

Informatyka stosowana, studia dzienne, II st.

semestr II

Eksploracja danych internetowych

2020/2021

Prowadzący: dr inż. Krzysztof Myszkowski

poniedziałek, 10:00

Ćwiczenie 2: Analiza atrybutów i klasterowa domen internetowych

Autorzy:

Paweł Jeziorski 234066

Karol Podlewski 234106

Spis treści

1. Cel	3
2. Przygotowanie danych	3
2.1. Pozyskiwanie danych	3
2.2. Przetworzenie pliku html	3
3. Metody przeprowadzonych badań	5
3.1. Wstęp	5
3.2. Analiza atrybutów	5
3.3. Klasteryzacja algorytmem EM	5
4. Wyniki analizy dla domeny ftims	6
4.1. Analiza atrybutów	7
4.2. Analiza klastrowa	15
5. Wyniki analizy dla domeny UŁ	19
5.1. Analiza atrybutów	19
5.2. Analiza klastrowa	24
6. Dyskusja	27
6.1. Analiza atrybutów	27
6.2. Analiza klastrowa	28
7. Podsumowanie	29

1. Cel

Celem zadania było przeprowadzenie analizy klastrowej dokumentów. Badaną domenę należało przekształcić w plik tekstowy zawierający odpowiednie dane. Należało przeanalizować atrybuty dokumentów w zależności od reprezentacji i ustawionych parametrów w programie Weka. Następnie należało dokonać klasteryzacji tych dokumentów dla każdego analizowanego przypadku.

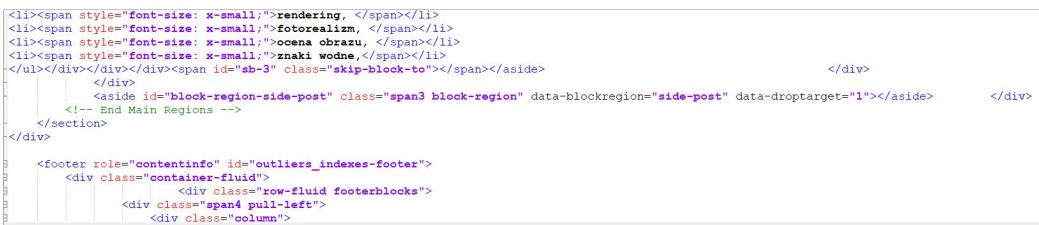
Domeny wykorzystane do analizy to:

- <https://ftims.p.lodz.pl> *
 - <https://www.uni.lodz.pl> *
- * ostatni dostęp 19.12.2020

2. Przygotowanie danych

2.1. Pozyskiwanie danych

WebSphinx to robot internetowy, czyli program, który zbiera informacje o strukturze, stronach i treściach znajdujących się w internecie. Za pomocą tej aplikacji pobrano strony internetowe do pliku html.



```
<li><span style="font-size: x-small;">>rendering, </span></li>
<li><span style="font-size: x-small;">>fotorealizm, </span></li>
<li><span style="font-size: x-small;">>scena obrazu, </span></li>
<li><span style="font-size: x-small;">>znaki wodne, </span></li>
</ul></div></div><div><span id="sb-3" class="skip-block-to"></span></div>
</div>
<!-- End Main Regions -->
<aside id="block-region-side-post" class="span3 block-region" data-blockregion="side-post" data-droppart="1"></aside>
</div>
</div>
<footer role="contentinfo" id="outliers_indexes-footer">
<div class="container-fluid">
<div class="row-fluid footerblocks">
<div class="span4 pull-left">
<div class="column">
```

Rysunek 1. Fragment pliku html otrzymanego w wyniku pracy programu WebSphinx (<https://ftims.p.lodz.pl>)

2.2. Przetworzenie pliku html

Otrzymany plik należało oczyścić z elementów charakterystycznych dla formatu html, utrudniających analizę słów. Następnie trzeba było odpowiednio go przekształcić, aby otrzymać plik arff akceptowalny przez program Weka.

Do pierwszego etapu użyto krótkiego skryptu w języku Python implementującego bibliotekę html2text. Dzięki temu otrzymano plik w formacie complainText.

Kolejny krok dotyczył wyodrębnienia poszczególnych dokumentów z poprzednio uzyskanego pliku i konwersji do formatu arff. W związku z tym przygotowano skrypt w języku Python, który umożliwia pogrupowanie kolejnych podstron i zapisanie ich do pliku. Plik wynikowy wymaga zachowania odpowiedniego schematu, tj. nagłówków oraz atrybutów (Rysunek 3)

```
import pandas as pd

df = pd.DataFrame(columns = ['Title', 'data'])

line_back = ""
term = "Page "
page_counter = 0
data = ""
title = ""

file = open('concatenate_text.txt', encoding="utf8")
for line in file:
    line.strip().split('/n')
    if term in line:
        if title != "":
            title = 'Page ' + str(page_counter)
            page_counter += 1
            df.loc[-1] = [title, data] # adding a row
            df.index = df.index + 1 # shifting index
            df = df.sort_index() # sorting by index
            out = line.partition("[") [2].partition("]") [0]
            if out == "":
                out = line_back.partition("[") [2].partition("]") [0]
            title = out
            data = ""
        else:
            data += line
            line_back= line
file.close()
df = df[::-1]
df.to_csv("Output.csv", sep=',', index=False, header=False)
print(df)

arff.dump('Out.arff',
          df.values,
          relation='ftims',
          names=df.columns)
```

Rysunek 2. Skrypt grupujący dokumenty i zapisujący je do formaty arff

1 @relation ftims
2 @attribute Title string
3 @attribute data string
4 @data
5 'Page 0','\n[Przejdz_ do g9\xla\x02Bwnej zawarto9:ci] (https://ftims.p.lodz.pl#maincontent)\n\n[1](https://ftims.p.
6 'Page 1','\n[Przejdz_ do g9\xla\x02Bwnej zawarto9:ci] (http://ftims.p.lodz.pl#maincontent)\n\nWikamp\n\n * _\n

Rysunek 3. Zawartość pliku arff

3. Metody przeprowadzonych badań

3.1. Wstęp

W przypadku realizacji zapytania związanego z wyszukiwaniem określonych słów znalezienie rzetelnej informacji jest dość trudne. Zazwyczaj lista wynikowa jest znaczących rozmiarów i nie niesie za sobą wielu istotnych wartości. Do przedstawienia miarodajnych rezultatów potrzeba więcej informacji dotyczących badanych dokumentów, takich jak częstotliwość czy miejsce wystąpienia jego składowych. Aby to osiągnąć należy odpowiednio przygotować atrybuty. Formalną reprezentację dokumentów nazywamy termami.

3.2. Analiza atrybutów

Aby osiągnąć powyższy cel należy dokonać modyfikacji atrybutów. W obecnej formie każdy z nich jest typu string. Dlatego, pierwszy z nich przekształcono do typu nominalnego. Wykorzystano filtr StringToNominal w programie Weka. W przypadku drugiego atrybutu zastosowano model przestrzeni wektorowej. Przedstawia on dane w postaci wektora w wielowymiarowej przestrzeni. W tym przypadku użyto filtra StringToWordVector. Metoda ta dostarcza wielu możliwych konfiguracji, jednakże na potrzeby eksperymentów wykorzystanych zostanie tylko kilka z nich - TFTransform, IDFTransform czy OutputWordCounts. Zasady działania oraz sposób wykorzystania zostały opisane w dalszej części sprawozdania.

