

Politechnika Warszawska

W Y D Z I A Ł M A T E M A T Y K I
I N A U K I N F O R M A C Y J N Y C H



Praca dyplomowa magisterska

na kierunku Matematyka
w specjalności Statystyka Matematyczna i Analiza Danych

Statystyczne metody i narzędzia oceny trudności tekstu w języku
polskim.

Karolina Marcinkowska

Numer albumu 245453

promotor

dr hab. inż. Przemysław Biecek

WARSZAWA 2018

.....
.....

podpis promotora

.....
.....

podpis autora

Streszczenie

Statystyczne metody i narzędzia oceny trudności tekstu w języku polskim.

Celem pracy jest budowa metodologii oraz narzędzi, które pomoże w lokalizacji potencjalnie trudnych słów dla dzieci w wieku szkolnym. Przez trudność rozumiana jest między innymi częstość występowania danego słowa w bazie tekstów dziecięcych, jak również jego długość oraz rozkład poszczególnych części mowy.

Pierwszym etapem pracy jest analiza literatury oraz przegląd dostępnych narzędzi dla języka polskiego. Następnie zostanie zbudowany reprezentatywny korpus tekstów.

Bazując na własnych wynikach oraz wynikach z literatury, opracowana zostanie metoda oceny trudności tekstu. Dzięki temu, autor np. scenariusza wykładu, będzie mógł dane słowo zamienić na inne – prostsze, lub spróbować tę samą wiedzę przekazać w inny sposób. Metodologia zostanie przedstawiona w formie aplikacji, która będzie korzystała ze stworzonej bazy tekstów (słownika). Baza będzie składała się między innymi z ogólnodostępnych tekstów dla dzieci w formie elektronicznej, a także materiałów udostępnionych przez Centrum Nauki Kopernik.

Zbudowana metodologia oparta będzie na długości (liczba sylab) oraz różnorodności wyrazów. W tym celu wykorzystane zostaną takie indeksy jak np. FOG, ale także metody klasyfikacji nienadzorowanej.

Słowa kluczowe: analiza trudności języka, statystyka obliczeniowa, segmentacja

Abstract

Statistical methods and tools in assessing documents readability.

The aim of this thesis is to build the tool for locating difficult words for children between 7 and 13 years old. The definition of word difficulty is based on how frequent the given word occurs in the database of texts written for children as well as on text's length and distribution of parts of speech.

The first step is to review already presented results and tools for Polish language. Then, we will build our own database of texts and provide methods for assessing text difficulty. Thanks to these methods, authors of exercise scenarios or lectures will be able to change some difficult words in it to easier one what will result in conveying the same knowledge in more understandable way.

Built methods will be presented in an application which will use the database prepared beforehand. The database will consist of open for general use texts for children and materials provided by Centrum Nauki Kopernik.

The methods will be based on the length of the words and their variety. To measure these features we will use such indexes as FOG but also methods of unsupervised learning.

Keywords: text difficulty analysis, computational statistics, segmentation

Warszawa, dnia

Oświadczenie

Oświadczam, że pracę magisterską pod tytułem „Statystyczne metody i narzędzia oceny trudności tekstu w języku polskim.”, której promotorem jest dr hab. inż. Przemysław Biecek, wykonałam samodzielnie, co poświadczam własnoręcznym podpisem.

.....

Spis treści

Wstęp	11
1. Indeksy lingwistyczne oraz narzędzia analizujące trudność tekstu dla języka polskiego	12
1.1. Indeksy lingwistyczne określające złożoność tekstu	12
1.1.1. Współczynnik mglistości FOG	12
1.1.2. Flesch Reading Ease Score (FRES)	13
1.1.3. Indeks Dalle'go-Chall'a	13
1.1.4. Indeks Pisarka	14
1.2. Narzędzia do analizy złożoności tekstu dla języka polskiego	15
1.2.1. Jasnopis.pl	15
1.2.2. Logios.pl	16
2. Algorytmy uczenia nienadzorowanego	18
2.1. Ocena liczby skupień	18
2.2. Miary odmiенноści	21
2.3. K-średnich	23
2.4. Metody hierarchiczne	24
2.5. Miary jakości podziału	24
2.5.1. Miary wewnętrzne	25
2.5.2. Miary stabilności	26
3. Grupowanie tekstu pod względem trudności	28
3.1. Baza tekstów oraz jej wstępne przetworzenie	28
3.2. Eksploracja danych	30
3.2.1. Wstępna analiza danych – wyznaczenie liczby klas	31
3.2.2. Grupowanie oraz ocena jego jakości	34
3.2.3. Charakterystyka grup obserwacji wyznaczonych przez model k-średnich	37
4. TrudneSłówka – opis aplikacji	40

Wstęp

Żyjemy w świecie cyfryzacji i automatyzacji, gdzie komputery i maszyny wykonują za nas czynności, które jeszcze niedawno robiliśmy sami. Jedną z bardziej żmudnych i czasochłonnych z nich jest weryfikacja poprawności i zrozumiałości tekstu pisanej, czyli tak zwana moderacja. Teksty pisane są najpowszechniejszą formą komunikacji w urzędach, szkołach, czy wielu zakładach pracy i o ile dla nas, dorosłych, większość tekstów jest zrozumiała, o tyle dziecko nie będzie w stanie przeczytać ze zrozumieniem np. fragmentu Kodeksu Pracy. Pojawia się zatem pytanie, czy jesteśmy w stanie ocenić, w sposób automatyczny, przejrzystość i zrozumiałość tekstu, który otrzymujemy? Jeśli tak, to w jaki sposób należy te cechy mierzyć i czy informacja, że napisany przez nas tekst jest trudny w odbiorze, pomoże nam lepiej dostosować go do odbiorcy?

Problemem tym zajęli się Broda i inni w [9], jednak skupili się na tekstach użytkowych dla dorosłych. Ciekawym tematem wydaje się jednak podobna analiza przeprowadzona tylko dla tekstów skierowanych do dzieci, ponieważ dla nich, nawet krótki tekst, może okazać się niezrozumiały, jeżeli zawiera słowa, których dziecko wcześniej nie znało lub które zostały użyte w nowym kontekście.

W pierwszym rozdziale pracy dokonano przeglądu indeksów lingwistycznych, określających trudność tekstu, a także opisano istniejące już narzędzia do analizy trudności tekstu dla języka polskiego. Rozdział drugi przedstawia przegląd algorytmów uczenia nienadzorowanego. Trzeci rozdział opisuje eksplorację danych – tutaj została przedstawiona konstrukcja bazy danych oraz zbadano jej własności, a także dobrano odpowiedni model do określenia trudności tekstu. Rozdział czwarty zawiera opis narzędzia do analizy trudności tekstu oraz uzyskane na jego podstawie wyniki.

Celem niniejszej pracy jest zatem budowa metodologii, a w konsekwencji narzędzia, które można będzie wykorzystać do analizy zrozumiałości i czytelności tekstu pisanej dla dzieci w wieku szkolnym. Praca powstała we współpracy z Panią Katarzyną Potęgą z Centrum Nauki Kopernik.

1. Indeksy lingwistyczne oraz narzędzia analizujące trudność tekstu dla języka polskiego

W tym rozdziale zostaną omówione indeksy lingwistyczne, określające złożoność tekstu, a także opisane narzędzia badające jego złożoność dla języka polskiego.

1.1. Indeksy lingwistyczne określające złożoność tekstu

Badania nad złożonością, a co za tym idzie czytelnością tekstu, mają swój początek w XIX wieku w Stanach Zjednoczonych ([10]). Zauważono wówczas, iż poziom zrozumienia tekstu zależy od liczby słów i długości zdań w nim zawartych, jednak dopiero w połowie XX wieku powstały analityczne formuły, z których kilka zostanie przedstawionych poniżej. Większość z nich odnosi się do języka angielskiego, jednak zostały one również przystosowane do języka polskiego, np. indeks FOG.

Podstawowymi miarami złożoności tekstu są między innymi: średnia długość zdania (słowa) wyrażona w słowach (sylbach), liczba zdań, słów, akapitów w tekście, stosunek liczby przyimotników do rzeczowników lub rzeczowników do czasowników oraz wiele innych. Odpowiednie zestawienie tych miar pozwala na bardziej przejrzystą miarę czytelności tekstu.

1.1.1. Współczynnik mglistości FOG

Najbardziej popularną miarą jest współczynnik mglistości FOG, którego autorem jest Robert Gunnig ([11]). Jego wartość jest wyrażona jako liczba lat edukacji potrzebna do zrozumienia danego tekstu. Oryginalnie sformułowany dla języka angielskiego, dostosowany został do języka polskiego. Wyraża się wzorem:

$$FOG = 0.4 \times \left(\frac{\#\text{słów}}{\#\text{zdań}} + 100 \times \frac{\#\text{długich słów}}{\#\text{słów ogółem}} \right) \quad (1.1)$$

W języku angielskim za słowo długie uważa się słowo, które w formie słownikowej ma co najmniej trzy sylaby, w języku polskim zaś – cztery lub więcej. Tabela 1.1 przedstawia interpretację

1.1. INDEKSY LINGWISTYCZNE OKREŚLAJĄCE ZŁOŻONOŚĆ TEKSTU

wartości indeksu. Przykładowo, jeżeli dany tekst ma wartość indeksu $FOG = 7.5$, to będzie on zrozumiały dla osób, które kończą szkołę podstawową.

wartość	interpretacja
1 – 8	język bardzo prosty, zrozumiały dla uczniów szkoły podstawowej
9 – 12	język dość prosty, zrozumiały dla uczniów liceum
13 – 15	język dość trudny, zrozumiały na poziomie studiów licencjackich
16 – 17	język trudny, zrozumiały na poziomie studiów magisterskich
18 i więcej	język bardzo trudny, zrozumiały na poziomie studiów doktoranckich, specjalistyczny

Tablica 1.1: Przedziały określające klasy trudności ([11])

1.1.2. Flesch Reading Ease Score (FRES)

Indeks FRES jest sformułowany tylko dla języka angielskiego. Jego wartość wyznacza się na podstawie średniej liczby słów w zdaniu oraz średniej liczby sylab w słowie. Przyjmuje wartości z zakresu 0 – 100 i im mniejsza wartość, tym tekst jest bardziej trudny do zrozumienia.

$$FRES = 206.835 - 1.015 \times \left(\frac{\#\text{słów}}{\#\text{zdań}} - 84.6 \times \frac{\#\text{sylab}}{\#\text{słów}} \right) \quad (1.2)$$

Tabela 1.2 przedstawia wartości indeksu wraz z interpretacją dla amerykańskiego systemu nauczania podaną w nawiasach.

wartość	interpretacja
90 – 100	tekst zrozumiały dla uczniów w wieku 11 lat (V klasa)
80 – 90	tekst zrozumiały dla uczniów w wieku 12 lat (VI klasa)
70 – 80	tekst zrozumiały dla uczniów w wieku 13 lat (VII klasa)
60 – 70	tekst zrozumiały dla uczniów w wieku 14 – 15 lat (VIII-IX klasa)
50 – 60	tekst zrozumiały dla uczniów w wieku 16 – 18 lat (X-XII klasa, liceum)
30 – 50	tekst zrozumiały dla studentów
0 – 30	tekst zrozumiały dla absolwentów uczelni wyższych

Tablica 1.2: Wartości indeksu FRES wraz z ich interpretacją ([8])

1.1.3. Indeks Dalle'go-Chall'a

Indeks skonstruowany dla języka angielskiego, pozwala na ocenę czytelności i zrozumienia tekstu, konstruowany w oparciu o 3000 najpopularniejszych słów, które uczeń IV klasy szkoły

1. INDEKSY LINGWISTYCZNE ORAZ NARZĘDZIA ANALIZUJĄCE TRUDNOŚĆ TEKSTU DLA JĘZYKA POLSKIEGO

podstawowej zna i rozumie. Klasyfikacja dokonana została według angielskiego systemu nauczania, w Polsce – będzie to dziecko około 10-tego roku życia.

$$DC = 0.1579 \times \left(100 \times \frac{\#\text{trudnych słów}}{\#\text{słów}} \right) + 0.0496 \times \left(\frac{\#\text{słów}}{\#\text{zdań}} \right) + 3.6365 \quad (1.3)$$

Tabele 1.3 zawiera wartości indeksu wraz z jego interpretacją według amerykańskiego systemu nauczania ([5]).

