana jest jako `&#347;`. Jeżeli znajdziesz w swojej bazie jakieś komórki, w których treść zaczyna się od & i kończy się średnikiem (;), możesz zmienić sposób wyświetlania takich znaków na czytelny. Zrobisz to, klikając w danej kolumnie *Edit cells*/*Common transforms*/*Unescape HTML entities*. W bazie dotacji unijnych możesz sprawdzić w ten sposób kolumnę *Skrócony opis*. Okazuje się, że program zamienił 466 komórek, wykrywając m.in. takie znaki jak cudzysłów, który wcześniej był zapisany w formie kodu HTML-a jako `&rdquo;`.

Kolejną możliwością, jaką dają transformacje, jest zmiana wielkości liter w komórkach tekstowych (np. z małych na duże i odwrotnie). Aby zmienić tekst komórki na duże litery, kliknij daną kolumnę, a następnie *Edit cells*/*Common transforms*/*To uppercase*. Wykorzystanie odpowiednich metod zależy od tego, w jaki sposób chcesz przekształcić teksty w poszczególnych kolumnach.

Więcej możliwości edycji komórek daje z pewnością użycie wyrażeń regularnych. Jeżeli zaczniesz pracować z Open Refine, warto, byś zapoznał się z możliwościami, jakie daje wykorzystanie właśnie wyrażeń regularnych. Więcej informacji na temat General Refine Expression Language znajdziesz pod linkiem *https://github.com/OpenRefine/OpenRefine/wiki/General-Refine-Expression-Language*.

5.1.6. Usuwanie niewłaściwych danych

Wiesz już, jak filtrować dane i jak wykorzystać wyszukiwanie fasetowe. Wykrycie duplikatów i powtarzających się elementów w treści komórek to dopiero część Twojej pracy. W dalszej kolejności powinieneś pozbyć się tych niepotrzebnych komórek. I tutaj zaczyna się to, co nazywamy czyszczeniem danych, czyli usunięcie konkretnych komórek z bazy danych. Pamiętaj, że usuwanie komórek powinno się odbywać po uprzednim wykorzystaniu metod *facet* i filtrów, w przeciwnym razie narazisz się na usunięcie wszystkich komórek.

Wróć teraz do wcześniejszej bazy muzeum z Sydney i wykonaj te same kroki. W kolumnie *Registration Number* wybierz *Facet*/*Customized facets*/*Duplicates facet*. W bocznym panelu z wartością *true* znajduje się 667 rekordów. Jak już wiesz, wśród tych rekordów są także puste pola, których jednak nie chcesz się pozbyć. Musisz je więc wyodrębnić. W tym celu kliknij *Registration Number*/*Facet*/*Customized facets*/*Facet by blank*. Gdy zaznaczysz *false*, przekonasz się, że teraz rekordów zduplikowanych jest 37. Nadszedł czas, aby je usunąć. W nagłówku pierwszej kolumny *All* kliknij *Edit rows*/*Remove all matching rows*. Jak widzisz w komunikacie, właśnie usunąłeś 37 wierszy.

