Metody eksploracji danych
Laboratorium 4

Klasyfikacja dokumentów tekstowych
Naiwny model Bayesa
Drzewa decyzyjne

Zbiory danych

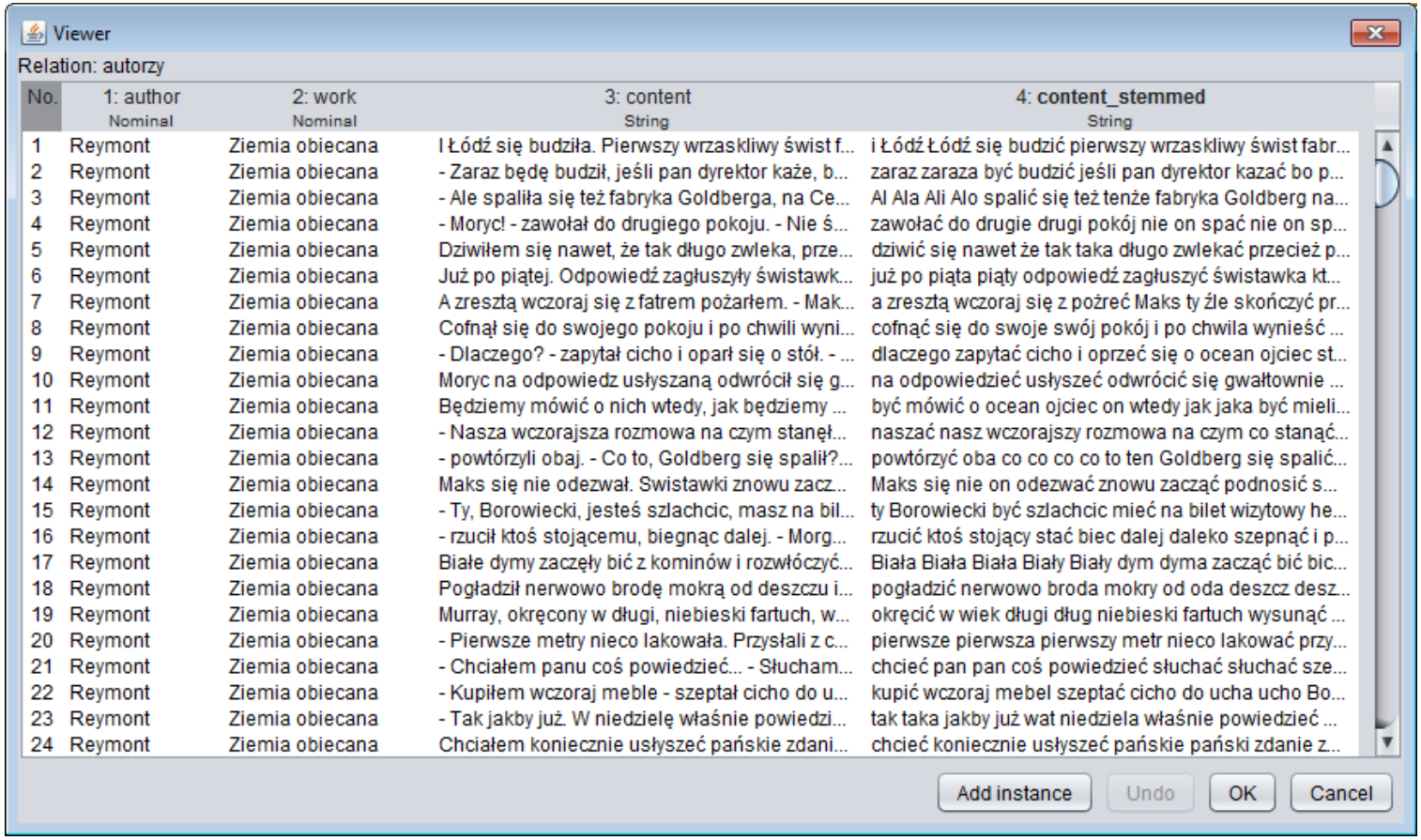
- Podczas ćwiczeń będziemy przetwarzali dane tekstowe pochodzące z 5 książek z przełomu XIX i XX wieku
  1. Reymont: Ziemia Obiecana
  2. Żuławski: Na srebrnym globie
  3. Sienkiewicz: W pustyni i w puszczy
  4. Sienkiewicz: Rodzina Połanieckich
  5. Żeromski: Syzyfowe prace
- Zawartość książek została podzielona na zdania i utworzono 8 zbiorów dokumentów:
  - złożonych z 10, 5, 3, 2 i 1 zdań
  - obejmujących treść wszystkich książek (five-books*.arff)
  - obejmujących treść pierwszych dwóch książek (two-books*.arff)
- Każdy element zbioru danych zawiera informacje o autorze, książce (work), treść (content) oraz formy podstawowe wyrazów, tzw. lematy: content_stemmed
- Zbiory są zapisane w frmacie UTF-8