wartość	interpretacja
< 4.9	tekst zrozumiały dla uczniów IV klasy i niższej
5.0 – 5.9	tekst zrozumiały dla uczniów klasy V lub VI
6.0 – 6.9	tekst zrozumiały dla uczniów klasy VII lub VIII
7.0 – 7.9	tekst zrozumiały dla uczniów klasy IX lub X
8.0 – 8.9	tekst zrozumiały dla uczniów klasy XI lub XII
9.0 – 9.9	tekst zrozumiały dla uczniów klasy XII i wyżej

Tablica 1.3: Wartości indeksu Dall'ego-Chall'a wraz z interpretacją ([5])

1.1.4. Indeks Pisarka

Walery Pisarek był pionierem w dziedzinie badań nad analizą złożoności tekstów w języku polskim. Opracowany przez niego wzór uwzględnia średnią długość zdania, mierzoną liczbą wyrazów oraz procent wyrazów trudnych (tj. dłuższych niż czterosylabowe w formie słownikowej), przez co można go interpretować jako polski odpowiednik indeksu FRES. Tabela 1.4 zawiera wartości indeksu wraz z ich interpretacją.

$$P_{NL} = \frac{1}{2} \sqrt{\left(\frac{\#\text{słów}}{\#\text{zdań}} \right)^2 + \left(\frac{\#\text{trudnych słów}}{\#\text{słów}} \right)^2} \quad (1.4)$$

wartość	interpretacja
4 – 7	teksty bardzo łatwe
7.1 – 10	teksty łatwe
10.1 – 13	teksty średnie
13.1 – 16	teksty trudne
16.1 – 20	teksty bardzo trudne

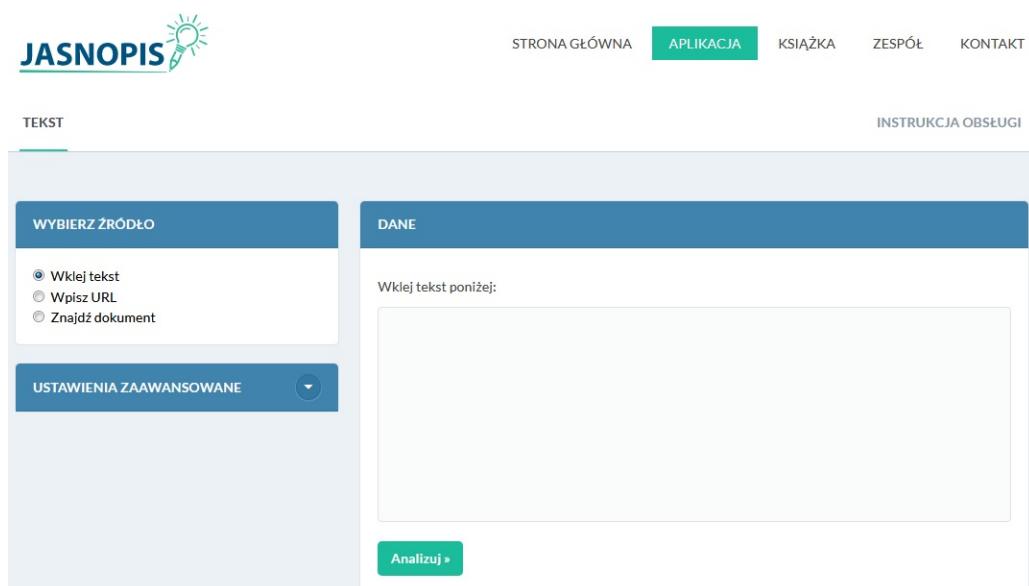
Tablica 1.4: Przedziały wartości indeksu Pisarka określające trudność tekstu ([15])

1.2. Narzędzia do analizy złożoności tekstu dla języka polskiego

Zostaną teraz krótko przedstawione dwa ogólnodostępne narzędzia, badające przejrzystość i zrozumiałość tekstów w języku polskim.

1.2.1. Jasnopis.pl

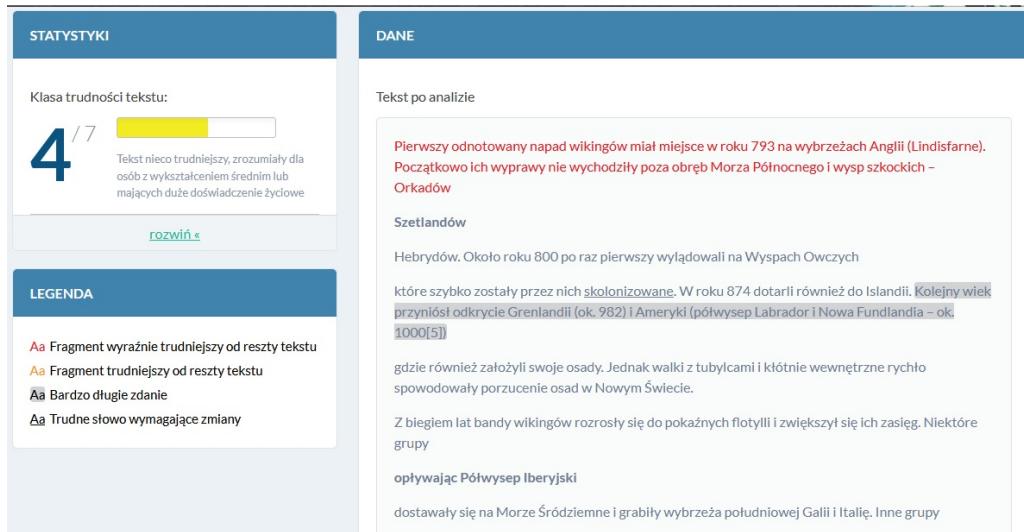
Projekt badawczy, realizowany w latach 2012 – 2014 przez Uniwersytet SWPS w Warszawie. Efektem pracy zespołu naukowców jest darmowe narzędzie informatyczne, które analizuje każdy tekst użytkowy (za wyjątkiem tekstów artystycznych, np. wierszy), podkreśla słowa i fragmenty potencjalnie trudne w odbiorze, a także zwraca zestaw statystyk dotyczących danego tekstu, takich jak indeks FOG, indeks Pisarka, liczba zdań, akapitów, słów, trudnych słów i inne (więcej w [7]). Autorzy za słowo trudne uznają takie, które w formie słownikowej ma cztery lub więcej sylab oraz nie należy do grupy słów najczęściej używanych.



Rysunek 1.1: Jasnopis.pl, ekran wejściowy

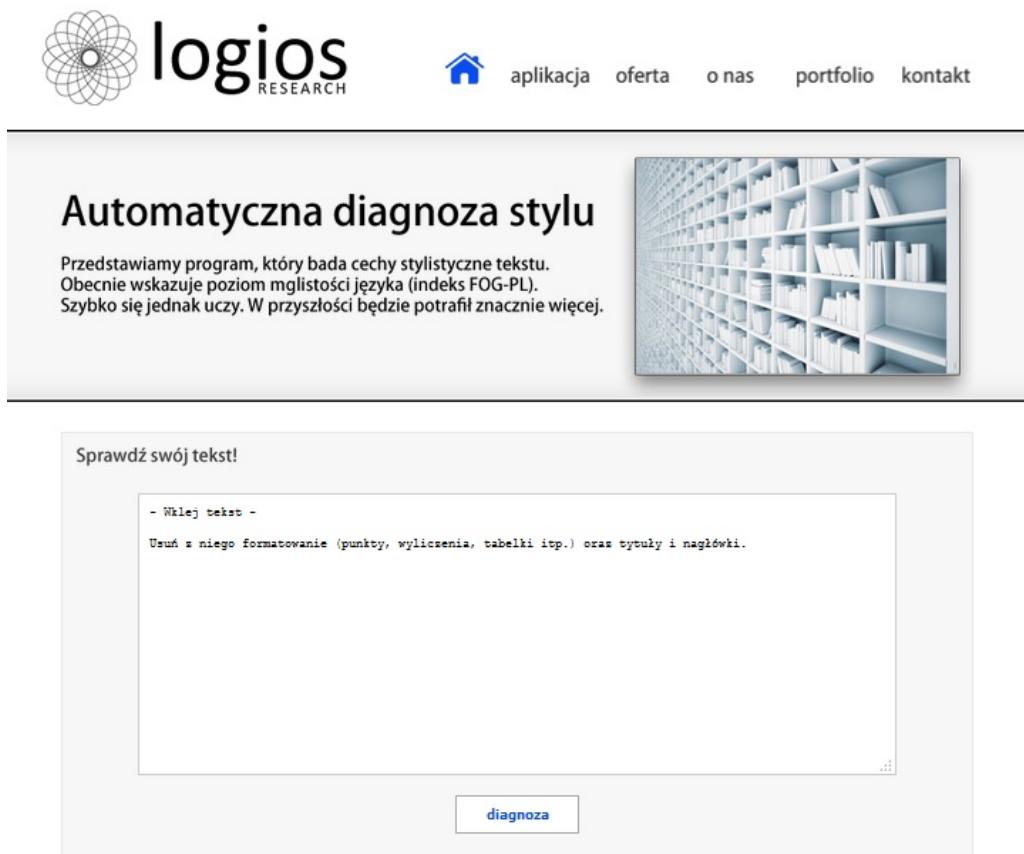
Rysunek 1.1 przedstawia ekran startowy aplikacji Jasnopis. Po lewej stronie mamy możliwość wyboru źródła tekstu, a po prawej znajduje się miejsce na analizowany przez nas tekst. Poniżej zaś (rys. 1.2), pokazano przykładowe wyniki analizy, na które składają się między innymi: klasa trudności tekstu wraz z opisem oraz zaznaczone fragmenty potencjalnie trudne w odbiorze.

1. INDEKSY LINGWISTYCZNE ORAZ NARZĘDZIA ANALIZUJĄCE TRUDNOŚĆ TEKSTU DLA JĘZYKA POLSKIEGO



Rysunek 1.2: Jasnopis.pl, wyniki dla fragmentu tekstu dotyczącego Wikingów ([18])

1.2.2. Logios.pl



Rysunek 1.3: Logios.pl, ekran wejściowy

Prosta aplikacja internetowa, wyznaczająca dla danego tekstu wartość indeksu FOG dostosowanego do języka polskiego. Oblicza ile lat nauki jest potrzebne, aby zrozumieć podany przez

1.2. NARZĘDZIA DO ANALIZY ZŁOŻONOŚCI TEKSTU DLA JĘZYKA POLSKIEGO

użytkownika tekstu.

W celu otrzymania wyników wystarczy wkleić wybrany tekst w okno dialogowe, a następnie nacisnąć przycisk 'diagnoza'. Po kilku sekundach na ekranie pojawia się krótki komunikat wystosowany do użytkownika, podający wartość indeksu FOG wraz z jego interpretacją. Przykładowa analiza została przedstawiona na rysunkach 1.3 oraz 1.4.



Rysunek 1.4: Logios.pl, wynik działania programu

2. Algorytmy uczenia nienadzorowanego

Celem pracy jest zaklasyfikowanie danego tekstu pod względem jego trudności i zrozumiałosci. W tym celu zostanie utworzona baza, zawierająca teksty skierowane do dzieci w wieku 7 – 13 lat. Jej konstrukcja oraz analiza zostały opisane w dalszej części pracy. Baza ta nie zawiera jednak zmiennej, określającej dwie wyżej wymienione cechy, czyli dysponujemy zbiorem bez zmiennej odpowiedzi. Aby móc stwierdzić, czy dany tekst jest trudny i zrozumiały, należy utworzyć zmienną grupującą teksty i właśnie w tym celu wykorzystane zostaną metody uczenia nienadzorowanego.

