Laboratorium <i>ADiO</i>									
Rok akademicki	Termin	Rodzaj studiów	Kierunek	Prowadzący	Sekcja				
2019/2020	czwartek 8:30-10:45	SSI	INF	dr inż. Łukasz Wróbel	2				

Sprawozdanie z ćwiczenia numer 1

Data wykonania ćwiczenia: 2019-10-24

Data oddania sprawozdania: 2019-11-06

Temat:

Wstępne przetwarzanie danych

Autor:

Bartłomiej Krasoń

Synonim:

CIĘ świta

DANE

Wybrałem zespoły akcent i boys. Wygenerowany plik TekstyPiosenek.csv dołączam w wysłanym na platformę archiwum.

SYNONIM

Wylosowany synonim: CIE świta.

Proces generowania synonimu:

```
Process
                                                                                 res
) inp
                                                                                 res (
                              Select by Random
<?xml version="1.0" encoding="UTF-8"?><process version="9.4.001">
 <context>
  <input/>
  <output/>
  <macros/>
 </context>
 <operator activated="true" class="process" compatibility="9.4.001" expanded="true"</pre>
name="Process">
  <parameter key="logverbosity" value="init"/>
  <parameter key="random_seed" value="-1"/>
  <parameter key="send_mail" value="never"/>
  <parameter key="notification email" value=""/>
  <parameter key="process_duration_for_mail" value="30"/>
  <parameter key="encoding" value="SYSTEM"/>
  cess expanded="true">
   <operator activated="true" class="retrieve" compatibility="9.4.001" expanded="true"</pre>
height="68" name="Retrieve" width="90" x="45" y="136">
    <parameter key="repository_entry" value="//Local Repository/data/TekstyPiosenek"/>
   </operator>
   <operator activated="true" class="select by random" compatibility="9.4.001" expanded="true"</pre>
height="82" name="Select by Random" width="90" x="246" y="136">
    <parameter key="use_fixed_number_of_attributes" value="true"/>
    <parameter key="number_of_attributes" value="2"/>
    <parameter key="use_local_random_seed" value="false"/>
    <parameter key="local_random_seed" value="1992"/>
   </operator>
   <connect from_op="Retrieve" from_port="output" to_op="Select by Random" to_port="example"</pre>
set input"/>
   <connect from op="Select by Random" from port="example set output" to port="result 1"/>
   <portSpacing port="source input 1" spacing="0"/>
```

```
<portSpacing port="sink_result 1" spacing="0"/>
  <portSpacing port="sink_result 2" spacing="0"/>
  </process>
  </operator>
</process>
```

PRZYGOTOWANIA

Wnioski:

Discretize by Binning: Wartości atrybutów zmieniają postać z numerycznej na nominalną (przedziały liczbowe). Zmieniając parametr 'number of bins' wpływamy na ilość generowanych przedziałów. Przedziały są wyznaczane w taki sposób że ich zakres jest prawie równy.

Discretize by Entorpy: wartości atrybutów zmieniają jak wyżej postać z numerycznych na nominalną, z tym że granice przedziałów są wybierane tak aby zminimalizować entropię (nieuporządkowanie) w poszczególnych przedziałach. Ponadto dyskretyzacja przez entropię odrzuca z automatu "bezużyteczne atrybuty", czyli takie, dla których został wyznaczony jeden przedział wartości. W rezultacie otrzymujemy zbiór z zredukowaną ilością atrybutów regularnych. W naszym przypadku zredukowano z 7171 do 11 atrybutów regularnych.

Detect outlier (LOF): dodaje każdemu rekordowi (przykładowi) kolumnę z wartością "local outlier factor", która wyznaczana jest na podstawie porównania lokalnej gęstości obiektu z lokalnymi gęstościami jego sąsiadów. W rezultacie możemy zidentyfikować regiony o podobnej gęstości oraz punkty które mają znacznie niższą gęstość niż sąsiedzi (większy współczynnik) - wtedy obiekty te należy traktować jako przykłady odstające. Jako że korzystamy z funkcji euklidesowego wyznaczania dystansów, dane uprzednio musimy znormalizować.

Weights by correlation: oblicza wagę dla każdego atrybutu wyznaczając wartość korelacji wejściowego zestawu przykładów w odniesieniu do atrybutu "label". Select by weight pozwala przefiltrować zbiór atrybutów, wybierając spośród nich te najważniejsze na podstawie wyznaczonych do nich wag.

Decision Tree: przed wykonaniem drzewa decyzyjnego, poddajemy dane dyskretyzacji przez entropię, aby odrzucić bezużyteczne atrybuty w celu przyśpieszenia wyznaczenia reguł decyzyjnych blokiem "Rule inducion". Po wykonaniu otrzymujemy wagi atrybutów, result test jak poprzednio oraz ukształtowane drzewo decyzyjne, które pokazuje nam na podstawie jakich wartości wybranych atrybutów możemy uważać że piosenka jest autorstwa danego zespołu.