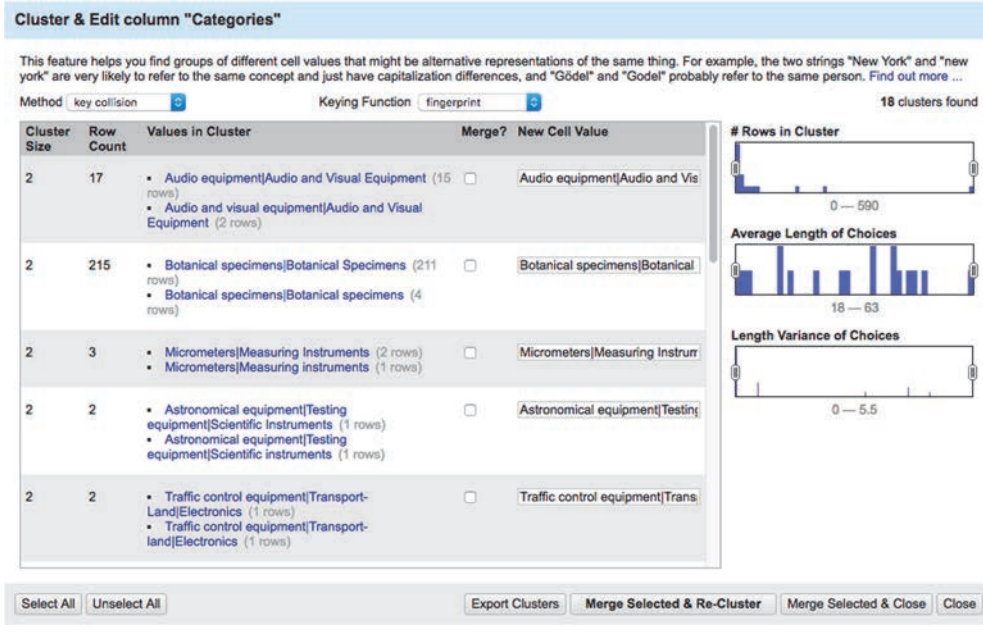
5.1.7. Grupowanie danych, czyli klastrowanie

W ostatnim przykładzie zajmiesz się problemem, który często pojawia się w bazach danych, zwłaszcza gdy masz do czynienia z danymi publicznymi. Tego typu bazy zazwyczaj są tworzone przez wiele osób, dlatego często można napotkać różne nazwy dotyczące tych samych elementów. W kategoriach wydatków możesz natknąć się np. na: *Wydatki na usługi zdrowotne | zdrowie*, *Wydatki na usługi zdrowotne | Zdrowie* i *Wydatki na zdrowie*. Jak można się domyślić, wszystkie dane objęte tymi kategoriami dotyczą tego samego rodzaju wydatku. Problem ten rozwiązuje tzw. klastrowanie. To nic innego jak grupowanie obiektów o podobnych właściwościach. Open Refine daje Ci możliwość automatycznego poradzenia sobie z problemem grupowania.

Wykorzystasz teraz bazę muzeum z Sydney. Aby sprawdzić, czy podobna sytuacja występuje w tej bazie, sprawdź kolumnę kategorii dzieł. Do klastrowania kolumny użyj metody *cluster*. Znajdziesz ją w *Edit cells*/*Cluster and edit*. Zostanie wyświetlony panel z pogrupowanymi kategoriami. Program Open Refine stosuje algorytmy skonstruowane z wielu metod pozwalających wykryć podobieństwa w odpowiednich kategoriach.

Przeanalizuj ramkę z rysunku 5.4. *Cluster Size* oznacza liczbę różnych oznaczeń tych samych kategorii, jaką znalazł program. *Row Count* pokazuje, ile znaleziono wierszy z wszystkimi nazwami kategorii. W *Values in Cluster* wypisane są wykryte przez program rodzaje kategorii

(w naszym przypadku program wykrył po dwa rodzaje oznaczeń tej samej kategorii dla wszystkich kategorii). W polu *Merge?* możesz zaznaczyć *checkbox* — wówczas nazwy kategorii zmienią się w treść znajdującą się w *New Cell Value*. Następnie wybierz *Merge Selected & Re-Cluster* (jeżeli nie chcesz zamykać okna i np. przyjrzeć się jeszcze raz możliwości klastrowania) lub *Merge & Close* (jeżeli chcesz dokonać klastrowania i zamknąć okno). Przy grupowaniu podobnych elementów powinieneś działać ostrożnie, aby się nie okazało, że połączyłeś dwa zupełnie różne obiekty.



Rysunek 5.4. Grupowanie umożliwia wydobycie danych, które zostały np. niewłaściwie opisane.
Źródło: materiały własne

Open Refine z pewnością daje duże możliwości czyszczenia danych, a jego wielka zaleta to fakt, że jest darmowy. Program można rozwijać, dodając do niego kolejne rozszerzenia, lub można (w przypadku wyrażeń regularnych) zastosować język programowania.