2.1. Ocena liczby skupień

Pierwszym etapem pracy ze zbiorem, który ma zostać pogrupowany, jest sprawdzenie czy oraz w jakiej liczbie występują w nim naturalne skupiska obserwacji. Jednym ze sposobów jest skorzystanie ze statystyki Hopkinaса ([1]), która mierzy prawdopodobieństwo, że dane pochodzą z rozkładu jednostajnego. Założymy, że dany jest zbiór D . Statystyka Hopkinaса jest wyznaczana następująco:

1. wylosuj w sposób jednostajny m ($m << n$) obserwacji ze zbioru D , ozn. p_1, \dots, p_m ,
2. $\forall i = 1, \dots, m$ $p_i \in D$ znajdź najbliższego sąsiada, ozn. p_j i wyznacz odległość pomiędzy tymi punktami: $x_i = \text{dist}(p_i, p_j)$,
3. wygeneruj m -elementowy zbiór $random_D = (q_1, \dots, q_m)$ z rozkładu jednostajnego z taką samą wariancją jak wyjściowy zbiór D ,
4. $\forall i = 1, \dots, m$ $q_i \in random_D$ znajdź najbliższego sąsiada, ozn. q_j i wyznacz odległość pomiędzy tymi punktami $y_i = \text{dist}(q_i, q_j)$,
5. wartość statystyki Hopkinaса wyznacz ze wzoru:

$$H = \frac{\bar{y}}{\bar{x} + \bar{y}}, \quad (2.1)$$

2.1. OCENA LICZBY SKUPIEŃ

$$\text{gdzie } \bar{x} = \frac{1}{m} \sum_{i=1}^m x_i, \bar{y} = \frac{1}{m} \sum_{i=1}^m y_i.$$

Z definicji statystyki wynika, że jeżeli $\bar{x} \sim \bar{y}$, to $H \approx 0.5$, a więc dane pochodzą z rozkładu jednostajnego i trudno jest dopatrywać się w nich naturalnego podziału na skupienia.

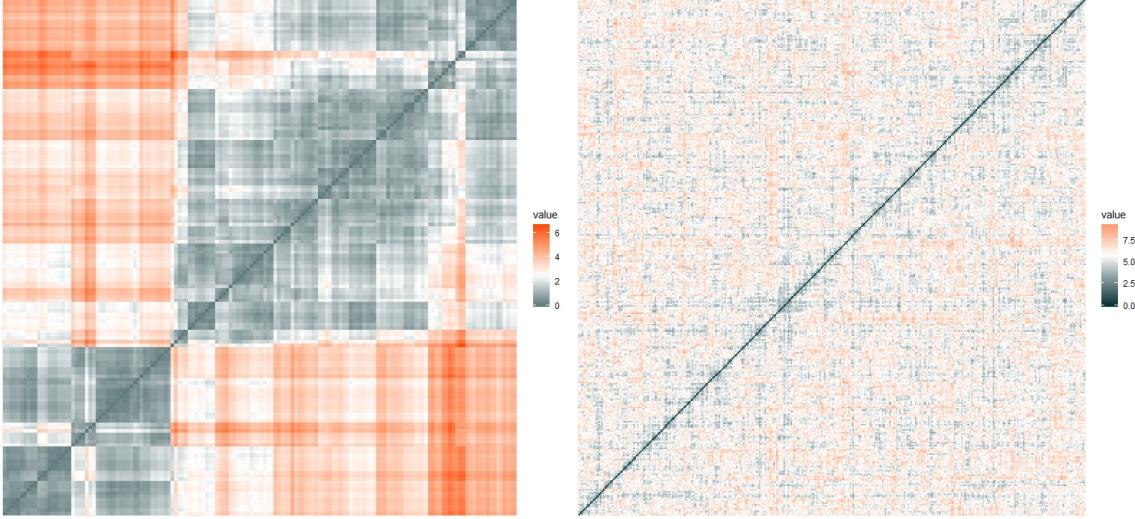
Inną metodą, służącą do oceny, czy w danych występują naturalne skupiska obserwacji, jest tak zwana VAT (Visual Assessment of (Cluster) Tendency, [3]). Założymy, że dana jest macierz odmiенноści pomiędzy obserwacjami ze zbioru, ozn. R . Algorytm opisany poniżej ma na celu zamianę kolejności wierszy i kolumn macierzy tak, aby obserwacje najbardziej do siebie podobne, czyli o podobnych wartościach odmiенноści, znajdowały się obok siebie.

Zanim zostanie krótko opisany algorytm, należy przyjąć pewne oznaczenia. Niech $I, J \subset K = \{1, \dots, n\}$, gdzie n to liczba obserwacji w zbiorze, a więc macierz odmiенноści R jest $n \times n$. Niech $\arg \min_{p \in I, q \in J} \{R_{pq}\}$ będzie zbiorem uporządkowanych par $(i, j) \in I \times J$, takich, że $R_{ij} = \arg \min_{p \in I, q \in J} \{R_{pq}\}$. Analogicznie definiujemy argument maksimum.

Elementy macierzy R są sortowane w następujący sposób:

1. niech $K = 1, \dots, n$, $I = J = \emptyset$, $P[0] = (0, \dots, 0)$,
2. wybierz $(i, j) \in \arg \min_{p \in K, q \in K} \{R_{pq}\}$, a następnie niech $P(1) = i$, $I = \{i\}$, $J = K \setminus \{i\}$,
3. dla $l = 2, \dots, n$ wybierz $(i, j) \in \arg \min_{p \in I, q \in J} \{R_{pq}\}$, a następnie przyjmij $P(l) = j$, $I \leftarrow I \cap \{i\}$, $J \leftarrow J \setminus \{j\}$,
4. ostatnim krokiem jest budowa uporządkowanej macierzy odmiенноści $\tilde{R} := [\tilde{R}_{ij}]_{i,j=1,\dots,n}$, gdzie $\tilde{R}_{ij} = R_{P(i)P(j)}$, którą następnie można przedstawić na wykresie.

Poniższy wykres przedstawia wykres VAT dla zbioru Iris. Po lewej stronie znajduje się wykres wygenerowany na podstawie oryginalnych danych. Wzdłuż diagonali, zaznaczonej niebieską linią, można zauważać ciemne kwadraty. Ich liczba pokazuje między innymi, ile skupień można wyodrębnić w danych. Wykres po prawej stronie nie pokazuje żadnej struktury, co nie powinno być niespodzianką, ponieważ został wygenerowany z danych pochodzących z rozkładu jednostajnego.



Rysunek 2.1: Wykres VAT, zbiór Iris, a). dane oryginalne, b). dane wygenerowane losowo z rozkładu jednostajnego

Kolejnym krokiem, jaki należy wykonać w tego typu analizie, jest ustalenie liczby skupień, na jakie zostaną podzielone dane. Wyboru można dokonać na podstawie kilu metod. Założymy, że dysponujemy n -elementową próbą o wartościach w przestrzeni \mathbb{R}^p i naszym celem jest podział próby na K skupień. Można wówczas zdefiniować całkowitą sumę kwadratów odległości pomiędzy parami punktów, następująco:

$$T = B + W, \text{ gdzie} \quad (2.2)$$

$$W = \frac{1}{2} \sum_{k=1}^K \sum_{i,j \in C(i)=k} dist(x_i, x_j) \quad (2.3)$$

$$B = \frac{1}{2} \sum_{k=1}^K \sum_{\substack{i \in C(i)=k, \\ j \notin C(i)=k}} dist(x_i, x_j) \quad (2.4)$$

Wielkość W oznacza sumę kwadratów pomiędzy punktami wewnętrz tego samego skupienia, B – pomiędzy różnymi skupieniami, natomiast K oznacza liczbę klas. Poniżej zostały opisane krótko metody, które mogą być wykorzystane do określenia liczby podgrup, na jakie można podzielić dane:

- statystyka odstępu

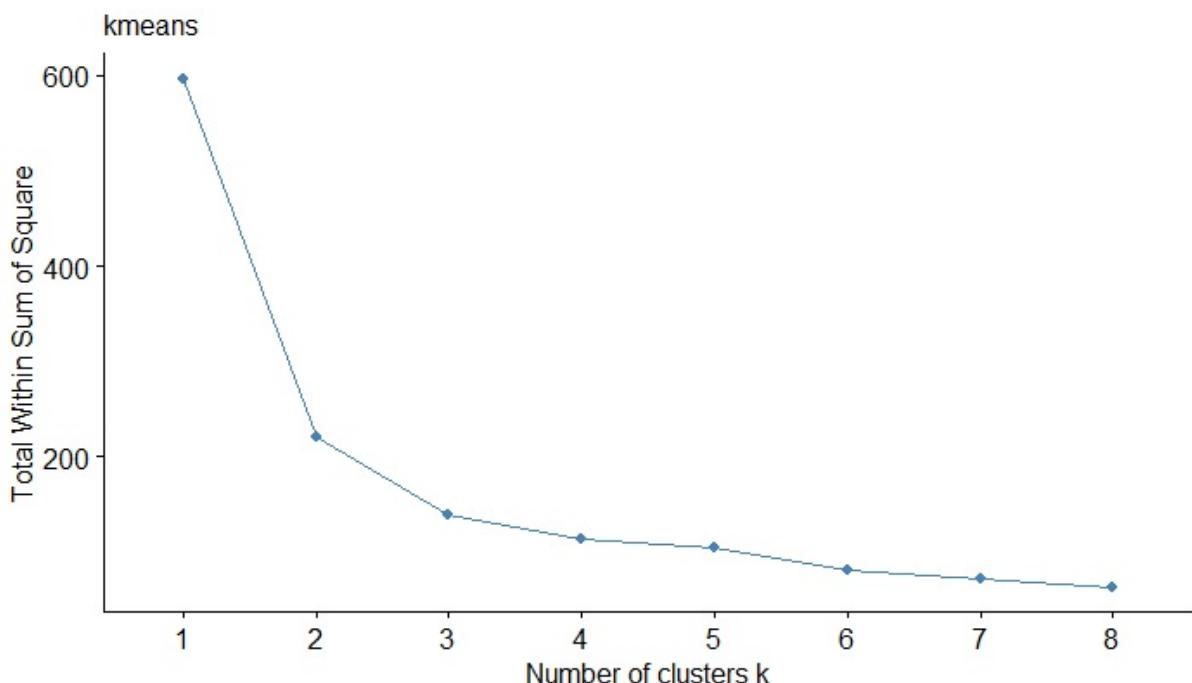
Zaprezentowana przez Tibshirani'ego i innych ([16]) metoda opiera się na porównaniu wartości funkcji $G(k) = \log W_k$ dla różnej liczby klas k z analogicznymi wartościami dla próby wygenerowanej z rozkładu jednostajnego na kostce opisującej wyjściowy zbiór danych. Optymalną liczbę klas K^* otrzymuje się w następujący sposób:

2.2. MIARY ODMIENNOŚCI

$$K^* = \arg \min_{p \in I, q \in J} \{k : s_{k+1} \leq G(k) - G(k+1)\} \quad (2.5)$$

- metoda krzywej usypiskowej

Dla różnych wartości parametru k wyznaczana jest całkowita suma odległości obserwacji od środków swoich skupień: $WSS = \sum_{k=1}^K \sum_{i \in C(k)} dist(x_i, m_k)$, gdzie m_k to środek k -tego skupienia (średnia wektorowa po współrzędnych), $k = 1, \dots, K$. Następnie generowany jest wykres WSS w zależności od k :



Rysunek 2.2: Wykres usypiskowy dla zbioru Iris, metoda k-średnich

Miejsce, w którym następuje wypłaszczenie wykresu, w tym przypadku jest to $K = 3$, uznaje się za optymalną liczbą klas.

2.2. Miary odmienności

Przed omówieniem metod klasyfikacji, zostaną krótko omówione miary, określające odmiennosć pary podgrup w zbiorze, a co za tym idzie par wektorów (obserwacji). Pojęcie odmienności jest nieco słabsze niż odległość w klasycznym sensie, często jednak spełniające aksjomaty odległości. Odmiennosć pomiędzy dwoma elementami zbioru $x_i, x_j \in D$ będziemy oznaczać d_{ij} .

Definicja 2.1. Założymy, że dany jest zbiór D o wymiarach $n \times p$. Odległością $dist(x, y)$ nazywiemy funkcję, która dla dowolnej pary $x, y \in D$ jest funkcją symetryczną, nieujemną i spełniającą nierówność trójkąta.

