Eksploracyjna analiza tekstu w R

Zadania do wykonania

Przeprowadź eksploracyjną analizę zbioru dokumentów tekstowych według poniższych wymagań:

1. W badaniach wykorzystaj 20 dokumentów tekstowych podobnej długości (nie za krótkich, ale również nie za długich) pochodzących z 5 różnych dziedzin/tematyk (po 4 dokumenty z każdej tematyki). Przynajmniej jedna tematyka powinna się wyróżniać od pozostałych, a dwie tematyki powinny być do siebie 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 wystąpić słowa, żeby były wzięte pod uwagę w macierzy częstości
    d. przeprowadź próby redukcji wymiarów macierzy częstości przy użyciu analizy głównych składowych oraz dekompozycji według wartości osobliwych
    e. dokonaj analizy skupień dokumentów; potraktuj liczbę tematyk jako potencjalną liczbę skupień, ale przeprowadź eksperymenty również dla większej i mniejszej liczby skupień
    f. przeprowadź analizę korzystając z metody ukrytej alokacji Dirichlet'a
    g. dla każdego dokumentu wyznacz słowa/frazy kluczowe korzystając z 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]
3. Dokumenty tekstowe zapisz w katalogu o nazwie NumerGrupyProjektowej_Dokumenty, a następnie spakuj ten katalog do archiwum .zip o analogicznej nazwie.
4. Wszystkie polecenia R zawrzyj w jednym skrypcie i zapisz go pod nazwą NumerGrupyProjektowej_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. Wydrukowany dokument sprawozdania przynieś w terminie ustalonym z prowadzącym. Wystarczy wydruk w jakości roboczej. Kartki dokumentu połącz za pomocą zszywacza lub spinacza biurowego (bez koszulek, bindowania, listw itp.).