Procesy:

Normalize:

```
Process
                                                                                res
) inp
                                                                                res (
                                 Normalize
        Retrieve
                                        exa
                                         ori
                                        pre
<?xml version="1.0" encoding="UTF-8"?><process version="9.4.001">
 <context>
  <input/>
  <output/>
  <macros/>
 </context>
 <operator</pre>
              activated="true"
                                  class="process"
                                                     compatibility="9.4.001"
                                                                               expanded="true"
name="Process">
  <parameter key="logverbosity" value="init"/>
  <parameter key="random_seed" value="-1"/>
  <parameter key="send mail" value="never"/>
  <parameter key="notification_email" value=""/>
  <parameter key="process_duration_for_mail" value="30"/>
  <parameter key="encoding" value="SYSTEM"/>
  cess expanded="true">
   <operator
               activated="true"
                                   class="retrieve"
                                                     compatibility="9.4.001"
                                                                               expanded="true"
height="68" name="Retrieve" width="90" x="45" y="136">
    <parameter key="repository_entry" value="//Local Repository/data/TekstyPiosenek"/>
   </operator>
               activated="true"
                                  class="normalize"
                                                      compatibility="9.4.001"
                                                                               expanded="true"
   <operator
height="103" name="Normalize" width="90" x="246" y="136">
    <parameter key="return preprocessing model" value="false"/>
    <parameter key="create_view" value="false"/>
    <parameter key="attribute_filter_type" value="all"/>
    <parameter key="attribute" value=""/>
    <parameter key="attributes" value=""/>
    <parameter key="use_except_expression" value="false"/>
    <parameter key="value_type" value="numeric"/>
    <parameter key="use_value_type_exception" value="false"/>
    <parameter key="except_value_type" value="real"/>
    <parameter key="block_type" value="value_series"/>
    <parameter key="use block type exception" value="false"/>
    <parameter key="except block type" value="value series end"/>
    <parameter key="invert_selection" value="false"/>
    <parameter key="include special attributes" value="false"/>
    <parameter key="method" value="Z-transformation"/>
    <parameter key="min" value="0.0"/>
    <parameter key="max" value="1.0"/>
```

Discritize by Binning:

```
Process
) inp
<?xml version="1.0" encoding="UTF-8"?><process version="9.4.001">
 <context>
  <input/>
  <output/>
  <macros/>
 </context>
 <operator</pre>
              activated="true"
                                  class="process"
                                                     compatibility="9.4.001"
                                                                                expanded="true"
name="Process">
  <parameter key="logverbosity" value="init"/>
  <parameter key="random seed" value="-1"/>
  <parameter key="send mail" value="never"/>
  <parameter key="notification email" value=""/>
  <parameter key="process_duration_for_mail" value="30"/>
  <parameter key="encoding" value="SYSTEM"/>
  cprocess expanded="true">
   <operator
                activated="true"
                                   class="retrieve"
                                                      compatibility="9.4.001"
                                                                                expanded="true"
height="68" name="Retrieve" width="90" x="45" y="136">
    <parameter key="repository_entry" value="//Local Repository/data/TekstyPiosenek"/>
   </operator>
   <operator activated="true" class="discretize_by_bins" compatibility="9.4.001" expanded="true"</pre>
height="103" name="Discretize" width="90" x="313" y="136">
    <parameter key="return preprocessing model" value="false"/>
    <parameter key="create_view" value="false"/>
    <parameter key="attribute filter type" value="all"/>
    <parameter key="attribute" value=""/>
    <parameter key="attributes" value=""/>
    <parameter key="use_except_expression" value="false"/>
    <parameter key="value_type" value="numeric"/>
```

```
<parameter key="use_value_type_exception" value="false"/>
    <parameter key="except_value_type" value="real"/>
    <parameter key="block_type" value="value_series"/>
    <parameter key="use_block_type_exception" value="false"/>
    <parameter key="except_block_type" value="value_series_end"/>
    <parameter key="invert_selection" value="false"/>
    <parameter key="include_special_attributes" value="false"/>
    <parameter key="number_of_bins" value="2"/>
    <parameter key="define_boundaries" value="false"/>
    <parameter key="range name type" value="long"/>
    <parameter key="automatic_number_of_digits" value="true"/>
    <parameter key="number of digits" value="3"/>
   </operator>
   <connect from_op="Retrieve" from_port="output" to_op="Discretize" to_port="example set</pre>
input"/>
   <connect from_op="Discretize" from_port="example set output" to_port="result 1"/>
   <portSpacing port="source_input 1" spacing="0"/>
   <portSpacing port="sink_result 1" spacing="0"/>
   <portSpacing port="sink_result 2" spacing="0"/>
  </process>
 </operator>
</process>
```