Przykłady odmiенноści dla różnych typów danych:

1. dane ilościowe:

- odległość euklidesowa:

$$d_{ij} = \sqrt{\sum_{l=1}^p (x_i^{(l)} - x_j^{(l)})^2} \quad (2.6)$$

- odległość Mińskowskiego:

$$d_{ij} = \left(\sum_{l=1}^p |x_i^{(l)} - x_j^{(l)}|^q \right)^{\frac{1}{q}} \quad (2.7)$$

2. dane jakościowe binarne:

- niech $a = |\{x_i = 1 \wedge x_j = 1\}|$, $b = |\{x_i = 1 \wedge x_j = 0\}|$, $c = |\{x_i = 0 \wedge x_j = 1\}|$, $d = |\{x_i = 0 \wedge x_j = 0\}|$, wówczas miarą Jaccarda nazywamy:

$$d_{ij} = \frac{b + c}{a + b + c} \quad (2.8)$$

- miara Czekanowskiego:

$$d_{ij} = 1 - \frac{2a}{2a + b + c} \quad (2.9)$$

3. dane jakościowe o więcej niż dwóch poziomach:

$$d_{ij} = 1 - \frac{|\{x_i = x_j\}|}{p} \quad (2.10)$$

4. dane mieszane – odmiennaść Gowera:

$$d_{ij} = 1 - s_{ij} \quad \text{gdzie} \quad s_{ij} = \sum_{l=1}^p s_{ijk} / \sum_{l=1}^p \delta_{ijk} \quad (2.11)$$

s_{ij} jest współczynnikiem podobieństwa, a s_{ijk} to współczynnik podobieństwa dla k-tej zmiennej, wyznaczany następująco:

- dla zmiennych ilościowych

$$s_{ijk} = 1 - \frac{|x_i^{(k)} - x_j^{(k)}|}{\max_i x_i^{(k)} - \min_i x_i^{(k)}} \quad (2.12)$$

2.3. K-ŚREDNICH

- dla zmiennych jakościowych

$$s_{ijk} = \begin{cases} 1 & \text{jeżeli } x_i^{(k)} = x_j^{(k)} \\ 0 & \text{w przeciwnym wypadku.} \end{cases} \quad (2.13)$$

W przypadku zmiennych binarnych, gdy obie współrzędne są równe 0, czyli cecha nie występuje, $s_{ijk} = 0$. Natomiast δ_{ijk} opisuje możliwość porównania dwóch obserwacji dla k-tej zmiennej. Oznacza to, iż odmienność Gowera uwzględnia wystąpienie braków w danych. Jeżeli $\delta_{ijk} = 0$, to należy przyjąć $s_{ijk} = 0$.

2.3. K-średnich

Jedna z najpopularniejszych metod klasyfikacji bez nadzoru, zaliczana do grupy metod kombinatorycznych. Pozwala na podział zbioru danych na ustaloną z góry liczbę podgrup K . Celem jest wyznaczenie środków podgrup (skupień) w taki sposób, aby odległość sklasyfikowanych punktów od środków swoich skupień była jak najmniejsza, to znaczy, aby obserwacje wewnętrz klasy były do siebie jak najbardziej podobne. Jest to równoważne minimalizacji następującego wyrażenia:

$$W = \sum_{i=1}^K \sum_{C(i)=k} dist(x_i, m_k) \quad (2.14)$$

gdzie $dist(\cdot, \cdot)$ to odległość euklidesowa pomiędzy dwoma punktami, $C(i) = k$ – funkcja, przypisująca numerowi obserwacji dane skupienie, $m_k = \frac{1}{|n_k|} \sum_{C(i)=k} x_i$ – średnia wektorowa z elementów należących do danej klasy, $i = 1, \dots, n$, $k = 1, \dots, K$. Algorytm klasyfikacji przebiega następująco:

1. dany jest zbiór D wymiaru $n \times p$. Jako K pierwszych środków skupień przyjmuje się K losowo wybranych obserwacji spośród n , ozn. m_1, \dots, m_K
2. każdą obserwację x_i , $i = 1, \dots, n$ przyporządkowuje się do klasy, od której środka najbliższej się dana obserwacja znajduje: $C(i) = \arg \min_{k=1, \dots, K} dist(x_i, m_k)$
3. kolejnym krokiem jest aktualizacja współrzędnych środków skupień m_k , $k = 1, \dots, K$.
4. kroki 1 – 3 należy powtarzać do momentu, kiedy różnica wartości funkcji celu (2.14) pomiędzy i-tym a $i - 1$ -szym krokiem będzie mała lub zostanie wykonana maksymalna liczba iteracji.

Wynikiem działania algorytmu jest wektor, którego i-ta współrzędna opisuje przynależność i-tej obserwacji do jednej z K grup. Z uwagi na fakt, iż algorytm ten zbiega zawsze, choć nieko-

niecznie do rozwiązania optymalnego, należy go zastosować wielokrotnie, wybierając za każdym razem inny wektor początkowy m_1, \dots, m_K .

2.4. Metody hierarchiczne

Kolejna grupa metod klasyfikacji nienadzorowanej. W odróżnieniu od metody k-średnich, nie trzeba przy nich zakładać z góry liczby klas. Zamiast tego, należy określić, tzw. odmienności między skupieniami, które są oparte na różnicach pomiędzy parami obserwacji z dwóch różnych grup. Metody hierarchiczne dzielą się na dwie grupy:

- aglomeracyjne – procedura startuje przy n skupieniach (każda pojedyncza obserwacja to skupienie), następnie dwa skupienia najmniej odmienne od siebie zostają połączone. Krok ten jest powtarzany do uzyskania jednego skupienia, którym jest cały zbiór;
- oparte na dzieleniu zbioru – procedura startuje z jednego skupienia (cały zbiór), następnie zostaje ono podzielone na dwa podzbiory o największej odmienności. Krok ten jest wykonywany dla każdego skupienia i powtarzany do momentu uzyskania n skupień, gdzie każda obserwacja jest oddzielnym skupieniem. Metoda bardziej złożona obliczeniowo w porównaniu do aglomeracyjnej.

Dla metod hierarchicznych wyróżnia się trzy sposoby łączenia (dzielenia) dwóch skupień:

1. odmienność najbliższego sąsiada $\min d_{ij}$,
2. odmienność najdalszego sąsiada $\max d_{ij}$,
3. odmienność średnia $\bar{d} = \frac{1}{n_i n_j} \sum_{i,j} d_{ij}$.

Ciekawym podejściem wydaje się połączenie dwóch wcześniej opisanych metod. Podobnie, jak w klasycznej metodzie k-średnich, należy ustalić liczbę skupień K . Przy pomocy metod hierarchicznych dzieli się dane na K skupień. Następnie, dla tak otrzymanych skupień, wyznacza się ich środki i traktuje się je jako początkowe skupienia w metodzie k-średnich.

2.5. Miary jakości podziału

W przypadku klasyfikacji pod nadzorem, weryfikacja jakości dopasowania modelu jest prosta – wystarczy porównać dopasowane za pomocą modelu wartości zmiennej odpowiedzi z wartościami

2.5. MIARY JAKOŚCI PODZIAŁU

prawdziwymi, czyli znajdującymi się w zbiorze danych. Należy pamiętać, iż przy klasyfikacji bez nadzoru, nie dysponujemy zmienną odpowiedzi dla zbioru uczącego, toteż takie podejście nie będzie dobre. Poniżej zostały opisane metody, umożliwiające sprawdzenie jakości dopasowania modelu klasyfikacji nienadzorowanej.

2.5.1. Miary wewnętrzne

Miary wewnętrzne wykorzystują informację zawartą bezpośrednio w utworzonych skupieniach. Zalicza się do nich między innymi:

- sylwetkę obserwacji – mierzy, na ile obserwacja jest dobrze zaklasyfikowana poprzez estymację średniej odległości pomiędzy skupieniami. Sylwetkę dla i -tej obserwacji oznaczamy przez S_i i wyznaczamy następująco:

1. założmy, że $x_i \in C_{k_0}$
2. wyznacz średnią odmiennosć pomiędzy i -tą obserwacją a pozostałymi punktami ze skupienia, w którym on się znajduje: $a_i = \frac{1}{n_i} \sum_{j: C(j)=k_0} d_{ij}$
3. dla pozostałych skupień wyznacz średnią odmiennosć i -tej obserwacji od danego skupienia, a następnie niech $b_i = \min_{C_k: k \neq k_0} d_{iC_k}$, gdzie d_{iC_k} jest średnią odległością i -tej obserwacji od k -tego skupiania, $k = 1, \dots, K$
4. $S_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}$, $i = 1 \dots n$.

Wartości bliskie 1 sugerują, iż obserwacja jest dobrze zaklasyfikowana, bliskie 0 – obserwacja pasuje do dwóch różnych klas jednocześnie, wartości ujemne zaś, iż dana obserwacja jest źle zaklasyfikowana. Wartość średniej sylwetki, która jest miarą wyznaczoną dla całego zbioru, można wykorzystać do określenia optymalnej liczby klas. W tym celu, należy dla różnych wartości K dokonać klasteryzacji, a następnie wyznaczyć $K^* = \operatorname{argmax}_K S_{avg}$, gdzie $S_{avg} = \frac{1}{n} \sum_{i=1}^n S_i$.

- indeks Dunn'a – jeden z popularniejszych wskaźników opisujących jakość zbudowanego modelu uczenia nienadzorowanego. Wyraża stosunek pomiędzy najmniejszą odległością pomiędzy dwoma obserwacjami z dwóch różnych skupień a największą odległością między skupieniami. Dla k -tego skupienia wyraża się wzorem:

$$DI_K = \frac{\min_{i \neq j, i, j=1, \dots, K} d(C_i, C_j)}{\max_{1 \leq k \leq K} \Delta_k} \quad (2.15)$$

gdzie

$$d(C_i, C_j) = \min_{x \in C_i, y \in C_j} dist(x_i, y_j),$$

$$\Delta_k = \max_{x, y \in C_k} dist(x, y).$$

Wartość indeksu Dunn'a dla całego zbioru jest minimum po wszystkich skupieniach z wartości DI_k :

$$DI = \min_{1 \leq k \leq K} DI_k. \quad (2.16)$$

2.5.2. Miary stabilności

Opierają się na porównaniu wyników klasyfikacji otrzymanych na podstawie danych pełnych i na danych z pominiętą jedną zmienną. We wszystkich opisanych niżej przypadkach, wyniki dla wszystkich zmiennych są uśredniane. Im otrzymane wartości są mniejsze, tym rozważany model jest uznawany za stabilniejszy.

- APN – odsetek obserwacji nie nakładających się. Zlicza obserwacje, które należą do różnych klas w zależności od tego, czy rozważany był model oparty na pełnych danych, czy na danych z pominiętą jedną zmienną, co formalnie można zapisać jako:

$$APN(K) = \frac{1}{MN} \sum_{i=1}^N \sum_{l=1}^M \left(1 - \frac{n(C^{i,0} \cap C^{i,l})}{n(C^{i,0})} \right) \quad (2.17)$$

gdzie $C^{i,0}, C^{i,l}$ – numer skupienia dla i -tej obserwacji dla danych pełnych oraz dla danych z pominiętą l -tą zmienną, odpowiednio dla $i = 1, \dots, N, l = 1, \dots, M$

- AD – średnia odległość pomiędzy obserwacjami z tego samego skupienia w zależności od tego, czy rozważany był model oparty na pełnych danych, czy na danych z pominiętą jedną zmienną:

$$AD(K) = \frac{1}{MN} \sum_{i=1}^N \sum_{l=1}^M \frac{1}{n(C^{i,0})n(C^{i,l})} \left[\sum_{i \in C^{i,0}, j \in C^{i,l}} dist(x_i, x_j) \right] \quad (2.18)$$

- ADM – średnia odległość pomiędzy środkami skupień liczona dla obserwacji, które należą do tego samego skupienia w zależności od tego, czy rozważany był model oparty na pełnych danych, czy na danych z pominiętą jedną zmienną:

$$ADM(K) = \frac{1}{MN} \sum_{i=1}^N \sum_{l=1}^M dist(\bar{x}_{C^{i,l}}, \bar{x}_{C^{i,0}}) \quad (2.19)$$

gdzie $\bar{x}_{C^{i,l}}, \bar{x}_{C^{i,0}}$ oznacza odpowiednio średnie z obserwacji należących do tej samej klasy co i -ta obserwacja dla modelu bez l -tej zmiennej i modelu pełnego, odpowiednio.

2.5. MIARY JAKOŚCI PODZIAŁU

- FOM – mierzy średnią wariancję wewnątrz klasy dla pominiętej zmiennej, gdzie klasyfikacja jest oparta o model pełny:

$$FOM(l, K) = \sqrt{\frac{1}{N} \sum_{k=1}^K \sum_{i \in C_k(l)} dist(x_{i,l}, \bar{x}_{C_k(l)})} \quad (2.20)$$

3. Grupowanie tekstu pod względem trudności

W tym rozdziale omówione zostaną kolejne kroki, jakie zostały podjęte w celu stworzenia bazy, na podstawie której zbudowano model wykorzystany w aplikacji, jak i ścieżka wyboru samego modelu.

