

Dominik Wróbel

Inżynieria oprogramowania i systemów

Informatyka, II stopień, 2018/19

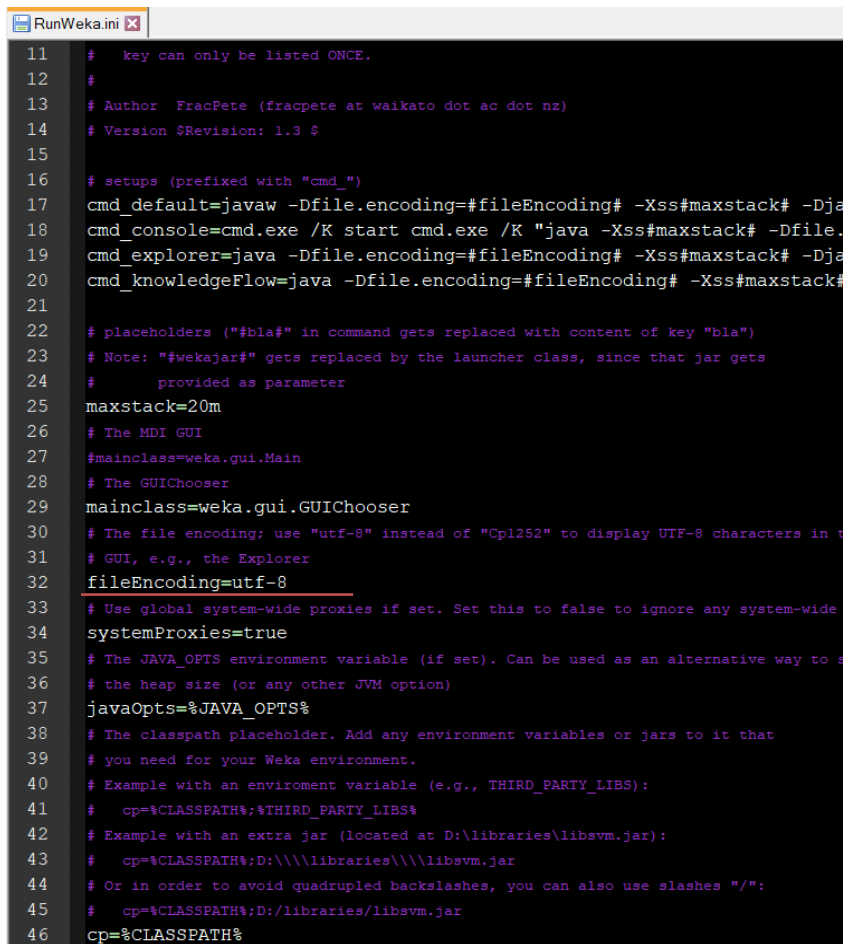
Metody eksploracji danych

Laboratorium 4 – 09.04.2019

Klasyfikacja dokumentów tekstowych, Naiwny model Bayesa, Drzewa decyzyjne

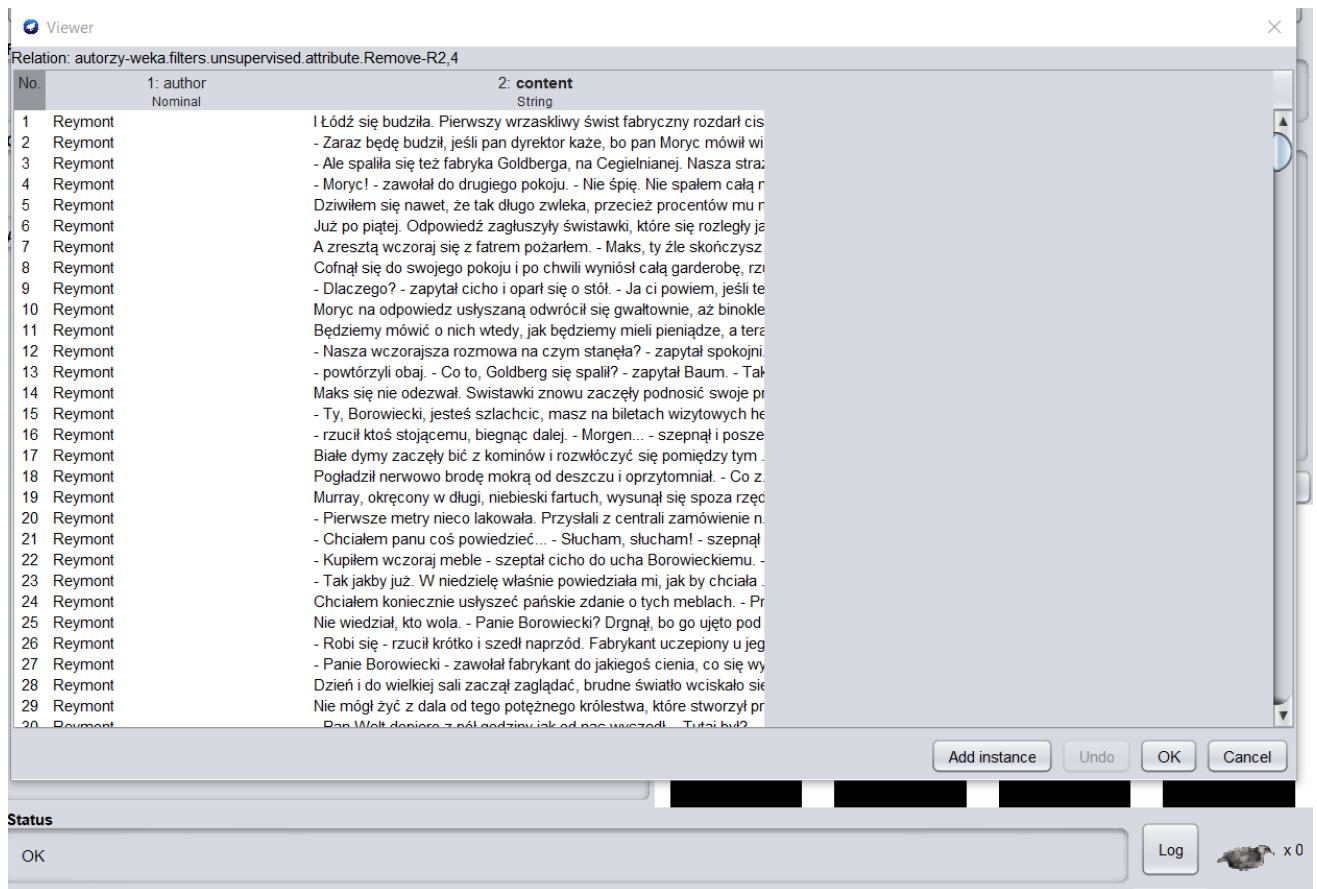
4.1 Zbiór *five-books-all-1000-10-stem.arff*