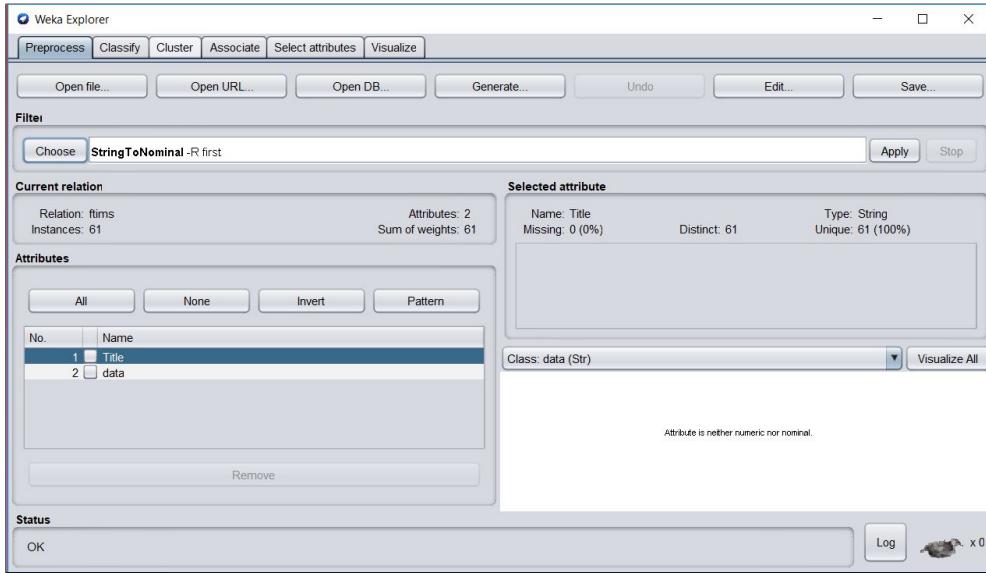
3.3. Klasteryzacja algorytmem EM

Algorytm EM (ang. Expectation–Maximization) nazywany maksymalizacją oczekiwaniń to iteracyjna metoda szacowania maksymalnego prawdopodobieństwa. Opiera się on na dwóch krokach: wyznacza wartości spodziewane prawdopodobieństwa, a następnie oblicza rozkład parametrów oraz wiarygodność zmiennych. Proces ten trwa tak długo, aż zostanie osiągnięta maksymalna wartość wiarygodności, bądź wynik iteracji się ustabilizuje.

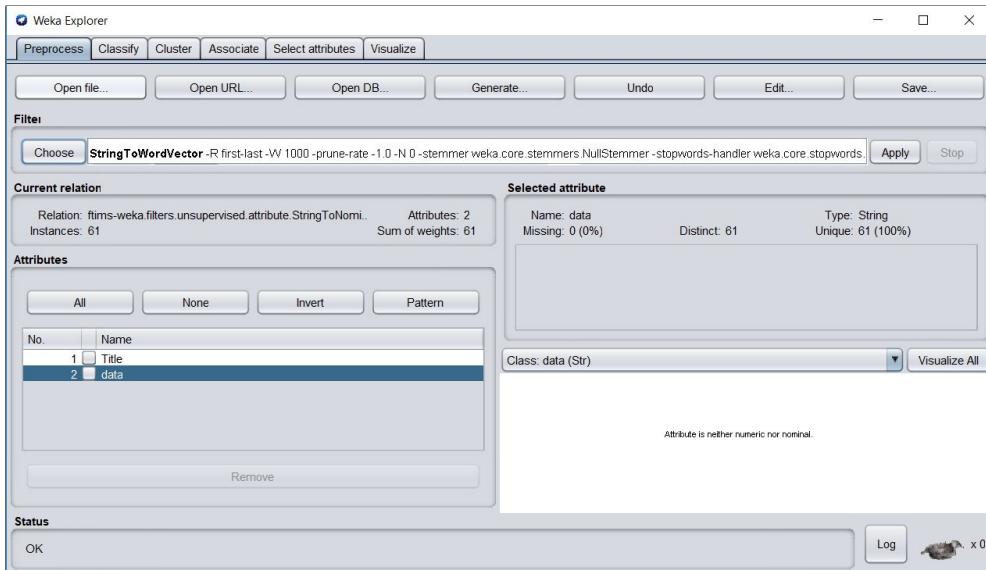
Dobrze sprawdza się w analizowanym przypadku ze względu na brak konieczności podawania ilości klastrów.

4. Wyniki analizy dla domeny ftims

Zimportowane dane do programu Weka, należy przekonwertować do odpowiedniego typu. Do tego celu użyto dwóch, wcześniej wspomnianych filtrów: StringToNominal i StringToWordVector. Na Rysunku 3 przedstawiono ustawienia, które pozostaną niezmienne podczas eksperymentów, zaś konfiguracja na Rysunku 5 będzie odpowiednio modyfikowana w kolejnych badaniach.



Rysunek 4. Konfiguracja filtra StringToNominal dla pierwszego atrybutu (Program Weka)



Rysunek 5. Konfiguracja filtra StringToWordVector dla drugiego atrybutu (Program Weka)

4.1. Analiza atrybutów

■ Przypadek 0.

Brak ustawionych parametrów filtra dało wynik w postaci macierzy obecności termów w danym dokumencie (Rysunek 6). Jeżeli element występuje widnieje wartość '1.0', w przeciwnym wypadku '0.0'.

Tabela 1. Parametry zastosowanego filtra StringToWordVector dla przypadku 0.

Parametr	Wartość
outputWordCounts	false
StopwordsHandler	false
TFTransform	false
IDFTransform	false

Rysunek 6. Wyniki analizy domeny dla Przypadku 0.

■ Przypadek 1. - outputWordCount

Parametr outputWordCount pozwala zliczyć wystąpienia termu w danym dokumencie. Spośród widocznych elementów przedstawionych na Rysunku 7 najczęściej występującym termem jest znak *, tj znak poprzedzający listy. W każdym widocznym dokumencie występuje ponad sto razy. Z kolei najrzadziej występujące to te opatrzone numerem 11 czy 12. Dla każdego prezentowanego dokumentu widnieje wówczas wartość 0.

Tabela 2. Parametry zastosowanego filtra StringToWordVector dla przypadku 1.

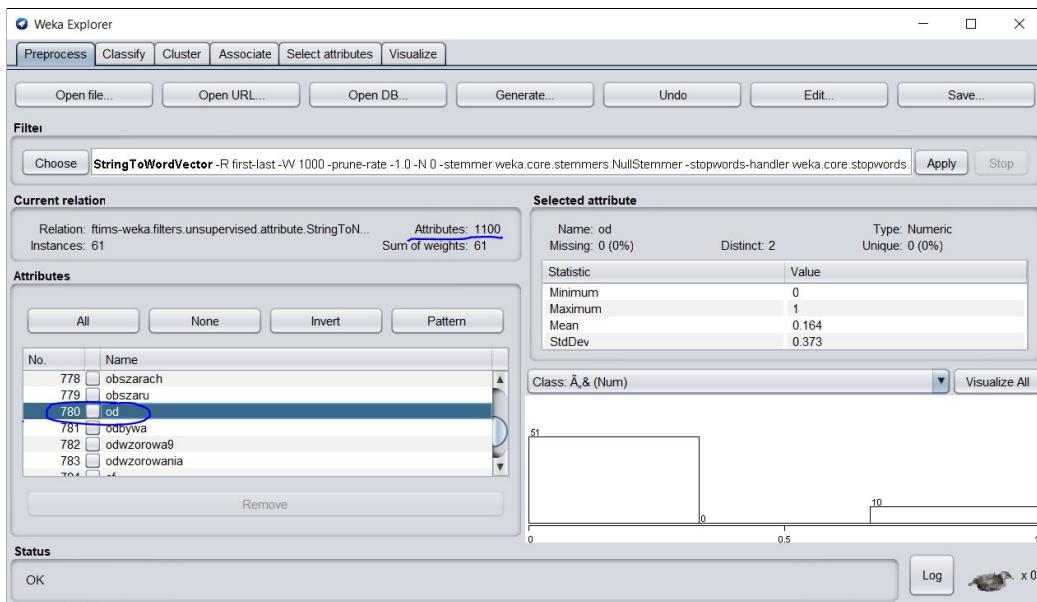
Parametr	Wartość
outputWordCounts	true
StopwordsHandler	false
TFTransform	false
IDFTransform	false