3.1. Baza tekstów oraz jej wstępne przetworzenie

W budowie modelu wykorzystano własnoręcznie skonstruowaną bazę 351 tekstów różnego typu i różnej długości przeznaczonych dla dzieci w wieku 7 – 13 lat:

- książki oraz lektury szkolne w liczbie 117 tekstów, pobrane ze strony WolneLektury.pl,
- napisy do filmów animowanych skierowanych do młodych widzów – 104 teksty, pobrane ze strony opensubtitles.org,
- 130 scenariuszy do zajęć udostępnionych przez Centrum Nauki Kopernik.

Tabela 3.1 zawiera opis miar wyznaczonych dla każdego tekstu. Zmienna **trudnosc** określa numeryczną wartość złożoności tekstu. Jest wyliczona na podstawie nieco uproszczonej formuły wyznaczonej w [9] – z przyczyn technicznych pominięte zostały zmienne takie jak: średnia długość paragrafu, odsetek dopełniaczy, czasowników bezosobowych oraz cztero- lub więcej sylabowych słów ze słownika Imiołczyka.

Ponadto, biorąc pod uwagę fakt, iż mamy do czynienia z tekstami dla dzieci, którym długie słowa często sprawiają trudność, przyjęto następującą definicję słowa długiego:

Definicja 3.1. Słowo uznajemy za długie, jeżeli ma cztery lub więcej sylab w formie odmienionej.

3.1. BAZA TEKSTÓW ORAZ JEJ WSTĘPNE PRZETWORZENIE

$$\begin{aligned}
 \text{trudnosc} = & -1.479 + 0.02708 \times \text{asl} + 0.02909 \times \text{ods_dlugich} + 0.0248 \times \text{ods_rzecz} + \\
 & + 0.04793 \times \text{ods_rzecz_dl} - 0.03267 \times \text{ods_czas} + 0.04752 \times \text{ods_czas_dl} + \\
 & + 0.03114 \times \text{ods_przym} + 0.06377 \times \text{ods_przym_dl} + 0.1585 \times \text{rzecz_czas} + \\
 & + 0.9057 \times \text{awl} + 0.1938 \times \text{rzecz_osc} + 0.0299 \times \text{ods_rzecz_odczas}
 \end{aligned} \tag{3.1}$$

zmienna	wyjaśnienie
czasownik	liczba czasowników
liczebnik	liczba liczebników
przymiotnik	liczba przymiotników
przysłowiek	liczba przysłówków
rzeczownik	liczba rzeczowników
spojnik	liczba spójników
zaimek	liczba zaimków
inne	liczba pozostałych części mowy
FOG	wartość indeksu FOG
liczba_zdan	liczba zdań
asl	średnia liczba słów w zdaniu
ods_dlugich	odsetek wyrazów długich wg. definicji 3.1
ods_rzecz	odsetek rzeczowników
ods_rzecz_dl	odsetek długich rzeczowników wg. definicji 3.1
ods_czas	odsetek czasowników
ods_czas_dl	odsetek długich przymiotników wg. definicji 3.1
ods_przym	odsetek przymiotników
ods_przym_dl	odsetek długich przymiotników wg. definicji 3.1
rzecz_czas	stosunek rzeczowników do czasowników
awl	średnia długość słowa wyrażona w sylabach
rzecz_osc	odsetek rzeczowników zakończonych na '-ość'
ods_rzecz_odczas	odsetek rzeczowników odczasownikowych
trudnosc	wartość opisująca trudność tekstu
typ	rodzaj badanego tekstu - książka, film, inne
liczba_trudnych	liczba słów trudnych wg. definicji 3.2
odsetek_trudnych	odsetek trudnych słów

Tablica 3.1: Cechy wyznaczone dla każdego tekstu

Dodatkowo, dla całej bazy utworzono słownik unikalnych słów w niej występujących oraz posortowano je malejąco według częstości występowania. Przyjęto, iż 10 000 słów o największej częstości występowania, to słowa najbardziej powszechnie, które zna każde dziecko. Dalej, ustalono progi następujące: 10 001 – 20 000 – słowa mniej popularne, 20 000 i dalej – słowa rzadkie.

Definicja 3.2. Słowo uznajemy za trudne, jeżeli należy do grupy słów mniej popularnych lub rzadkich.

Z uwagi na fakt, iż język polski jest językiem silnie fleksyjnym, do wstępnej obróbki tekstu wykorzystano istniejący już słownik Polimorfologik ([17]). Zawiera on wszystkie możliwe odmiany danego słowa wraz z jego opisem gramatycznym i formą bezokolicznikową, co umożliwia sprowadzenie badanego tekstu do formy podstawowej oraz wyznaczenie wskaźników opisanych w tabeli 3.1. Dokładny opis form gramatycznych można znaleźć w [2].

forma odmieniona	forma słownikowa	opis gramatyczny
podeszwa	podeszwa	subst:sg:nom:f
podeszwach	podeszwa	subst:pl:loc:f
podeszwami	podeszwa	subst:pl:inst:f
podeszwą	podeszwa	subst:sg:inst:f
podeszwę	podeszwa	subst:sg:acc:f

Tablica 3.2: Przykładowe rekordy ze słownika Polimorfologik ([17])

Przewiduje się, iż cechami, które będą miały wpływ na przynależność danego tekstu do konkretnej grupy tekstów, będzie między innymi długość tekstu, wskaźnik trudności oraz odsetek słów trudnych. Jednak, aby potwierdzić te przypuszczenia, w dalszej części pracy została przeprowadzona dokładna analiza przygotowanych danych.

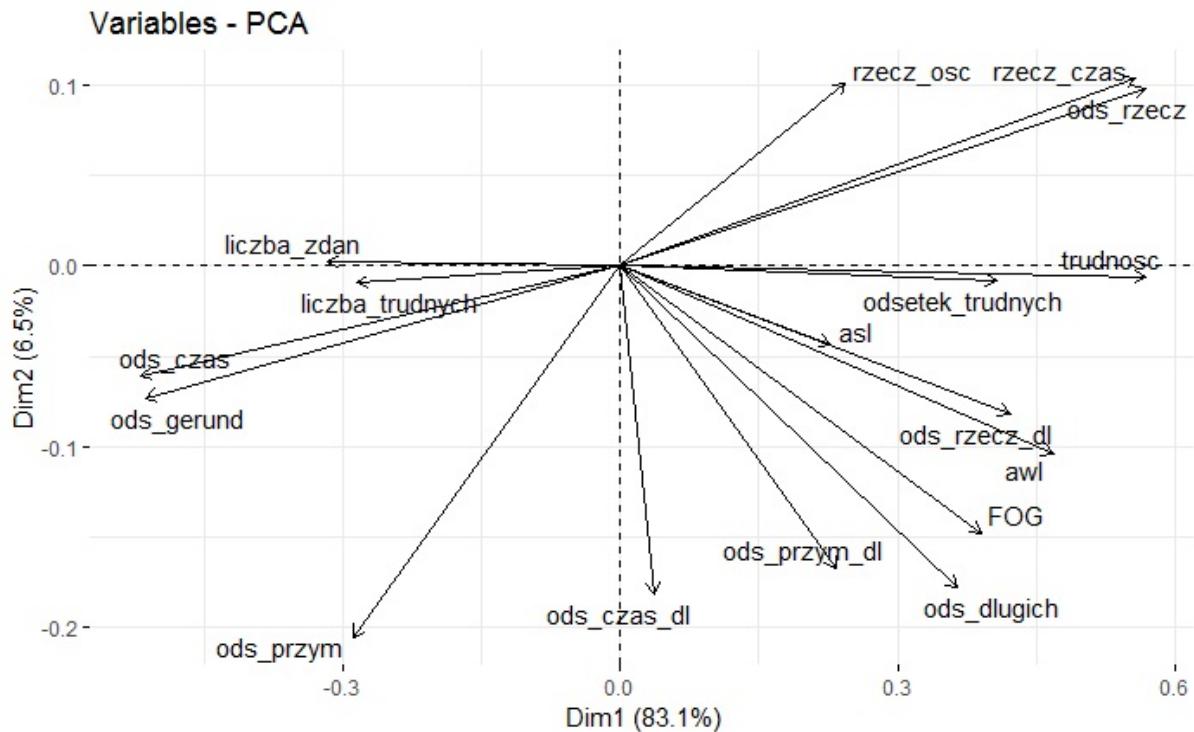
3.2. Eksploracja danych

Wszystkie obliczenia, jak i projekt aplikacji, zostały zrealizowane przy pomocy pakietu R. Zanim została zaprojektowana aplikacja, należało ustalić, w jaki sposób rozróżnić teksty pod względem trudności oraz innych cech. Dla zbudowanej już bazy danych, dokonano jej analizy, wykorzystując metody opisane w rozdziale 2.

3.2. EKSPLORACJA DANYCH

3.2.1. Wstępna analiza danych – wyznaczenie liczby klas

Ramka danych składała się z 17 zmiennych, które opisano w tabeli 3.1 – pominięto zmienne dotyczące liczności poszczególnych części mowy na ze względu na fakt, iż wartości nie były porównywalne między sobą – dopiero po uwzględnieniu długości tekstu można było wyciągnąć odpowiednie wnioski co rozkładu części mowy w danym tekście.

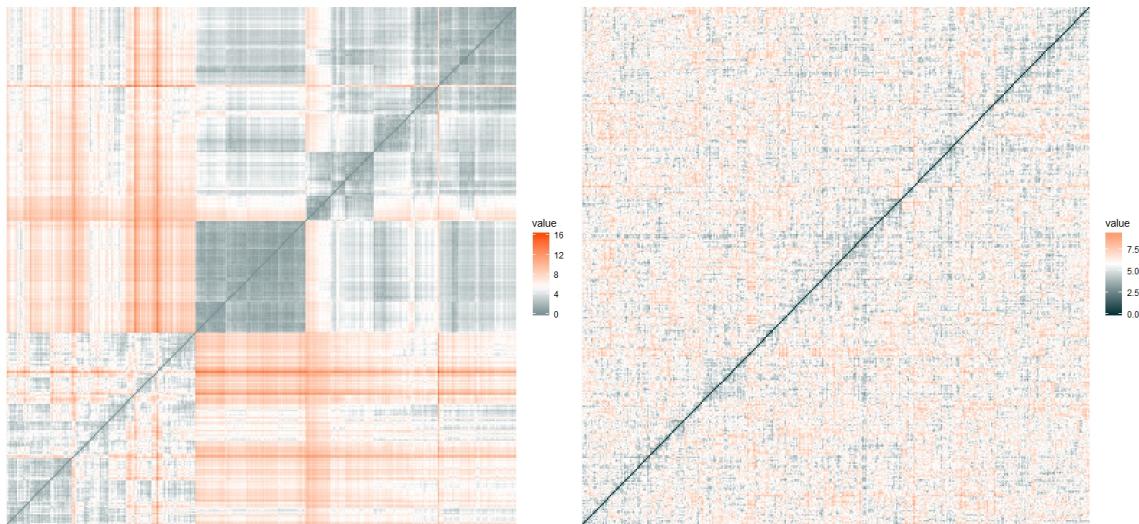


Rysunek 3.1: Macierz korelacji zmiennych użytych do budowy modelu

Rysunek 3.1 przedstawia wizualizację macierzy korelacji dla zmiennych, wykorzystanych w budowie modelu. Dzięki temu, łatwo zauważać, które zmienne są ze sobą skorelowane – obrazuje to długość strzałki, na przykład odsetek rzeczowników oraz stosunek rzeczowników do czasowników ('ods_rzecz' i 'rzecz_czas') są silnie ze sobą skorelowane (0.96), co nie jest niczym zaskakującym, ponieważ obie te cechy odnoszą się do występowania rzeczowników w tekście.

Pierwszym krokiem w celu pogrupowania tekstów było ustalenie, czy w danych występują naturalne podgrupy obserwacji. Najpierw wyznaczono statystykę Hopkinса (2.1): $H = 0.166$. Jest to wartość znacznie mniejsza niż 0.5, więc możemy przyjąć, iż w danych występują naturalne skupienia. Następnie przeanalizowano wykres VAT. Natężenie koloru odpowiada wartości odmiенноści pomiędzy obiektami – niebieski oznacza brak odmiенноści, czerwony – dużą odmiennosć.