W pliku konfiguracyjnym RunWeka.ini ustawiono kodowanie plików na utf-8.



```
11 # key can only be listed ONCE.
12 #
13 # Author  FracPete (fracpete at waikato dot ac dot nz)
14 # Version $Revision: 1.3 $
15
16 # setups (prefixed with "cmd_")
17 cmd_default=javaw -Dfile.encoding=#fileEncoding# -Xss#maxstack# -Dja
18 cmd_console=cmd.exe /K start cmd.exe /K "java -Xss#maxstack# -Dfile.
19 cmd_explorer=java -Dfile.encoding=#fileEncoding# -Xss#maxstack# -Dja
20 cmd_knowledgeFlow=java -Dfile.encoding=#fileEncoding# -Xss#maxstack#
21
22 # placeholders ("#bla#" in command gets replaced with content of key "bla")
23 # Note: "#wekajar#" gets replaced by the launcher class, since that jar gets
24 # provided as parameter
25 maxstack=20m
26 # The MDI GUI
27 #mainclass=weka.gui.Main
28 # The GUIChooser
29 mainclass=weka.gui.GUIChooser
30 # The file encoding; use "utf-8" instead of "Cpl252" to display UTF-8 characters in t
31 # GUI, e.g., the Explorer
32 fileEncoding=utf-8
33 # Use global system-wide proxies if set. Set this to false to ignore any system-wide
34 systemProxies=true
35 # The JAVA_OPTS environment variable (if set). Can be used as an alternative way to s
36 # the heap size (or any other JVM option)
37 javaOpts=%JAVA_OPTS%
38 # The classpath placeholder. Add any environment variables or jars to it that
39 # you need for your Weka environment.
40 # Example with an enviroment variable (e.g., THIRD_PARTY_LIBS):
41 # cp=%CLASSPATH%;%THIRD_PARTY_LIBS%
42 # Example with an extra jar (located at D:\\libraries\\libsvm.jar):
43 # cp=%CLASSPATH%;D:\\\\libraries\\\\libsvm.jar
44 # Or in order to avoid quadrupled backslashes, you can also use slashes "/":
45 # cp=%CLASSPATH%;D:/libraries/libsvm.jar
46 cp=%CLASSPATH%
```

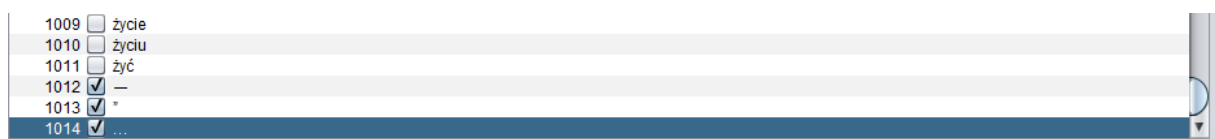
Z pliku usunięto atrybuty work i content_stemmed i wyświetlono otrzymane dane.