Discretize by Entorpy:

```
Process
                                                                                  res
) inp
                                                                                  res (
<?xml version="1.0" encoding="UTF-8"?><process version="9.4.001">
 <context>
  <input/>
  <output/>
  <macros/>
 </context>
 <operator</pre>
              activated="true"
                                  class="process"
                                                      compatibility="9.4.001"
                                                                                 expanded="true"
name="Process">
  <parameter key="logverbosity" value="init"/>
  <parameter key="random_seed" value="-1"/>
  <parameter key="send_mail" value="never"/>
  <parameter key="notification_email" value=""/>
```

```
<parameter key="process_duration_for_mail" value="30"/>
  <parameter key="encoding" value="SYSTEM"/>
  cprocess expanded="true">
   <operator</pre>
                activated="true"
                                   class="retrieve"
                                                     compatibility="9.4.001"
                                                                               expanded="true"
height="68" name="Retrieve" width="90" x="45" y="136">
    <parameter key="repository_entry" value="//Local Repository/data/TekstyPiosenek"/>
   </operator>
   <operator
                  activated="true"
                                       class="discretize by entropy"
                                                                         compatibility="9.4.001"
expanded="true" height="103" name="Discretize (2)" width="90" x="380" y="34">
    <parameter key="return preprocessing model" value="false"/>
    <parameter key="create view" value="false"/>
    <parameter key="attribute filter type" value="all"/>
    <parameter key="attribute" value=""/>
    <parameter key="attributes" value=""/>
    <parameter key="use_except_expression" value="false"/>
    <parameter key="value_type" value="numeric"/>
    <parameter key="use_value_type_exception" value="false"/>
    <parameter key="except_value_type" value="real"/>
    <parameter key="block_type" value="value_series"/>
    <parameter key="use_block_type_exception" value="false"/>
    <parameter key="except block type" value="value series end"/>
    <parameter key="invert_selection" value="false"/>
    <parameter key="include_special_attributes" value="false"/>
    <parameter key="remove useless" value="true"/>
    <parameter key="range name type" value="long"/>
    <parameter key="automatic_number_of_digits" value="true"/>
    <parameter key="number_of_digits" value="-1"/>
   <operator activated="false" class="discretize_by_bins" compatibility="9.4.001" expanded="true"</pre>
height="103" name="Discretize" width="90" x="246" y="187">
    <parameter key="return_preprocessing_model" value="false"/>
    <parameter key="create_view" value="false"/>
    <parameter key="attribute filter type" value="all"/>
    <parameter key="attribute" value=""/>
    <parameter key="attributes" value=""/>
    <parameter key="use except expression" value="false"/>
    <parameter key="value_type" value="numeric"/>
    <parameter key="use value type exception" value="false"/>
    <parameter key="except value type" value="real"/>
    <parameter key="block_type" value="value_series"/>
    <parameter key="use_block_type_exception" value="false"/>
    <parameter key="except_block_type" value="value_series_end"/>
    <parameter key="invert selection" value="false"/>
    <parameter key="include special attributes" value="false"/>
    <parameter key="number_of_bins" value="2"/>
    <parameter key="define boundaries" value="false"/>
    <parameter key="range name type" value="long"/>
    <parameter key="automatic number of digits" value="true"/>
    <parameter key="number_of_digits" value="3"/>
   </operator>
   <operator activated="false" class="normalize" compatibility="9.4.001"</pre>
                                                                               expanded="true"
height="103" name="Normalize" width="90" x="447" y="238">
```

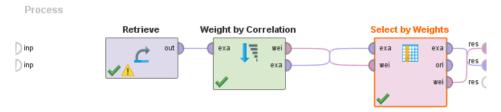
```
<parameter key="return_preprocessing_model" value="false"/>
    <parameter key="create_view" value="false"/>
    <parameter key="attribute filter type" value="all"/>
    <parameter key="attribute" value=""/>
    <parameter key="attributes" value=""/>
    <parameter key="use_except_expression" value="false"/>
    <parameter key="value_type" value="numeric"/>
    <parameter key="use value type exception" value="false"/>
    <parameter key="except value type" value="real"/>
    <parameter key="block type" value="value series"/>
    <parameter key="use_block_type_exception" value="false"/>
    <parameter key="except block type" value="value series end"/>
    <parameter key="invert_selection" value="false"/>
    <parameter key="include_special_attributes" value="false"/>
    <parameter key="method" value="Z-transformation"/>
    <parameter key="min" value="0.0"/>
    <parameter key="max" value="1.0"/>
    <parameter key="allow_negative_values" value="false"/>
   </operator>
   <connect from_op="Retrieve" from_port="output" to_op="Discretize (2)" to_port="example set</pre>
input"/>
   <connect from_op="Discretize (2)" from_port="example set output" to_port="result 1"/>
   <portSpacing port="source input 1" spacing="0"/>
   <portSpacing port="sink result 1" spacing="0"/>
   <portSpacing port="sink_result 2" spacing="0"/>
  </process>
 </operator>
</process>
```