Rysunek po prawej stronie ukazuje dane losowe wygenerowane na podstawie posiadanych danych – w ogóle nie można zauważać w nich struktury, w przeciwieństwie do rysunku po lewej stronie, który powstał na podstawie oryginalnych danych. Wykres potwierdza zatem wnioski wysnute na podstawie statystyki Hopkinса – w danych występują naturalne podgrupy obserwacji.



Rysunek 3.2: Wykres VAT a). dane oryginalne, b).dane losowe

Następnie, podjęto się wyznaczenia liczby potencjalnych klas, na jakie dane będą podzielone. W procesie grupowania zostaną wykorzystane trzy metody – k-średnich, metody hierarchiczne ze średnią funkcjąłączającą oraz połączenie tych dwóch metod, zwane dalej hierarchicznym k-średnich (tutaj też wykorzystana została średnia funkcjałączająca). Następnie, spośród dopasowanych modeli, wybrany zostanie model najlepiej pasujący do danych. Do ustalenia liczby klas wykorzystano statystykę odstępu, metodę sylwetki oraz metodę wykresu usypiskowego. Rysunek 3.3 przedstawia wykresy, na podstawie których podjęto decyzje o wyborze liczby klas.

- k-średnich:

1. statystyka odstępu – $K = 8$, jednak widać, iż od $K = 6$ wykres się mocno wypłaszcza
2. metoda sylwetki – $K = 2$ – przypadek dwuklasowy jest odrzucony automatycznie ze względu na jego interpretację dla rozważanego problemu
3. metoda wykresu usypiskowego – wypłaszczenie następuje dla $K = 5, 6$

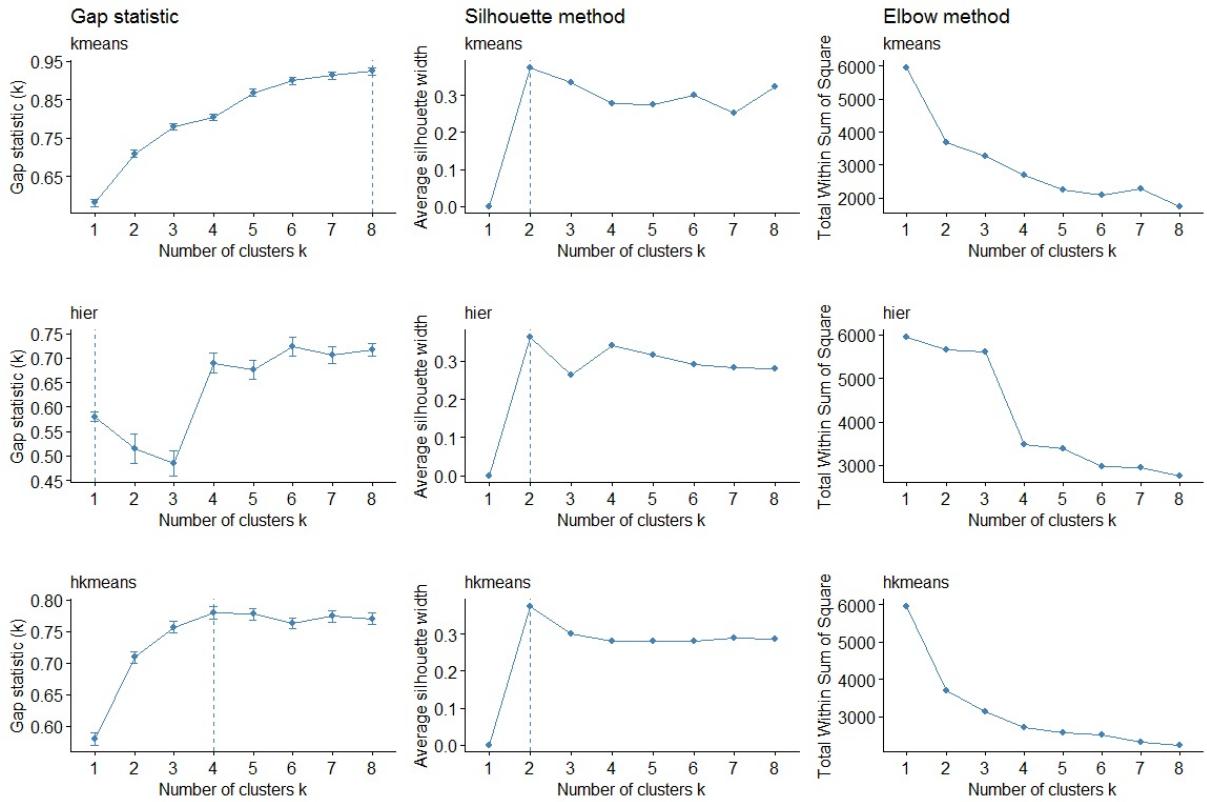
- metoda hierarchiczna, średnia funkcjałączająca:

1. statystyka odstępu – $K = 1$ przypadek odrzucony arbitralnie
2. metoda sylwetki – tutaj podobnie jak dla k-średnich odrzucany zostaje przypadek dwuklasowy

3.2. EKSPLORACJA DANYCH

3. metoda wykresu usypiskowego – liczba klas $K = 4, 5$

- hierarchiczne k-średnich:
 1. statystyka odstępu – $K = 4$
 2. metoda sylwetki – odrzucony zostaje przypadek $K = 2$, podobnie jak w pozostałych przypadkach
 3. metoda wykresu usypiskowego – wypłaszczenie następuje dla $K = 4, 5$.



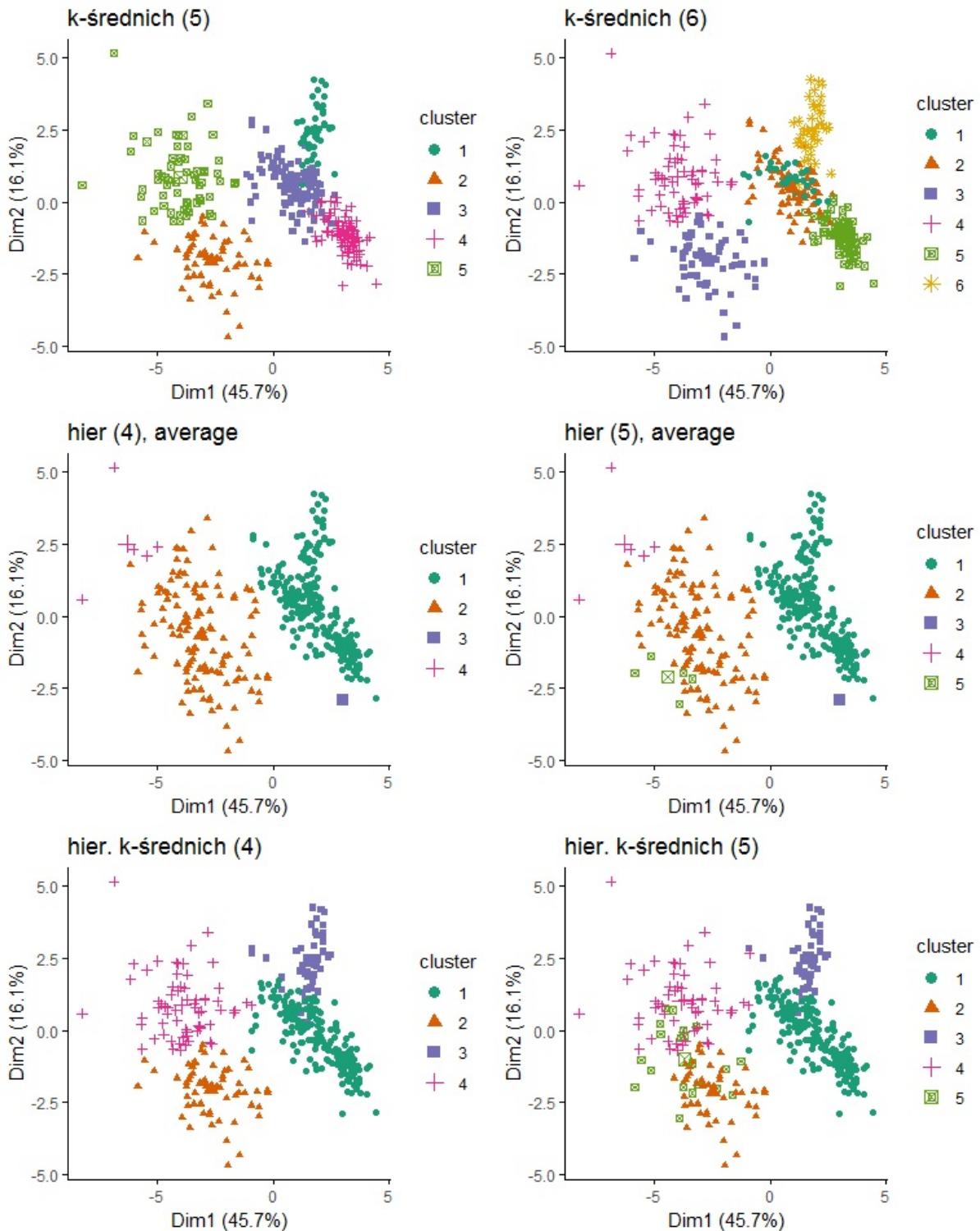
Rysunek 3.3: Ocena optymalnej liczby skupień: a). statystyka odstępu, b). metoda sylwetki, c). metoda wykresu usypiskowego

Z uwagi na fakt, iż dla różnych modeli otrzymano różne wartości liczby klas, dopasowano następujące modele:

- k-średnich dla $K = 5, 6$
- metoda hierarchiczna ze średnią funkcją łączającą dla $K = 4, 5$
- hierarchiczne k-średnich ze średnią funkcją $K = 4, 5$.

3.2.2. Grupowanie oraz ocena jego jakości

Poniżej zostały zamieszczone wykresy obrazujące podział na klasy. Wykres jest przedstawiony na podstawie dwóch pierwszych głównych składowych, otrzymanych przy pomocy metody PCA.



Rysunek 3.4: Rozkład obserwacji w klasach dla wybranych modeli

3.2. EKSPLORACJA DANYCH

- k-średnich:
 1. $K = 5$ – wyznaczone skupienia wydają się dobrze rozgraniczać dane, mimo iż znajdują się blisko siebie
 2. $K = 6$ – obserwacje z klasy 1 i 2 nakładają się na siebie. Jeżeli dokładniej się przyjrzeć, to widać, iż są to obserwacje z klasy 3 dla modelu z $K = 5$.
- modele hierarchiczne dla $K = 4, 5$ to tak naprawdę te same modele, różniące się jedynie liczbą klas. W obu modelach dochodzi do sytuacji, gdzie występuje skupienie jednoelementowe (klasa 3 w obu przypadkach). Można by pokusić się o złączenie klas 1 i 3, czego skutkiem będzie otrzymanie modeli o 3 i 4 klasach, odpowiednio. Pomyśl ten nie przyniósł zbytnio korzyści – liczności klas nadal okazały się mocno niezrównoważone.
- hierarchiczne k-średnich:
 1. $K = 4$ – w miarę dobrze rozgranicza klasy
 2. $K = 5$ – podobnie, jak w przypadku modelu k-średnich z $K = 6$, występują nakładające się na siebie skupienia.

Podsumowując, najlepiej sprawuje się model k-średnich z $K = 5$ i hierarchiczne k-średnich z $K = 4$. Powyższe obserwacje należy jednak poprzedzić bardziej formalnymi wskaźnikami. W dalszych rozważaniach pominięte zostają modele k-średnich z $K = 6$ i hierarchiczne k-średnich z $K = 5$, ze względu na przypadki nakładających się na siebie klas obserwacji.

Dla każdego z czterech rozważanych dalej modeli, wyliczono następujące miary jakości: średnią wartość sylwetki oraz indeks Dunn'a.