Część tych możliwości z pewnością znajdziesz też w innych programach. Przy mniejszych zbiorach danych wystarczy Ci Google Sheets lub Excel (ewentualnie z manualnym czyszczeniem). Tak czy inaczej pokazane wyżej elementy występują niemal w każdej tabeli, tzn. na pewno znajdziesz niepotrzebne puste miejsca, powtarzające się wiersze, dublujące się kategorie itd. Czasami jednak trzeba będzie połączyć dwie tabele dotyczące tego samego zagadnienia, a sformatowane w zupełnie inny sposób, i tu już Open Refine okaże się nieoceniony.

5.2. Szukanie wzorców, czyli analiza

Jak wspominałem na początku książki, etap pobierania, czyszczenia i agregowania danych zajmuje nawet do 90% czasu tworzenia materiału. Może się okazać, że będziesz obrabiać przez 5 dni jeden zbiór danych, po to, by w godzinę przygotować wykres słupkowy. To jednak esencja pracy dziennikarza danych. To tutaj będziesz odkrywać historie: w liczbach.

Aby je jednak odkryć, musisz wiedzieć, jak szukać i gdzie szukać. W liczbach mogą się kryć różnego rodzaju zależności, których poszukujesz. Jak je znaleźć? Cóż, musisz swoje zbiory przeanalizować, dokonać ich analizy ilościowej. Tymi zagadnieniami zajmuje się statystyka. I to jest moment, przy którym dziennikarzy i dziennikarki ogarnia paniczny strach. Już spieszę z pocieszeniem. To, czego potrzebujesz ze statystyki, aby być dziennikarzem danych, to raptem kilka pojęć i wzorów. W większości przypadków Twoja praca będzie opierać się na tzw. danych zastanych. Oczywiście możesz zaproponować w redakcji przeprowadzenie ankiet, wykorzystać crowdsourcing itp., wówczas jednak proponuję konsultacje z osobą zajmującą się statystyką i socjologią. Przejdźmy zatem do zagadnień, które powinny być częścią warsztatu każdego dziennikarza, czyli podstawowych pojęć ze statystyki.

5.2.1. Odrobina statystyki

Kluczowym pojęciem niewątpliwie jest zmienna. Zanim rozpoczniesz pracę, musisz zobaczyć, z czym masz do czynienia. Co możesz wówczas odkryć? Z pewnością to, jakie zmienne występują w Twoim zbiorze. W tabeli głosowań radnych możesz znaleźć takie zmienne jak płeć, przynależność partyjna, liczba oddanych głosów, lata itd.

Jak można, Twoim zdaniem, pogrupować te zmienne? W dużym uproszczeniu: na te, które da się „zmierzyć", i na te, których „zmierzyć" się nie da. Zmienna ma jedną kluczową właściwość: jest albo ciągła, albo nieciągła (dyskretna, skokowa). Zmienną ciągłą będzie zatem „liczba oddanych głosów" (tak samo wzrost, kwota dochodu itp.), natomiast zmiennymi nieciągłymi są pozostałe zmienne (także np.: wykształcenie, ocena, stanowisko).