Detect outlier (LOF):

```
Process
                                       Detect Outlier (LOF)
         Retrieve
) inp
                                 ori 
                                                                         res (
<?xml version="1.0" encoding="UTF-8"?><process version="9.4.001">
 <context>
  <input/>
  <output/>
  <macros/>
 </context>
 <operator</pre>
              activated="true"
                                   class="process"
                                                      compatibility="9.4.001"
                                                                                 expanded="true"
name="Process">
  <parameter key="logverbosity" value="init"/>
  <parameter key="random seed" value="2001"/>
  <parameter key="send mail" value="never"/>
  <parameter key="notification email" value=""/>
  <parameter key="process duration for mail" value="30"/>
  <parameter key="encoding" value="SYSTEM"/>
  cess expanded="true">
```

```
activated="true"
                                   class="retrieve"
                                                     compatibility="9.4.001"
                                                                               expanded="true"
   <operator
height="68" name="Retrieve" width="90" x="45" y="34">
    <parameter key="repository entry" value="//Local Repository/TekstyPiosenek"/>
   </operator>
   <operator activated="true"</pre>
                                  class="normalize"
                                                      compatibility="9.4.001"
                                                                               expanded="true"
height="103" name="Normalize" width="90" x="179" y="34">
    <parameter key="return_preprocessing_model" value="false"/>
    <parameter key="create view" value="false"/>
    <parameter key="attribute filter type" value="all"/>
    <parameter key="attribute" value=""/>
    <parameter key="attributes" value=""/>
    <parameter key="use except expression" value="false"/>
    <parameter key="value_type" value="numeric"/>
    <parameter key="use_value_type_exception" value="false"/>
    <parameter key="except_value_type" value="real"/>
    <parameter key="block_type" value="value_series"/>
    <parameter key="use_block_type_exception" value="false"/>
    <parameter key="except_block_type" value="value_series_end"/>
    <parameter key="invert selection" value="false"/>
    <parameter key="include_special_attributes" value="false"/>
    <parameter key="method" value="Z-transformation"/>
    <parameter key="min" value="0.0"/>
    <parameter key="max" value="1.0"/>
    <parameter key="allow negative values" value="false"/>
   </operator>
   <operator activated="true" class="detect_outlier_lof" compatibility="9.4.001" expanded="true"</pre>
height="82" name="Detect Outlier (LOF)" width="90" x="313" y="34">
    <parameter key="minimal points lower bound" value="10"/>
    <parameter key="minimal_points_upper_bound" value="20"/>
    <parameter key="distance function" value="euclidian distance"/>
   </operator>
   <connect from_op="Retrieve" from_port="output" to_op="Normalize" to_port="example set</pre>
input"/>
   <connect from op="Normalize" from port="example set output" to op="Detect Outlier (LOF)"</pre>
to port="example set input"/>
   <connect from op="Detect Outlier (LOF)" from port="example set output" to port="result 1"/>
   <portSpacing port="source_input 1" spacing="0"/>
   <portSpacing port="source input 2" spacing="0"/>
   <portSpacing port="sink result 1" spacing="0"/>
   <portSpacing port="sink_result 2" spacing="0"/>
  </process>
 </operator>
</process>
```

Weight by Correlation, Select by Weights:



```
<?xml version="1.0" encoding="UTF-8"?><process version="9.4.001">
 <context>
  <input/>
  <output/>
  <macros/>
 </context>
              activated="true"
                                  class="process"
                                                     compatibility="9.4.001"
                                                                               expanded="true"
 <operator
name="Process">
  <parameter key="logverbosity" value="init"/>
  <parameter key="random seed" value="2001"/>
  <parameter key="send mail" value="never"/>
  <parameter key="notification email" value=""/>
  <parameter key="process duration for mail" value="30"/>
  <parameter key="encoding" value="SYSTEM"/>
  cprocess expanded="true">
                activated="true"
                                   class="retrieve"
                                                     compatibility="9.4.001"
   <operator
                                                                               expanded="true"
height="68" name="Retrieve" width="90" x="112" y="34">
    <parameter key="repository_entry" value="//Local Repository/TekstyPiosenek"/>
   </operator>
   <operator
                  activated="true"
                                       class="weight_by_correlation"
                                                                         compatibility="9.4.001"
expanded="true" height="82" name="Weight by Correlation" width="90" x="246" y="34">
    <parameter key="normalize weights" value="true"/>
    <parameter key="sort weights" value="true"/>
    <parameter key="sort direction" value="ascending"/>
    <parameter key="squared_correlation" value="false"/>
   </operator>
   <operator activated="true" class="select_by_weights" compatibility="9.4.001" expanded="true"</pre>
height="103" name="Select by Weights" width="90" x="447" y="34">
    <parameter key="weight_relation" value="top p%"/>
    <parameter key="weight" value="1.0"/>
    <parameter key="k" value="10"/>
    <parameter key="p" value="0.002"/>
    <parameter key="deselect unknown" value="true"/>
    <parameter key="use absolute weights" value="true"/>
   </operator>
               from op="Retrieve"
                                     from port="output"
                                                            to op="Weight
                                                                                    Correlation"
   <connect
                                                                              by
to_port="example set"/>
   <connect from op="Weight by Correlation" from port="weights" to op="Select by Weights"</pre>
to port="weights"/>
   <connect from_op="Weight by Correlation" from_port="example set" to_op="Select by Weights"</pre>
to_port="example set input"/>
   <connect from_op="Select by Weights" from_port="example set output" to_port="result 2"/>
   <connect from op="Select by Weights" from port="weights" to port="result 1"/>
   <portSpacing port="source input 1" spacing="0"/>
   <portSpacing port="source_input 2" spacing="0"/>
   <portSpacing port="sink_result 1" spacing="0"/>
   <portSpacing port="sink result 2" spacing="0"/>
   <portSpacing port="sink result 3" spacing="0"/>
  </process>
 </operator>
</process>
```