| No. | 1: author Nominal | 2: content String |
|-----|----------------------|--|
| 1 | Reymont | I Łódź się budziła. Pierwszy wrzaskliwy świst fabryczny rozdarł cis |
| 2 | Reymont | - Zaraz będę budził, jeśli pan dyrektor każe, bo pan Moryc mówił wi |
| 3 | Reymont | - Ale spaliła się też fabryka Goldberga, na Cegielnianej. Nasza straz |
| 4 | Reymont | - Moryc! - zawołał do drugiego pokoju. - Nie śpię. Nie spałem całą n |
| 5 | Reymont | Dziwiłem się nawet, że tak długo zwleka, przecież procentów mu r |
| 6 | Reymont | Już po piątej. Odpowiedź zagłuszyły świstawki, które się rozległy ja |
| 7 | Reymont | A zresztą wczoraj się z fatrem pożarłem. - Maks, ty źle skończysz |
| 8 | Reymont | Cofnął się do swojego pokoju i po chwili wyniósł całą garderobę, rzi |
| 9 | Reymont | - Dlaczego? - zapytał cicho i oparł się o stół. - Ja ci powiem, jeśli te |
| 10 | Reymont | Moryc na odpowiedź usłyszaną odwrócił się gwałtownie, aż binokle |
| 11 | Reymont | Będziemy mówić o nich wtedy, jak będziemy mieli pieniądze, a tera |
| 12 | Reymont | - Nasza wczorajsza rozmowa na czym stała? - zapytał spokojni |
| 13 | Reymont | - powtórzyli obaj. - Co to, Goldberg się spalił? - zapytał Baum. - Tak |
| 14 | Reymont | Maks się nie odezwał. Świstawki znowu zaczęły podnosić swoje pr |
| 15 | Reymont | - Ty, Borowiecki, jesteś szlachcic, masz na biletach wizytowych he |
| 16 | Reymont | - rzucił ktoś stojącemu, biegnąc dalej. - Morgen... - szepnął i posze |
| 17 | Reymont | Białe dymy zaczęły bić z kominów i rozwłóczyć się pomiędzy tym |
| 18 | Reymont | Pogładził nerwowo brodę mokrą od deszczu i oprzytomniał. - Co z |
| 19 | Reymont | Murray, okręcony w długi, niebieski fartuch, wysunął się spoza rzęd |
| 20 | Reymont | - Pierwsze metry nieco lakowała. Przysłali z centrali zamówienie n |
| 21 | Reymont | - Chciałem panu coś powiedzieć... - Słucham, słucham! - szepnął |
| 22 | Reymont | - Kupiłem wczoraj meble - szeptał cicho do ucha Borowieckiemu. - |
| 23 | Reymont | - Tak jakby już. W niedzielę właśnie powiedziała mi, jak by chciała |
| 24 | Reymont | Chciałem koniecznie usłyszeć pańskie zdanie o tych meblach. - Pr |
| 25 | Reymont | Nie wiedział, kto wola. - Panie Borowiecki? Drgnął, bo go ujęto pod |
| 26 | Reymont | - Robi się - rzucił krótko i szedł naprzód. Fabrykant uczipiony u jeg |
| 27 | Reymont | - Panie Borowiecki - zawołał fabrykant do jakiegoś cienia, co się wy |
| 28 | Reymont | Dzień i do wielkiej sali zaczął zaglądać, brudne światło wciskało się |
| 29 | Reymont | Nie mógł żyć z dala od tego potężnego królestwa, które stworzył pr |
| 30 | Reymont | Pan Wolt dopiero z pół godziny jak od nas uwodził... Tutaj był? |

Następnie dla atrybutu content zastosowano filtr StringToWordVector z odpowiednio ustawionymi opcjami. W wyniku otrzymano atrybuty numeryczne reprezentujące częstość występowania określonych słów w zdaniach.

Z atrybutów usunięto te, które nie są słowami, takie jak pauzy, kropki itp.



| | | |
|------|-------------------------------------|-------|
| 1009 | <input type="checkbox"/> | życie |
| 1010 | <input type="checkbox"/> | życiu |
| 1011 | <input type="checkbox"/> | żyć |
| 1012 | <input checked="" type="checkbox"/> | — |
| 1013 | <input checked="" type="checkbox"/> | * |
| 1014 | <input checked="" type="checkbox"/> | ... |

Po podziale zdań, dla atrybutu 'author' zastosowano naiwny klasyfikator Bayesa. W wyniku otrzymano parametry:

Time taken to build model: 1.45 seconds

=== Stratified cross-validation ===
=== Summary ===

| | | |
|----------------------------------|-----------|-----------|
| Correctly Classified Instances | 4145 | 93.2508 % |
| Incorrectly Classified Instances | 300 | 6.7492 % |
| Kappa statistic | 0.8932 | |
| Mean absolute error | 0.0342 | |
| Root mean squared error | 0.1683 | |
| Relative absolute error | 10.8281 % | |
| Root relative squared error | 42.3696 % | |
| Total Number of Instances | 4445 | |

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|-------|----------|----------|-------------|
| | 0,955 | 0,044 | 0,905 | 0,955 | 0,929 | 0,897 | 0,993 | 0,984 | Reymont |
| | 0,909 | 0,010 | 0,895 | 0,909 | 0,902 | 0,892 | 0,995 | 0,967 | Żuławski |
| | 0,950 | 0,032 | 0,968 | 0,950 | 0,959 | 0,918 | 0,994 | 0,994 | Sienkiewicz |
| | 0,792 | 0,012 | 0,873 | 0,792 | 0,831 | 0,814 | 0,985 | 0,915 | Żeromski |
| Weighted Avg. | 0,933 | 0,032 | 0,933 | 0,933 | 0,932 | 0,899 | 0,993 | 0,981 | |

=== Confusion Matrix ===

| a | b | c | d | <-- classified as |
|------|-----|------|-----|-------------------|
| 1303 | 11 | 30 | 21 | a = Reymont |
| 22 | 358 | 11 | 3 | b = Żuławski |
| 60 | 26 | 2141 | 26 | c = Sienkiewicz |
| 55 | 5 | 30 | 343 | d = Żeromski |

W **macierzy pomyłek** największe wartości znajdują się na przekątnej, jest to liczba poprawnie sklasyfikowanych słów dla danego autora. Pozostałe liczby w wierszu oznaczają niepoprawną klasyfikację i przypisanie atrybutu do innego autora. Najwięcej pomyłek w klasyfikacji zaszło pomiędzy Sienkiewiczem, a Reymontem (Sienkiewicz został uznany za Reymonta 60 razy, a Reymont za Sienkiewicza 30 razy).