	1: Titel	2: #	3: ##	4: ####	5: #####	6: #####	7: &	8: &cy;	9: *	10: **	11: ***FTIMSS**	12: **Instytut	13: **PÁz**	14: **Przydatn	15: **Zakl>xlaad	16: **dr	17: **e-mail	18: **mgr	19: **praw	20: **A	21: +48	22: -	:	
Nominal	Nominal	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal
Page 0	2.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 1	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 2	1.0	2.0	1.0	3.0	5.0	0.0	0.0	108.0	6.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 3	1.0	2.0	0.0	0.0	4.0	0.0	0.0	0.0	108.0	10.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 4	0.0	8.0	0.0	0.0	0.0	0.0	0.0	0.0	108.0	4.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 5	1.0	1.0	4.0	0.0	0.0	0.0	0.0	0.0	114.0	5.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 6	0.0	2.0	0.0	0.0	2.0	0.0	0.0	0.0	110.0	4.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 7	0.0	2.0	2.0	0.0	0.0	0.0	0.0	0.0	146.0	4.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 8	0.0	3.0	0.0	0.0	3.0	0.0	0.0	0.0	150.0	7.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 9	0.0	2.0	0.0	4.0	3.0	0.0	0.0	0.0	108.0	4.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 10	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	108.0	4.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 11	0.0	3.0	1.0	0.0	0.0	0.0	0.0	0.0	108.0	15.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	6.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 12	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	109.0	4.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 13	0.0	2.0	0.0	0.0	0.0	1.0	0.0	0.0	121.0	11.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
Page 14	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	122.0	7.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 15	1.0	1.0	0.0	10.0	10.0	0.0	0.0	0.0	143.0	4.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	5.0	0.0	0.0	0.0	0.0
Page 16	1.0	1.0	0.0	10.0	0.0	0.0	0.0	0.0	143.0	4.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 17	1.0	1.0	0.0	3.0	0.0	0.0	0.0	0.0	122.0	4.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 18	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	103.0	4.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 19	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	108.0	4.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 20	1.0	1.0	2.0	0.0	0.0	0.0	0.0	0.0	111.0	5.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 21	1.0	1.0	4.0	2.0	0.0	0.0	0.0	0.0	119.0	4.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 22	2.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	106.0	4.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 23	1.0	1.0	6.0	17.0	0.0	0.0	0.0	0.0	185.0	4.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	14.0	0.0	0.0	0.0
Page 24	1.0	1.0	5.0	14.0	1.0	0.0	0.0	0.0	162.0	4.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	16.0
Page 25	1.0	1.0	7.0	6.0	0.0	0.0	0.0	0.0	139.0	4.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	11.0
Page 26	1.0	1.0	4.0	12.0	0.0	0.0	0.0	0.0	142.0	6.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	8.0
Page 27	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	116.0	4.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6.0
Page 28	1.0	1.0	4.0	9.0	0.0	0.0	0.0	0.0	129.0	8.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	26.0

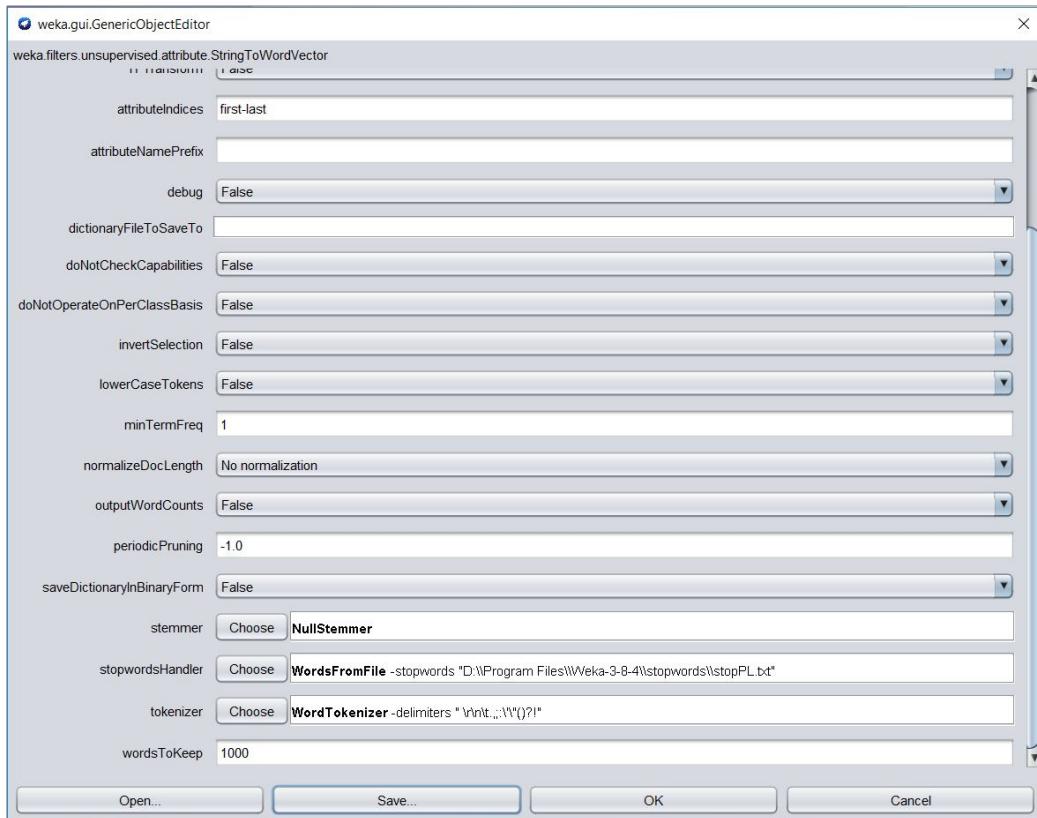
Tabela 3. Parametry zastosowanego filtra StringToWordVector dla przypadku 2.

Parametr	Wartość
outputWordCounts	false
StopwordsHandler	true
TFTransform	false
IDFTransform	false

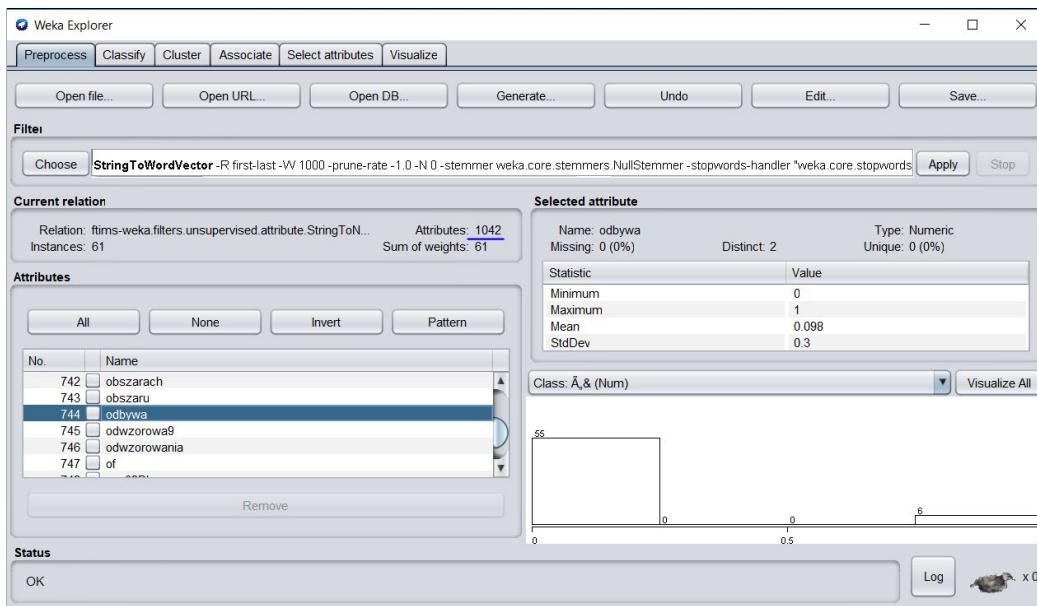
Rysunek 18 przedstawia statystyki atrybutów przed zastosowaniem mechanizmu StopwordsHandler. Po aktywowaniu tego filtra widać zmianę ilość atrybutów oraz brak wcześniejsza zaobserwowanego terminu ("od") Rysunek 10.



