HISTORYKA. Studia Metodologiczne
T. 54, 2024 PL ISSN 0073-277X
DOI 10.24425/hsm.2024.153701
s. 175–194

Andrzej Radomski
Uniwersytet Marii Curie-Skłodowskiej
ORCID: 0000-0002-1735-605X

ANALIZA EKSPLORACYJNA I WIZUALIZACJA DANYCH ZA POMOCĄ PAKIETU GGPLOT2 W R, CZYLI PODSTAWY CYFROWEGO WARSZTATU BADACZA KULTURY I HISTORII

EXPLORATORY ANALYSIS AND DATA VISUALIZATION USING THE GGPLOT2 PACKAGE IN R - THE BASICS OF A DIGITAL CULTURAL AND HISTORICAL RESEARCHER'S WORKSHOP

Abstract

The article presents the basics of the research toolkit of the digital humanist. Digital humanities is a new research trend that has emerged in the 21st century. The main object of research of digital humanists is digital culture and digitized artifacts of the past. Digital humanists are increasingly studying large data sets – known as big data. This can include textual, media and audio data. In order for them to become the subject of cognition, they should then be converted to numerical form. ICT tools are used to study digital creations. With their help it is possible to: a) retrieve data from databases and digital archives, b) prepare them for study, c) analyse them, d) use visualization methods and prepare research reports. The paper shows examples of such research conducted using digital tools. For this purpose, the R programming language, a special development environment, were used: R Studio and the programming package: ggplot2 were used.

Keywords: digital culture, digital humanities, digital history, big data, research workshop, exploratory analysis, visualization

Słowa kluczowe: kultura cyfrowa, humanistyka cyfrowa, historia cyfrowa, big data, warsztat badawczy, analiza eksploracyjna, wizualizacja

Świat kultury – zarówno ten współczesny, jak i ten historyczny – już jest tradycyjnym przedmiotem badania różnych dyscyplin humanistycznych. I niewiele w tym aspekcie zmieniło pojawienie się w ostatnich dwóch dekadach XXI wieku nowych tendencji w humanistyce – opatrywanych jako kolejne zwroty, nurty badawcze czy studia. Mimo że niektóre z tych nowych orientacji badawczych kierują swe zainteresowania ku problemom z pogranicza świata ludzkiego i nie-ludzkiego (np. humanistyka „ekologiczna" czy „rzeczy"), zdecydowana większość badaczy i badaczek humanistycznych penetruje poznawczo różne aspekty kultury – zarówno tej w rzeczywistości fizycznej, jak i wirtualnej.

Jedną z najważniejszych cech świata współczesnego jest jego dygitalny charakter. Niemalże wszystkie ludzkie praktyki dotknęła cyfrowa rewolucja – także wytwory kultur historycznych. Od kilkudziesięciu lat dygitalizuje się na wielką skalę kulturowe artefakty pochodzące z nieistniejących już światów. Owa masowa dygitalizacja otwiera przed badaczami i badaczkami humanistycznymi zupełnie nowe możliwości poznawcze. Archeolodzy, historycy, badacze dziejów sztuki, kulturoznawcy, literaturoznawcy czy medioznawcy otrzymali do dyspozycji ogromny materiał empiryczny, stwarzający zupełnie nowe możliwości badawcze. Aby w pełni wykorzystać ten dobrostan, potrzebne są nowe kompetencje warsztatowe. Te dotychczasowe były stworzone z myślą o badaniu świata – nazwijmy go analogowym – i w ogromnej większości nie nadają się do badania świata cyfrowego, zarówno współczesnego, jak i tego historycznego, przekonwertowanego do postaci cyfrowej (chodzi oczywiście o konwersję wytworów). Zatem ci wszyscy, którzy chcieliby penetrować poznawczo cyfrowe materiały (np. źródła), powinni rozważyć zastosowanie nowych metod i narzędzi. Właśnie wybranym aspektom tego zagadnienia zostanie poświęcona niniejsza „wypowiedź".

Tytuł niniejszego artykułu wskazuje już na kluczowe pojęcia charakteryzujące podstawy warsztatu cyfrowego badacza kultury i historii. Zostaną one rozwinięte i uzupełnione kilkoma dodatkowymi jeszcze, aby czytelnik lub czytelniczka otrzymali przynajmniej zarys tej bogatej problematyki związanej z badaniami charakterystycznymi dla cyfrowej nauki. Moja uwaga skupi się na humanistyce cyfrowej (a w jej ramach historii cyfrowej) i jej dominującym nurcie nazywanym eksploracyjnym.

Kultura cyfrowa i powstanie humanistyki cyfrowej

Każda epoka historyczno-kulturowa ma charakterystyczne dla siebie cechy i wartości. O świecie współczesnym (pierwszych dekad XXI wieku) mówi się najczęściej, że jest światem cyfrowym lub kulturą cyfrową. Wskazuje się, że został urządzony – a mówiąc jeszcze ściślej, że został zaprogramowany – przez technologie ICT[1].

Z punktu widzenia rozważanej tu problematyki ważne jest to, że w świecie cyfrowym produkuje się ogromne ilości informacji określanych jako *big data*. Cy-

[1] Jest to skrót od angielskiego Information and Communication Technology.

frowe *big data* tworzone są także z materiałów analogowych (zwykle w wyniku dygitalizacji zasobów archiwalnych). Powoduje to, że współcześni badacze i badaczki ze wszystkich dyscyplin humanistycznych mają do dyspozycji olbrzymie ilości informacji i, co ważne, najczęściej w wolnym dostępie (formuła Open Access)[2].

Uczestnicy i uczestniczki poszczególnych praktyk kultury cyfrowej zostawiają po swojej działalności ogromne ilości informacji czy danych (tzw. źródła *born-digital*), bez uwzględnienia których niemożliwe jest poznanie współczesnego świata. Statystyczne próbki losowe, wyrywkowe obserwacje czy ważne teksty kultury (uchodzące wręcz za kanoniczne) już nie wystarczają do badania współczesnej, cyfrowej kultury i jej społeczeństw. Podobnie jest w przypadku badania światów z minionych epok historycznych. Tu również przyrost informacji źródłowych, a także literatury naukowej (np. historiograficznej, językoznawczej czy kulturoznawczej) ma charakter wręcz wykładniczy.

Zasadniczym problemem współczesnych analogowych badaczy i badaczek – i to niezależnie od tego, jaką dyscyplinę humanistyczną reprezentują – jest niemożność przetworzenia wciąż wzrastającej liczby informacji. Główną przyczyną tego stanu rzeczy jest to, że metody używane przez wspomnianych badaczy zostały stworzone w czasach, kiedy nie istniały jeszcze praktyki cyfrowe[3]. Inaczej wówczas wyglądała praca archeologa, historyka czy antropologa. Najczęściej dokonywano „ręcznej" interpretacji źródeł, wyników obserwacji czy w ogóle różnych artefaktów.

W odpowiedzi na potrzebę poradzenia sobie z wyzwaniami, jakie stawia przez badaczami cyfrowa kultura, powstała humanistyka cyfrowa. Wyrodziła się ona z informatyki humanistycznej, ukonstytuowanej jeszcze w połowie wieku XX. Określenie „humanistyka cyfrowa" pojawiło się w roku 2004[4]. Od początku istnienia spierano się o jej przedmiot i metody badania. Ze względu na to, że humanistyka cyfrowa obejmuje wiele zjawisk, trudno jest ją zdefiniować. Jak pisze Rafael Alvarado, nie mamy wspólnie podzielanej definicji, jeśli przez definicję rozumiemy spójny zestaw teoretycznych zainteresowań i metod. Zamiast tego mamy genealogię, sieci podobieństw rodzinnych, zainteresowania metodologiczne i preferowane narzędzia, historię ludzi, którzy zdecydowali się nazywać siebie cyfrowymi humanistami[5]. Z kolei Danuta Smołucha uważa, że może się

---

[2] Dobrym wprowadzeniem na gruncie polskim do świata kultury cyfrowej oraz roli narzędzi ICT jest monografia Magdaleny Szpunar: Magdalena Szpunar, *Kultura algorytmów* (Kraków: Wyd. UJ, 2019).

[3] Co prawda amerykańscy kliometryści używali metod komputerowych do analizy źródeł już od lat 50. XX wieku, lecz wydajność ówczesnych komputerów była bardzo mała i dalece niewystarczająca do analizy większych zbiorów danych.

[4] W tym bowiem roku ukazała się monografia: Susan Schreibman, Ray Siemens i John Unsworth, *A Companion to Digital Humanities* (Hoboken: Wiley-Blackwell, 2004), w której pojawiła się nazwa humanistyka cyfrowa oraz zarysowano przedmiot i metody tego nurtu badawczego.