Wskaźnik recall to stosunek liczby poprawie zakwalifikowanych słów do wszystkich słów dla danego autora. Wskaźnik ten jest najniższy dla autorów o najmniejszej liczbie atrybutów (Żuławski, Żeromski). Dla tych autorów znaleziono najmniej słów.

Wskaźnik precision informuje w ilu procentach przypadków klasyfikator zwrócił poprawny wynik. Ponownie wskaźnik ten jest najniższy dla autorów o najmniejszej liczbie słów w danych. Najlepszą wartość wskaźnika uzyskał Sienkiewicz, dla którego liczba słów wejściowych jest największa.

Wskaźnik F-Measure to średnia harmoniczna wskaźników recall oraz precision. Wielkość ta niesie informację o jakości klasyfikacji, jest ona bliższa 1 dla bardziej poprawnych klasyfikacji. Z danych wynika, że najbardziej poprawną klasyfikację uzyskano dla Sienkiewicza (wskaźniki recall i precision zwracają zbliżoną wartość, co świadczy o stosunkowo poprawnym działaniu klasyfikatora), a najniższą dla Żeromskiego.

4.2 Drzewo decyzyjne

W tym zadaniu zastosowano klasyfikację przy użyciu drzew decyzyjnych, ponownie użyto pliku five-books-all-1000-10-stem.arff po przetworzeniu. Zastosowano klasyfikator J48. W wyniku działania klasyfikatora otrzymano dane:

Time taken to build model: 46.27 seconds

=== Stratified cross-validation ===
=== Summary ===

| | | |
|----------------------------------|-----------|-----------|
| Correctly Classified Instances | 3864 | 86.9291 % |
| Incorrectly Classified Instances | 581 | 13.0709 % |
| Kappa statistic | 0.7932 | |
| Mean absolute error | 0.0685 | |
| Root mean squared error | 0.2425 | |
| Relative absolute error | 21.6961 % | |
| Root relative squared error | 61.0221 % | |
| Total Number of Instances | 4445 | |

=== Detailed Accuracy By Class ===

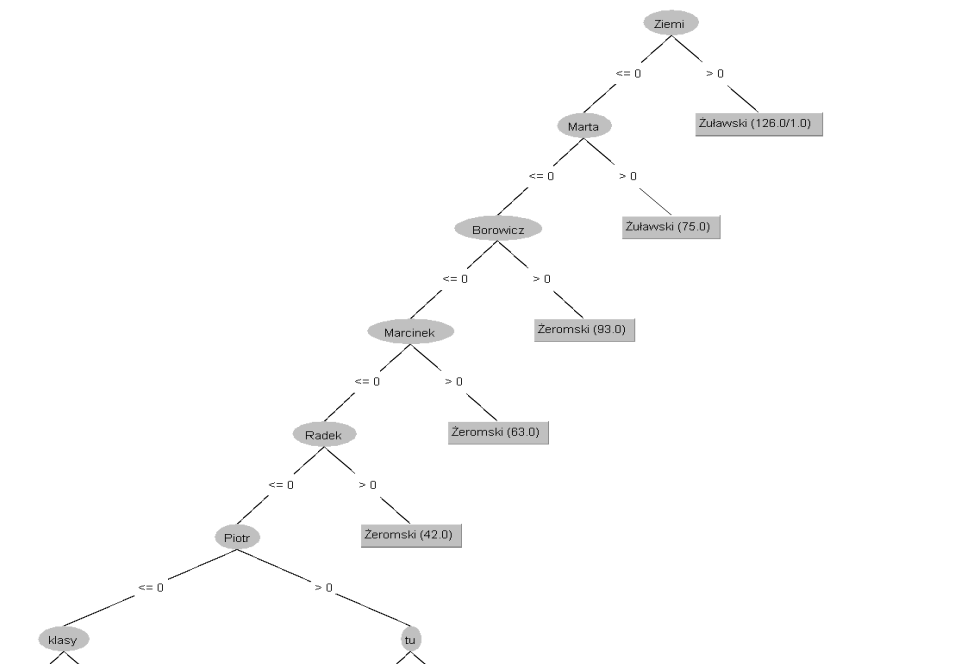
| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|-------|----------|----------|-------------|
| | 0,887 | 0,072 | 0,844 | 0,887 | 0,865 | 0,804 | 0,932 | 0,861 | Reymont |
| | 0,744 | 0,022 | 0,767 | 0,744 | 0,755 | 0,732 | 0,905 | 0,665 | Żuławski |
| | 0,918 | 0,064 | 0,936 | 0,918 | 0,927 | 0,854 | 0,947 | 0,937 | Sienkiewicz |
| | 0,672 | 0,032 | 0,695 | 0,672 | 0,683 | 0,650 | 0,881 | 0,637 | Żeromski |
| Weighted Avg. | 0,869 | 0,060 | 0,869 | 0,869 | 0,869 | 0,808 | 0,932 | 0,860 | |