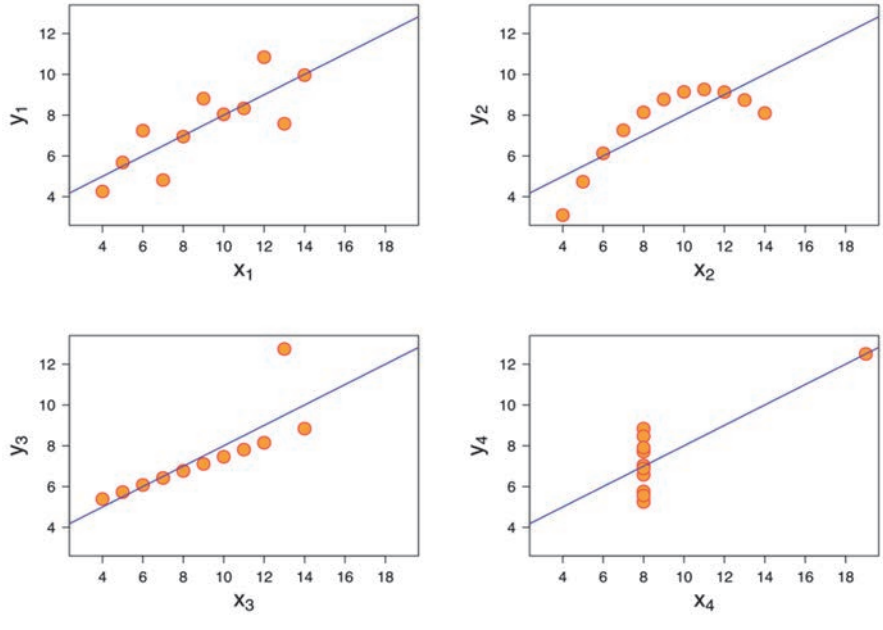
Zmienne z kolei będą Ci narzucać dobór skali, na jakiej będziesz pracować. Ogólnie rzecz biorąc, są cztery skale statystyczne: nominalna, porządkowa, interwałowa i ilorazowa. Często skale te dzieli się na nominalną, porządkową i skale ilościowe:

- **Skala nominalna.** W przypadku tej skali obiekty różnią się od siebie i nie są w żaden sposób ze sobą powiązane. Jest to np. lista miast w Polsce czy lista posłów na Sejm (poseł z jednej partii różni się od posła z innej partii). Nie jesteśmy w stanie oszacować tej różnicy, np. stwierdzić matematycznie, że poseł partii X jest lepszy od posła partii Y.
- **Skala porządkowa.** W tej skali obiekty różnią się od siebie i nie jesteśmy w stanie oszacować tej różnicy (że poseł partii X jest lepszy od posła partii Y), jednak jesteśmy w stanie uszeregować odpowiednie zmienne (np.: minister, wiceminister, podsekretarz stanu).

- **Skale ilościowe.** Te skale dotyczą zmiennych, które różnią się od siebie, można je uszeregować, ale także można je ocenić, a dokładnie ich różnicę (np. ile razy „za" w obecnej kadencji głosowała posłanka X w porównaniu do posłanki Y).

Widząc po raz pierwszy zbiór danych, warto mu się przyjrzeć pod kątem tego, jakiego rodzaju zmienne zawiera. Czy są uporządkowane, czy nie (czy skala nominalna czy porządkowa), a może da się je uporządkować i dokładnie porównać (skale ilościowe)?

Zrozumienie, z jakimi danymi masz do czynienia, ułatwi Ci dobór odpowiedniego wykresu. W przypadku skal ilościowych na początek należy pochylić się nad wykresami liniowymi (trendy z datami), wszelkiego rodzaju histogramami czy wykresami pudełkowymi. Jeżeli chodzi o skale nominalną i porządkową, najpierw warto sprawdzić wykresy kołowe i słupkowe. Dzięki tej wiedzy i praktyce będziesz w stanie bardzo szybko dobrać odpowiedni wykres do tabeli i sprawdzić pierwsze zależności i informacje, które mogą się za nimi kryć. Na takim etapie, zanim zaczniesz dokonywać analizy, możesz już spróbować zwizualizować swój zbiór na prostym wykresie zgodnie z przyjętą skalą. Możesz także odwrócić ten proces i rozpocząć analizę od szybkich prostych wizualizacji. Eksploracja danych poprzez wizualizację to w końcu odkrywanie czegoś, czego nie zauważysz na pierwszy rzut oka, warto więc wykorzystać szybkie eksploracje. Jeden z pierwszych przykładów znajdziesz na rysunku 5.5.



Rysunek 5.5. Francis Anscombe, twórca tych wizualizacji, stworzył cztery grupy danych, które miały te same cechy statystyczne, tj. średnią, wariancję, a nawet współczynnik korelacji. Jednak po zwizualizowaniu każdej z nich osobno okazało się, że znacząco się różnią na wykresie. Eksperyment ten pokazał po raz pierwszy, jak ważna jest wizualna analiza danych.