[5] Rafael C. Alvarado, „Blog posts", w *Debates in the Digital Humanities*, red. Matthew Gold (Minneapolis: University Minnesota Press, 2012), https://dhdebates.gc.cuny.edu/read/untitled-88c11800-9446-469b-a3be-3fdb36bfbd1e/section/c513af64-8f99-4e02-9869-babc1cecc451#p1b1 (dostęp: 12.03.2023).

nawet zdarzyć tak, że humanistyka cyfrowa nie doczeka się ostatecznej definicji, bo określenie „cyfrowa" po prostu straci sens, ponieważ każda nauka korzysta już w mniejszym lub większym stopniu z narzędzi i metod cyfrowych[6]. Jednakże na użytek dalszych rozważań będę się posługiwał następującym jej określeniem: przez humanistykę cyfrową można rozumieć zespół praktyk polegających na wykorzystaniu technologii ICT (metod i narzędzi) do gromadzenia, przetwarzania (analizowania, wizualizowania) i publikowania cyfrowych danych pochodzących z różnych kultur (współczesnych i historycznych). Przedmiotem zainteresowania tak rozumianej humanistyki będzie kultura cyfrowa, czyli zarówno ta współczesna, jak i analogowa (tzn. wszystkie historyczne), którą chcemy badać za pomocą narzędzi i metod ICT.

W ramach humanistyki cyfrowej można wyznaczyć wiele pól zainteresowań. Z tego też względu wyróżniam w jej ramach trzy zasadnicze obszary działalności: a) humanistykę dygitalizacyjną – zajmującą się konwersją analogowych artefaktów (np. źródeł archeologicznych czy historycznych) na format cyfrowy oraz tworzeniem cyfrowych archiwów i baz danych udostepniających zdygitalizowane wytwory badaczom, b) generatywną – tworzącą wielkie projekty cyfrowe typu Wirtualny Rzym[7] czy Photogrammar[8] i umieszczającą je w Internecie, c) eksploracyjną – zajmującą się badaniem zdygitalizowanych źródeł lub materiałów typu *born-digital*[9].

W ostatnich latach na czoło praktyk cyfrowych humanistów i humanistek (w ramach historii cyfrowej, językoznawstwa cyfrowego, literaturoznawstwa cyfrowego czy kulturoznawstwa) wysuwa się działalność eksploracyjna. W jej ramach operuje się na wielkich zbiorach zakodowanych do postaci cyfrowej: setkach powieści, tysiącach czasopism i niezliczonych korpusach innych tekstów, których objętość liczy się w miliardach czy bilionach wyrazów. Cechą charakterystyczną jest wykorzystywanie do badań nowoczesnych technologii informatycznych[10]. Za ich pomocą cyfrowe humanistki i cyfrowi humaniści agregują dane, formatują je do postaci numerycznej – zrozumiałej dla komputera – dokonują analizy maszynowej, wizualizują i przygotowują do publikacji, często w Internecie.

Jednym z najbardziej znanych przykładów tego typu badań są te realizowane przez zespół Lva Manovicha. Polegają one na reprezentowaniu kultury przez dane – zarówno te historyczne, jak i współczesne, wydobyte na przykład z portali społecznościowych. Współczesne technologie ICT oferują wyjątkowe możliwości prowadzenia analizy obliczeniowej dużych zbiorów danych kulturowych i historycznych w porównaniu z metodami jakościowymi stosowanymi w huma-

6 Danuta Smołucha, *Humanistyka cyfrowa w badaniach kulturowych. Analiza zjawiska na wybranych przykładach kulturowych* (Kraków: Wydawnictwo Naukowe Akademii Ignatianum, 2023), 9.

7 https://www.romereborn.virginia.edu/ (dostęp: 23.05.2024).

8 https://photogrammar.org/maps (dostęp: 24.05.2024).

9 Są to materiały powstałe już w formie cyfrowej, czyli w świecie praktyk cyfrowej kultury, przede wszystkim w sieci.

10 Kamil Szubański, „Naukowiec o przewadze, jaką daje humanistyka cyfrowa", Nauka w Polsce, 2.12.2018, https://naukawpolsce.pl/aktualnosci/news%2C31908%2Cnaukowiec-o-przewadze-jaka-daje-humanistyka-cyfrowa.html (dostęp: 8.03.2023).

nistyce analogowej i naukach społecznych. Jak pisze Manovich, wielu badaczy publikuje coraz więcej artykułów, w których analizują wzorce w ogromnych zbiorach danych kulturowych. Ponadto coraz częściej uwaga badaczy i badaczek skupia się na analizie różnych zjawisk w poszczególnych okresach historycznych: fotografia mody z ostatnich kilkudziesięciu lat, dwudziestowieczna muzyka popularna czy dziewiętnastowieczna literatura[11].

Eksploracja danych medialnych to tylko część badań realizowanych w ramach wspomnianego nurtu humanistyki cyfrowej. Równolegle mamy do czynienia z eksploracyjną analizą dużych korpusów tekstualnych, które to badania mutują z wcześniejszego paradygmatu analizy tekstu w analitykę danych tekstualnych. Połączone są one z wizualizacją i nazywane *distant reading* (czytanie dużych korpusów) jako opozycja do *close reading* (szczegółowego czytania pojedynczego tekstu czy źródła)[12].

Kolejnym polem badawczym w analizach cyfrowych humanistek i humanistów z zastosowaniem dużych zasobów danych – tym razem grafowych – są badania sieci społecznych. Prekursorem na tym polu był Uniwersytet Stanforda, w którym zrealizowano projekt: „Mapping the Republic of Letters"[13]. Uczeni z tego ośrodka zebrali 55 000 listów napisanych przez ponad sześć tysięcy filozofów, literatów, publicystów i naukowców działających i tworzących w epoce Oświecenia. Zostały one umieszczone na interaktywnej mapie dostępnej w sieci. Mapa ówczesnej Europy przedstawiała, kto, z kim i kiedy prowadził korespondencję. Klikając na dane nazwisko, mogliśmy uzyskać sieć kontaktów, jakie ta osoba utrzymywała z innymi za pośrednictwem wspomnianej korespondencji. Za pomocą trzech podstawowych trybów mogliśmy też śledzić „wędrówki" listów między państwami i miastami, korespondencję w określonych latach oraz wędrówki poszczególnych intelektualistów[14].

Narzędzia cyfrowe umożliwiają także łączenie kilku zbiorów danych w różnych formatach i mapowanie ich. Są to tak zwane *linked open data*. Dzięki tym metodom jest możliwe łączenie informacji zawartych w tekstach, zbiorach grafik, mediach czy nagraniach audio[15], które są umieszczane na mapach opartych na GIS (ang. *geographic information system*).

Tu na marginesie zauważmy, że badania prowadzone w nurcie eksploracyjnym humanistyki cyfrowej nie koncentrują się tylko i wyłącznie na *big data*, ale obejmują także mniejsze zbiory, a nawet pojedyncze dokumenty czy obrazy filmowe.

---

[11] Lev Manovich, „The Science of Culture? Social Computing, Digital Humanities, and Cultural Analytics", Manovich, 2015, http://manovich.net/index.php/projects/cultural-analytics-social-computing (dostęp: 10.03.2023).

[12] Alan Liu, „The State of the Digital Humanities: A Report and a Critique", 4.06.2016, https://escholarship.org/content/qt23h6v6x8/qt23h6v6x8_noSplash_0ef49e16691ca30b532e3cb6bc5cf080.pdf (dostęp: 10.03.2023), 27.

[13] Mapping the Republic of Letters, http://republicofletters.stanford.edu/ (dostęp: 24.05.2024).

[14] Roy Rosenzweig Center for History and New Media, „Visualization: Mapping the Republic of Letters", 18.08.2021, 3:33, https://www.youtube.com/watch?v=RyXzjiYNers (dostęp: 10.03.2023).

[15] Szubański, „Naukowiec".

Warsztat badacza cyfrowej kultury i historii

Humanistyka cyfrowa wypracowała nowy, charakterystyczny dla siebie warsztat badawczy. Jego podstawą nie jest (jak to było w humanistykach analogowych) teoria, ale narzędzia i dostosowane do nich metody. Teoria odgrywa tu rolę drugorzędną. W przypadku nurtu eksploracyjnego (gdyż on nas tutaj przede wszystkim interesuje) wykorzystuje się metody i narzędzia wypracowane przez informatyków, statystyków i nową dyscyplinę wiedzy, jaką jest Data Science.

