Projekt i implementacja aplikacji do kodowania odpowiedzi z pytań otwartych w badaniach ankietowych

1. Wstęp.

Przedmiotem niniejszej pracy jest opis wykonanej – w ramach pracy dyplomowej - aplikacji webowej służącej do kodowania odpowiedzi z pytań otwartych (wykorzystywanych w badaniach społecznych lub badaniach rynku do uwzględnienia tych odpowiedzi respondentów, które nie zostały zdefiniowane przez badacza na etapie projektowania kwestionariusza wywiadu), a także przedstawienie jej projektu, w tym dalszych możliwych kierunków rozwoju.

Tłem dla zrealizowanego pomysłu była praca przy kodowaniu odpowiedzi respondentów z badań rynku, która odbywała się z użyciem arkuszy kalkulacyjnych i to właśnie narzędzie było głównym punktem odniesienia. Ma to swoje konsekwencje – po pierwsze, starano się zachować pewną spójność i zgodność, jeśli chodzi o strukturę plików czy możliwość wymiany danych między aplikacją a arkuszem kalkulacyjnym; po drugie, skoncentrowano się na usprawnieniu procesu ręcznego kodowania, a nie zastąpieniu go przez kodowanie automatyczne, szczególny nacisk kładąc na wyszukiwanie (filtrowanie) oraz zarządzanie kluczem kodowym.

Od strony technicznej, aplikacja została przygotowana z wykorzystaniem języka R, a w jego ramach (głównie) frameworku (pakietu) *shiny*, który stanowi zbiór funkcji tłumaczonych na HTML, CSS i JavaScript. R to jeden z najbardziej oczywistych wyborów w kontekście celów aplikacji, ze względu na jego przeznaczenie do pracy z danymi (zwłaszcza w formie tabelarycznej) oraz względnie dużą popularność w obszarze nauk społecznych [A]. Podobnie, *shiny* to najpopularniejszy pakiet (w ramach R) umożliwiający udostępnianie użytkownikom rozwiązań bez konieczności pisania przez nich kodu [B].

Struktura pracy dyplomowej, oprócz rozdziałów poświęconych wybranej technologii oraz obszaru problemowego, jakim jest kodowanie pytań z odpowiedzi otwartych, obejmuje przedstawienie założeń aplikacji, jej funkcji, opis bazy danych, testów oraz funkcji, których nie zaimplementowano, a które znacznie poprawiłyby skuteczność kodowania odpowiedzi i przyśpieszyłyby je. Wybór technologii oraz konstrukcja samej aplikacji (w tym również bazy danych) tworzona była również z myślą o tych pominiętych funkcjonalnościach, dlatego wątek ten jest dość istotny, przy czym ograniczono go głównie – w myśl przyjętych założeń, na których miała opierać się budowa aplikacji – do problematyki dalszego usprawnienia procesu wyszukiwania odpowiedzi, w których może znajdować się dany kod (tj. do wyszukiwania odpowiedzi, które są do siebie na tyle podobne, że w ramach przyjętego klucza kodowego, traktowane są jako tożsame). Problem ten sprowadza się, oczywiście, do określenia podobieństwa łańcuchów znaków i jako taki jest niezwykle szerokim, interdyscyplinarnym zagadnieniem, który funkcjonuje pod terminem przetwarzania języka naturalnego (NLP).

Pracę zamyka podsumowanie, które stara się zebrać najważniejsze z poruszonych wątków, przedstawiając je w możliwie najbardziej skondensowanej, spójnej postaci.

