Katedra Systemów Informacyjnych

**Eksploracja zasobów internetowych**

**Kierunek Systemy teleinformatyczne, WE, II st., sem. II**

**Laboratorium nr 4**

**Temat: Przetwarzanie dokumentów tekstowych**

**Cel:**

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

1. **Zapoznaj się ze strukturą katalogu CCSU departments.**

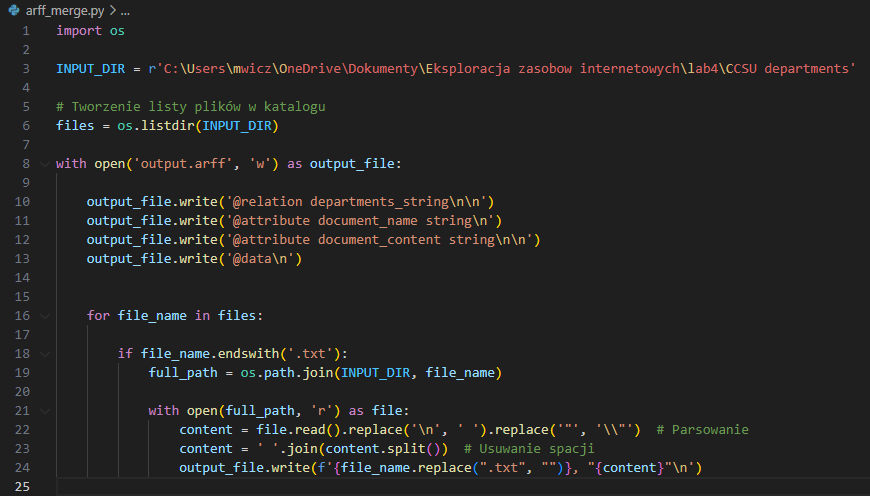
**Co zawiera ten katalog?**

*Katalog zawiera pliki tekstowe, z których każdy reprezentuje oddzielny dział lub wydział na uczelni CCSU. Każdy plik zawiera informacje dotyczące danego działu, w tym opisy programów studiów oferowanych przez ten dział, dane kontaktowe do kierownictwa wydziału oraz lokalizację i adres strony internetowej działu.*

**Jaka jest struktura zawartości dokumentów katalogu?**

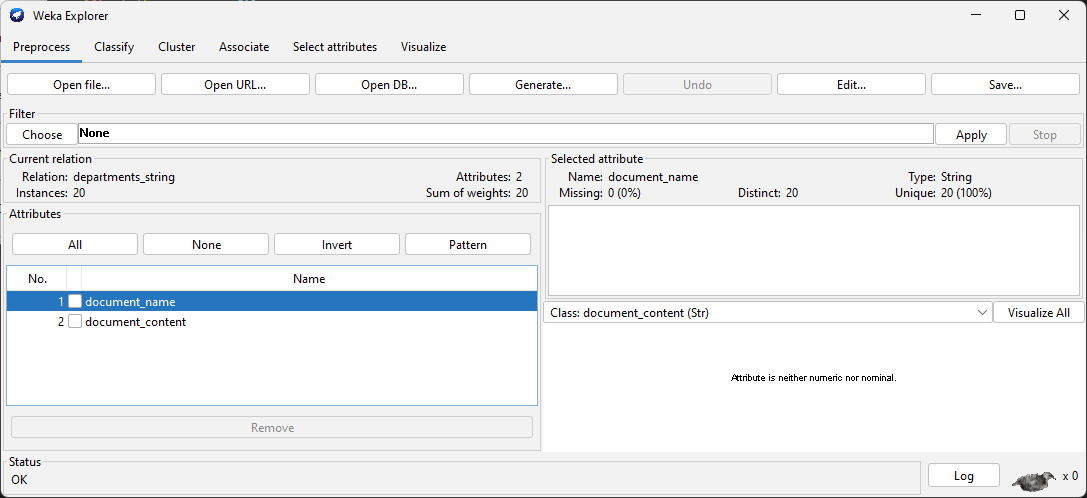
*Struktura zawartości dokumentów w katalogu "CCSU departments" jest podobna dla każdego pliku - rozpoczyna się od tytułu działu (np. "Art", "English"), a następnie zawiera szczegółowe opisy programów studiów, informacje o wydziale, kontakty do kierownictwa i lokalizacje.*

2.