Weka

- Aby prawidłowo interpretować zawartość plików UTF-8 należy:
  - uruchomić Weka za pomocą polecienia
    java -Dfile.encoding=utf-8 -jar weka.jar
  - lub zmienić zawartość pliku RunWeka.ini ustawiając:
    fileEncoding=utf-8
- Do przetworzenia pliku five-books-1000-1-stem.arff może być konieczne zwiększenie pamięci maszyny wirtualnej, np.:
  -Xmx2048M
- lub w pliku RunWeka.ini
  maxheap=2024M

4.1 Zbiór five-books-all-1000-10-stem.arff

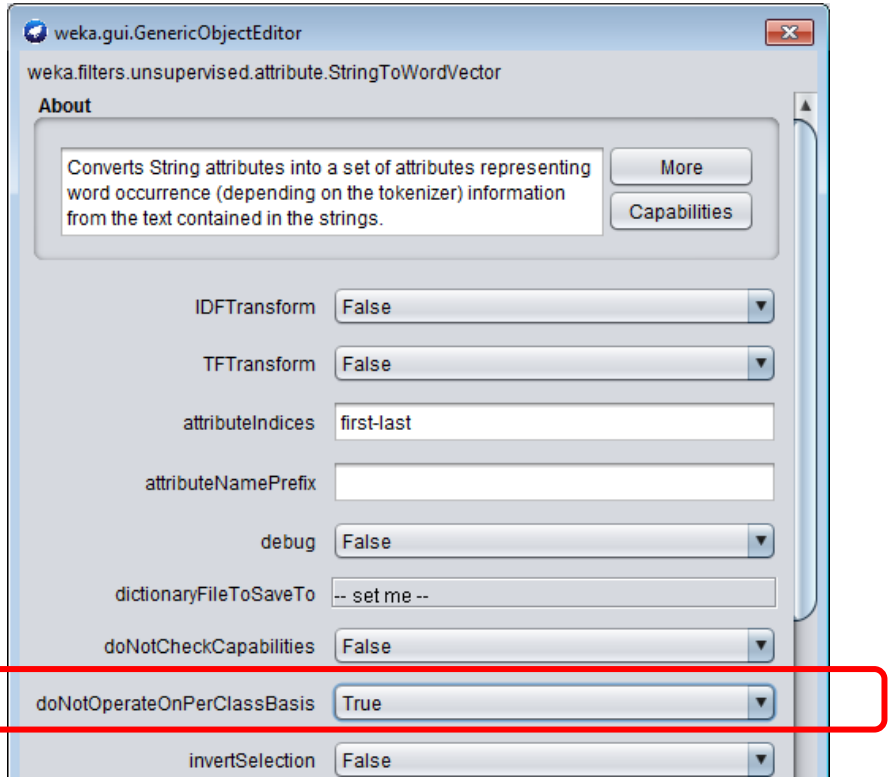
- W Weka Explorer załaduj plik five-books-all-1000-10-stem.arff
- Sprawdź, czy polskie teksty są prawidłowo wyświetlane