Rysunek 8. Statystki zbioru przed zastosowaniem mechanizmu StopwordsHandler



Rysunek 9. Ustawienia mechanizmu StopwordsHandler



Rysunek 10. Statystki zbioru po zastosowaniu mechanizmu StopwordsHandler

■ Przypadek 3. - TFTransform

TFTransform odnosi się do terminu TF (ang. term frequency) oznaczającego częstotliwość wystąpienia danego termu w obrębie dokumentu. Na Rysunku 11 przedstawiono obliczoną miarę TF.

Ten sposób reprezentacji posiada pewne luki. Termy o wysokiej częstotliwości, które występują w wielu dokumentach stają się bezużyteczne przez swoją uniwersalność i powszechność. Jeżeli są obecne w wielu dokumentach powoduje to problemy z ich rozróżnieniem. Stąd ważenie częstością termów wskazuje lokalne znaczenie termów w danym dokumencie.

Tabela 4. Parametry zastosowanego filtra StringToWordVector dla przypadku 3.

Parametr	Wartość
outputWordCounts	false
StopwordsHandler	false
TFTransform	true
IDFTransform	false

Rysunek 11. Wyniki analizy domeny dla Przypadku 3.

■ Przypadek 4. - IDFTransform

IDFTransform nawiązuje do IDF (ang. inverse document frequency), czyli odwrotna częstość dokumentów. Celem tej transformacji jest uwzględnienie lokalnej częstości termów jak i jego znaczenia w kontekście całej puli dokumentów. Przekłada się to na zmniejszenie wagę termów, które pojawiają się w wielu dokumentach.

Jest to stosunek liczby wszystkich dokumentów do liczby dokumentów zawierających dany term. Oznacza to, że im rzadziej występuje dany term tym większą osiąga wartość IDF.

Tabela 5. Parametry zastosowanego filtra StringToWordVector dla przypadku 4.

Parametr	Wartość
outputWordCounts	false
StopwordsHandler	false
TFTransform	false
IDFTransform	true

Dla termu numer 9., który występował bardzo licznie i w każdym dokumencie (Rysunek 7 i 15) wartość IDF jest równa zero.

Z kolei term 6. występował licznie, ale w niewielkiej liczbie dokumentów, wtedy wartość IDF jest wysoka (1.91)

Viewer		Relation: firms-weka.filters.unsupervised.attribute.StringToNominal-Rfist-weka.filters.unsupervised.attribute.StringToWordVector-R2-W1000-prune-rate-1.0-i-NO-stemmerweka.core stemmers.NullStemmer-stopwords-handlereweka.core stop																				
1: Title	2: #	3: ##	4: ###	5: #####	6: #####	7: &	8: &cy;	9: *	10: **	11: **FTMS**	12: **Instytut	13: **P9A5**	14: **Przydatne	15: **Zak8taad	16: **dr	17: **e-mail	18: **ngr	19: **prof	20: **A	21: +8	22:	
Page 0	0.26...	0.03...	0.0	0.39...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 1	0.26...	0.03...	0.0	0.39...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 2	0.0	0.03...	0.37...	0.39...	1.9136...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 3	0.26...	0.03...	0.0	0.0	1.9136...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 4	0.0	0.03...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 5	0.26...	0.03...	0.37...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 6	0.0	0.03...	0.0	0.0	1.9136...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 7	0.0	0.03...	0.37...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 8	0.0	0.03...	0.0	0.0	0.0	0.0	1.9136...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 9	0.0	0.03...	0.0	0.0	0.0	0.0	1.9136...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 10	0.0	0.03...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 11	0.0	0.03...	0.37...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 12	0.0	0.03...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 13	0.0	0.03...	0.0	0.0	0.0	2.02...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 14	0.26...	0.03...	0.37...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 15	0.26...	0.03...	0.0	0.39...	1.9136...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 16	0.26...	0.03...	0.0	0.39...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 17	0.26...	0.03...	0.0	0.39...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 18	0.0	0.03...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 19	0.26...	0.03...	0.37...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 20	0.26...	0.03...	0.37...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 21	0.26...	0.03...	0.37...	0.39...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 22	0.26...	0.03...	0.0	0.39...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 23	0.26...	0.03...	0.37...	0.39...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.72...
Page 24	0.26...	0.03...	0.37...	0.39...	1.9136...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 25	0.26...	0.03...	0.37...	0.39...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 26	0.26...	0.03...	0.37...	0.39...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 27	0.0	0.03...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 28	0.26...	0.03...	0.27...	0.29...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Rysunek 12. Wyniki analizy domeny dla Przypadku 4.

■ Przypadek 5. - TFTransform i IDFTransform

Transformacje TF i IDF można połączyć. Wówczas następuje pomnożenie obu wartości.

Dzięki temu możemy znaleźć terminy występujące często w małej liczbie dokumentów (wartości maksymalne) np. term 16 czy 18 (Rysunek 13). Z kolei niskie wyniki otrzymamy dla tych pojawiających się w prawie każdym dokumencie, tj. term 3 lub 9.

Tabela 6. Parametry zastosowanego filtra StringToWordVector dla przypadku 5.

Parametr	Wartość
outputWordCounts	false
StopwordsHandler	false
TFTransform	true
IDFTransform	true

	1: Title	2: #	3: ##	4: ###	5: #####	6: #####	7: &	8: &cy;	9: *	10: **	11: **FTIMS**	12: *Instytut	13: **P9A5j**	14: *Przydatne	15: **ZakonTaa	16: *dr	17: *e-mail	18: *mgr	19: *prof	20: **Å	21: +48	22:
Nominal	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
Page 0	0.18...	0.02...	0.0	0.27...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01*
Page 1	0.18...	0.02...	0.0	0.27...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01*
Page 2	0.0	0.02...	0.25...	0.27...	1.3264...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01*
Page 3	0.18...	0.02...	0.0	0.0	1.3264...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01*
Page 4	0.0	0.02...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01*
Page 5	0.18...	0.02...	0.25...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01*
Page 6	0.0	0.02...	0.0	0.0	1.3264...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01*
Page 7	0.0	0.02...	0.25...	0.0	0.0	1.3264...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01*
Page 8	0.0	0.02...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01*
Page 9	0.0	0.02...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01*
Page 10	0.0	0.02...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01*
Page 11	0.0	0.02...	0.25...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01*
Page 12	0.0	0.02...	0.25...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01*
Page 13	0.0	0.02...	0.0	0.0	0.0	0.0	1.40...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01*
Page 14	0.18...	0.02...	0.25...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01*
Page 15	0.18...	0.02...	0.0	0.0	0.27...	1.3264...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01*
Page 16	0.18...	0.02...	0.0	0.0	0.27...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01*
Page 17	0.18...	0.02...	0.0	0.0	0.27...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01*
Page 18	0.0	0.02...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01*
Page 19	0.18...	0.02...	0.25...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01*
Page 20	0.18...	0.02...	0.25...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01*
Page 21	0.18...	0.02...	0.25...	0.27...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01*
Page 22	0.18...	0.02...	0.0	0.0	0.27...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01*
Page 23	0.18...	0.02...	0.25...	0.27...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01*
Page 24	0.18...	0.02...	0.25...	0.27...	1.3264...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01*
Page 25	0.18...	0.02...	0.25...	0.27...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01*
Page 26	0.18...	0.02...	0.25...	0.27...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01*
Page 27	0.0	0.02...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01*
Page 28	0.19...	0.02...	0.25...	0.27...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01*

Rysunek 13. Wyniki analizy domeny dla Przypadku 5.