Jak zasygnalizowano w poprzednim fragmencie, humanistyka cyfrowa i jej poddziedziny (historia cyfrowa czy, powiedzmy, kulturoznawstwo cyfrowe) operują na dużych zbiorach danych. Tym, co szczególnie odróżnia praktyki humanistyki cyfrowej od podejścia klasycznego (pisze Adam Pawłowski[16]), jest automatyczne przetwarzanie dużych zestawów danych:

> […] zauważa, że w humanistyce cyfrowej pojawiło się nieznane w podejściu tradycyjnym pojęcie „informacyjnej linii produkcyjnej" (angielski termin to *work flow*). Jak mówi, rozszerzyło ono system generowania wiedzy, w którym dotychczas badacz interpretował tekst lub teksty, wykorzystując jako jedyne narzędzie analizy swój umysł.
> W humanistyce cyfrowej praca ma najczęściej charakter zespołowy, obejmuje wiele etapów, które właśnie składają się na ową „linię produkcyjną". Badacz i jego kompetencje analityczne pojawią się głównie na etapie projektowania i wnioskowania. Istotne w tym procesie jest wytworzenie danych w formacie czytelnym dla komputera oraz dodanie metadanych. Dopiero w momencie, kiedy tekst lub grafikę wzbogacimy o metadane, możliwe staje się efektywne przetwarzanie[17].

Na wspomnianą przed momentem „linię produkcyjną" składa się szereg operacji: pobieranie danych, „oczyszczanie", analiza, wizualizacja czy przygotowanie raportów z badań. Do ich przeprowadzenia musimy mieć odpowiednie narzędzia i metody. Te opracowane przez humanistykę analogową nie nadają się do analizy czy wizualizacji danych wielkoskalowych, dlatego cyfrowi humaniści i cyfrowe humanistki szeroko wykorzystują metody i narzędzia wypracowane w ramach Data Science.

Data Science to interdyscyplinarna dziedzina, która zajmuje się analizą i wykorzystaniem danych w celu odkrywania wzorców, wykrywania trendów oraz opracowywania problemów biznesowych i naukowych. Za pomocą procedur wypracowanych w ramach Data Science analizuje się dane tekstowe, liczbowe, obrazowe i dźwiękowe pochodzące z różnych źródeł, na przykład z Internetu, urządzeń mobilnych, czujników, baz danych i oczywiście ze zdygitalizowanych materiałów z przeszłości.

W ramach Data Science używa się różnych narzędzi i metod, takich jak: statystyka, uczenie maszynowe, sztuczna inteligencja, programowanie czy wizualizacja. Specjaliści w ramach Data Science zajmują się między innymi a) analizą

---

16 Kierownik Pracowni Humanistyki Cyfrowej w Uniwersytecie Wrocławskim.
17 Szubański, „Naukowiec".

danych w celu uzyskania cennych informacji (tzw. *data mining*), b) modelowaniem danych i tworzeniem modeli predykcyjnych, c) projektowaniem eksperymentów i ich testowaniem, d) budowaniem aplikacji bazujących na danych, na rzykład systemów rekomendacyjnych, e) analizą sentymentu i nastrojów na przykład w mediach społecznościowych, f) personalizacją treści i reklam.

Podstawą powyższych operacji jest dysponowanie odpowiednim środowiskiem pracy. Jego głównym składnikiem (w przypadku pracy z danymi) jest zintegrowane środowisko programistyczne (z ang. *Integrated Development Environment*, IDE). Typowe IDE stanowi bardzo duże ułatwienie w pracy zarówno programisty, jak i badacza lub badaczki wykorzystujących zaawansowane narzędzia informatyczne do pracy naukowej. IDE bowiem integruje ze sobą edytory kodu z wieloma innymi narzędziami, takimi jak: kompilatory, debuggery (do wykrywania błędów), narzędzia do testowania kodu i środowisko uruchomieniowe, a także możliwość przygotowania publikacji[18].

W przypadku praktyki naukowej podstawowym IDE jest to oparte na R. R to język programowania[19] i określenie na IDE – tylko pod nazwą R Studio. Język R został stworzony w roku 1993 na bazie języka S. Od początku był on projektowany z myślą o zastosowaniu w nauce[20]. Język R (kolejne jego wersje) jest wydawany na licencji GNU General Public Licence[21], można go więc pobrać – razem z R Studio – za darmo z repozytorium CRAN[22]. Oprócz desktopowej wersji R Studio, do dyspozycji badaczy jest również wersja „chmurowa", umożliwiająca także realizację projektów zespołowych[23].

Za pomocą R można przeprowadzić zaawansowane obliczenia statystyczne, analizy danych przy użyciu technik uczenia maszynowego, wizualizacje wyników czy opracować raporty sieciowe przy zastosowaniu znaczników R Markdown i platformy Shiny. Mocną zaletą R jest dobra współpraca z relacyjnymi bazami danych, a także bogaty zestaw pakietów programistycznych ułatwiających pisanie skryptów z kodem. Podobne możliwości oferuje też Python, który również jest szeroko używany do analizy danych[24], a także do trenowania sieci neuronowych.

R jest przykładem wysokopoziomowego, funkcyjnego języka programowania. Znaczy to, że za jego pomocą można pisać złożone i elastyczne funkcje, które nawet w pojedynczej linii kodu pozwalają na wykonywanie dużej ilości pracy[25].

---

18 W pewnych przypadkach – np. do prostych analiz czy wizualizacji – można używać klasycznych programów w rodzaju Tableau, Exploratory, Power BI czy Flourish, z zastrzeżeniem, że posiadają one pewne ograniczenia.

19 Został on skonstruowany przez Roberta Gentelmana i Rossa Ihakę z Uniwersytetu w Auckland.

20 Początkowo była to bioinformatyka, a później i inne dyscypliny.

21 Jest to licencja wolnego i otwartego oprogramowania stworzona przez Richarda Stallmana i Ebena Moglena w roku 1989 na potrzeby projektu GNU (wolny system operacyjny).

22 The Comprehensive R Archive Network, https://cran.r-project.org/ (dostęp: 25.05.2024).

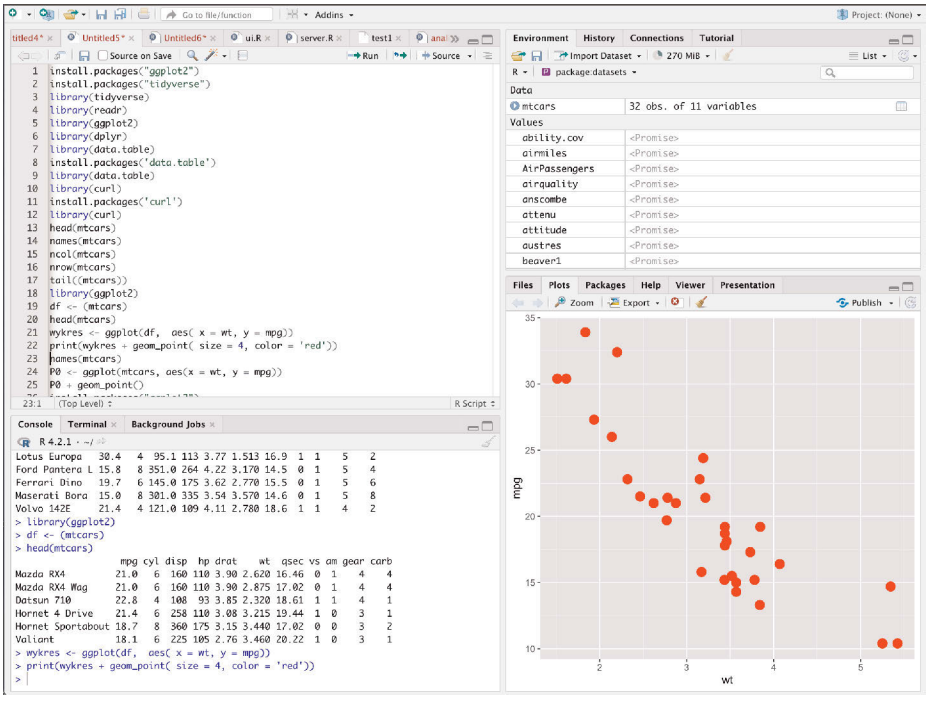
23 Teraz nazywa się ona: Posit Cloud, https://posit.cloud/ (dostęp: 25.05.2024).

24 Podstawowymi bibliotekami programistycznymi Pythona w analizie danych są: Numpy, Pandas i Matplotlib.

25 Colin Gillespie, Robin Lovelace, *Wydajne programowanie w R*, przeł. Krzysztof Kapustka (Warszawa: Wyd. APN, 2018), 6.

Nie wymaga on kompilacji. Po napisaniu danej instrukcji/linijki kodu od razu otrzymujemy wynik na ekranie komputera. R jest dostępny na wszystkie systemy operacyjne. Pobieramy go z repozytorium CRAN.