No.	1: author (Nominal)	2: work (Nominal)	3: content (String)	4: content_stemmed (String)
1	Reymont	Ziemia obiecana	I Łódź się budziła. Pierwszy wrzaskliwy świst f...	i Łódź Łódź się budzić pierwszy wrzaskliwy świst fabr...
2	Reymont	Ziemia obiecana	- Zaraz będę budził, jeśli pan dyrektor każe, b...	zaraz zaraza być budzić jeśli pan dyrektor kazać bo p...
3	Reymont	Ziemia obiecana	- Ale spaliła się też fabryka Goldberga, na Ce...	Al Ala Ali Alo spalić się też tenże fabryka Goldberg na...
4	Reymont	Ziemia obiecana	- Moryc! - zawołał do drugiego pokoju. - Nie ś...	zawołać do drugie drugi pokój nie on spać nie on sp...
5	Reymont	Ziemia obiecana	Dziwiłem się nawet, że tak długo zwleka, prze...	dziwić się nawet że tak taka długo zwlekać przecież p...
6	Reymont	Ziemia obiecana	Już po piątej. Odpowiedź zagłuszyły świstawk...	już po piąta piąty odpowiedź zagłuszyć świstawka kt...
7	Reymont	Ziemia obiecana	A zresztą wczoraj się z fatrem pożarłem. - Mak...	a zresztą wczoraj się z pożreć Maks ty źle skończyć pr...
8	Reymont	Ziemia obiecana	Cofnął się do swojego pokoju i po chwili wyni...	cofnąć się do swoje swój pokój i po chwila wynieść ...
9	Reymont	Ziemia obiecana	- Dlaczego? - zapytał cicho i oparł się o stół. - ...	dlaczego zapytać cicho i oprzeć się o ocean ojciec st...
10	Reymont	Ziemia obiecana	Moryc na odpowiedz usłyszaną odwrócił się g...	na odpowiedzieć usłyszeć odwrócić się gwałtownie ...
11	Reymont	Ziemia obiecana	Będziemy mówić o nich wtedy, jak będziemy ...	być mówić o ocean ojciec on wtedy jak jaka być mieli...
12	Reymont	Ziemia obiecana	- Nasza wczorajsza rozmowa na czym stanął...	naszać nasz wczorajszy rozmowa na czym co stanąć...
13	Reymont	Ziemia obiecana	- powtórzyli obaj. - Co to, Goldberg się spalił?...	powtórzyć oba co co co co to ten Goldberg się spalić...
14	Reymont	Ziemia obiecana	Maks się nie odezwał. Swistawki znowu zacz...	Maks się nie on odezwać znowu zacząć podnosić s...
15	Reymont	Ziemia obiecana	- Ty, Borowiecki, jesteś szlachcic, masz na bil...	ty Borowiecki być szlachcic mieć na bilet wizytowy he...
16	Reymont	Ziemia obiecana	- rzucił ktoś stojącemu, biegnąc dalej. - Morg...	rzucić ktoś stojący stać biec dalej daleko szepnąć i p...
17	Reymont	Ziemia obiecana	Białe dymy zaczęły bić z kominów i rozwłóczyć...	Biała Biała Biała Biały Biały dym dyma zacząć bić bic...
18	Reymont	Ziemia obiecana	Pogładził nerwowo brodę mokrą od deszczu i...	pogładzić nerwowo broda mokry od oda deszcz desz...
19	Reymont	Ziemia obiecana	Murray, okręcony w długi, niebieski fartuch, w...	okręcić w wiek długi dług niebieski fartuch wysunąć ...
20	Reymont	Ziemia obiecana	- Pierwsze metry nieco lakowała. Przysłali z c...	pierwsze pierwsza pierwszy metr nieco lakować przy...
21	Reymont	Ziemia obiecana	- Chciałem panu coś powiedzieć... - Słucham...	chcieć pan pan coś powiedzieć słuchać słuchać sze...
22	Reymont	Ziemia obiecana	- Kupiłem wczoraj meble - szeptał cicho do u...	kupić wczoraj mebel szeptać cicho do ucha ucho Bo...
23	Reymont	Ziemia obiecana	- Tak jakby już. W niedzielę właśnie powiedzi...	tak taka jakby już wat niedziela właśnie powiedzieć ...
24	Reymont	Ziemia obiecana	Chciałem koniecznie usłyszeć pańskie zdani...	chcieć koniecznie usłyszeć pańskie pański zdanie z...

Przetwarzanie wstępne

- Usuń atrybuty work i content_stemmed
- Wybierz filtr StringToWordVector jego działanie jest opisane na końcu **wykładu 4** natomiast idea zastosowania w tekście wykładu 5 (około slajdu 15)
- Zastosuj go dla atrybutu 2 zmieniając opcję doNotOperateOnPerClassBasis



Po zastosowaniu filtru powinno pojawić się ponad 1000 atrybutów numerycznych reprezentujących liczby wystąpień słów.

Klasyfikacja

- W zakładce Classify wybierz atrybut reprezentujacy klasę (author) i Naiwny model Bayesa.
- Dla zaoszczędzenia czasu wybierz 5-fold cross validation
- Przeprowadź klasyfikację i odczytaj wyniki
- Zinterpretuj wartości Confusion Matrix oraz Precision, Recall i F-measure

	TP Rate	FP Rate	Precision	Recall	F-Measure
	0,979	0,009	0,980	0,979	0,980
	0,924	0,006	0,933	0,924	0,929
	0,977	0,023	0,977	0,977	0,977
	0,896	0,013	0,884	0,896	0,890
Weighted Avg.	0,965	0,016	0,965	0,965	0,965

=== Confusion Matrix ===

a	b	c	d	<-- classified as
1337	13	6	9	a = Reymont
22	364	5	3	b = Żuławski
1	12	2201	39	c = Sienkiewicz
4	1	40	388	d = Żeromski