=== Confusion Matrix ===

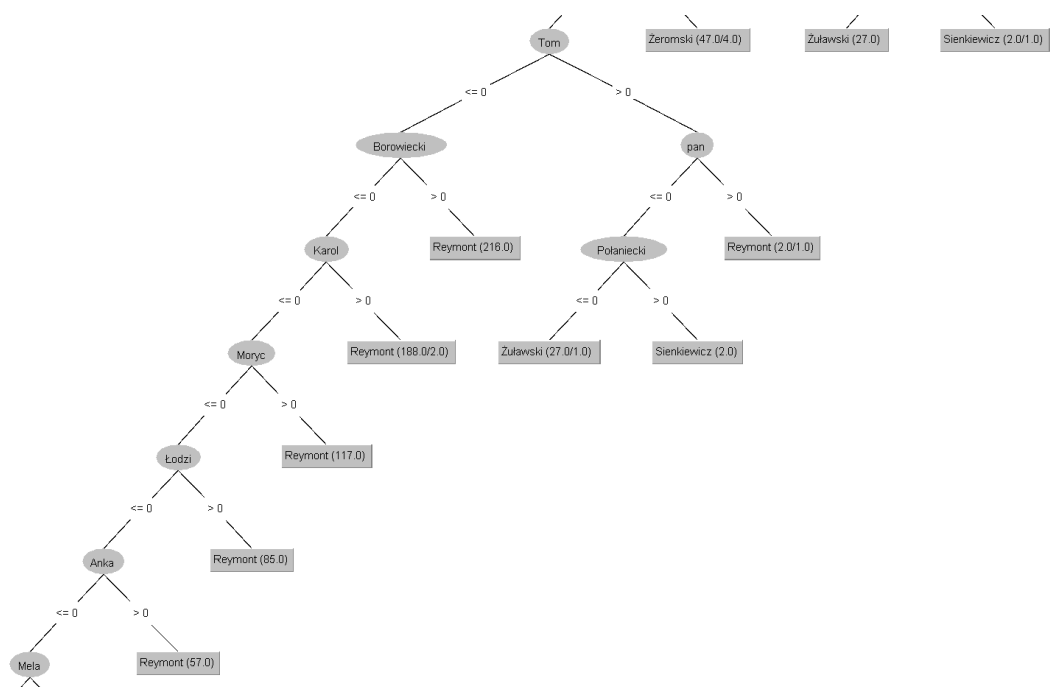
| a | b | c | d | <-- classified as |
|------|-----|------|-----|-------------------|
| 1211 | 38 | 64 | 52 | a = Reymont |
| 62 | 293 | 22 | 17 | b = Żuławski |
| 86 | 39 | 2069 | 59 | c = Sienkiewicz |
| 75 | 12 | 55 | 291 | d = Żeromski |

Po zastosowaniu drzewa decyzyjnego ponownie najniższe wskaźniki uzyskano dla Żuławskiego oraz Żeromskiego. Klasyfikacja dała nieco gorsze wyniki niż w przypadku zastosowaniu Naive Bayes, średnie wartości wszystkich wskaźników są niższe od tych wygenerowanych w zadaniu poprzednim. Ponownie najlepsze wartości wskaźników udało się uzyskać dla Sienkiewicza (największa liczba atrybutów).

Początkowy fragment drzewa:



Ciąg dalszy drzewa:



W wygenerowanym drzewie charakterystyczne jest podejmowania decyzji o klasyfikacji **na podstawie słów będących nazwami własnymi (najczęściej imiona, nazwiska, nazwy miast)**. Nie jest to zaskakujące, ponieważ takie nazwy są zazwyczaj różne dla różnych powieści.

Czas wykonania klasyfikacji jest znacznie dłuższy w przypadku zastosowania drzew decyzyjnych niż w przypadku użycia Naive Bayes. Wynika to ze złożoności obliczeniowej stosowanych algorytmów i operacji, które wykorzystują. Naive Bayes jest algorytmem prostszym, zawiera mniej obliczeń.

4.3 Ocena wykorzystania lematów

W tym zadaniu ponownie zastosowano naiwny klasyfikator Bayesa dla pliku five-books-all-1000-10-stem.arff, tym razem z pliku usunięto jednak parametr content, a pozostawiono content_stemmed. Wyniki uzyskane w tym podejściu to

```
Time taken to build model: 1.59 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      4172           93.8583 %
Incorrectly Classified Instances    273           6.1417 %
Kappa statistic                    0.9033
Mean absolute error                 0.0308
Root mean squared error             0.1651
Relative absolute error             9.7416 %
Root relative squared error         41.5457 %
Total Number of Instances          4445

=== Detailed Accuracy By Class ===
```

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|-------|----------|----------|-------------|
| | 0,955 | 0,035 | 0,923 | 0,955 | 0,939 | 0,911 | 0,994 | 0,987 | Reymont |
| | 0,924 | 0,014 | 0,863 | 0,924 | 0,892 | 0,882 | 0,996 | 0,965 | Żuławski |
| | 0,943 | 0,028 | 0,972 | 0,943 | 0,957 | 0,915 | 0,993 | 0,994 | Sienkiewicz |
| | 0,875 | 0,011 | 0,896 | 0,875 | 0,886 | 0,873 | 0,991 | 0,951 | Żeromski |
| Weighted Avg. | 0,939 | 0,028 | 0,940 | 0,939 | 0,939 | 0,907 | 0,994 | 0,985 | |

```

=== Confusion Matrix ===

  a   b   c   d  <-- classified as
1304  11  34  16 |  a = Reymont
  23 364   7   0 |  b = Żuławski
  56  44 2125  28 |  c = Sienkiewicz
  30   3   21 379 |  d = Żeromski

```

Dla tego samego pliku dla atrybutu content przy zastosowaniu klasyfikatora NaiveBayes otrzymano:

| Precision | Recall | F-Measure |
|-----------|--------|-----------|
| 0,905 | 0,955 | 0,929 |
| 0,895 | 0,909 | 0,902 |
| 0,968 | 0,950 | 0,959 |
| 0,873 | 0,792 | 0,831 |
| 0,933 | 0,933 | 0,932 |

Porównując wyniki, widoczne jest, że użycie danych w bardziej ogólnej formie przyniosło poprawę wyników, tylko w czterech przypadkach uzyskano mniejszą wartość wskaźnika, w pozostałych jest ona większa.