Rule Induction, Decision Tree:

```
res
inp
                                                                          res
                                                                          res
                                                                          res (
                         Discretize
                                         Rule Induction
                                                           Decision Tree
         Retrieve
                                                 mod
                                exa
                                 ori 📗
                                                  exa
                                 pre |
<?xml version="1.0" encoding="UTF-8"?><process version="9.4.001">
 <context>
  <input/>
  <output/>
  <macros/>
 </context>
 <operator activated="true" class="process" compatibility="9.4.001" expanded="true"</pre>
name="Process">
  <parameter key="logverbosity" value="init"/>
  <parameter key="random seed" value="-1"/>
  <parameter key="send mail" value="never"/>
  <parameter key="notification_email" value=""/>
  <parameter key="process_duration_for_mail" value="30"/>
  <parameter key="encoding" value="SYSTEM"/>
  cess expanded="true">
   <operator activated="true" class="retrieve" compatibility="9.4.001" expanded="true"</pre>
height="68" name="Retrieve" width="90" x="45" y="136">
    <parameter key="repository entry" value="//Local Repository/TekstyPiosenek"/>
   </operator>
   <operator activated="true" class="discretize by entropy" compatibility="9.4.001"</pre>
expanded="true" height="103" name="Discretize" width="90" x="179" y="136">
    <parameter key="return_preprocessing_model" value="false"/>
    <parameter key="create view" value="false"/>
    <parameter key="attribute filter type" value="all"/>
    <parameter key="attribute" value=""/>
    <parameter key="attributes" value=""/>
    <parameter key="use_except_expression" value="false"/>
    <parameter key="value type" value="numeric"/>
    <parameter key="use value type exception" value="false"/>
    <parameter key="except value type" value="real"/>
    <parameter key="block_type" value="value_series"/>
    <parameter key="use block type exception" value="false"/>
    <parameter key="except block type" value="value series end"/>
    <parameter key="invert_selection" value="false"/>
    <parameter key="include_special_attributes" value="false"/>
    <parameter key="remove useless" value="true"/>
    <parameter key="range_name_type" value="long"/>
```

```
<parameter key="automatic_number_of_digits" value="true"/>
    <parameter key="number_of_digits" value="-1"/>
   </operator>
   <operator activated="true" class="rule induction" compatibility="9.4.001" expanded="true"</pre>
height="82" name="Rule Induction" width="90" x="313" y="136">
    <parameter key="criterion" value="information_gain"/>
    <parameter key="sample_ratio" value="0.9"/>
    <parameter key="pureness" value="0.9"/>
    <parameter key="minimal prune benefit" value="0.25"/>
    <parameter key="use local random seed" value="false"/>
    <parameter key="local_random_seed" value="1992"/>
   </operator>
   <operator activated="true" class="concurrency:parallel decision tree" compatibility="9.4.001"</pre>
expanded="true" height="103" name="Decision Tree" width="90" x="447" y="136">
    <parameter key="criterion" value="gain_ratio"/>
    <parameter key="maximal_depth" value="10"/>
    <parameter key="apply_pruning" value="true"/>
    <parameter key="confidence" value="0.1"/>
    <parameter key="apply prepruning" value="true"/>
    <parameter key="minimal_gain" value="0.01"/>
    <parameter key="minimal leaf size" value="2"/>
    <parameter key="minimal size for split" value="4"/>
    <parameter key="number_of_prepruning_alternatives" value="3"/>
   </operator>
   <connect from op="Retrieve" from port="output" to op="Discretize" to port="example set</pre>
input"/>
   <connect from_op="Discretize" from_port="example set output" to_op="Rule Induction"</pre>
to port="training set"/>
   <connect from_op="Rule Induction" from_port="exampleSet" to_op="Decision Tree"</pre>
to port="training set"/>
   <connect from op="Decision Tree" from port="model" to port="result 1"/>
   <connect from_op="Decision Tree" from_port="exampleSet" to_port="result 2"/>
   <connect from op="Decision Tree" from port="weights" to port="result 3"/>
   <portSpacing port="source input 1" spacing="0"/>
   <portSpacing port="sink result 1" spacing="0"/>
   <portSpacing port="sink result 2" spacing="0"/>
   <portSpacing port="sink_result 3" spacing="0"/>
   <portSpacing port="sink result 4" spacing="0"/>
  </process>
 </operator>
</process>
```