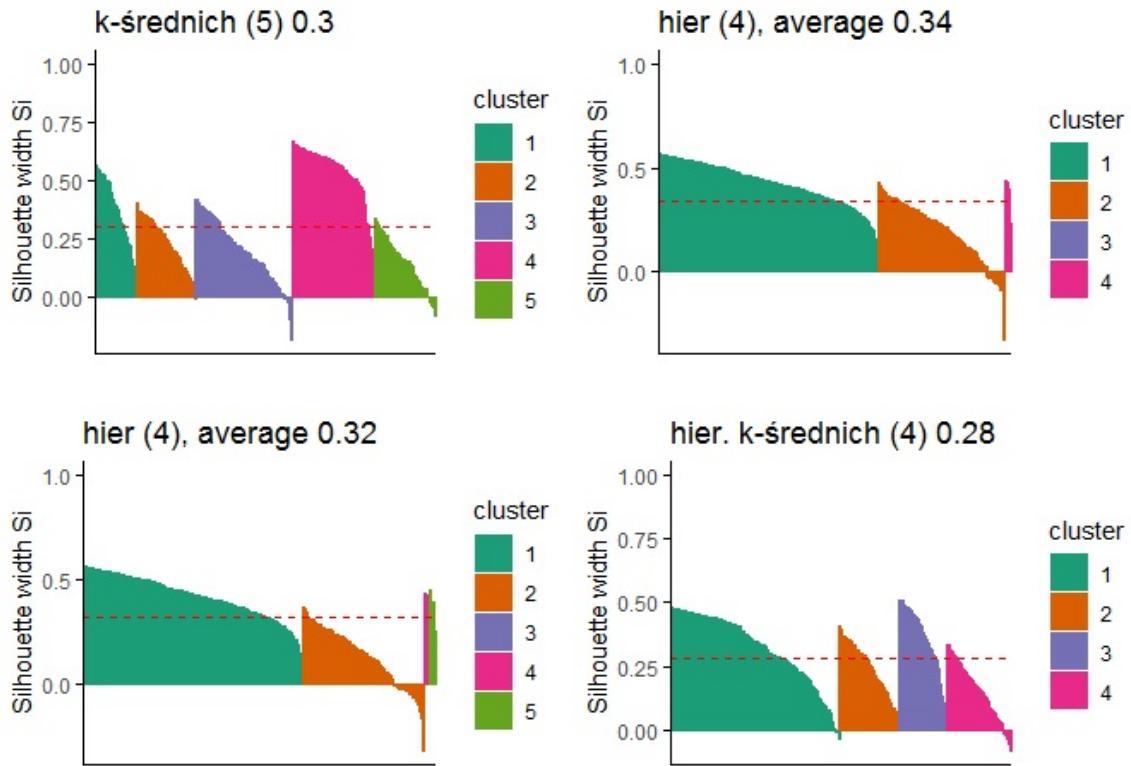
wskaźnik	k-średnich $K = 5$	hier. $K = 4$	hier. $K = 5$	hier. k-średnich $K = 4$
sylwetka	0.30	0.34	0.32	0.28
indeks Dunn'a	0.10	0.20	0.20	0.10

Tablica 3.3: Wewnętrzne miary jakości dopasowania

Klasyczne modele hierarchiczne posiadają najwyższe wartości indeksu Dunn'a, co sugeruje najlepsze dopasowanie spośród rozważanych modeli. Model hierarchiczny z $K = 4$ ma najwyższą wartość sylwetki, jednak należy pamiętać, iż uzyskane modele hierarchiczne dają bardzo niezrównoważone wyniki pod względem liczności klas oraz klas jednoelementowe.

Model k-średnich z $K = 5$, który na wykresie obrazującym podział na skupienia wyglądał dość dobrze, posiada najmniejszą wartość indeksu Dunn'a, co sugeruje raczej kiepskie dopasowanie tego modelu. Podobnie jest z modelem hierarchicznym k-średnich.

Poniżej znajdują się dodatkowo wykresy średnich sylwetek dla wszystkich czterech modeli:



Rysunek 3.5: Wykresy sylwetek dla wybranych modeli

Model k-średnich z $K = 5$ i hierarchiczny model k-średnich z $K = 4$ mają niższe średnie sylwetki niż modele hierarchiczne, jednak posiadają mniej obserwacji o ujemnych wartościach sylwetki, co jest równoważne z mniejszą liczbą obserwacji błędnie zaklasyfikowanych.

Średnie sylwetki dla modeli hierarchicznych (klasycznych) są najwyższe, jednak mają największe odsetki obserwacji błędnie zaklasyfikowanych. Ponadto, widać tutaj bardzo dobrze niezrównoważenie klas – prawie wszystkie obserwacje znajdują się tylko w dwóch z nich. Dlatego też, zwykłe modele hierarchiczne nie będą brane pod uwagę przy wyborze ostatecznego modelu.

k-średnich $K = 5$	hier. $K = 4$	hier. $K = 5$	hier. k-średnich $K = 4$
4.56%	5.13%	8.83%	3.70%

Tablica 3.4: Procent błędnie zaklasyfikowanych obserwacji, wyznaczony na podstawie wartości sylwetek poszczególnych obserwacji

Zostały również wyznaczone indeksy mówiące o stabilności dla zaproponowanych modeli. Poniżej znajdują się wyniki uzyskane przy pomocy pakietu `clValid`. Porównywano modele dla danych oryginalnych z modelami dla danych z pominiętą jedną zmienną.

3.2. EKSPLORACJA DANYCH

wskaźnik	k-średnich $K = 5$	hier. k-średnich $K = 4$
APN	0.04	0.06
AD	3.26	3.73
ADM	0.17	0.35
FOM	0.65	0.74

Tablica 3.5: Miary stabilności

Dla każdego wskaźnika zwykła metoda k-średnich ma przewagę nad hierarchiczną metodą k-średnich, a więc jest bardziej stabilna niż hierarchiczna. Zatem model k-średnich z $K = 5$ uznaje się za ostateczny i to właśnie on będzie wykorzystany w aplikacji.

3.2.3. Charakterystyka grup obserwacji wyznaczonych przez model k-średnich

Przyjrzyjmy się jeszcze dokładniej grupom wyznaczonym przez model k-średnich. Analiza ta jest niezbędna dla użytkownika aplikacji, który jako jedną z informacji zwrotnych otrzyma krótką charakterystykę analizowanego tekstu (patrz rozdz. 4). Poniżej znajdują się wykresy skrzynkowe dla każdej zmiennej użytej do budowy modelu.

Na ich podstawie zostały przygotowane krótkie charakterystyki poszczególnych klas. W nawiasach zostały podane numery paneli, opisujących konkretne zmienne na wykresie 3.6:

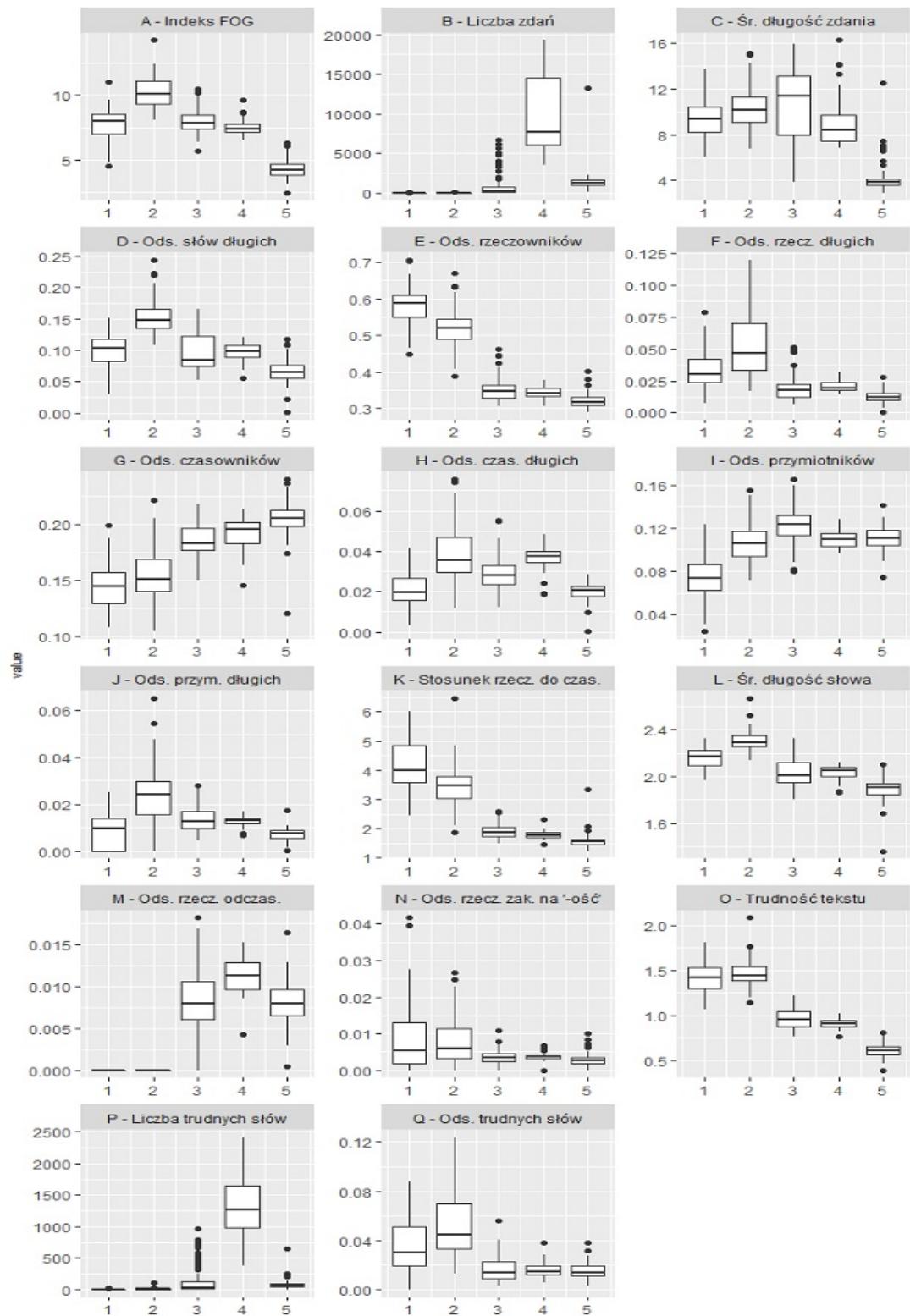
1. klasa 1 – bardzo krótkie teksty (B) o zdaniach średniej długości (C), trudne (O), mimo iż liczba słów trudnych jest bardzo mała (P), to w odniesieniu do długości tekstów, stanowią one znaczy odsetek (Q). Wartość indeksu FOG w obrębie grupy wynosi średnio 8 (A). Zawierają bardzo dużo rzeczowników (E) i mało czasowników (G). Klasę tę uznaje się jako klasę tekstów trudnych.
2. klasa 2 – bardzo krótkie teksty (B), zbudowane z długich zdań (C), trudne (O), odsetek długich (D) i trudnych (Q) słów jest wysoki. Wartość indeksu FOG w obrębie grupy wynosi średnio 10 (A), zatem jest to klasa tekstów trudnych.
3. klasa 3 – teksty średniej długości (B), ale o długich zdaniach (C) i średnim poziomie trudności (O), mały odsetek słów trudnych (Q). Teksty należące do tej klasy zawierają bardzo dużo przymiotników (I). Wartość indeksu FOG w obrębie grupy wynosi średnio 8 (A). Jest to klasa tekstów średnio trudnych.
4. klasa 4 – długie teksty (B), o zdaniach średniej długości (C), łatwe (O), mimo największej liczby słów trudnych (P), ich odsetek jest niski (Q), co pokazuje, iż nie wpływają one na

3. GRUPOWANIE TEKSTU POD WZGLEDEM TRUDNOŚCI

trudność w odbiorze tekstu. Wartość indeksu FOG w obrębie grupy wynosi średnio 7 (A). Klasę tę uznaje się jako klasę tekstów średnio trudnych.

5. klasa 5 – średniej długości teksty (B), zbudowane z bardzo krótkich zdań (C), łatwe (O), niski odsetek słów trudnych (Q). Wartość indeksu FOG w obrębie grupy wynosi średnio 5 (A). Klasa tekstów łatwych.

3.2. EKSPLORACJA DANYCH



Rysunek 3.6: Rozkłady poszczególnych zmiennych w klasach, model k-średnich $K = 5$

4. TrudneSłówka – opis aplikacji

W ostatniej części pracy zostało opisane praktyczne zastosowanie zbudowanego modelu. Przy pomocy pakietu 'Shiny' z pakietu R została przygotowana aplikacja, dzięki której możliwa jest weryfikacja potencjalnie trudnych słów w zadanym przez użytkownika tekście.

