README

Szymon Talaga

10.12.2014

[**English follows**]

Repozytorium zawiera dane, skrypty i własne funckje, które są niezbędne do zreplikowania wyników badań, które przeprowadziłem w ramach swojej pracy magisterskiej pisanej na Wydziale Psychologii Uniwersytetu Warszawskiego pod opieką profesor Marii Lewickiej. Dodatkowo, znajdują się w nim również raporty (w formatach .pdf/.html i .Rmd), które w sposób syntetyczny, ale opisowy i wyczerpujący przedstawiają wszytkie najważniejsze etapy skalowania i redukcji danych oraz właściwych analiz służących weryfikacji postawionych w pracy hipotez.

Wszystkie dane są udostępnione zarówno w formie surowej jak i w postaci przygotowanej do poszczególnych etapów analizy. Jedyna zmiana w danych surowych to ich pełna anonimizacja. Ponadto niektóre etapy transformacji i rekodowania danych nie są w pełni udokumentowane. Dotyczy to przede wszystkim klasyfikacji i grupowania surowych danych tekstowych dotyczących odwiedzanych przez respondentów miejsc. Wynika to z faktu, że ilość tych danych była bardzo duża (ponad 2 tys. wskazań na miejsca w danych surowych) i były one zupełnie nieustrukturyzowane przez co musiały być rekodowane w dużym stopniu ręcznie i przy użyciu setek wyrażeń regularnych, co oczywiście sprawiało, że pełne udokumentowanie wszystkich transformacji było zadaniem ponad siły jednej osoby. Więcej informacji na temat rekodowania surowych danych ze wskazaniami na miejsca w opisie danych dotyczących miejsc.

Mapa repozytorium

- **Katalog główny (MA)** - zawiera plik z kwestionariuszem użytym do zebrania danych (Kwestionariusz.pdf) oraz plik projektu Rstudio o nazwie MA.Rproj. Ponadto zawiera w sobie katalogu z poszczególnymi elementami:
  - **rawData** - tutaj znaleźć można wszystkie pliki z danymi surowymi oraz pliki pomocniczne użyte przy rekodowaniu i klasyfikowaniu odpowiedzi z pytań otwartych. Zanjduje sę tutaj również skrypt ("RelabellingAndBasiDataTrans.R"), który został użyty do przekształcnia surowego zbioru "raw244.csv" w zbiór "DatInd231.csv", który znaleźć można w katalogu "MainData". Skrypt zawiera również opis kolejnych czynności i transformacji.
  - **MainData** - tutaj znajdują się wszystkie zbiory danych po kolejnych etapach rekodowana i transformacji. Wszystkie z nich zapisane są jako obiekty R (format .RData), ponieważ ten format zapewnia utrzymanie poprawnej kolejności poziomów zmiennych jakościowych. Niemniej jednak najważniejsze zbiory danych zapisane są również w formacie .csv w celu zapewnienia możliwości replikacj wyników również przy użyciu innych niż język R narzędzi. Dokładny opis kolejności powstawania zbiorów danych i relacji między nimi w sekcji "Generowanie Danych".
  - **HelperFunctionsMisc** - ten foleder zawiera skrypty z różnymi dodatkowymi funkcjami, które napisałem, żeby ułatwić sobi pracę oraz zwiększyć czytelność kodu.
  - **Imputation** - tutaj znaleźć można skrypty służące podstawaniu brakujących odpowiedzi. Znajduje się tu również skrypt z funkcjami pomocniczymi, których używałem przy podstawianiu a także obiekt .RData zawierający całą przestrzeń zmiennych związanych z finałowym etapem podstawiania brakujących danych przy pomocy algorytmu MICE. Służy to temu, żeby nie trzeba było za każdym razem od nowa dokonywać wszystkich obliczeń, ponieważ algorytm MICE jest dosyć kosztowny obliczeniowo, przez co cały proces może zająć nawet do kilkunastu minut.

Generowanie Danych

UWAGA: poprawne wykonanie wszelkich skryptów zawartych w tym repozytorium jest zależne od struktury środowiska całego projektu. Innymi słowy struktura katalogów oraz lokalizacja poszczególnych plików musi pozostać niezmieniona.

Dane na temat respondentów

Aby wygenerować ostateczny zbiór danych z poziomu danych surowych należy wykonać kolejno następujące kroki:

1. Wykonaj skrypt “rawData/RelabellingAndBasiDataTrans.R”
2. Wykonaj skrypt “Imputation/BasiImpute.R”
3. Wykonaj skrypt “Imputation/LimitData.R”
4. Wykonan skrypt “Imputation/ImputeGenMICE.R”