1. Wczytaj plik Department-string\_bc.arff za pomocą **Open file** w zakładce **Preprocess**.

Program pokaże też kilka statystyk dokumentu.



1. 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**).

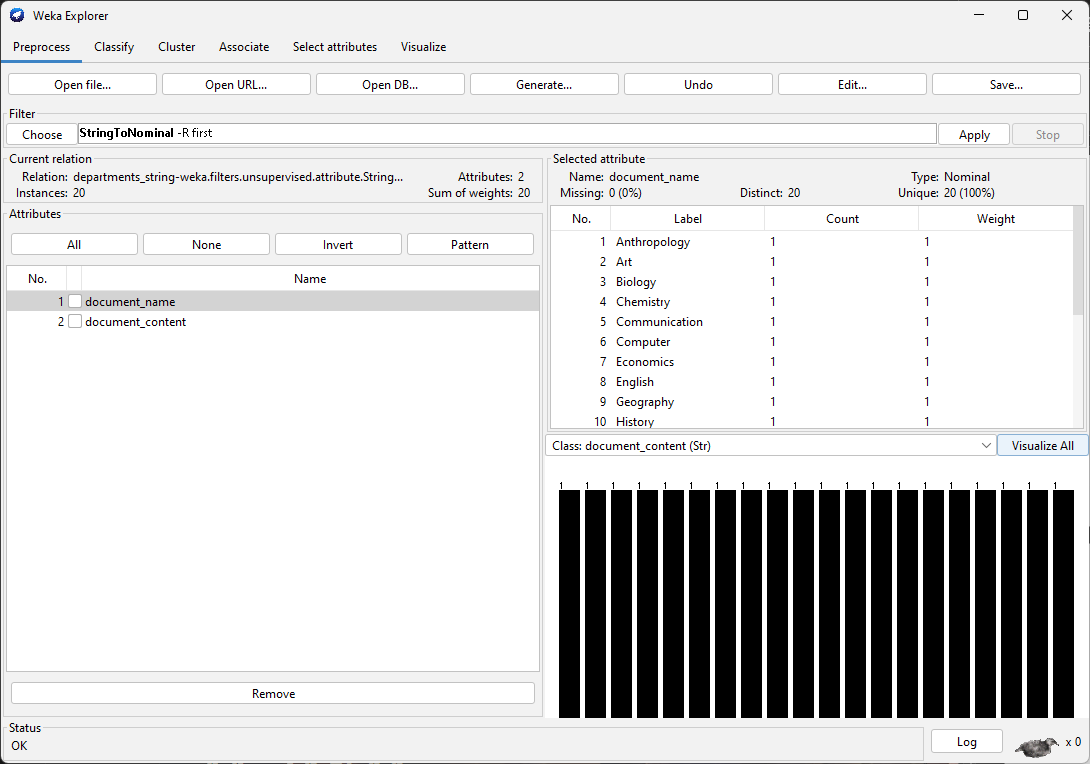
Zastosowanie filtra wymaga użycie przycisku **Apply**.

Co teraz uzyskałeś?

Jak wygląd twoja statystyka dotyczące dokumentu?