DYWERSJA

Przed wykonaniem zadania przygotowałem zbiór zawierający teksty piosenek wszystkich artystów:

```
Process
         Process Documents from Files
                                           Store
inp
                                                                       res (
inp
<?xml version="1.0" encoding="UTF-8"?><process version="9.4.001">
 <context>
  <input/>
  <output/>
  <macros/>
 </context>
 <operator activated="true" class="process" compatibility="9.4.001" expanded="true"</pre>
name="Process">
  <parameter key="logverbosity" value="init"/>
  <parameter key="random seed" value="2001"/>
  <parameter key="send mail" value="never"/>
  <parameter key="notification email" value=""/>
  <parameter key="process duration for mail" value="30"/>
  <parameter key="encoding" value="SYSTEM"/>
  cprocess expanded="true">
   <operator activated="true" class="text:process document from file" compatibility="8.2.000"</pre>
expanded="true" height="82" name="Process Documents from Files" width="90" x="112" y="34">
    key="text directories">
     <parameter key="akcent" value="C:\rep\VII sem\ADiO\Lab1\disco-polo\akcent"/>
     <parameter key="boys" value="C:\rep\VII sem\ADiO\Lab1\disco-polo\boys"/>
     <parameter key="shazza" value="C:\rep\VII sem\ADiO\Lab1\disco-polo\shazza"/>
    </list>
    <parameter key="file pattern" value="*"/>
    <parameter key="extract text only" value="true"/>
    <parameter key="use_file_extension_as_type" value="true"/>
    <parameter key="content type" value="txt"/>
    <parameter key="encoding" value="UTF-8"/>
    <parameter key="create_word_vector" value="true"/>
    <parameter key="vector_creation" value="TF-IDF"/>
    <parameter key="add_meta_information" value="true"/>
    <parameter key="keep text" value="false"/>
    <parameter key="prune method" value="none"/>
    <parameter key="prune_below_percent" value="3.0"/>
    <parameter key="prune above percent" value="30.0"/>
    <parameter key="prune below rank" value="0.05"/>
    <parameter key="prune above rank" value="0.95"/>
    <parameter key="datamanagement" value="double sparse array"/>
    <parameter key="data_management" value="auto"/>
    cess expanded="true">
```

```
<operator activated="true" class="text:tokenize" compatibility="8.2.000" expanded="true"</pre>
height="68" name="Tokenize" width="90" x="179" y="34">
      <parameter key="mode" value="non letters"/>
      <parameter key="characters" value=".:"/>
      <parameter key="language" value="English"/>
      <parameter key="max_token_length" value="3"/>
     </operator>
     <connect from port="document" to op="Tokenize" to port="document"/>
     <connect from op="Tokenize" from port="document" to port="document 1"/>
     <portSpacing port="source document" spacing="0"/>
     <portSpacing port="sink_document 1" spacing="0"/>
     <portSpacing port="sink document 2" spacing="0"/>
    </process>
   </operator>
   <operator activated="true" class="store" compatibility="9.4.001" expanded="true" height="68"</pre>
name="Store" width="90" x="313" y="34">
    <parameter key="repository_entry" value="//Local Repository/TekstyPiosenekAll"/>
   </operator>
   <connect from port="input 1" to op="Process Documents from Files" to port="word list"/>
   <connect from_op="Process Documents from Files" from_port="example set" to_op="Store"</pre>
to port="input"/>
   <portSpacing port="source input 1" spacing="0"/>
   <portSpacing port="source input 2" spacing="0"/>
   <portSpacing port="sink_result 1" spacing="0"/>
  </process>
 </operator>
</process>
```

Następnie, osobno dla każdego artysty stosując bloczek "Filter examples" i wybierając tylko przykłady oznaczone labelem dla konkretnego wykonawcy wyznaczyłem jego outlierów bloczkiem "Detect Outlier (LOF)", jednakże uprzednio normalizując dane blokiem "Normalize" ponieważ przy wyznaczaniu outlierów wykorzystałem obliczanie odległości euklidesowej. Następnie uzyskany result set wyfiltrowałem po wartości "outlier" malejąco i odczytałem pierwsze rekordy.