4.4 Przetwarzanie zbiorów danych

W tym zadaniu porównano działanie algorytmu klasyfikacji dla różnych plików wejściowych.

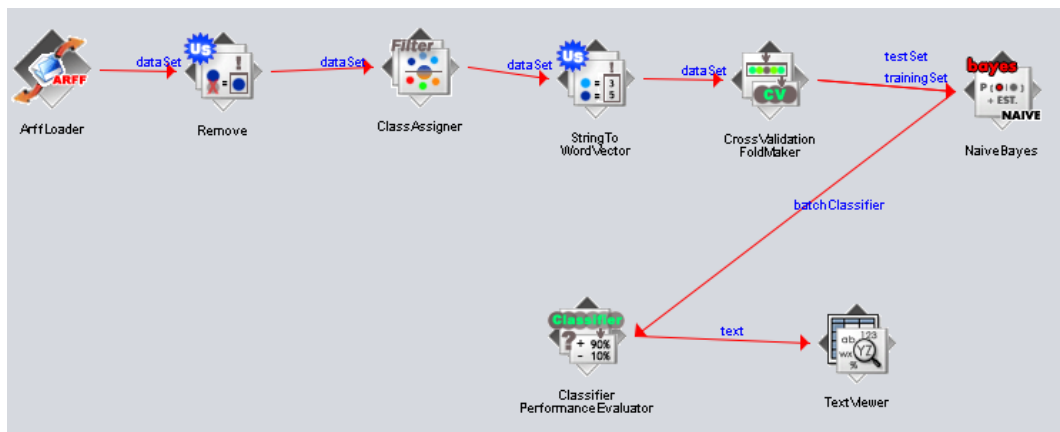
Autorzy od góry to: Reymont, Żuławski, Sienkiewicz, Żeromski.

| | |
|--|--|
| five-books-all-1000-10-stem.arff Precision Recall F-Measure 0,980 0,979 0,980 0,933 0,924 0,929 0,977 0,977 0,977 0,884 0,896 0,890 0,965 0,965 0,965 | two-books-all-1000-10-stem.arff Precision Recall F-Measure 0,986 0,990 0,988 0,966 0,952 0,959 <u>0,982 0,982 0,982</u> |
| five-books-all-1000-5-stem.arff Precision Recall F-Measure 0,949 0,906 0,927 0,737 0,823 0,778 0,939 0,944 0,941 0,730 0,735 0,733 0,904 0,901 0,902 | two-books-all-1000-5-stem.arff Precision Recall F-Measure 0,966 0,944 0,955 0,819 0,884 0,850 0,933 0,930 0,931 |
| five-books-all-1000-3-stem.arff Precision Recall F-Measure 0,902 0,812 0,855 0,584 0,721 0,645 0,882 0,904 0,893 0,570 0,554 0,562 0,832 0,826 0,827 | two-books-all-1000-3-stem.arff Precision Recall F-Measure 0,940 0,878 0,908 0,655 0,805 0,722 0,876 0,861 0,866 |
| five-books-all-1000-1-stem.arff Precision Recall F-Measure 0,768 0,613 0,682 0,352 0,354 0,353 0,689 0,883 0,774 0,383 0,065 0,111 <u>0,654 0,674 0,644</u> | two-books-all-1000-1-stem.arff Precision Recall F-Measure 0,855 0,850 0,852 0,490 0,498 0,494 0,773 0,772 0,772 |

W tabeli widoczny jest **spadek jakości klasyfikacji wraz ze spadkiem liczby zdań** zarówno dla dwóch jak i dla pięciu książek. Lepsze wyniki klasyfikacji udało się uzyskać gdy korzystano z dwóch książek niż z pięciu, klasyfikator jest w stanie wówczas podjąć bardziej trafną decyzję ze względu na mniejszą ilość danych. Najlepsze wskaźniki udało uzyskać się dla dwóch książek przy 10 zdaniach.

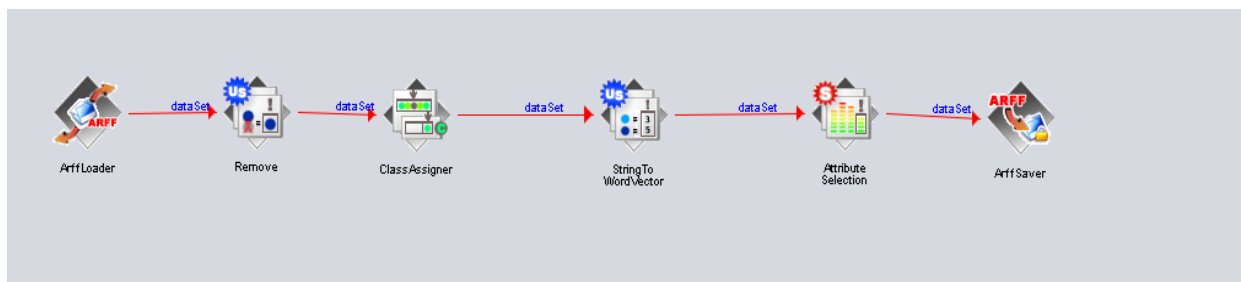
Narzędzia klasyfikacji tekstów mają duży potencjał, mogłyby być wykorzystywane do odgadywania autorów nieznanych tekstów na podstawie innych dzieł danego autora czy też do klasyfikacji zawartości stron przez wyszukiwarki.