■ Przypadek 6. - TFTransform, IDFTransform i outputWordCounts

Otrzymane wyniki z transformacji TF i IDF są potęgowane przez liczbę wystąpień termów. Efektem jest wyróżnienie wyższymi wartościami tych elementów bardziej pospolitych (term 18 - Dokument 5 i 11 - Rysunek 14)

Tabela 7. Parametry zastosowanego filtra StringToWordVector dla przypadku 6.

Parametr	Wartość
outputWordCounts	true
StopwordsHandler	false
TFTransform	true
IDFTransform	true

Rysunek 14. Wyniki analizy domeny dla Przypadku 6.

■ Przypadek 7. - Reprezentacja binarna

Dokument reprezentowany jest przez macierz binarną odpowiadającą wystąpieniu danego terminu. Wyniki (Rysunek 15) są tożsame z Rysunkiem 6, oraz interpretują dane z Rysunku 7, gdzie dla wartości większych od 0 ustawia "1". Taki sposób reprezentacji może być pomocny przy klasyfikacji dokumentów. W przypadku słów kluczowych nie jest to idealne rozwiązanie.

Tabela 8. Parametry zastosowanego filtra StringToWordVector dla przypadku 7.

Parametr	Wartość
outputWordCounts	false
StopwordsHandler	false
TFTransform	false
IDFTransform	false

Viewer

Relation: flms-weka.filters.unsupervised.attribute.StringToNominal-Rfirst-weka.filters.unsupervised.attribute.StringToWordVector-R2-W1000-prune-rate-1.0-NO-stemmerweka.core.stemmers.NullStemmer-stopwords-handlerveka.core.stopwords

1: Title	2: #_binarized	3: ##_binarized	4: ###_binarized	5: ####_binarized	6: #####_binarized	7: &_binarized	8: &cy_binarized	9: _binarized	10: **_binarized	11: **FTIMS**_binarized	12: **Instytut**_binarized	13: **P9A5**_binarized
Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal
Page 0	1	1	0	1	0	0	0	1	1	0	0	0
Page 1	1	1	0	1	0	0	0	1	1	0	0	0
Page 2	0	1	1	1	1	0	0	1	1	0	0	0
Page 3	1	1	0	0	1	0	0	1	1	0	0	0
Page 4	0	1	0	0	0	0	0	1	1	0	0	0
Page 5	1	1	0	0	0	0	0	1	1	0	0	0
Page 6	0	1	0	0	1	0	0	1	1	0	0	0
Page 7	0	1	1	0	0	0	0	1	1	0	0	0
Page 8	0	1	0	0	1	0	0	1	1	0	0	0
Page 9	0	1	0	1	1	0	0	1	1	0	0	0
Page 10	0	1	0	0	0	0	0	1	1	0	0	0
Page 11	0	1	1	0	0	0	0	1	1	0	0	0
Page 12	0	1	0	0	0	0	0	1	1	0	0	0
Page 13	0	1	0	0	0	0	1	0	1	1	0	0
Page 14	1	1	0	0	0	0	0	1	1	0	0	0
Page 15	1	1	0	1	1	0	0	1	1	0	0	0
Page 16	1	0	1	0	0	0	0	1	1	0	0	0
Page 17	1	1	0	1	0	0	0	1	1	0	0	0
Page 18	0	0	0	0	0	0	0	1	1	0	0	0
Page 19	1	1	0	0	0	0	0	1	1	0	0	0
Page 20	1	1	0	0	0	0	0	1	1	0	0	0
Page 21	1	1	1	1	0	0	0	0	1	1	0	0
Page 22	1	1	0	1	0	0	0	0	1	1	0	0
Page 23	1	1	1	1	0	0	0	1	1	0	0	0
Page 24	1	1	1	1	1	0	0	1	1	0	0	0
Page 25	1	1	1	1	0	0	0	0	1	1	0	0
Page 26	1	1	1	1	0	0	0	0	1	1	0	0
Page 27	0	1	0	0	0	0	0	0	1	1	0	0
Page 28	1	1	1	0	0	0	0	1	1	0	0	0

Add instance Undo OK Cancel

Rysunek 15. Wyniki analizy domeny dla Przypadku 7.

4.2. Analiza klastrowa

■ Przypadek 0.

Tabela 9. Parametry zastosowanego filtra StringToWordVector dla przypadku 0.

Parametr	Wartość
outputWordCounts	false
StopwordsHandler	false
TFTransform	false
IDFTransform	false

Tabela 10. Wyniki klasteryzacji dla przypadku 0.

klaster	Wartość
0	23 (38%)
1	16 (26%)
2	22 (36%)

Wartość prawdopodobieństwa : 2587.13578

■ Przypadek 1. - outputWordCount

Tabela 11. Parametry zastosowanego filtra StringToWordVector dla przypadku 1.

Parametr	Wartość
outputWordCounts	true
StopwordsHandler	false
TFTransform	false
IDFTransform	false

Tabela 12. Parametry zastosowanego filtra StringToWordVector dla przypadku 1.

Klaster	Wartość
0	44 (72%)
1	17 (28%)

Wartość prawdopodobieństwa : -533.66714

■ Przypadek 2. - Stop lista

Tabela 13. Parametry zastosowanego filtra StringToWordVector dla przypadku 2.

Parametr	Wartość
outputWordCounts	false
StopwordsHandler	true
TFTransform	false
IDFTransform	false

Tabela 14. Wynik analizy klastrowej dla przypadku 2.

Klaster	Wartość
0	16 (26%)
1	22 (36%)
2	23 (38%)

Wartość prawdopodobieństwa : 2403.668538

■ Przypadek 3. - TFTransform

Tabela 15. Parametry zastosowanego filtra StringToWordVector dla przypadku 3.

Parametr	Wartość
outputWordCounts	false
StopwordsHandler	false
TFTransform	true
IDFTransform	false

Tabela 16. Wynik analizy klastrowej dla przypadku 3.

Klaster	Wartość
0	16 (26%)
1	22 (36%)
2	23 (38%)

Wartość prawdopodobieństwa : -2920.08239

■ Przypadek 4. - IDFTransform

Tabela 17. Parametry zastosowanego filtra StringToWordVector dla przypadku 4.

Parametr	Wartość
outputWordCounts	false
StopwordsHandler	false
TFTransform	false
IDFTransform	true

Tabela 18. Wynik analizy klastrowej dla przypadku 4.

Klaster	Wartość
0	16 (26%)
1	22 (36%)
2	23 (38%)

Wartość prawdopodobieństwa : 2744.62806

■ Przypadek 5. - TFTransform i IDFTransform

Tabela 19. Parametry zastosowanego filtra StringToWordVector dla przypadku 5.

Parametr	Wartość
outputWordCounts	false
StopwordsHandler	false
TFTransform	true
IDFTransform	true

Tabela 20. Wynik analizy klastrowej dla przypadku 5.

Klaster	Wartość
0	16 (26%)
1	22 (36%)
2	23 (38%)

Wartość prawdopodobieństwa : 3079.46799

■ Przypadek 6. - TFTransform, IDFTransform i outputWordCounts

Tabela 21. Parametry zastosowanego filtra StringToWordVector dla przypadku 6.

Parametr	Wartość
outputWordCounts	true
StopwordsHandler	false
TFTransform	true
IDFTransform	true

Tabela 22. Wynik analizy klastrowej dla przypadku 6.

Klaster	Wartość
0	44 (72%)
1	17 (28%)

Wartość prawdopodobieństwa : 2515.94447

■ Przypadek 7. - Reprezentacja binarna

Tabela 23. Parametry zastosowanego filtra StringToWordVector dla przypadku 7.