4.2 Drzewo decyzyjne

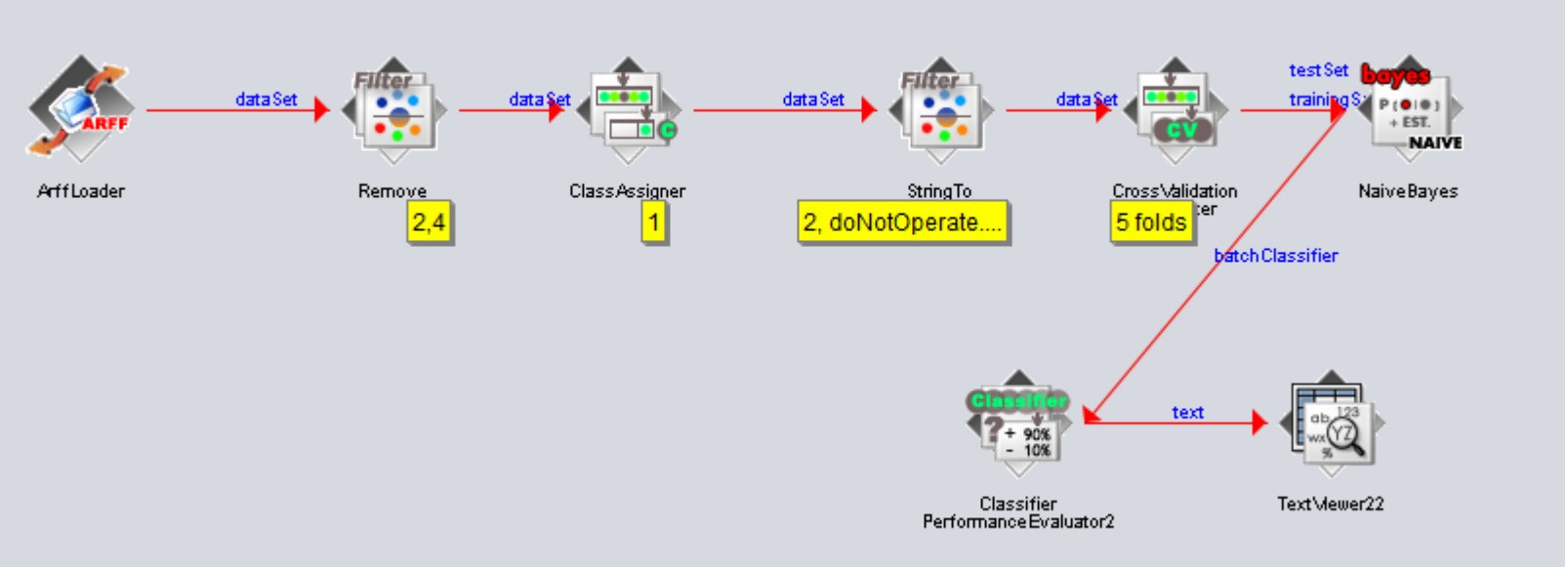
- Wybierz klasyfikator: J48 (drzewo decyzyjne). Drzewa decyzyjne są omówione w **wykładzie 5**.
- Przeprowadź klasyfikację i oceń rezultaty działania klasyfikatora (macierz pomyłek, precision, recall i F-measure) – patrz **wykład 4**
- Kliknij prawym klawiszem na rezultaty (trees J48) i wybierz opcję Visualize tree.
- W oknie użyj opcji Fit to screen i Auto Scale
- Oceń jakie atrybuty (słowa) zostały użyte, aby rozróżnić dokumenty będące fragmentami książek różnych autorów.
- Porównaj czasy wykonania klasyfikatorów NaiveBayes i J48

4.3 Ocena wykorzystania lematów

- Powtórnie otwórz plik five-books-all-1000-10-stem.arff
- Usuń atrybuty work i content
- Zastosuj filtr StringToWordVector dla content_stemmed
- Wybierz klasyfikator NaiveBayes
- Uruchom (5-fold cross vaidation) i porównaj wyniki z poprzednimi

4.4 Przetwarzanie zbiorów danych

- Zbuduj KnowledgeFlow, jak poniżej…



- i przetwarzaj kolejne zbiory danych:
  five-books-all-1000-n-stem.arff
  two-books-all-1000-n-stem.arff
  dla n=10,5,3,1…
- Zbierz wyniki w tabeli pokazującej zależność miar precision, recall i F-measure od n
- Sformułuj wnioski i przedstaw pomysły dotyczące potencjalnych zastosowań narzędzi klasyfikacji tekstów, np. do artykułów, komentarzy w Internecie, postów …