Dane zostały wygenerowane na postawie KnowledgeFlow:

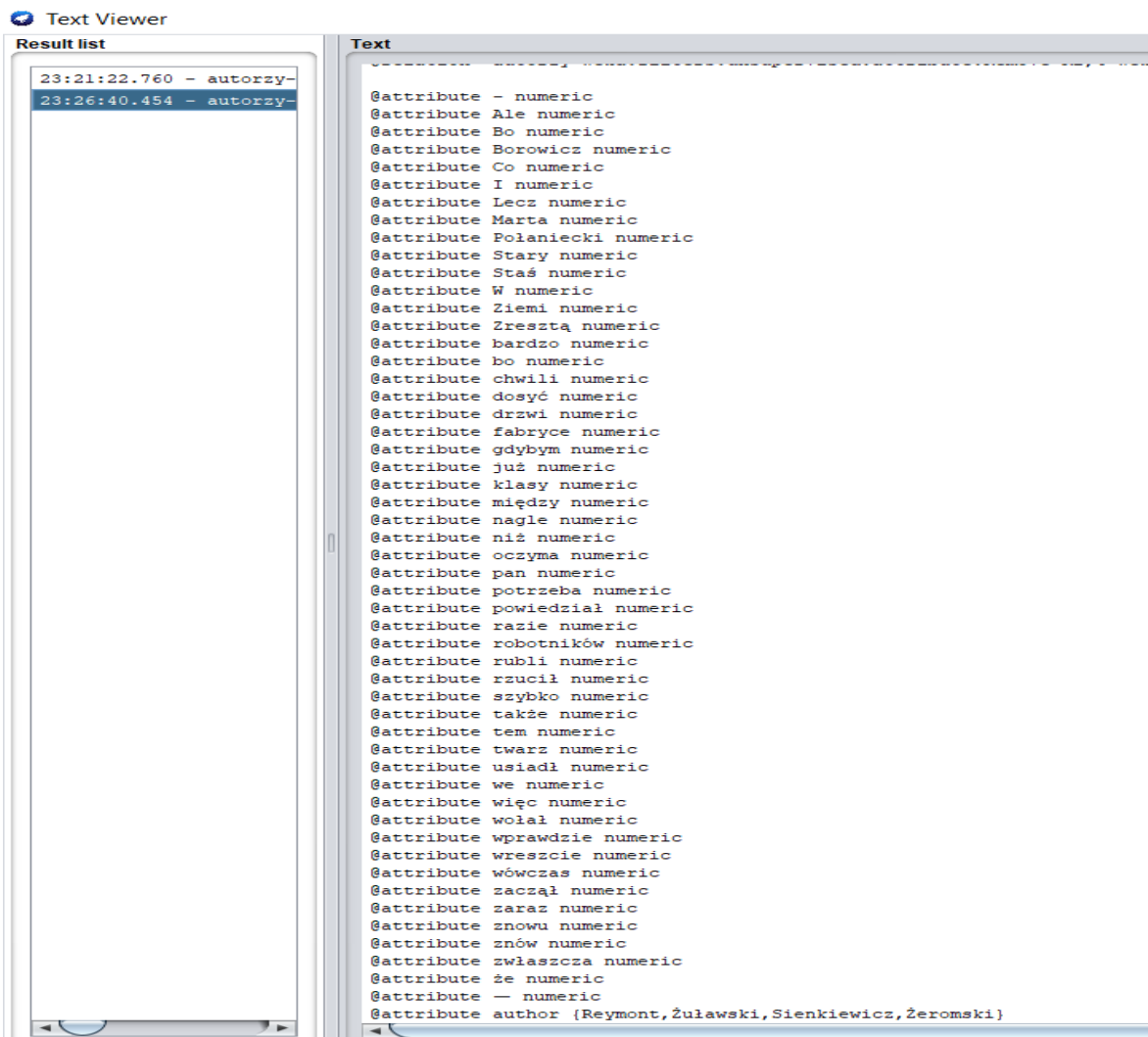


4.5 Redukcja liczby atrybutów

W tym zadaniu zredukowano liczbę atrybutów dla klasyfikatora w celu porównania wyników dla zbioru zredukowanego i wyjściowego. Do porównania wybrano plik *five-books-all-1000-10-stem.arff*. Redukcję wykonano przy użyciu Knowledge Flow:



W wyniku działania otrzymano plik z pozostawionymi słowami:



Pozostawione słowa to często nazwy własne występujące w tekście (te, które są charakterystyczne dla danego tekstu).

4.6 Porównaj wyniki dla NB i J48

W tym zadaniu porównano wyniki uzyskane dla danych po selekcji z zastosowaniem klasyfikatorów Naive Bayes oraz J48 i dane bez selekcji atrybutów.