Parametr	Wartość
outputWordCounts	false
StopwordsHandler	false
TFTransform	false
IDFTransform	false

Tabela 24. Wynik analizy klastrowej dla przypadku 7.

Klaster	Wartość
0	4 (7%)
1	3 (5%)
2	19 (31%)
3	19 (31%)
4	16 (26%)

Wartość prawdopodobieństwa : -212.27242

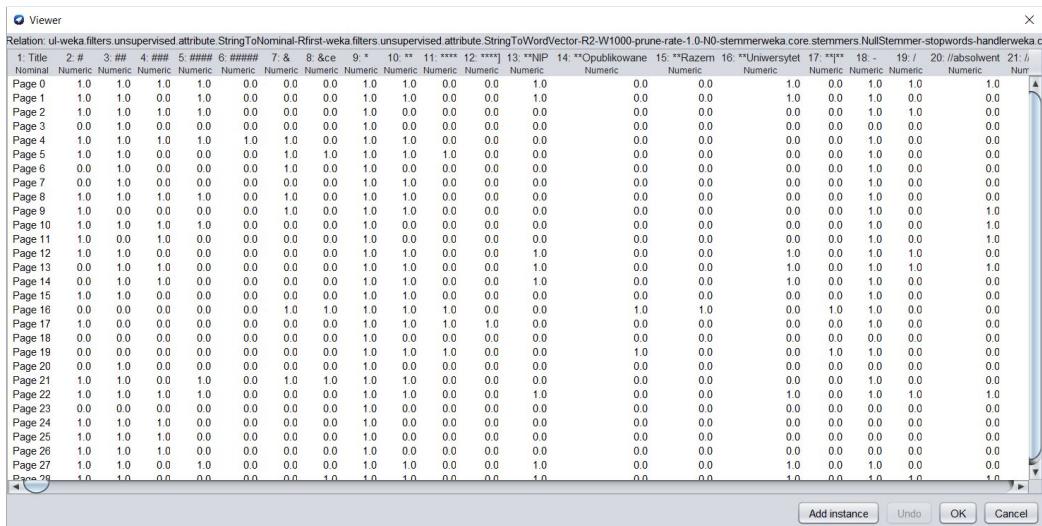
5. Wyniki analizy dla domeny UŁ

5.1. Analiza atrybutów

■ Przypadek 0.

Tabela 25. Parametry zastosowanego filtra StringToWordVector dla przypadku 0.

Parametr	Wartość
outputWordCounts	false
StopwordsHandler	false
TFTransform	false
IDFTransform	false



Rysunek 16. Wyniki analizy domeny dla Przypadku 0.

■ Przypadek 1. - outputWordCount

Tabela 26. Parametry zastosowanego filtra StringToWordVector dla przypadku 1.

Parametr	Wartość
outputWordCounts	true
StopwordsHandler	false
TFTransform	false
IDFTransform	false

1:Title	2:#	3:##	4:###	5:####	6:#####	7:&c	8:&ce	9:*	10:**	11:***	12:****	13:**NIP	14:**Opublikowane	15:**Razem	16:**Universytet	17:****	18:-	19:/	20://absolwent	21:/bis	Nu
Nominal	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric												
Page 0	6.0	1.0	11.0	18.0	0.0	0.0	0.0	164.0	2.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	12.0	2.0	5.0		
Page 1	1.0	1.0	0.0	0.0	0.0	0.0	0.0	65.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0		
Page 2	5.0	1.0	2.0	1.0	0.0	0.0	0.0	102.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	1.0	0.0		
Page 3	0.0	1.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
Page 4	1.0	4.0	2.0	3.0	26.0	1.0	0.0	149.0	11.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
Page 5	2.0	3.0	0.0	0.0	0.0	1.0	1.0	79.0	2.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	6.0	0.0	0.0		
Page 6	0.0	13.0	0.0	0.0	0.0	1.0	0.0	96.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.0	0.0	0.0		
Page 7	0.0	5.0	0.0	0.0	0.0	0.0	0.0	22.0	21.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0		
Page 8	1.0	1.0	1.0	14.0	0.0	1.0	0.0	63.0	4.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	7.0	0.0	0.0		
Page 9	1.0	0.0	0.0	0.0	0.0	4.0	0.0	44.0	23.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6.0	0.0	1.0		
Page ...	2.0	1.0	2.0	2.0	0.0	0.0	0.0	68.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	48.0		
Page ...	2.0	0.0	4.0	0.0	0.0	0.0	0.0	77.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	47.0		
Page ...	2.0	2.0	0.0	0.0	0.0	0.0	0.0	82.0	5.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	3.0	4.0	0.0		
Page ...	0.0	8.0	2.0	0.0	0.0	0.0	0.0	150.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	4.0	1.0	5.0		
Page ...	0.0	6.0	2.0	0.0	0.0	0.0	0.0	31.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	3.0	0.0	0.0		
Page ...	2.0	14.0	0.0	0.0	0.0	0.0	0.0	57.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	10.0	0.0	0.0		
Page ...	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	14.0	38.0	8.0	0.0	0.0	2.0	6.0	0.0	3.0	1.0	0.0		
Page ...	1.0	0.0	0.0	0.0	0.0	0.0	0.0	69.0	6.0	3.0	8.0	0.0	0.0	0.0	0.0	0.0	5.0	0.0	0.0		
Page ...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
Page ...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	12.0	16.0	9.0	0.0	0.0	2.0	0.0	0.0	3.0	9.0	0.0	0.0		
Page ...	0.0	1.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
Page ...	1.0	11.0	0.0	1.0	0.0	8.0	1.0	44.0	11.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0		
Page ...	1.0	1.0	1.0	5.0	0.0	0.0	0.0	135.0	2.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	6.0	1.0	5.0		
Page ...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
Page ...	1.0	1.0	37.0	0.0	0.0	0.0	0.0	49.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
Page ...	1.0	1.0	48.0	0.0	0.0	0.0	0.0	60.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
Page ...	1.0	1.0	25.0	0.0	0.0	0.0	0.0	37.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
Page ...	1.0	1.0	0.0	8.0	0.0	0.0	0.0	65.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0		
Page ...	1.0	1.0	0.0	0.0	0.0	0.0	0.0	123.0	2.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	5.0		

Rysunek 17. Wyniki analizy domeny dla Przypadku 1.

■ Przypadek 2. - Stop lista

Funkcja StopwordsHandler wykorzystuje wcześniej utworzony plik tekstuowy zawierający popularne stop słów dla języka polskiego.

Link do źródła: <https://github.com/bielis/stopwords/blob/master/polish.stopwords.txt>

Tabela 27. Parametry zastosowanego filtra StringToWordVector dla przypadku 2.

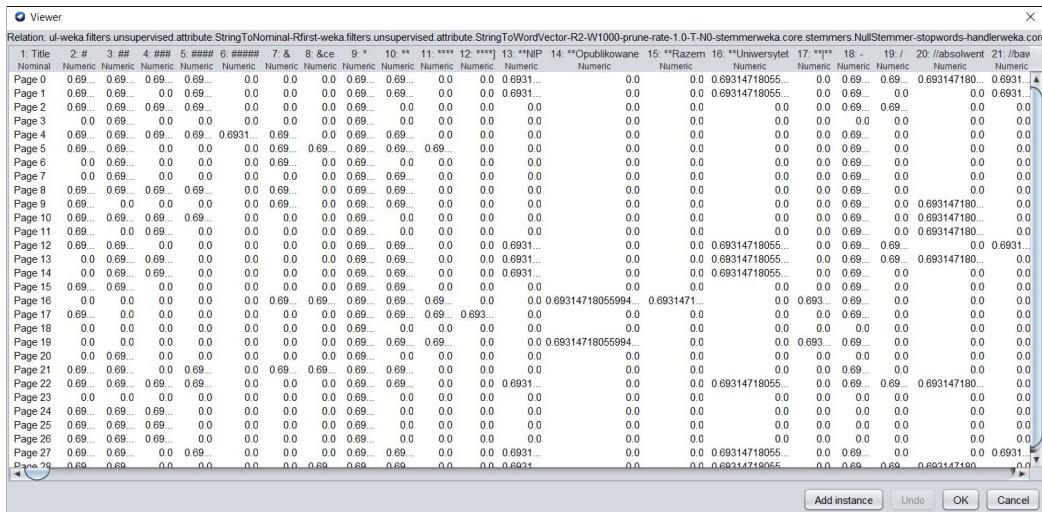
Parametr	Wartość
outputWordCounts	false
StopwordsHandler	true
TFTransform	false
IDFTransform	false