R Studio ma elastyczny interfejs. Jego wygląd standardowy składa się z czterech podstawowych paneli. W lewym górnym panelu znajduje się edytor tekstowy do pisania kodu, w lewym dolnym – konsola, na której między innymi pojawia się kod wynikowy (np. rezultaty obliczeń). Prawy górny panel prezentuje informacje o przestrzeni roboczej, historii poleceń, plikach w bieżącym folderze oraz kontrolę wersji Git. Natomiast prawy dolny zawiera wykresy i informacje o pakietach[26].



Rys. 1. Wygląd przestrzeni roboczej R Studio (opr. własne).

Na powyższym rysunku widzimy fragment kodu, który wyświetla dane ze zbioru „mtcars", który jest wbudowanym zbiorem danych w języku R, zawierającym informacje o różnych samochodach. Wizualizacja pokazuje zależności między masą (wt) a milami na galon (mpg) samochodów w tym zbiorze danych.

Za pomocą R Studio możemy wykonać szereg operacji związanych z analizą danych. Najczęściej wykorzystujemy do tego potężny pakiet: tidiverse (z repozytorium CRAN). Tidiverse składa się z kilku mniejszych pakietów: dplyr do zbierania danych, ich filtrowania i sortowania; readr, za pomocą którego wczy-

26 Jared P. Lander, *R dla każdego. Zaawansowane analizy i grafika statystyczna*, przeł. Marek Włodarz (Warszawa: Wyd. APN, 2018), 20.

tujemy dane z różnych formatów (np. Excela czy csv); tidyr do analizy danych; oraz ggplot2 do wizualizacji.

Od wielu lat ggplot2 jest podstawowym pakietem do tworzenia wykresów i wizualizacji danych[27]. Jest on popularny i powszechnie stosowany, szczególnie w dziedzinie nauki o danych, ze względu na jego potężne możliwości statystyczne i graficzne. Pakiety takie jak tidyverse i ggplot2 znacznie rozszerzają użyteczność R w nauce o danych, umożliwiając transformację i wizualizację danych oraz usprawnienie przepływu pracy.

Jeśli chodzi o wizualizację, to jest ona (w kontekście Data Science) rozumiana jako metoda badawcza, sposób (narzędzie) prezentacji wyników badań. Pakiet ten bazuje na tak zwanej gramatyce warstw, co oznacza, że pozwala na tworzenie wykresów poprzez łączenie różnych elementów, takich jak dane, zmienne, estetyki i geometrie. Za pomocą ggplot2 stworzymy różne wykresy: liniowe, słupkowe, kołowe, punktowe, histogramy czy mapy cieplne. Ponadto oferuje on możliwość dodania dodatkowych funkcji, w tym takich jak zmiany kolorów, stylów, etykiet, a także dodawania animacji oraz interaktywności.

Przykłady

W tej części zaprezentujemy dwa przykłady analizy i wizualizacji danych w R Studio. Będą to przykłady historyczne. Pierwszy zostanie pokazany z użyciem zbioru danych babynames. Jest to zbiór zawierający imiona nadawane dzieciom w USA w latach 1880–2017. Można go pobrać z repozytorium GITHUB (platforma przechowująca kody, różnorodne dane oraz oferująca środowisko programistyczne)[28] lub z Repozytorium CRAN.

Typowa praca z danymi składa się z kilku etapów. W pierwszym ładujemy do IDE (w naszym przypadku jest to oczywiście R Studio) intersujący nas zbiór danych. W przypadku zbioru babynames mamy do czynienia z danymi ustrukturyzowanymi, tj. zapisanymi w postaci tabelarycznej (Excel, ramki danych, ang. *data frame* itp.). W przeważającej części pracujemy na tak zwanych surowych danych. W naszym przypadku są one akurat przygotowane już do analizy. Ale po kolei.

Rozpoczynamy standardowo od wgrania niezbędnych pakietów (inaczej bibliotek). W przypadku poniższej analizy będzie to pakiet tidyverse zawierający między innymi ggplot2. Ten megapakiet pozwala na zautomatyzowanie większości operacji. Za pomocą pakietu readr wgrywamy nasz zbiór danych badawczych babynames.

Kolejnym etapem jest zazwyczaj przejrzenie załadowanych danych. Ów przegląd ma na celu obejrzenie ich zawartości: kolumn i rzędów. Kolumny zawierają kategorie (w naszym przypadku są to: rok, płeć, imię, ilość i wartość procentowa) oraz dane obserwacyjne, czyli imiona dzieci. Przeglądanie danych ma na celu

---

27 Został stworzony przez Hadleya Wickhama.

28 Hadley Wickham, „Babynames", GitHub, https://github.com/hadley/babynames (dostęp: 25.05.2024).

jeszcze jeden ważny aspekt, a mianowicie zorientowanie się, czy wszystkie one są w odpowiednim formacie i czy poszczególne rekordy nie zawierają jakichś luk, które na przykład trzeba będzie uzupełnić bądź ewentualnie usunąć. Ten etap nazywa się preprocessingiem. Zbiór babynames nie zawiera akurat żadnych defektów, ponieważ został już wcześniej przejrzany i przygotowany do dalszych operacji.

W następnym kroku można interesujące nas dane przefiltrować (w tym momencie wykorzystywany jest pakiet dplyr) lub, powiedzmy, połączyć na przykład dwie kolumny, jeśli zachodzi taka potrzeba.

Następnie możemy przejść do „właściwej" analizy (eksploracyjnej) – w zależności od potrzeb i zainteresowań. Analiza eksploracyjna polega na dążeniu do odkrywania wzorców, zależności czy trendów. Stanowi zwykle wstęp do bardziej szczegółowych badań czy pytań eksplanacyjnych. W przypadku badania dużych zbiorów danych (a już na pewno w przypadku *big data*) nieodzowną częścią eksploracji jest wizualizacja, wówczas można bowiem zobaczyć wśród „morza" obserwacji owe hipotetyczne trendy czy relacje. W preludium zaczynamy od statystyki opisowej (miary tendencji centralnej[29], typu średnia, mediana czy odchylenie standardowe).

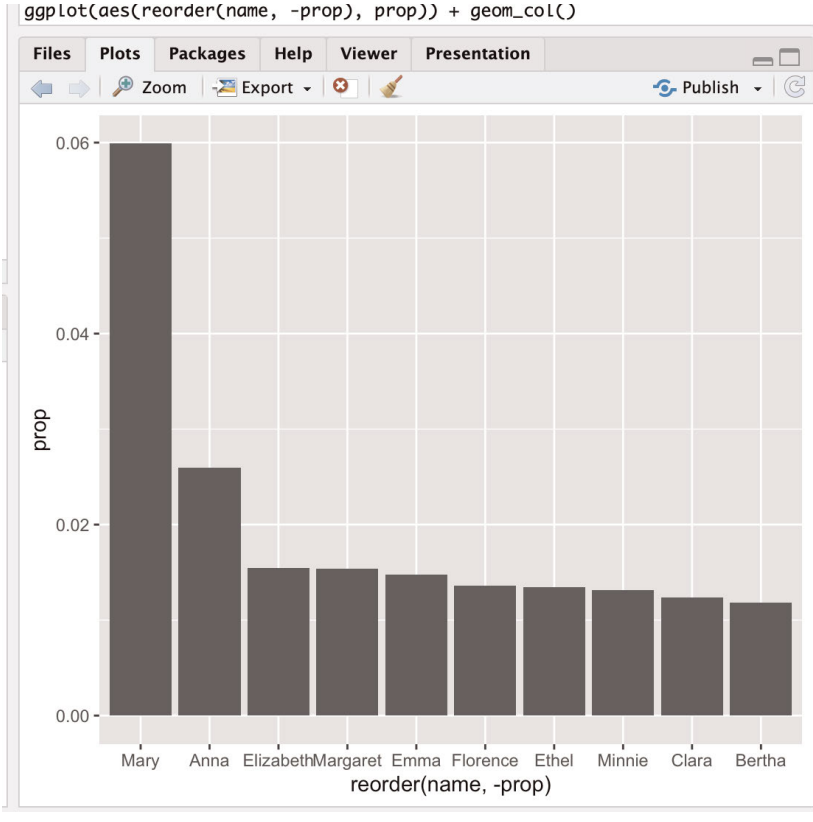
I wreszcie przechodzimy do wizualizacji naszych danych. W przypadku pakietu ggplot2 musimy ustalić kilka parametrów: po pierwsze, zdecydować, jakie wartości będą na osi OY, a jakie na osi OX (dlatego tak ważny jest początkowy przegląd danych, zaraz po ich załadowaniu do komputera). Jest to tak zwana estetyka wykresu (funkcja z przedrostkiem „aes"), a dalej typ wykresu (może to być wykres punktowy, kołowy, liniowy, histogram itp.). Tutaj wykorzystujemy funkcję z przedrostkiem „geom". Dalej możemy dodawać następne funkcje związane z daną wizualizacją (np. mapy cieplne czy legendę).

