Eksploracja zasobów internetowych

Laboratorium

Temat: Przetwarzanie dokumentów tekstowych

Cel:

- Zapoznanie się ze sposobem przetwarzania danych tekstowych
- Reprezentacje dokumentów tekstowych
- Poszukiwanie podobieństw w dokumentach tekstowych

Z wykonania laboratorium przygotuj sprawozdanie umieszczając w nim odpowiedzi na poszczególne pytania oraz komentarze istotne dla wykonania poszczególnych poleceń.

1. Zapozna się ze strukturą katalogu CCSU departments.
   Co zawiera ten katalog?
   Jaka jest struktura zawartości dokumentów katalogu?
2.
   a) Zobacz jak wygląd plik w formacie arff – formacie odpowiednim dla programu WEKA Pomocnym dla ciebie będzie przykładowy dokument programu WEKA o nazwie weather.nominal.arff
   b) Złącz zawartości dokumentów z katalogu i utwórz dokument arff odpowiedni dla programu WEKA (z tego programu skorzystamy celem przygotowania odpowiedniej struktury danych).
      Na początku każdej linii dodaj nazwę dokumentu oraz nagłówek pliku na jego początku.

   @relation departments_string

   @attribute document_name string
   @attribute document_content string
   @data Anthropology, " Anthropology consists of four …
   .."

   c) Jeśli nie udało się Tobie przygotować takiego pliku – zapoznaj się z zawartością pliku Department-string.arff
   d) Plik Department-string.arff używa dwóch atrybutów: document_name oraz document_content. Jest tam też dodatkowy atrybut class

A.

3. Wczytaj plik Department-string_bc.arff za pomocą **Open file** w zakładce **Preprocess**.
   Program pokaże też kilka statystyk dokumentu.

4. Wybierz filtr **StringToNominal** (znajduje się w grupie narzędzi unsupervised) i zastosuj do pierwszego atrybutu - document_name (aby to uczynić ustaw w parametrach filtru **attributeRange=first**).
   Zastosowanie filtra wymaga użycie przycisku **Apply**.

   Co teraz uzyskałeś?
   Jak wygląd twoja statystyka dotyczące dokumentu?

5. Wybierz filtr **StringToWordVector** z opcją **outputWord_Counts=true** oraz z opcją **tokenizer=AlphabeticTokenizer**

   Co otrzymałeś? Zobacz jak wygląd struktura danych wybierając Edit.

   Jest to macierz document-term.
   Pamiętasz co ona reprezentuje?
   Zapoznaj się z wykresami rozkładów termów.
6. Użyj filtru NumericToBinary.
   Co teraz możesz powiedzieć o uzyskanych statystykach (wykresach)?
   Znajdź najbardziej specyficzne termy dla każdego z dokumentów.
7. Możesz teraz usunąć zbędne termy?
   Jakie termy usuniesz?
   Zobacz po usunięciu termów jak wygląd zawartość danych (Edit) i zapisz dane do pliku csv nadając mu nazwę Department_terms.csv

B.

8. Wczytaj raz jeszcze plik Department-string_bc.arff za pomocą **Open file** w zakładce **Preprocess**.
9. Wybierz filtr **StringToNominal** (znajduje się w grupie narzędzi unsupervised) i zastosuj do pierwszego atrybutu - document_name (aby to uczynić ustaw w parametrach filtru **attributeRange=first**).
10. Wybierz filtr **StringToWordVector** z opcją **outputWord_Counts=true** oraz z opcją **tokenizer=AlphabeticTokenizer** oraz z opcją **IDFTransform=true**
11. Przeanalizuje dane przetworzone. Co one teraz przedstawiają? Jak jest różnica pomiędzy tą uzyskaną strukturą a strukturą binarną?
    Usuń termy, które nie wprowadzają istotnej informacji oraz zapisz dane do pliku csv nadając mu nazwę Department_TFIDF.csv

C.

12. Teraz wczytaj plik Department_TFIDF.csv do programu Rattle i dokonaj analizy struktury podobieństw dokumentów. Ile grup dokumentów podobnych obserwujemy w zbiorze analizowanych dokumentów?
13. Co może mieć wpływ na podobieństwa pomiędzy dokumentami? Dokonaj modyfikacji zbioru danych celem wyraźnej identyfikacji podobieństw pomiędzy dokumentami.

D.

14. Opracuj skrypt w języku **Python** dla potrzeb przetwarzania dokumentów tekstowych obejmujący funkcjonalności zarysowane wcześniejszymi punktami poleceń.

Umieść w skrypcie stosowne komentarze, które wyjaśnią znaczenie i przeznaczenie poszczególnych jego części.

15. Do sprawozdania dołącz opracowany sktypt.

E.

16. Przeprowadź analizę podobieństw (grupowania) dokumentów z folderu Top-100-websites. Analizę przeprowadź korzystając z poznanych narzędzi lub opracowanego skryptu.
    W sprawozdaniu zawrzyj odpowiedzi przynajmniej na następujące pytania:
    - Ile dokumentów będziesz przetwarzał?
    - Ile termów zidentyfikowałeś do popisu dokumentów?
    - Na jakich danych przeprowadzisz analizę?
    - Ile grup dokumentów podobnych zidentyfikowałeś?
    - Jakiej metody (metod) użyłeś do identyfikacji podobieństw?
    - Czy wszystkie metody wskazują na taką samą liczbę podobieństw?
    - Co może mieć wpływ na liczbę dokumentów podobnych?
    - Czy analizę przeprowadziłeś wykorzystując wszystkie atrybuty? Jeśli nie to jakie atrybuty wybrałeś do analizy i dlaczego?
    - Co możesz powiedzieć o grupach dokumentów, jakie zostały zidentyfikowane?

E.

17. Przeprowadza analizę dokumentów folderu Health-News-Tweets.
    Każdy plik tego folderu jest powiązany z jednym kontem agencji informacyjnej na Twitterze. Na przykład bbchealth.txt jest powiązany z wiadomościami zdrowotnymi BBC. Każdy wiersz zawiera identyfikator tweeta|datę i godzinę|tweet. Separatorem jest „|”.
    Analizę przeprowadź korzystając z poznanych narzędzi lub opracowanego skryptu.

    W sprawozdaniu zawrzyj odpowiedzi przynajmniej na następujące pytania:
    - Ile dokumentów będziesz przetwarzał?
    - Ile termów zidentyfikowałeś do popisu dokumentów?
    - Na jakich danych przeprowadzisz analizę?
    - Ile grup dokumentów podobnych zidentyfikowałeś?
    - Jakiej metody (metod) użyłeś do identyfikacji podobieństw?
    - Czy wszystkie metody wskazują na taką samą liczbę podobieństw?
    - Co może mieć wpływ na liczbę dokumentów podobnych?
    - Czy analizę przeprowadziłeś wykorzystując wszystkie atrybuty? Jeśli nie to jakie atrybuty wybrałeś do analizy i dlaczego?
    - Co możesz powiedzieć o grupach dokumentów, jakie zostały zidentyfikowane?