1:Title	2:#	3:##	4:###	5:####	6:#####	7:&c	8:&ce	9:*	10:**	11:***	12:****	13:**NIP	14:**Opublikowane	15:**Razem	16:**Universytet	17:****	18:-	19:/	20://absolwent	21:/bis	Nu
Nominal	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric												
Page 0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0
Page 1	1.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0
Page 2	1.0	1.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0
Page 3	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 4	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Page 5	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Page 6	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Page 7	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Page 8	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Page 9	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Page 10	1.0	1.0	1.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0
Page 11	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 12	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0
Page 13	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0
Page 14	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
Page 15	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
Page 16	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	0.0	0.0	1.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0
Page 17	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
Page 18	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 19	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0
Page 20	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Page 21	1.0	1.0	0.0	1.0	0.0																

■ Przypadek 3. - TFTransform

Tabela 28. Parametry zastosowanego filtra StringToWordVector dla przypadku 3.

Parametr	Wartość
outputWordCounts	false
StopwordsHandler	false
TFTransform	true
IDFTransform	false



Rysunek 19. Wyniki analizy domeny dla Przypadku 3.

■ Przypadek 4. - IDFTransform

Tabela 29. Parametry zastosowanego filtra StringToWordVector dla przypadku 4.

Parametr	Wartość
outputWordCounts	false
StopwordsHandler	false
TFTransform	false
IDFTransform	true

Rysunek 20. Wyniki analizy domeny dla Przypadku 4.

■ Przypadek 5. - TFTransform i IDFTransform

Tabela 30. Parametry zastosowanego filtra `StringToWordVector` dla przypadku 5.

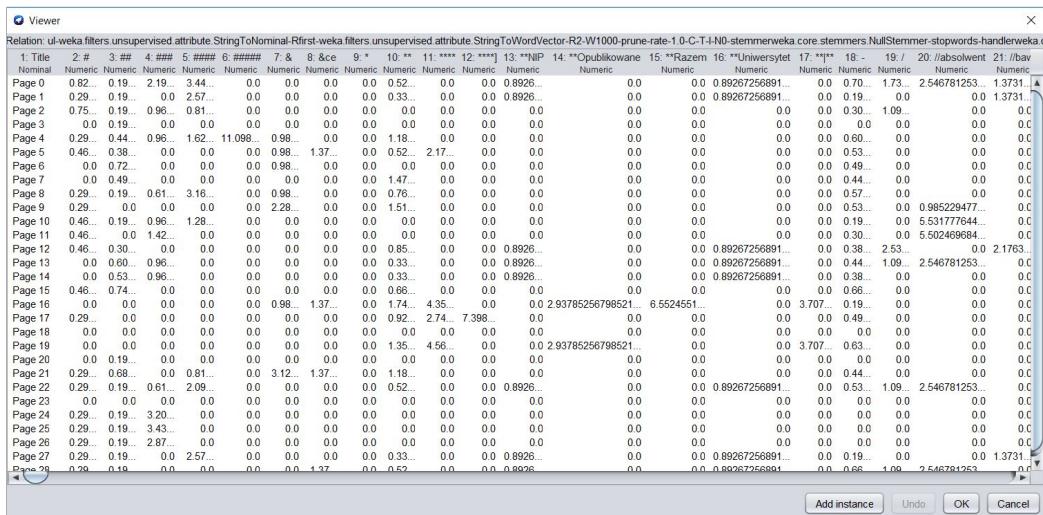
Parametr	Wartość
outputWordCounts	false
StopwordsHandler	false
TFTransform	true
IDFTransform	true

Rysunek 21. Wyniki analizy domeny dla Przypadku 5.

■ Przypadek 6. - TFTransform, IDFTransform i outputWordCounts

Tabela 31. Parametry zastosowanego filtra StringToWordVector dla przypadku 6.

Parametr	Wartość
outputWordCounts	true
StopwordsHandler	false
TFTransform	true
IDFTransform	true



Rysunek 22. Wyniki analizy domeny dla Przypadku 6.

■ Przypadek 7. - Reprezentacja binarna

Tabela 32. Parametry zastosowanego filtra StringToWordVector dla przypadku 7.

Parametr	Wartość
outputWordCounts	false
StopwordsHandler	false
TFTransform	false
IDFTransform	false

Viewer

Relation: ul-weka.filters.unsupervised.attribute.StringToNominal-Rfrst-weka.filters.unsupervised.attribute.StringToWordVector-R2-W1000-prune-rate-1.0-N0-stemmer-weka.core.stemmers.NullStemmer-stopwords-handleneweka.core.stopwords.NullM1-tokenizerweka.core.tostring

	1: Title	2: #_binarized	3: ##_binarized	4: ###_binarized	5: #####_binarized	6: #####_binarized	7: &_binarized	8: &ce_binarized	9: *_binarized	10: **_binarized	11: ***_binarized	12: ****_binarized	13: *****_binarized	14: **Opublikowane_binarized	15: Nominal
Page 0	1	1	1	1	0	0	0	1	1	0	0	1	0	0	0
Page 1	1	1	1	1	0	0	0	1	0	0	0	0	0	0	0
Page 2	1	1	1	1	0	0	0	1	0	0	0	0	0	0	0
Page 3	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0
Page 4	1	1	1	1	1	1	0	1	1	1	1	1	0	0	0
Page 5	1	1	0	0	0	0	1	1	1	1	1	1	0	0	0
Page 6	1	1	0	0	0	0	1	1	0	0	0	0	0	0	0
Page 7	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0
Page 8	1	1	1	1	0	1	0	1	1	0	0	0	0	0	0
Page 9	1	0	0	0	0	1	0	1	1	0	0	0	0	0	0
Page 10	1	1	1	1	0	0	0	0	1	0	0	0	0	0	0
Page 11	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0
Page 12	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0
Page 13	0	1	0	0	0	0	0	0	1	1	0	0	1	0	0
Page 14	0	1	1	0	0	0	0	0	1	1	0	0	1	0	0
Page 15	1	1	0	0	0	0	0	0	1	1	0	0	0	0	0
Page 16	0	0	0	0	0	1	1	1	1	1	0	0	1	1	1
Page 17	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0
Page 18	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0
Page 19	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
Page 20	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0
Page 21	1	0	1	0	0	1	1	1	1	0	0	0	0	0	0
Page 22	1	1	1	0	0	0	0	0	1	1	0	0	0	1	0
Page 23	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
Page 24	1	1	1	0	0	0	0	0	1	0	0	0	0	0	0
Page 25	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0
Page 26	1	1	1	0	0	0	0	0	1	0	0	0	0	0	0
Page 27	1	1	0	0	0	0	0	1	1	0	0	1	0	0	0
Page 28	1	1	0	0	0	0	0	1	1	0	0	0	1	0	0

Add instance Undo OK Cancel

Rysunek 23. Wyniki analizy domeny dla Przypadku 7.

5.2. Analiza klastrowa

■ Przypadek 0.

Tabela 33. Parametry zastosowanego filtra StringToWordVector dla przypadku 0.