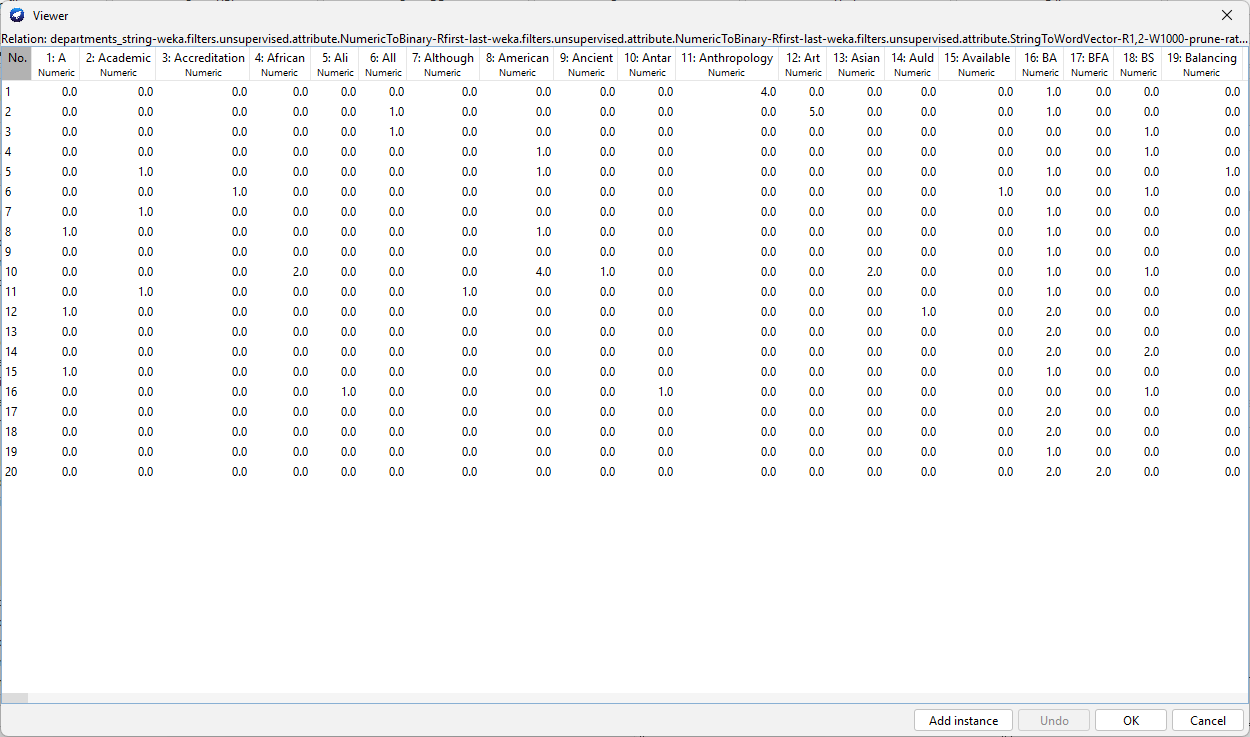
20 instancji – przedmioty na studiach

2 atrybuty – nazwa przedmiotu, zawartość(opis)



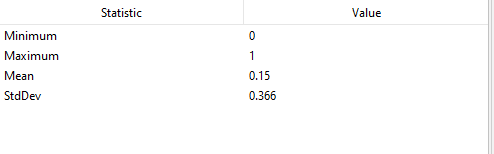
1. Wybierz filtr **StringToWordVector** z opcją **outputWord\_Counts=true** oraz z opcją **tokenizer=AlphabeticTokenizer**

Co otrzymałeś? Zobacz jak wygląd struktura danych wybierając Edit.