Weźmy z naszego „ćwiczebnego" zbioru babynames imiona żeńskie i spróbujmy ukazać ich popularność w zakresie od roku 1880 do początków XXI wieku, a następnie pokazać, jak owa popularność zmieniała się w czasie – co jest kanonem pracy na przykład historyka czy badacza dziejów kultury. Po napisaniu odpowiedniej instrukcji przy wykorzystaniu pakietu ggplot2 otrzymujemy następujący rezultat:

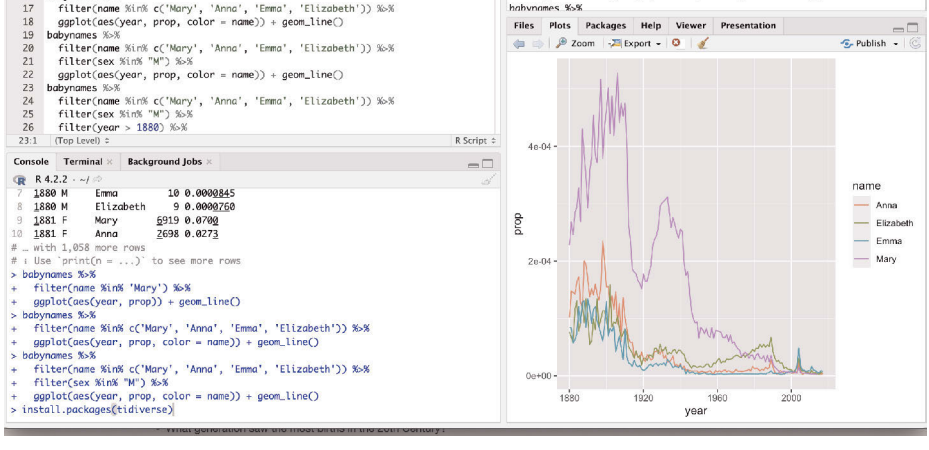
Na osi OX widzimy, jakie imiona nadawano najczęściej dziewczynkom od roku 1880 do początków XXI wieku. Oś OY z kolei zawiera udział procentowy każdego imienia, czyli najpopularniejsze imię z tego zbioru (w tym wypadku jest to Mary) stanowi 6% wszystkich imion, jakie zostały nadane w USA od końca XIX wieku. Możemy pójść dalej i spróbować pokazać, jak zmieniała się popularność poszczególnych wyróżnionych tu 10 imion w czasie. Po napisaniu kolejnych linijek kodu za pomocą ggplot2 wyświetla się nam następujący rezultat:

---

[29] Miary tendencji centralnej reprezentują punkt centralny lub typową wartość zbioru danych. Pozwalają one na zrozumienie, gdzie większość wartości w rozkładzie się znajduje. Ułatwiają zrozumienie ogólnych trendów i wzorców w danych. Miary tendencji centralnej mogą być także używane do porównywania różnych zestawów danych, na przykład mogą one pomóc w zrozumieniu, czy jeden zestaw danych ma wyższą średnią wartość niż inny. Miary tendencji centralnej są często używane jako punkt wyjścia do wielu testów statystycznych i modeli predykcyjnych.



Rys. 2. Wizualizacja 10 najpopularniejszych imion żeńskich w USA w latach 1880–2017 (opr. własne).



Rys. 3. Wizualizacja linii zmiany trendu popularności poszczególnych imion żeńskich w USA w czasie (opr. własne).

Na rysunku nr 3 widzimy wizualizację ukazującą, jak zmieniała się popularność kilku imion żeńskich w czasie. Wykres liniowy przedstawia coś, co nazywa się linią trendu. Widzimy więc, że mimo początkowej dużej dominacji, w kolejnych dziesięcioleciach XX wieku systematycznie zmniejszała się liczba dziewczynek, którym nadano imię Mary, a na przełomie XX i XXI wieku na czoło wysunęła się Emma (choć z niewielką przewagą). To samo możemy zrobić z imionami chłopięcymi[30].

To są przykłady wstępnych, eksploracyjnych analiz i wizualizacji. Mogą one być kontynuowane (przez zadawanie bardziej szczegółowych pytań), a także być punktem wyjścia do badań o bardziej jakościowym już charakterze. Możemy na przykład zadać pytanie o powody spadku popularności jednych imion, a wzrostu innych.

Przejdźmy teraz do drugiego przykładu, ilustrującego bardziej zaawansowane badania. Będzie to analiza danych ze zbioru titanic. Jest on dostępny na platformie Kaggle i na GitHub. Zawiera on dane pasażerów statku Titanic, który uległ słynnej katastrofie w roku 1912. Podczas tragicznego rejsu na statku znajdowało się 2224 pasażerów. W wyniku katastrofy zginęły 1502 osoby. Celem analizy będzie w tym wypadku oszacowanie szans na przeżycie, jakie mieli wszyscy pasażerowie. Trzeba jednak wziąć pod uwagę fakt, że szanse na przeżycie były nierówne, ponieważ w pierwszej kolejności na szalupy ratunkowe były umieszczane kobiety i małe dzieci (do 12 roku życia). Niestety Titanic posiadał ograniczoną liczbę takich szalup.

W celu oszacowania szans na przeżycie wszystkich pasażerów autor tej analizy zastosował na Kaggle[31] uczenie maszynowe – jego wersję klasyczną opartą na statystyce. Konkretnie zastosowano tutaj model regresji logistycznej i k-NN (k-najbliższych sąsiadów, ang. *k-Nearest Neighbors*). Klasyczne uczenie maszynowe wykorzystuje techniki statystyczne i probabilistyczne do budowania modeli, które potrafią przewidywać wyniki na podstawie danych wejściowych. W tym podejściu modele uczą się na podstawie danych treningowych, a następnie używają tych informacji do dokonywania predykcji dla nowych, nieznanych danych.

Podstawowymi technikami stosowanymi w uczeniu maszynowym opartym na statystyce są:

a) Regresja: polega na szukaniu zależności między zmiennymi wejściowymi a zmienną wyjściową. Modele regresji wykorzystują różne algorytmy, takie jak regresja liniowa czy regresja logistyczna.
b) Klasyfikacja: polega na przypisywaniu obiektów do określonych klas na podstawie ich cech. Modele klasyfikacji oparte na statystyce wykorzystują różne algorytmy, takie jak maszyny wektorów nośnych, drzewa decyzyjne czy sieci neuronowe.
c) Analiza skupień: polega na dzieleniu obiektów na grupy, w taki sposób, aby obiekty w jednej grupie były podobne do siebie, a jednocześnie różniły się od obiektów w innych grupach.

---

30 Cały kod jest dostępny pod następującym adresem: https://posit.cloud/content/5688158.

31 dwiuzila, „Titanic: Complete Analysis in R - Top 4%", Kaggle, https://www.kaggle.com/code/dwiuzila/titanic-complete-analysis-in-r-top-4 (dostęp: 23.03.2023).

d) Analiza głównych składowych: polega na redukcji wymiarowości danych poprzez identyfikację najważniejszych zmiennych i zredukowanie liczby wymiarów.

W uczeniu maszynowym opartym na statystyce istotnym elementem jest również walidacja modelu, czyli sprawdzenie, jak dobrze model działa na danych testowych, które nie były wykorzystane w procesie uczenia.

Regresja logistyczna jest jednym z algorytmów uczenia maszynowego wykorzystywanym w klasyfikacji binarnej, tj. przypisywaniu obiektów do jednej z dwóch klas na podstawie ich cech. Algorytm ten jest szczególnie przydatny w przypadku, gdy zmienna wyjściowa przyjmuje wartości 0 lub 1, a zmienna wejściowa może przyjmować różne wartości.

Regresja logistyczna wykorzystuje funkcję logistyczną, która zamienia wynik modelu liniowego na wartość z przedziału (0–1). Ta wartość interpretowana jest jako prawdopodobieństwo przynależności obiektu do jednej z klas. Jeśli wartość przekracza ustalony próg (np. 0,5), to obiekt jest przypisywany do klasy 1, w przeciwnym razie do klasy 0.

Proces uczenia regresji logistycznej polega na znalezieniu wartości parametrów modelu, które minimalizują funkcję straty. W przypadku regresji logistycznej funkcją straty jest logarytmiczna funkcja straty wykorzystywana do oszacowania błędu modelu[32].