Przykład na zespole akcent: (pozostałe analogiczne z odpowiednią zmianą filtrowania)

```
<macros/>
 </context>
 <operator activated="true" class="process" compatibility="9.4.001" expanded="true"</pre>
name="Process">
  <parameter key="logverbosity" value="init"/>
  <parameter key="random_seed" value="2001"/>
  <parameter key="send_mail" value="never"/>
  <parameter key="notification email" value=""/>
  <parameter key="process duration for mail" value="30"/>
  <parameter key="encoding" value="SYSTEM"/>
  cess expanded="true">
   <operator activated="true" class="retrieve" compatibility="9.4.001" expanded="true"</pre>
height="68" name="Retrieve" width="90" x="45" y="85">
    <parameter key="repository_entry" value="//Local Repository/TekstyPiosenekAll"/>
   </operator>
   <operator activated="true" class="filter_examples" compatibility="9.4.001" expanded="true"</pre>
height="103" name="Filter Examples" width="90" x="179" y="85">
    <parameter key="parameter_expression" value=""/>
    <parameter key="condition class" value="custom filters"/>
    <parameter key="invert_filter" value="false"/>
    key="filters list">
     <parameter key="filters entry key" value="label.equals.akcent"/>
    <parameter key="filters logic and" value="true"/>
    <parameter key="filters_check_metadata" value="true"/>
   </operator>
   <operator activated="true" class="normalize" compatibility="9.4.001" expanded="true"</pre>
height="103" name="Normalize" width="90" x="313" y="85">
    <parameter key="return_preprocessing_model" value="false"/>
    <parameter key="create view" value="false"/>
    <parameter key="attribute filter type" value="all"/>
    <parameter key="attribute" value=""/>
    <parameter key="attributes" value=""/>
    <parameter key="use_except_expression" value="false"/>
    <parameter key="value type" value="numeric"/>
    <parameter key="use_value_type_exception" value="false"/>
    <parameter key="except_value_type" value="real"/>
    <parameter key="block type" value="value series"/>
    <parameter key="use block type exception" value="false"/>
    <parameter key="except_block_type" value="value_series_end"/>
    <parameter key="invert_selection" value="false"/>
    <parameter key="include_special_attributes" value="false"/>
    <parameter key="method" value="Z-transformation"/>
    <parameter key="min" value="0.0"/>
    <parameter key="max" value="1.0"/>
    <parameter key="allow_negative_values" value="false"/>
   </operator>
   <operator activated="true" class="detect outlier lof" compatibility="9.4.001" expanded="true"</pre>
height="82" name="Detect Outlier (LOF)" width="90" x="447" y="85">
    <parameter key="minimal_points_lower_bound" value="10"/>
    <parameter key="minimal_points_upper_bound" value="20"/>
    <parameter key="distance_function" value="euclidian distance"/>
```

</operator>

<connect from_op="Retrieve" from_port="output" to_op="Filter Examples" to_port="example
set input"/>

<connect from_op="Filter Examples" from_port="example set output" to_op="Normalize"
to port="example set input"/>

<connect from_op="Normalize" from_port="example set output" to_op="Detect Outlier (LOF)"
to_port="example set input"/>

<connect from_op="Detect Outlier (LOF)" from_port="example set output" to_port="result 1"/>
<portSpacing port="source_input 1" spacing="0"/>

<portSpacing port="sink result 1" spacing="0"/>

<portSpacing port="sink_result 2" spacing="0"/>

</process>

</operator>

</process>

Wyniki:

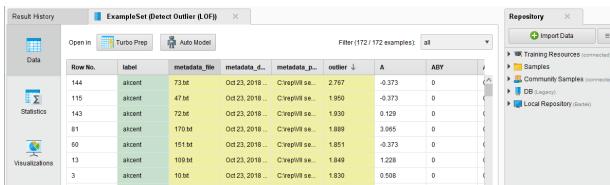
Pierwsze 3 podejrzane pliki dla:

Akcent: 73, 47, 72, (151 pozycja 5ta)

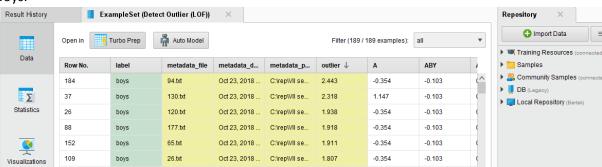
Boys: 94, 130, 120

Shazza: 8, 5, 67 (7 pozycja 4ta, 28 pozycja 9ta)

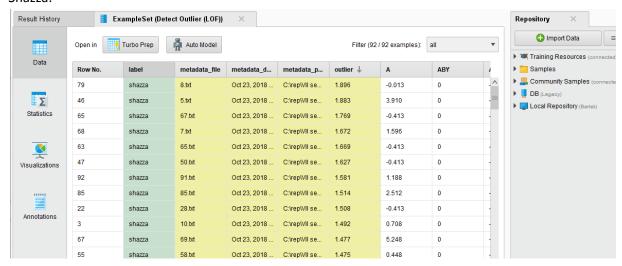
Potwierdzenie: Akcent:



Boys:



Shazza:



NA FRONCIE

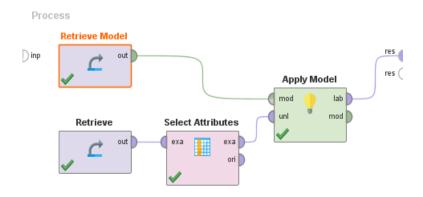
Skorzystałem ze zbioru wygenerowanego już dla zadania DYWERSJA.

