Eksploracyjna analiza tekstu w R

Zadania do wykonania

Przeprowadź eksploracyjną analizę własnego zbioru dokumentów tekstowych według poniższych wymagań:

1. W badaniach wykorzystaj 20 dokumentów tekstowych **w języku polskim** podobnej długości (nie za krótkich, ale również nie za długich) pochodzących z 5 różnych dziedzin/tematów (po 4 dokumenty z każdego tematu). Przynajmniej jeden temat powinien się wyróżniać od pozostałych, a dwa tematy powinny być do siebie bardzo zbliżone.
2. Wszystkie dokumenty zapisz w formacie .txt z kodowaniem znaków UTF-8
3. Korzystając z bibliotek R poznanych na zajęciach kolejno:
    a. utwórz korpus dokumentów
    b. poddaj korpus dokumentów wstępnemu przetwarzaniu
    c. utwórz kilka różnych macierzy częstości zmieniając wagę oraz maksymalną i minimalną liczbę dokumentów w których mogą/muszą wystąpić słowa (użyj 3 różnych par wartości), żeby były wzięte pod uwagę w macierzy częstości; wykorzystaj te macierze częstości w dalszych analizach
    d. przeprowadź próby redukcji wymiarów macierzy częstości przy użyciu analizy głównych składowych oraz dekompozycji według wartości osobliwych
    e. dokonaj analizy skupień dokumentów; dla metod hierarchicznych dobierz liczbę skupień na podstawie długości wiązań, a dla metod niehierarchicznych potraktuj liczbę tematów jako potencjalną liczbę skupień, ale przeprowadź eksperymenty również dla większej i mniejszej liczby skupień
    f. przeprowadź analizę tematyk korzystając z metody ukrytej alokacji Dirichlet'a; potraktuj liczbę tematów jako potencjalną liczbę tematyk, ale przeprowadź eksperymenty również dla większej i mniejszej liczby tematyk
    g. dla każdego dokumentu wyznacz słowa/frazy kluczowe korzystając z 2-3 różnych metod
4. Na podstawie uzyskanych wyników przygotuj sprawozdanie w którym znajdą się:
    a. opis utworzonego zbioru dokumentów z ich podstawowymi statystykami
    b. opis przeprowadzonych eksperymentów
    c. opis wyników eksperymentów wraz z wnioskami odnoszącymi się do konkretnego badanego zbioru dokumentów

 Opracowała: Katarzyna Wójcik

Dodatkowe wymagania

1. Jako stronę tytułową sprawozdania wykorzystaj plik dostępny na platformie e-learningowej uzupełniony danymi swojego zespołu projektowego. Podaj za jakie czynności odpowiadał każdy członek zespołu i jaki był jego/jej udział procentowy w wykonaniu zadania.
2. Plikowi nadaj nazwę według wzoru: NumerGrupyProjektowej_Projekt.[doc|docx|odt] (np. 1_Projekt.docx), **a następnie zapisz, wyeksportuj lub wydrukuj plik do formatu .pdf nie zmieniając samej nazwy pliku** (np. 1_Projekt.pdf).
3. Dokumenty tekstowe zapisz w katalogu o nazwie NumerGrupyProjektowej_Dokumenty, a następnie spakuj ten katalog do archiwum .zip o analogicznej nazwie (np. 1_Dokumenty.zip).
4. Wszystkie polecenia R zawrzyj w jednym skrypcie i zapisz go pod nazwą NumerGrupyProjektowej_Kod.R (np. 1_Kod.R).
5. Prześlij 3 pliki (sprawozdanie, archiwum z dokumentami oraz skrypt R) przez platformę e-learningową w terminie wskazanych we właściwym zadaniu (aktywności). UWAGA! W zespołach 2-osobowych wystarczy jeśli jedna osoba prześle pliki przez platformę e-learningową.
6. Przynieś wydrukowane sprawozdanie (1 kopia na grupę projektową) na najbliższe zajęcia (dla studiów stacjonarnych) lub na najbliższy zjazd (dla studiów niestacjonarnych). Wydruk może być roboczej jakości, monochromatyczny, pomniejszony do 2 stron na 1.