Model k-NN jest jednym z algorytmów uczenia maszynowego, wykorzystywanym do klasyfikacji i regresji. Algorytm ten działa, bazując na podobieństwie obiektów w przestrzeni cech. W przypadku klasyfikacji algorytm k-NN szuka k najbliższych sąsiadów danego obiektu ze zbioru treningowego, a następnie przypisuje obiekt do klasy, do której należy większość z tych sąsiadów. W przypadku regresji algorytm k-NN szuka k najbliższych sąsiadów danego obiektu i przypisuje mu wartość wyjściową, będącą średnią wartością wyjściową tych sąsiadów. Odległość między obiektami mierzona jest w przestrzeni cech, na podstawie których model został wytrenowany. W zależności od typu danych wejściowych, mogą to być miary euklidesowe, manhattan czy miary cosinusowe[33].

Jedną z zalet algorytmu k-NN jest to, że nie wymaga on trenowania modelu, a jedynie przeszukania zbioru treningowego pod kątem najbliższych sąsiadów. Algorytm ten jest stosowany w wielu dziedzinach, takich jak rozpoznawanie obrazów, diagnozowanie chorób czy analiza tekstu.

Wspomniana analiza szans na przeżycie poszczególnych pasażerów Titanica składała się z następujących kroków głównych. Pierwszym krokiem było załadowanie niezbędnych pakietów w R Studio. Oprócz dobrze już nam znanego pakietu tidyverse (z ggplot2), wykorzystano inne: caret, MLmetrics, car, rpart i class. Pakiet caret to narzędzie do budowy modeli klasyfikacji i regresji logistycznej, pakiet MLmetrics to zestaw narzędzi i funkcji, które służą do oceny

---

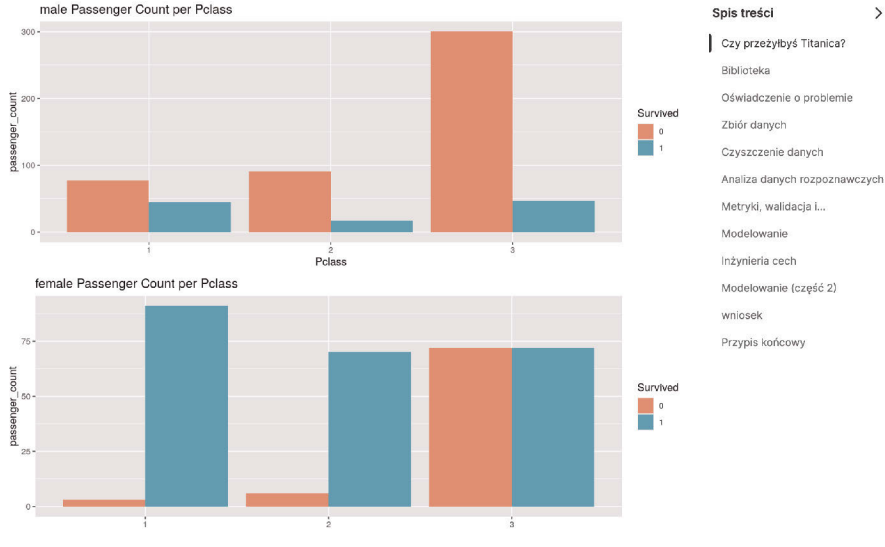
[32] Josh Patterson, Adam Gibson, *Deep Learning*, przeł. Marek Watrak (Gliwice: Helion, 2018), 47–48.

[33] Drew Conway, John White, *Uczenie maszynowe dla programistów*, przeł. Przemysław Szermiota (Gliwice: Helion, 2015), 219–221.

jakości modeli uczenia maszynowego (ML), pakiet car (ang. *Companion to Applied Regression*) służy do analizy regresji w R, oferuje wiele narzędzi do wizualizacji i diagnostyki modeli regresji, a także funkcje do wykonywania testów statystycznych i analizy wariancji, a pakiet rpart (ang. *Recursive Partitioning and Regression Trees*) służy do tworzenia drzew decyzyjnych w R, pozwala na budowanie modeli opartych na klasach, regresji i przetrwaniu, a także na generowanie drzew decyzyjnych dla różnych problemów. Z kolei pakiet class służy do klasyfikacji danych w R, oferuje wiele algorytmów klasyfikacji, w tym klasyfikator k-najbliższych sąsiadów (k-NN), drzewa decyzyjne, klasyfikator Bayesa oraz maszyny wektorów nośnych (SVM). Pakiet umożliwia również tworzenie macierzy pomyłek i obliczanie różnych metryk wydajności klasyfikatora.

Po zainstalowaniu pakietów załadowano dane ze zbioru titanic (w formacie. csv). Następne kroki to:

a) Przygotowanie zbioru do analizy eksploracyjnej: „czyszczenie" i uzupełnianie brakujących wartości (chodziło o wartości z etykietą „Cabin").
b) Przygotowanie zbioru treningowego i testowego, założenie, że wszyscy mężczyźni zginęli, a kobiety przeżyły. Okazało się, że model przewidywał (w zbiorze danych testowych) prawdopodobieństwo śmierci lub przeżycia na poziomie 76%.
c) Analiza i wizualizacja szans przeżycia pasażerów obu płci (z uwzględnieniem pobytu w jednej z trzech klas na Titanicu) dokonana za pomocą ggplot2. Przedstawia się ona następująco:



Rys. 4. Wizualizacja ukazująca pasażerów Titanica, którzy uratowali się lub zginęli w katastrofie, z podziałem na płeć i klasę, którą podróżowali. Zrzut z ekranu. Źródło: https://www.kaggle.com/code/dwiuzila/titanic-complete-analysis-in-r-top4 (dostęp: 23.03.2023).

Na rysunku widzimy, że tylko niewielka część pasażerów płci męskiej zdołała się uratować. Najwięcej zginęło mężczyzn podróżujących trzecią klasą. Odmiennie było z pasażerkami płci żeńskiej, z których większość się uratowała. Najwięcej kobiet zatonęło z klasy trzeciej (połowa).

d) Końcowym, lecz najważniejszym etapem analizy było zbudowanie modelu, który by przewidywał prawdopodobieństwo przeżycia pasażerów i to niezależnie od rodzaju kabiny (klasy), w której się znajdowali. Weryfikacja dokładności modelu odbywa się na podstawie nowych danych ze zbioru: dane.test.csv. Zbudowano w tym celu dwa modele: regresję logistyczną i k-najbliższych sąsiadów (k-NN), aby porównać, który z tych dwóch modeli daje dokładniejsze prognozy. Model oparty na regresji logistycznej to technika statystyczna używana do modelowania prawdopodobieństwa wystąpienia pewnego zdarzenia (np. przeżycia lub śmierci) na podstawie jednej lub więcej zmiennych predykcyjnych. W zbiorze danych titanic zmienne predykcyjne mogą obejmować wiek pasażera, klasę biletu, płeć itp. Model k-NN (k-najbliższych sąsiadów) to technika uczenia maszynowego, która przewiduje klasę obiektu (w tym przypadku, czy pasażer przeżył, czy nie) na podstawie k-najbliższych sąsiadów w przestrzeni cech. W kontekście katastrofy Titanica cechy mogą obejmować wiek pasażera, płeć, klasę biletu, liczbę rodzeństwa/małżonków na pokładzie, liczbę rodziców/dzieci na pokładzie, czy pasażer podróżował samotnie itp. Dla danego pasażera model k-NN znajduje k-najbliższych sąsiadów (inni pasażerowie o podobnym wieku, płci, klasie biletu itp.) w przestrzeni cech i przypisuje klasę, która jest najczęściej reprezentowana wśród tych sąsiadów. Na przykład, jeśli dla danego pasażera większość z k-najbliższych sąsiadów przeżyła, model k-NN przewiduje, że ten pasażer również przeżył. Jeśli większość z k-najbliższych sąsiadów nie przeżyła, model przewiduje, że ten pasażer również nie przeżył. Wartość k jest parametrem, który można dostosować w celu optymalizacji wydajności modelu na danych testowych. Trzeba jednak pamiętać, że model k-NN i ten oparty na regresji logistycznej jest jedynie modelem predykcyjnym i nie może dostarczyć „prawdziwych" przyczynowych wniosków o przetrwaniu katastrofy Titanica. To znaczy, że choć model może przewidzieć, czy dany pasażer przeżył na podstawie jego cech, nie możemy na podstawie tych przewidywań stwierdzić, że te cechy były przyczyną przeżycia lub śmierci pasażera – choć oczywiście mogło tak być!