1. Założenia projektowe.
   1. Kodowanie odpowiedzi z pytań otwartych jako niezbędny element procesu badawczego.

Kodowanie to proces, w którym danym fragmentom tekstu (wypowiedzi) przyporządkowuje się kategorie. Kategorie to zazwyczaj liczby wraz z określoną etykietą [C]. Proces ten jest konieczny wtedy, gdy w badaniach ankietowych zastosowano pytania otwarte lub półotwarte, tj. takie, w których wypowiedź respondenta jest wpisywana w całości. Z dużym prawdopodobieństwem bowiem, każda taka wypowiedź strukturalnie (gramatycznie) lub semantycznie jest odmienna – respondenci w badaniach posługują się różnymi sformułowaniami (popełniają również błędy) i, dodatkowo, nie zawsze to samo rozumieją pod danym sformułowaniem. Trudno sobie zatem wyobrazić jakąkolwiek analizę (nie tylko ilościową), w której nie doprowadzono by surowych danych do postaci zbioru wspólnych kategorii (kodów właśnie). Dane w postaci zakodowanej pozwalają następnie na różnego rodzaju analizy (np. analizę częstości występowania danych odpowiedzi) i wizualizacje.

Największym wyzwaniem podczas kodowania jest niewątpliwie tworzenie kodów – istnieje napięcie między tym, na ile kod powinien być szczegółowy, a nie ile ogólny. Żadna z tych skrajności nie jest pożądana – zbyt duża szczegółowość doprowadzi do braku możliwości uchwycenia różnic między respondentami (różnice będą tak małe, że staną się statystycznie nieistotne), podobnie zbyt duża ogólność. Przyjmuje się, że kody powinny odpowiadać celom badawczym – postawionym pytaniom szczegółowym i hipotezom. Nie powinny być także w całości oparte o uprzednio stworzony klucz kodowy (tj. zbiór kodów wraz z etykietami) – to w skrajnym ujęciu doprowadziłoby do absurdalnej sytuacji, kiedy swobodna wypowiedź respondenta musi być zamknięta w z góry narzucone ramy skonstruowane przez badacza. Kodowanie wymagałoby zatem wiedzy na temat pytań szczegółowych i hipotez badawczych przy jednoczesnym oparciu klucza kodowego na rzeczywistej treści wypowiedzi [C].

Tak opisaną procedurę trudno zautomatyzować. Tym trudniej, im mniej ścisłe jest określenie, *co* badacz ma zamiar sprawdzić. I tak np. o wiele prostsze (choć nietrywialne) wydaje się – myśląc o automatyzacji kodowania – sprawdzenie, czy respondenci mają pozytywny, neutralny czy negatywny sentyment wobec Mazur niż określenie, jakie skojarzenia mają respondenci z Mazurami. Z tych powodów oprogramowania do tak rozumianego kodowania odpowiedzi z pytań otwartych opierają głównie na przeszukiwaniu danych tekstowych [D]. Taki też był główny cel wykonanej aplikacji – miała ona ułatwić (przyspieszyć) proces ręcznego kodowania, a nie go zastąpić.

* 1. Przegląd oprogramowania wykorzystywanego do kodowania tekstu.

Bibliografia

A: Robinson D., *The Impressive Growth of R*, Stack Overflow Blog, 10.11.2017, dostęp: <https://stackoverflow.blog/2017/10/10/impressive-growth-r/> (14.12.2021).

B: King R., *Shiny vs. Dash: A Side-by-side Comparison*, R. King Data Consulting, 6.03.2019, dostęp: <https://www.rkingdc.com/blog/2019/3/6/shiny-vs-dash-a-side-by-side-comparison> (14.12.2021).

C: Haczkowska A., *Analiza trudności związanych z kodowaniem odpowiedzi na pytania otwarte i próba ich rozstrzygnięcia z pomocą psychologii poznawczej na przykładzie badania skojarzeń internautów z regionami Polski*, „Nauki o Zarządzaniu. Management Sciences”, 2012, 11, 2, s. 55-73.

D: Bryda G., *CAQDAS a badania jakościowe w praktyce*, „Przegląd Socjologii Jakościowej”, 2014, 10, 2, s. 12-39.

E: Kuckartz U., Rädiker S., *Analyzing Qualitative Data with MAXQDA. Text, Audio, and Video*, Springer, 2019.

F: Chandra Y., Shang L., *Qualitative Research Using R: A Systematic Approach*, Springer, 2019.

G: Huang R., *RQDA: R-based qualitative data analysis. R package version 0.2-8*, dostęp: <http://rqda.r-forge.r-project.org/> (15.12.2021).