Naive Bayes z selekcją:

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|-------|----------|----------|-------------|
| | 0,940 | 0,019 | 0,957 | 0,940 | 0,948 | 0,926 | 0,996 | 0,991 | Reymont |
| | 0,805 | 0,026 | 0,749 | 0,805 | 0,776 | 0,754 | 0,986 | 0,898 | Żuławski |
| | 0,953 | 0,041 | 0,960 | 0,953 | 0,957 | 0,913 | 0,994 | 0,994 | Sienkiewicz |
| | 0,774 | 0,027 | 0,755 | 0,774 | 0,764 | 0,738 | 0,980 | 0,867 | Żeromski |
| Weighted Avg. | 0,919 | 0,031 | 0,920 | 0,919 | 0,919 | 0,886 | 0,992 | 0,972 | |

=== Confusion Matrix ===

```
  a    b    c    d  <-- classified as
1283  65    0   17 |  a = Reymont
  56 317    6   15 |  b = Żuławski
   1  27 2148   77 |  c = Sienkiewicz
   1  14   83 335 |  d = Żeromski
```

Naive Bayes bez selekcji:

=== Summary ===

| | | |
|----------------------------------|-----------|-----------|
| Correctly Classified Instances | 4290 | 96.5129 % |
| Incorrectly Classified Instances | 155 | 3.4871 % |
| Kappa statistic | 0.9448 | |
| Mean absolute error | 0.0181 | |
| Root mean squared error | 0.1221 | |
| Relative absolute error | 5.7339 % | |
| Root relative squared error | 30.7404 % | |
| Total Number of Instances | 4445 | |

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|-------|----------|----------|-------------|
| | 0,979 | 0,009 | 0,980 | 0,979 | 0,980 | 0,971 | 0,999 | 0,998 | Reymont |
| | 0,924 | 0,006 | 0,933 | 0,924 | 0,929 | 0,922 | 0,998 | 0,981 | Żuławski |
| | 0,977 | 0,023 | 0,977 | 0,977 | 0,977 | 0,954 | 0,998 | 0,998 | Sienkiewicz |
| | 0,896 | 0,013 | 0,884 | 0,896 | 0,890 | 0,878 | 0,994 | 0,960 | Żeromski |
| Weighted Avg. | 0,965 | 0,016 | 0,965 | 0,965 | 0,965 | 0,949 | 0,998 | 0,993 | |

=== 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
```

W przypadku klasyfikatora NaiveBayes nie udało się uzyskać poprawy po zastosowaniu redukcji atrybutów. Wynika stąd, że redukcja atrybutów nie zawsze wpływa korzystnie na jakość klasyfikacji.

J48 z selekcją:

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|-------|----------|----------|-------------|
| | 0,964 | 0,042 | 0,910 | 0,964 | 0,936 | 0,908 | 0,983 | 0,949 | Reymont |
| | 0,645 | 0,015 | 0,806 | 0,645 | 0,717 | 0,697 | 0,958 | 0,782 | Żuławski |
| | 0,968 | 0,071 | 0,933 | 0,968 | 0,951 | 0,898 | 0,976 | 0,973 | Sienkiewicz |
| | 0,630 | 0,018 | 0,789 | 0,630 | 0,701 | 0,678 | 0,934 | 0,727 | Żeromski |
| Weighted Avg. | 0,906 | 0,052 | 0,901 | 0,906 | 0,901 | 0,862 | 0,973 | 0,925 | |

=== Confusion Matrix ===

```
  a    b    c    d  <-- classified as
1316  38    0   11 |   a = Reymont
 114 254   15   11 |   b = Żuławski
   8  12 2182   51 |   c = Sienkiewicz
   8  11  141  273 |   d = Żeromski
```

J48 bez selekcji:

Time taken to build model: 26.19 seconds

=== Stratified cross-validation ===

=== Summary ===

| | | |
|----------------------------------|-----------|-----------|
| Correctly Classified Instances | 4006 | 90.1237 % |
| Incorrectly Classified Instances | 439 | 9.8763 % |
| Kappa statistic | 0.8411 | |
| Mean absolute error | 0.0586 | |
| Root mean squared error | 0.2138 | |
| Relative absolute error | 18.5471 % | |
| Root relative squared error | 53.804 % | |
| Total Number of Instances | 4445 | |

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|-------|----------|----------|-------------|
| | 0,954 | 0,038 | 0,917 | 0,954 | 0,935 | 0,906 | 0,965 | 0,893 | Reymont |
| | 0,701 | 0,014 | 0,829 | 0,701 | 0,759 | 0,741 | 0,921 | 0,708 | Żuławski |
| | 0,957 | 0,081 | 0,924 | 0,957 | 0,940 | 0,877 | 0,956 | 0,940 | Sienkiewicz |
| | 0,628 | 0,022 | 0,758 | 0,628 | 0,687 | 0,660 | 0,886 | 0,618 | Żeromski |
| Weighted Avg. | 0,901 | 0,056 | 0,897 | 0,901 | 0,898 | 0,853 | 0,949 | 0,874 | |

=== Confusion Matrix ===

```
  a    b    c    d  <-- classified as
1302  42   11   10 |   a = Reymont
 104 276   12    2 |   b = Żuławski
  10  12 2156   75 |   c = Sienkiewicz
   4   3  154  272 |   d = Żeromski
```

W przypadku drzew decyzyjnych udało się uzyskać poprawę po zastosowaniu redukcji atrybutów. Redukcja ta miała również nieznaczny wpływ na skrócenie czasu działania algorytmu, co może mieć duże znaczenie przy dużych rozmiarach danych do przetworzenia.