20instancji

771 atrybutow



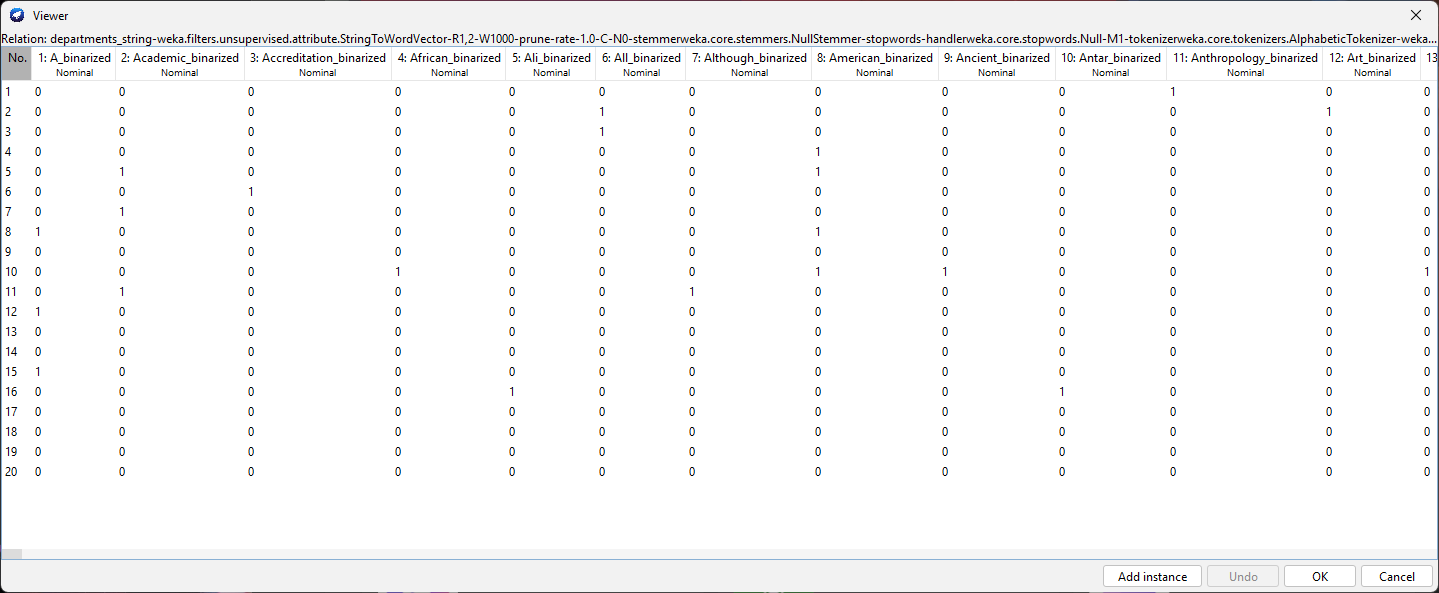
* Żadne z analizowanych słów nie występuje we wszystkich dokumentach
* W dokumentach najczęściej występujące słowo pojawia się maksymalnie raz w jednym dokumencie.
* To średnia liczba wystąpień słów w poszczególnych dokumentach. (Ponieważ Mean<1, średnio większość słów nie występuje we wszystkich dokumentach) ???
* Stopień zróżnicowania w liczbie wystąpień słów między poszczególnymi dokumentami. Srednia dużo mniejsza niż 1, max=1 a wiec odchylenie nie jest duże w kontekście min max.

Jest to macierz document-term. Pamiętasz co ona reprezentuje?

*Każdy wiersz reprezentuje dokument, a kazda kolumna wyrażenie które wystepuje w tych dokumentach.*

Zapoznaj się z wykresami rozkładów termów.

1. Użyj filtru NumericToBinary.



Co teraz możesz powiedzieć o uzyskanych statystykach (wykresach)?

Znajdź najbardziej specyficzne termy dla każdego z dokumentów. (identyfikacja słów, które pojawiają się tylko w danym dokumencie)

1. Możesz teraz usunąć zbędne termy?

Jakie termy usuniesz?

Skorzystalem z StringToWordVector oraz filtra stopwordHandler - Rainbow: Opcja ta usuwa standardowe stopwords, czyli często występujące słowa (np. "i", "oraz", "jest") itp.

Zobacz po usunięciu termów jak wygląd zawartość danych (Edit) i zapisz dane do pliku csv nadając mu nazwę Department\_terms.csv

B.

1. Wczytaj raz jeszcze plik Department-string\_bc.arff za pomocą **Open file** w zakładce

**Preprocess**.

1. 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**).
2. Wybierz filtr **StringToWordVector** z opcją **outputWord\_Counts=true** oraz z opcją **tokenizer=AlphabeticTokenizer** oraz z opcją **IDFTransform=true**
3. 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.

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

D.

1. 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.

1. Do sprawozdania dołącz opracowany sktypt.

E.

1. 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.

1. 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?