Po przetrenowaniu modelu (oczywiście automatycznym) okazało się, że pierwszy model przewidywał szanse przeżycia na poziomie 77% na danych testowych. Natomiast k-NN dał wynik 73%. Jeśli chodzi o inżynierię cech[34], to

[34] Inżynieria cech to proces wyodrębniania, tworzenia lub selekcji cech z danych wejściowych w celu poprawienia wydajności modeli uczenia maszynowego. Cechy to właściwości lub atrybuty, które opisują dane wejściowe. W przypadku uczenia maszynowego modele uczą się na podstawie zestawu cech, które opisują różne aspekty danych. Przykładowe cechy mogą

model oparty ona regresji logistycznej dał wynik 77%, a k-NN – 80%. Model k-NN z inżynierią cech jest najbardziej precyzyjny w odniesieniu do danych testowych, osiągając dokładność 0,80%[35].

Cyfrowe kompetencje

Zwrot cyfrowy w archeologii, historii, literaturoznawstwie, kulturoznawstwie czy medioznawstwie, jaki dokonał się w XXI wieku, doprowadził do powstania nowego paradygmatu nazywanego humanistyką cyfrową. Jego praktykowanie i refleksja nad specyfiką przekazów cyfrowych wymagają – jak trafnie dostrzegają uznani przedstawiciele historii cyfrowej w Polsce, Wiktor Werner i Adrian Trzoss – fachowych umiejętności, które zazwyczaj są trudno dostępne dla humanisty. Stąd dające się zauważyć rozdzielenie refleksji nad fenomenami cyfrowymi na część humanistyczną, w której zagadnienia techniczne są traktowane powierzchownie, i część informatyczną, słabo zakorzenioną w tradycji humanistycznej[36].

Cyfrowa kultura, w której żyjemy i która już zdominowała świat analogowy, wymaga do jej poznawania i badania nowych metod, narzędzi i organizacji pracy badawczej. To samo odnosi się i do zdygitalizowanych artefaktów z przeszłości i w ogóle źródeł cyfrowych. W przypadku nauk historycznych ważne jest rozróżnienie na tak zwane źródła *born-digital* (pierwotnie cyfrowe) oraz *reborn-digital*[37]. Wybrane przykłady takich metod i narzędzi (wraz z próbą ich aplikacji do konkretnych problemów badawczych) zostały przedstawione we wcześniejszych fragmentach niniejszego tekstu. Stanowią one podstawę warsztatu cyfrowego humanisty. Jakie zatem nowe kompetencje powinni posiadać ci wszyscy, którzy chcieliby uprawiać badania w zakresie cyfrowej historii, literaturoznawstwa, antropologii kulturowej i innych dyscyplin humanistycznych? Najogólniej rzecz biorąc, można je sprowadzić do czterech głównych grup.

---

obejmować długość tekstu, ilość słów kluczowych czy też parametry obrazu, takie jak jasność, kontrast czy rozmiar.

35 Tu warto zauważyć, ze zbiór titanic jest też często używany do uczenia i testowania algorytmów uczenia maszynowego takich jak drzewa decyzyjne czy lasy losowe, a także do przeprowadzania różnych testów statystycznych, np. chi-kwadrat, testy t, analiza wariancji (ANOVA) i tworzenia różnych wizualizacji danych. Więcej, zob. Iqbal Saker, *Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision Making and Application Perspective* (New York: Springer, 2021), 377.

36 Wiktor Werner, Adrian Trzoss, „Towards the Digital Historiography", *Historyka. Studia Metodologiczne* 50, (2020): 435–436.

37 Niels Brugger, któremu zawdzięczamy to rozróżnienie (jego prace koncentrują się na koncepcji historiografii sieciowej, historii cyfrowej i studiów nad Internetem), zauważa w pracy „When the Present Web is Later the Past: Web Historiography, Digital History, and Internet Studies", że sieć jest ważnym źródłem historycznym i że szczególną uwagę należy zwrócić na sieć w archiwach sieciowych – określaną jako materiał *reborn-digital* – ponieważ te źródła prawdopodobnie będą jedynym materiałem pozostawionym dla przyszłych historyków. Brügger zwraca uwagę na charakterystykę materiału *reborn-digital* w archiwach sieciowych i jak wpływa to na wykorzystanie tego materiału przez historyka, a także na możliwe zastosowanie cyfrowych narzędzi analitycznych do tego rodzaju materiału.

Pierwsza grupa to analiza danych. Stanowią one podstawę empiryczną badacza cyfrowego. W humanistykach analogowych były to źródła historyczne, dzieła literackie czy obserwacje (np. z terenu fizycznego lub wirtualnego). Podstawową umiejętnością była interpretacja takich źródeł i tekstów, często nazywana *close reading* (dokładne, szczegółowe czytanie). W wyniku aktów interpretacji konstruowano fakty, ustalano procesy literackie czy podstawy funkcjonowania rzeczywistości kulturowej, a następnie pisano narracje o dziejach, zjawiskach literackich czy rozwoju kultury. W przypadku humanistyki cyfrowej przedmiotem badania stają przede wszystkim dane. To, co dla historyków i historyczek jest źródłem (np. pisanym), dla literaturoznawców i literaturoznawczyń dziełem literackim, a dla antropologów i antropolożek obserwacją z życia jakiejś grupy kulturowej, to dla badaczy spod znaku cyfrowej humanistyki będzie potencjalnym zbiorem danych. Inaczej mówiąc, nie interesuje ich narracja na przykład kroniki, poematu czy notatka z badań terenowych, ale przekształcenie źródła historycznego, tekstu literackiego czy notatki w zestaw danych. Przekształcenie to nazywa się konwersją z postaci analogowej (ciągłej) na postać numeryczną (jednostek nieciągłych, dyskretnych). Dzięki temu otrzymujemy zbiór danych właśnie, w formacie czytelnym dla komputera. W przypadku danych ustrukturyzowanych (tabelarycznych) będzie to najczęściej format .csv. Dzięki temu na takich danych będziemy mogli przeprowadzać różnego typu analizy (eksploracyjne czy bardziej pogłębione), odkrywać różnego rodzaju relacje, korelacje czy trendy. Jest to szczególnie istotne w przypadku dużej ilości danych nazywanych *big data*.

Druga grupa kompetencji to wizualizacja wiedzy. Jest ona rozumiana jako: a) metoda badawcza, b) sposób prezentacji wyników badań. W humanistyce analogowej dominowały analizy tekstów (np. kronik, dokumentów, powieści itp.). Podstawową kompetencją była interpretacja tekstu, jego języka czy tekstualnego świata. Narracja o dziejach czy literaturze zawsze ma charakter pisany. Publikacja wyników badań ma formę artykułu w czasopiśmie lub książki (drukowanej czy ostatnio też elektronicznej). W przypadku humanistyki cyfrowej coraz większą rolę odgrywa wizualizacja. Jest to konsekwencja pracy z danymi, które zazwyczaj albo są w postaci tabelarycznej zawierającej tysiące czy setki tysięcy rekordów, albo zbiorem danych medialnych (nieustrukturyzowanych). Aby analiza takich danych mogła przynieść interesujące i przede wszystkim czytelne rezultaty, musi być ukazana w formie obrazowej, na przykład w postaci diagramów, linii trendu, szeregów czasowych czy histogramów (tak jak to było pokazane we wcześniejszych przykładach). Wiele projektów z humanistyki cyfrowej ma charakter wizualny: filmy, modele 3D, mapowanie danych w różnej postaci, czyli multimedialnych (tzw. *linked-data*), animacje wydarzeń itp. Stąd tak ważne jest operowanie „językiem" obrazu, czyli tworzenie wizualnych narracji, które następnie są publikowane w sieci (np. infografiki, *digital stories* czy interaktywne e-booki).

Trzecia grupa to umiejętności programistyczne. Większość wyłuszczonych tu czynności badawczych charakterystycznych dla humanistyki cyfrowej ma charakter zautomatyzowany – począwszy od pobierania danych (np. z baz danych

czy „zeskrobywania" ich z Internetu) po ich przygotowanie do dalszych badań: analizę, wizualizację czy tworzenie dashboardów (raportów z badań do publikacji w sieci). Służą do tego między innymi programy narzędziowe typu Tableau, Power Bi, Exploratory czy Flourish, jednak mają one ograniczone funkcjonalności i w większości przypadków są komercyjne. Takich ograniczeń nie mają badania prowadzone za pomocą odpowiednich środowisk, jak pokazane tu R Studio. Wymagają one jednak umiejętności programistycznych. Dla cyfrowych humanistów i humanistek podstawowym językiem programowania jest R oraz Python. Oba te języki mają wiele pakietów i bibliotek programistycznych dostosowanych do automatycznego pobierania danych, ich „czyszczenia", analizowania i wizualizacji. Umożliwiają też tworzenie modeli uczenia maszynowego, a także głębokiego uczenia, czyli sztucznej inteligencji.