Za pomocą kreatora "Auto Model" utworzyłem modele dla wszystkich typów. W moim przypadku najlepszym modelem okazał się model typu "Generalized Linear Model".

plik testowy	przewidywany zespół	rzeczywisty zespół	
1	akcent	akcent	
2	boys	boys	
3	boys	shazza	
4	shazza	akcent	
5	akcent	akcent	
6	akcent	boys	
7	akcent	shazza	
8	boys	shazza	
9	boys	boys	

Dokładność klasyfikacji: 4/9 * 100% = **44.44%**

Proces wykorzystania modelu "Generalized Linear Model" do testowego zbioru:



```
<?xml version="1.0" encoding="UTF-8"?><process version="9.4.001">
 <context>
  <input/>
  <output/>
  <macros/>
 </context>
 <operator activated="true" class="process" compatibility="9.4.001" expanded="true"</pre>
name="Process">
  <parameter key="logverbosity" value="init"/>
  <parameter key="random seed" value="2001"/>
  <parameter key="send mail" value="never"/>
  <parameter key="notification email" value=""/>
  <parameter key="process duration for mail" value="30"/>
  <parameter key="encoding" value="SYSTEM"/>
  cess expanded="true">
   <operator activated="true" class="retrieve" compatibility="9.4.001" expanded="true"</pre>
height="68" name="Retrieve" width="90" x="45" y="136">
    <parameter key="repository_entry" value="../TekstyPiosenekTest"/>
   </operator>
   <operator activated="true" class="select attributes" compatibility="9.4.001" expanded="true"</pre>
height="82" name="Select Attributes" width="90" x="179" y="136">
    <parameter key="attribute filter type" value="single"/>
    <parameter key="attribute" value="label"/>
    <parameter key="attributes" value=""/>
    <parameter key="use except expression" value="false"/>
    <parameter key="value_type" value="attribute_value"/>
    <parameter key="use value type exception" value="false"/>
    <parameter key="except_value_type" value="time"/>
    <parameter key="block_type" value="attribute_block"/>
    <parameter key="use block type exception" value="false"/>
    <parameter key="except block type" value="value matrix row start"/>
    <parameter key="invert_selection" value="true"/>
    <parameter key="include_special_attributes" value="true"/>
   </operator>
   <operator activated="true" class="retrieve" compatibility="9.4.001" expanded="true"</pre>
height="68" name="Retrieve Model" width="90" x="45" y="34">
    <parameter key="repository_entry" value="../AutoModel/Generalized Linear Model/Model"/>
   </operator>
   <operator activated="true" class="apply model" compatibility="9.4.001" expanded="true"</pre>
height="82" name="Apply Model" width="90" x="313" y="85">
    <list key="application parameters"/>
    <parameter key="create_view" value="false"/>
   </operator>
   <connect from op="Retrieve" from port="output" to op="Select Attributes" to port="example
set input"/>
   <connect from_op="Select Attributes" from_port="example set output" to_op="Apply Model"</p>
to port="unlabelled data"/>
   <connect from op="Retrieve Model" from port="output" to op="Apply Model"</pre>
to_port="model"/>
   <connect from_op="Apply Model" from_port="labelled data" to_port="result 1"/>
   <portSpacing port="source_input 1" spacing="0"/>
   <portSpacing port="sink_result 1" spacing="0"/>
```

```
<portSpacing port="sink_result 2" spacing="0"/>
  </process>
```

NOWA NADZIEJA

plik testowy	CIĘ świta	Deszczowy Zosia	Rozpływasz bajkach	uwielbiam dawnych	przewidywany zespół	rzeczywisty zespół
1	akcent	akcent	akcent	boys	akcent	akcent
2	boys	akcent	boys	boys	boys	boys
3	boys	boys	akcent	shazza	boys	shazza
4	shazza	boys	boys	shazza	shazza	akcent
5	akcent	shazza	akcent	akcent	akcent	akcent
6	akcent	boys	akcent	akcent	akcent	boys
7	akcent	boys	akcent	akcent	akcent	shazza
8	boys	shazza	boys	boys	boys	shazza
9	boys	shazza	boys	boys	boys	boys

Dokładność klasyfikacji niestety bez zmian: 4/9 * 100% = **44.44%** ⊗