Źródło: Wikipedia

BEZ DANYCH JESTEŚ JEDYNIE KOLEJNĄ OSOBĄ Z OPINIĄ...

Dziennikarstwo danych przeżywa dziś prawdziwy rozkwit. Dzieje się tak dlatego, że nasze życie w dużej mierze przeniosło się do internetu, a internet to... dane. Megabajty, gigabajty, terabajty danych. Misją współczesnego dziennikarza jest przedstawiać je społeczeństwu rzetelnie, a równocześnie atrakcyjnie, czyli w sposób zrozumiały, łatwy do przyswojenia. Nim się jednak owe dane pięknie zestawi, trzeba je znaleźć. Gdzie szukać? Jak je zdobyć? W jaki sposób opowiedzieć dane? Na takie pytania autor odpowiada w tej książce.

Nie przeczytasz w niej o „ładnych wykresach", bo wbrew pozorom to nie one są esencją dziennikarstwa danych i data storytellingu. Dowiesz się natomiast, gdzie biją źródła potrzebnych Ci informacji, jak je przetwarzać i analizować. Znajdziesz tu także wskazówki, w jaki sposób tworzyć dobre wizualizacje za pomocą prostych aplikacji dostępnych za darmo w internecie i jak kreować angażujące odbiorców *data stories*. Na koniec wejdziesz na wyższy poziom — nauczysz się prezentować dane z wykorzystaniem kodu programistycznego.

Przebicie się przez gigabajty informacji, przetworzenie ich i stworzenie materiału, który tłumaczy odbiorcy rzeczywistość, jest dziś działaniem obarczonym ogromnym wysiłkiem i jeszcze większym ryzykiem. Dlatego, jeżeli chcemy mieć przynajmniej cień nadziei na dobrze wykonaną pracę, warto sięgnąć po książkę Łukasza Żyły. *Dziennikarstwo danych i data storytelling* to także pozycja dla osób doświadczonych w tym zawodzie. Powód jest oczywisty, technologia zmieniła dziennikarstwo i w pędzie żywiołu, którym ono jest, łatwo popaść w bezpieczną i przez to złudną rutynę, a wtedy jesteśmy o krok od poważnego błędu.

Bartosz Kurek, były dziennikarz Polsatu, obecnie menedżer ds. public affairs w Philip Morris

Co wy tam tak naprawdę robicie? — to częste pytanie, kiedy mówię, że pracuję w dziale danych „Wyborczej". Teraz, zamiast wchodzić w szczegóły, będę mógł zacząć odpowiedź od słów: „Jest taka książka, warto przeczytać...", bo Łukasz w bardzo przystępny sposób tłumaczy, czym to się je. Kiedy czytałem tę książkę, wiele razy żałowałem, że czegoś takiego nie było, kiedy ja zaczynałem przygodę z danymi. Dzięki niej widzę, ile jeszcze powinienem się w tej dziedzinie nauczyć.

Dominik Uhlig, szef BIQdata.pl — działu danych „Gazety Wyborczej"

ŁUKASZ ŻYŁA — dziennikarz i specjalista od *data storytelling*, programista, magister prawa (specjalizuje się w prawie do informacji). Prezes Fundacji Media 3.0. Prowadzi zajęcia na kierunku informatyka społeczna w Akademii Górniczo-Hutniczej. Przeszkolił kilkudziesięciu dziennikarzy z takich mediów jak „Gazeta Wyborcza", Wirtualna Polska, Polska Press. Prowadzi serwis *datablog.pl*. Był członkiem grupy roboczej do spraw otwartości danych przy Ministerstwie Cyfryzacji.

onepress

Księgarnia internetowa:
http://onepress.pl

HELION SA
ul. Kościuszki 1c, 44-100 Gliwice
tel.: 32 230 98 63
onepress@onepress.pl

książkiklasybusiness

ebook dostępny na:
ebookpoint

ISBN 978-83-283-8312-8

Cena: 49,90 zł