Parametr	Wartość
outputWordCounts	false
StopwordsHandler	false
TFTransform	false
IDFTransform	false

Tabela 34. Wyniki klasteryzacji dla przypadku 0.

klaster	Wartość
0	18 (62%)
1	2 (7%)
2	2 (7%)
3	3 (10%)
4	4 (14%)

Wartość prawdopodobieństwa : 56.27762

■ Przypadek 1. - outputWordCount

Tabela 35. Parametry zastosowanego filtra StringToWordVector dla przypadku 1.

Parametr	Wartość
outputWordCounts	true
StopwordsHandler	false
TFTransform	false
IDFTransform	false

Tabela 36. Parametry zastosowanego filtra StringToWordVector dla przypadku 1.

Klaster	Wartość
0	1 (3%)
1	2 (10%)
2	25 (86%)

Wartość prawdopodobieństwa : -621.09288

■ Przypadek 2. - Stop lista

Tabela 37. Parametry zastosowanego filtra StringToWordVector dla przypadku 2.

Parametr	Wartość
outputWordCounts	false
StopwordsHandler	true
TFTransform	false
IDFTransform	false

Tabela 38. Wynik analizy klastrowej dla przypadku 2.

Klaster	Wartość
0	3 (10%)
1	22 (76%)
2	4 (14%)

Wartość prawdopodobieństwa : -11.73055

■ Przypadek 3. - TFTransform

Tabela 39. Parametry zastosowanego filtra StringToWordVector dla przypadku 3.

Parametr	Wartość
outputWordCounts	false
StopwordsHandler	false
TFTransform	true
IDFTransform	false

Tabela 40. Wynik analizy klastrowej dla przypadku 3.

Klaster	Wartość
0	18 (62%)
1	2 (7%)
2	2 (7%)
3	3 (10%)
4	4 (14%)

Wartość prawdopodobieństwa : 552.1696

■ Przypadek 4. - IDFTransform

Tabela 41. Parametry zastosowanego filtra StringToWordVector dla przypadku 4.

Parametr	Wartość
outputWordCounts	false
StopwordsHandler	false
TFTransform	false
IDFTransform	true

Tabela 42. Wynik analizy klastrowej dla przypadku 4.

Klaster	Wartość
0	18 (62%)
1	2 (7%)
2	2 (7%)
3	3 (10%)
4	4 (14%)

Wartość prawdopodobieństwa : -855.26944

■ Przypadek 5. - TFTransform i IDFTransform

Tabela 43. Parametry zastosowanego filtra StringToWordVector dla przypadku 5.

Parametr	Wartość
outputWordCounts	false
StopwordsHandler	false
TFTransform	true
IDFTransform	true

Tabela 44. Wynik analizy klastrowej dla przypadku 5.

Klaster	Wartość
0	18 (62%)
1	2 (7%)
2	2 (7%)
3	3 (10%)
4	4 (14%)

Wartość prawdopodobieństwa : -359.37746

■ Przypadek 6. - TFTransform, IDFTransform i outputWordCounts

Tabela 45. Parametry zastosowanego filtra StringToWordVector dla przypadku 6.

Parametr	Wartość
outputWordCounts	true
StopwordsHandler	false
TFTransform	true
IDFTransform	true

Tabela 46. Wynik analizy klastrowej dla przypadku 6.

Klaster	Wartość
0	3 (10%)
1	21 (72%)
2	1 (3%)
3	3 (10%)
4	1 (3%)

Wartość prawdopodobieństwa : 226.22116

■ Przypadek 7. - Reprezentacja binarna

Tabela 47. Parametry zastosowanego filtra StringToWordVector dla przypadku 7.

Parametr	Wartość
outputWordCounts	false
StopwordsHandler	false
TFTransform	false
IDFTransform	false

Tabela 48. Wynik analizy klastrowej dla przypadku 7.

Klaster	Wartość
0	22 (76%)
1	4 (14%)
2	3 (10%)

Wartość prawdopodobieństwa : 226.22116

6. Dyskusja

6.1. Analiza atrybutów

W przypadku analizy parametrów filtra warto zauważyć, że w trzech przypadkach (Przypadek 0, 1 i 7) wyniki były identyczne, bądź niosły podobną informację. Reprezentacja binarna jest tożsama dla wyników filtra bezparametrowego i przedstawia macierz obecności termów, przez co przekazują bardzo mało wiadomości. Wyniki z zastosowaniem opcji outputWordCount są rozszerzeniem informacji zawartych w ww. o liczbę wystąpień.

Mechanizm TFTransform sprawdza się dużo lepiej w przypadku wykorzystanych przez nas witryn. Pozwala wyciągnąć znacznie więcej wniosków. Jednak posiada pewną wadę. Im częściej dane słowo występuje w tekście tym bardziej traci na wartości, co ma negatywny efekt w przypadku wyszukiwania istotnych dokumentów.

IDFTransform pozwoliła na uwiarygodnienie elementów na tle całego zbioru. Im mniejsza liczba dokumentów, które zawierają dany term, tym wyższa wartość IDF.

Połączenie obu rozwiązań stanowi bardzo dobre i szeroko stosowane rozwiązanie. Pozwala na odfiltrowanie popularnych terminów. Wysoką wagę TFIDF uzyskuje się dzięki wysokiej częstotliwości termów (w danym dokumencie) i małej częstości występowania tego termu w całym zbiorze dokumentów. Im wyższa liczbowa wartość wagi, tym rzadszy term. Im mniejsza waga, tym bardziej powszechny.

Dzięki słowom o dużej wadze TFIDF, treści zawsze będą znajdować się na szczytowym wyniku wyszukiwania.

6.2. Analiza klastrowa

Dla każdej z domen można wyróżnić kilka rodzajów zbudowanych klastrów. Należy zwrócić uwagę, że kolejne kombinacje TFTransform i IDFTransform dawały identyczne wyniki. Podobna sytuacja nastąpiła dla przypadków 1 i 6, które oba mają aktywną opcję liczenia wystąpień termów.

Tabela 49. Podsumowanie charakterystyki wyników klasterowania dla domeny <https://ftims.p.lodz.pl>

Przypadek testowy	Charakterystyka klastrów
0, 2, 3, 4, 5	Trzy równomierne klastry
1, 6	Jeden duży, dwa małe
7	Łącznie siedem klastrów, dwa bardzo małe

Tabela 50. Podsumowanie charakterystyki wyników klasterowania dla domeny <https://www.uni.lodz.pl>

Przypadek testowy	Charakterystyka klastrów
0, 3, 4, 5	Jeden duży i cztery bardzo małe
1, 2, 6, 7	Jeden duży, dwa bardzo małe

Dla pierwszej domeny przeważa podział na trzy klastry. Warto zauważyć, że przypadek 6 połączył dwa z trzech poprzednich klastrów w jeden, przy zachowaniu wysokiej wartości prawdopodobieństwa. Może to oznaczać, że istnieje mała różnica pomiędzy tymi grupami, zaś parametr wordCount był tym przeważającym na korzyść dwóch. W większości przypadków występowała wysoka wartość prawdopodobieństwa (ang. likelihood), parametru algorytmu EM, który mówi: im wyższy tym lepsze dopasowanie modelu.

W przypadku drugiej domeny, w każdym teście istniał klaster, który zawierał ok. 70% wszystkich obserwacji. Wówczas pozostałe były bardzo małe.

Warto zaznaczyć, że w przypadku tej domeny większość przypadków cechowała niska wartość likelihood.

7. Podsumowanie

Wyszukiwarki internetowe wykorzystują wiele technik przeszukiwania sieci w celu znalezienia odpowiednich treści. Dzięki robotom internetowym jesteśmy w stanie zbadać zawartość i strukturę strony. Stosując odpowiednie techniki eksploracji danych możemy przeanalizować każdy dokument - w tym jego elementy, którym możemy nadać etykiety, które zawierają najbardziej istotne i wartościowe informacje. Wykorzystując analizę skupień można wydobyć tematy dokumentów oraz szybko wyszukać i przefiltrować informacje.