I wreszcie, czwarta grupa to nowa organizacja pracy i narzędzia z nią związane. W humanistyce analogowej zdecydowanie dominowała indywidualna praca, w zaciszu gabinetu, a głównym jej narzędziem był umysł badacza czy badaczki i ich umiejętności interpretacyjne. W humanistyce cyfrowej pracuje się coraz częściej nad wielkimi projektami wymagającymi współdziałania obok humanistów także informatyków, grafików czy specjalistów od zarządzania i marketingu[38]. Do tego celu utworzono dziesiątki Laboratoriów czy Centrów Humanistyki Cyfrowej. Praca nad projektami wymaga konkretnych cech i umiejętności. Są to: praca w zespole, myślenie projektowe (*design thinking*), wykorzystanie cyfrowych map myśli (np. Mindomo czy MindMup), posługiwanie się narzędziami do zarządzania treścią i projektami (np. Trello, Notion czy Asana), komunikacji (np. Slack), a także umiejętności „miękkie" (np. komunikacja, współpraca, przywództwo, kreatywność, motywowanie, empatia czy adaptacja).

Zaprezentowany powyżej warsztat cyfrowego badacza z dziedziny badań humanistycznych zawiera tylko najważniejsze narzędzia i metody. Każda z dyscyplin humanistycznych, oprócz wymienionych tu składników rzeczonego warsztatu, zawiera charakterystyczne dla siebie metody i oprogramowanie służące do analizy i wizualizacji różnych badanych problemów.

Bibliografia

Allington, Daniel, Sarah Brouillette, and David Golumbia. *Neoliberal Tools and Archives: A Political History of Digital Humanities.* Ann Arbor: University of Michigan Press, 2021.

Alvarado, Rafael C. „Blog posts". W *Debates in the Digital Humanities*, red. Matthew Gold. Minneapolis: Univeristy Minessota Press, 2012. https://dhdebates.gc.cuny.edu/read/untitled-

[38] Obok już wspomnianych projektów – Wirtualny Rzym i Photogrammar – przykładem mogą być: Programujący historyk (https://programminghistorian.org/) i jego polski odpowiednik realizowany przez CKC UW (https://humanistyka.dev/?fbclid=IwAR2FJq4eWIK6tQbi450-SYWmYYxyRF2pnzY42ReHRzYNGOsF5ZWr8snQo9Xc_aem_AXGcJr-slar5X1z5R6HJT90x6LdjNFhF7ST1pTay9XaY8UJNL2t287dVHXrbYd7GwEwnzHy4My-zf4T9nIqtnfD-N6) (dostęp: 25.05.2024), a także np. zbiór narzędzi przygotowanych przez CLARIN.PL (https://clarin-pl.eu/).

-88c11800-9446-469b-a3be-3fdb36bfbd1e/section/c513af64-8f99-4e02-9869-babc1cecc451#p1b1 (dostęp: 12.03.2023).

Berry, David M. *Understanding Digital Humanities*. New York: Palgrave Macmillan, 2012.

Brugger, Niels. *The Archived Web: Doing History in the Digital Age*. Cambridge, Massachusetts: The MIT Press, 2018.

Burdick, Anne et al. *Digital_Humanities*. Cambridge, Massachusetts: MIT Press, 2012.

Conway, Drew, John White. *Uczenie maszynowe dla programistów*, przeł. Przemysław Szermiota. Gliwice: Helion, 2015.

Debates in the Digital Humanities, red. Matthew Gold. Minneapolis: University of Minnesota Press, 2012.

Drucker, Johanna. *The Digital Humanities*. London and New York: Routledge, 2012.

Drucker, Johanna. *SpecLab: Digital Aesthetics and Projects in Speculative Computing*. Chicago: University of Chicago Press, 2009.

Gillespie, Colin, Robin Lovelace. *Wydajne programowanie w R*. Warszawa: Wyd. APN, 2018.

The Historical Web and Digital Humanities: The Case of National Web Domain, red. Niels Brugger, Ditte Laursen. Londyn: Routledge, 2019.

Lander, Jared P. *R dla każdego. Zaawansowane analizy i grafika statystyczna*, przeł. Marek Włodarz. Warszawa: Wyd. APN, 2018.

Kirschenbaum, Matthew G. *Mechanisms: New Media and the Forensic Imagination*. Cambridge Massachusetts: MIT Press, 2008.

Liu, Alan. *Local Transcendence: Essays on Postmodern Historicism and the Database*. Chicago: University of Chicago Press, 2008.

Liu, Alan. „The State of the Digital Humanities: A Report and a Critique". 4.06.2016. https://escholarship.org/content/qt23h6v6x8/qt23h6v6x8_noSplash_0ef49e16691ca30-b532e3cb6bc5cf080.pdf (dostęp: 10.03.2023).

Manovich, Lev. *Cultural Analytics*. Cambridge, Massachusetts: The MIT Press, 2020.

Manovich, Lev. „The Science of Culture? Social Computing, Digital Humanities, and Cultural Analytics". Manovich. 2015. http://manovich.net/index.php/projects/cultural-analytics--social-computing (dostęp: 10.03.2023).

Moretti, Franco. *Distant Reading*. Londyn: Verso, 2013.

Patterson, Josh, Adam Gibson. *Deep Learning*. Gliwice: Helion, 2018.

Posner, Miriam. „Humanities Data: A Necessary Contradiction". *Journal of Digital Humanities* 2, 3 (2013).

Presner, Todd. *Digital Humanities 2.0: A Report on Knowledge*. Cambridge, Massachusetts: University of California Press, 2016.

Ramsay, Stephen. *Reading Machines: Toward an Algorithmic Criticism*. Champaign: University of Illinois Press, 2011.

Risam, Roopika. *New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy*. Evanston: Northwestern University Press, 2018.

Rockwell, Geoffrey, Stéfan Sinclair. *Hermeneutica: Computer-Assisted Interpretation in the Humanities*. Cambridge, Massachusetts: MIT Press, 2016.

Schreibman, Susan, Ray Siemens, and John Unsworth. *A Companion to Digital Literary Studies*. Hoboken: Wiley-Blackwell, 2008.

Siemens, Ray, and Susan Schreibman. *A Companion to Digital Humanities*. Hoboken: Wiley-Blackwell, 2004.

Smołucha, Danuta. *Humanistyka cyfrowa w badaniach kulturowych. Analiza zjawiska na wybranych przykładach kulturowych*. Kraków: Wydawnictwo Naukowe Akademii Ignatianum, 2023.

Svensson, Patrik. „The Landscape of Digital Humanities". *Digital Humanities Quarterly* 4, 1 (2010).

Szpunar, Magdalena. *Kultura algorytmów*. Kraków: Wyd. UJ, 2019.
Szubański, Kamil. „Naukowiec o przewadze, jaką daje humanistyka cyfrowa". Nauka w Polsce. 2.12.2018. https://naukawpolsce.pl/aktualnosci/news%2C31908%2Cnaukowiec-o-przewadze-jaka-daje-humanistyka-cyfrowa.html (dostęp: 8.03.2023).
Terras, Melissa, Julianne Nyhan, and Edward Vanhoutte. *Defining Digital Humanities: A Reader*. Farnham: Ashgate Publishing Ltd., 2013.
Werner, Wiktor, Adrian Trzoss. „Towards the Digital Historiography". *Historyka. Studia Metodologiczne* 50 (2020): 435–436.
Underwood, Ted. *Distant Horizons: Digital Evidence and Literary Change*. Chicago: University of Chicago Press, 2019.
The Web as History: Using Web Archives to Understand the Past and the Presen, red. Niels Brugger, Ralph Schroeder. Londyn: UCL Press, 2017.
Wymer, Kathryn. *Introduction to Digital Humanities*. London: Routledge, 2021.
Vanhoutte, Edward. *Scholarly Editing in the Computer Age: Theory and Practice*. Ann Arbor: University of Michigan Press, 1997.

Zasoby internetowe

GitHub. https://github.com/hadley/babynames (dostęp: 25.05.2024).
Kaagle. https://www.kaggle.com/code/dwiuzila/titanic-complete-analysis-in-r-top-4 (dostęp: 23.03.2023).
Photogrammar. https://photogrammar.org/maps (dostęp: 24.05.2024).
Programujący historyk. https://programminghistorian.org/ (dostęp: 25.05.2024).
Repozytorium CRAN. https://cran.r-project.org/ (dostęp: 25.05.2024)
Republika Listów. http://republicofletters.stanford.edu/ (dostęp: 24.05.2024).
Wirtualny Rzym. https://www.romereborn.virginia.edu/ (dostęp: 23.05.2024).