TrudneSłówka to narzędzie wspierające pracę osób konstruujących teksty skierowane do dzieci w wieku 7–13 lat, mająca na celu wskazanie potencjalnie trudnych i niezrozumiałych w odbiorze słów, dzięki czemu będzie można dane słowo zamienić na prostsze.

Aplikacja została zamieszczona na serwerach Shiny i można ją znaleźć pod adresem:

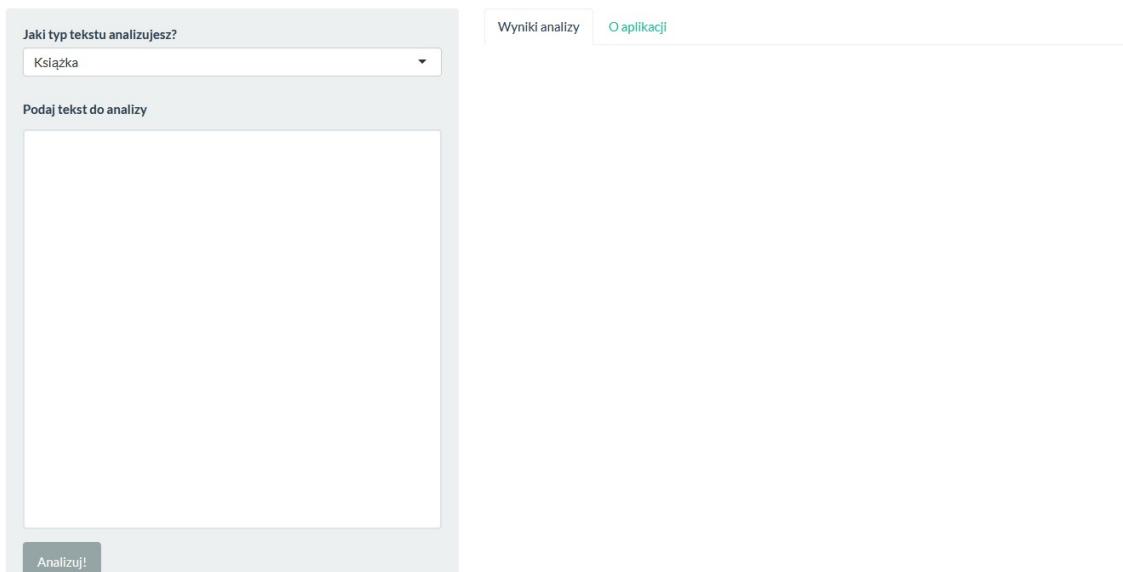
<https://trudneslowka.shinyapps.io/shinyapps/>. Po załadowaniu strony Użytkownik zobaczy prosty interfejs: po lewej stronie znajduje się lista z typami tekstów, jakie można analizować oraz pole tekstowe, w którym należy umieścić wybrany tekst. Do wyboru są trzy typy tekstów: książka (lub jej fragment), film animowany lub inny. Wyboru można dokonać z rozwijanej listy, znajdującej się ponad panelem zawierającym analizowany tekst. Naciśnięcie przycisku 'Analizuj!' spowoduje rozpoczęcie analizy tekstu.

Wyniki analizy są wyświetlane się po prawej stronie i zawierają, od góry:

1. przyporządkowanie do klasy trudności wraz z jej krótką charakterystyką, która powstała na podstawie analizy wykresu 3.6, patrz podrozdział 3.2.3,
2. oryginalny tekst z zaznaczonymi słowami potencjalnie trudnymi w odbiorze, gdzie natężenie koloru oznacza poziom trudności – żółty to słowa niepowszechnne, które mogą, ale nie muszą przeszkadzać w zrozumieniu tekstu; ciemnoczerwony kolor oznacza słowa bardzo rzadkie lub nie występujące w bazie danych, o którą oparta jest analiza, zatem są uznawane za słowa trudne i niezrozumiałe dla odbiorcy,
3. zestaw wykresów przedstawiających rozkłady dla takich zmiennych jak wartość indeksu FOG, średnia długości zdania oraz trudność tekstu. Czerwona linia oznacza pozycję analizowanego tekstu względem danego rozkładu, co pozwala określić, na ile analizowany tekst jest trudny,
4. ostatnim elementem outputu jest tabela z wyliczonymi statystykami dla analizowanego

tekstu. Zawiera wszystkie zmienne wykorzystane przy grupowaniu. Dodatkowo, w kolumnie **centyl** widnieje informacja, jak dana wartość statystyki odnosi się do bazy, na podstawie której zbudowano model. Ostatnia kolumna zawiera analogiczne statystyki, wyznaczone w obrębie typu tekstu, jaki podał na wejściu Użytkownik – nazwa tej kolumny będzie się zmieniać w zależności od wybranego typu tekstu. Istnieje możliwość posortowania wyników malejąco (rosnąco), co pozwala na łatwe odniesienie i sprawdzenie, co najbardziej wpływa na trudność i niezrozumiałość naszego tekstu.

TrudneSłówka - narzędzie do analizy trudności tekstów



Rysunek 4.1: TrudneSłówka – ekran startowy

Wykorzystanie aplikacji w praktyce zostanie zaprezentowane przy wykorzystaniu jednego ze scenariuszy, udostępnionych przez Centrum Nauki Kopernik, który nie został użyty budowie modelu – jest to scenariusz zajęć, który ma na celu zapoznać dzieci z falową naturą dźwięku.

Rysunki 4.2–4.4 przedstawiają wyniki dla wspomianego tekstu, który mimo, iż jest stosunkowo krótki, został zaklasyfikowany do grupy tekstów średnio trudnych. Dokładne statystyki prezentuje rysunek 4.3. Od razu rzuca się w oczy panel opisujący średnią długość zadania – analizowany tekst posiada je bardzo długie. Wysoki jest też indeks mglistości FOG. Więcej szczegółów można znaleźć na kolejnym rysunku 4.4. Łatwo zauważać, iż wymienione cechy posiadają bardzo wysokie wartości nie tylko w stosunku do bazy danych, ale też w obrębie wybranego typu tekstu.

4. TRUDNE SŁÓWKA – OPIS APLIKACJI

TrudneSłówka - narzędzie do analizy trudności tekstu

Jaki typ tekstu analizujesz?

Inne

Podaj tekst do analizy

z częsteczk, w której rozejda się drgania, np. powietrze (używasz tego zjawiska codziennie, kiedy rozmawiasz z drugą osobą, wasz dźwięk rozchodzi się właśnie przez powietrze), woda (możesz zanurkować na basenie lub w jeziorze i sprawdzić, czy da się tam porozmawiać z drugim człowiekiem) lub ciało stałe, np. rura (jeśli masz w domu wannę i mieszkanie w bloku, to jak zanurkujesz pod wodę, to czasem możesz usłyszeć sąsiadów, możesz też zbudować prosty model telefonu, jak jest to opisane poniżej). W przestrzeni kosmicznej nie ma częsteczek gazu (a tak naprawdę jest ich bardzo, bardzo mało), nie ma tam ośrodków, w których mógłby się rozchodzić dźwięk.

więc panuje tam absolutna cisza.

Pytania

Czy dźwięki były wyższe, czy niższe wraz z rosnącą długością rurki? Czy możesz znaleźć zależność między długością rurki a częstotliwością dźwięku?

Udało Ci się zagrać jakąś melodię? Udało Ci się zagrać duet na tym instrumencie?

Słowa kluczowe

Drgania, muzyka, ksylofon

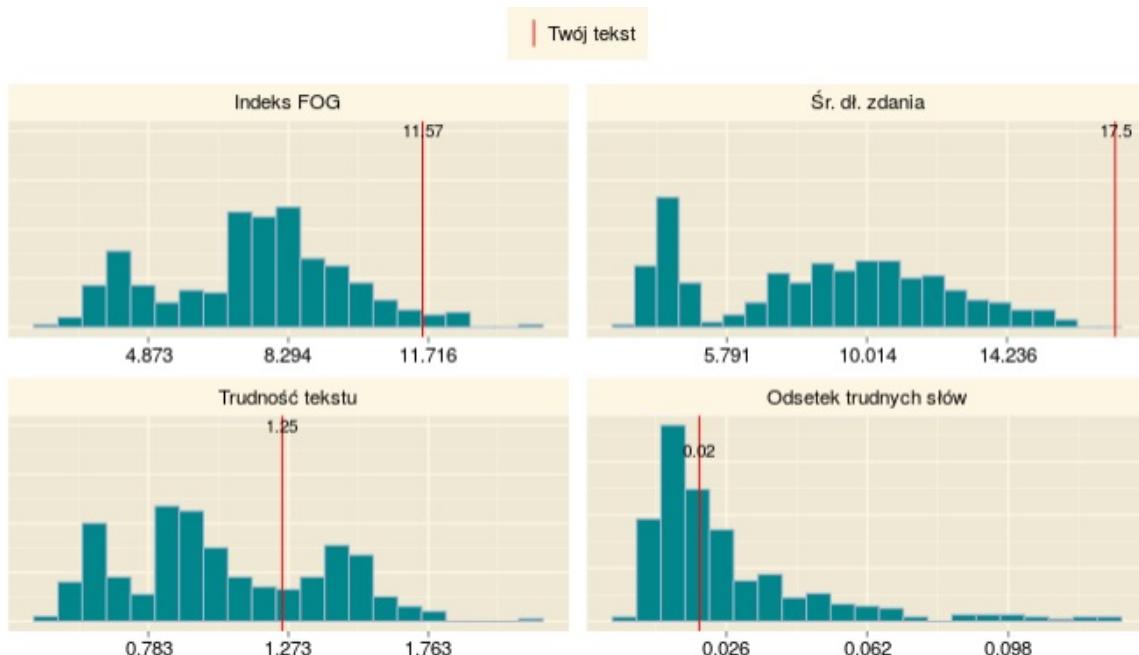
Analizuj!

Wyniki analizy O aplikacji

Oceniony poziom trudności - 3: teksty średniej długości, ale o długich zdaniach i średnim poziomie trudności, mały odsetek słów trudnych. Teksty należące do tej klasy zawierają bardzo dużo przyimków. Wartość indeksu FOG w obrębie grupy wynosi średnio 8. Jest to klasa tekstów średnio trudnych.

Temat scenariusza: Dźwięki muzyki Data dodania: 19.09.2017 Działanie: Fizyka Składniki • rurki pcw lub inne z twardego materiału, które można łatwo przeciąć • kolanka do rurek pcw lub innych, których używasz • łapka na muchy (z bardzo gęstą siatką albo całkowicie niedziurawą) / gietki klapki • kolorowe taśmki do obwiązywania rurek (opcjonalnie) • stroik (istnieją wersje online i aplikacje na telefon) lub kamerytony (dla osób o dobrym słuchu muzycznym) • szeroka taśma klejącej/ opaski zaciskowe (trytyki) Przebieg Etapy przeprowadzania doświadczenia 1. Uderz klapkiem w rurki o różnych długościach. 2. Sprawdź, czym różnią się dźwięki z rurek o różnej długości. Spróbuj zagrać na nich jakąś melodię. 3. Umieść je w odpowiednim stelażu, np. zbudowanym z innych rurek pcw albo desek za pomocą szerokiej taśmy klejącej lub opasek opasek zaciskowych. 4. Nastroi je - przytnij rurki tak, żeby dawały określone dźwięki! (skorzystaj z aplikacji na telefon lub, jeśli masz dobry słuch, możesz ją nastroić używając zestawu kamerytonów). Jeśli chcesz mieć dźwięki o innej wysokości, a nie starca ci już miejsca na dłuższą rurkę, możesz użyć kolanka, do wydłużenia rury (poprowadzenia jej w innym kierunku). 5. Kiedy już otrzymasz dźwięki o określonych wysokościach przewiąż rurki wstążką, żebyś wiedział, która rurka odpowiada za który dźwięk. 6. Ponownie spróbuj zagrać jakąś melodię. Ciekawostki Dźwięki, które słyszy człowiek to te, których częstotliwość mieści się od około 20 do 20 000 Hertzów [Hz] (od 25- 30 lat słuch zaczyna się pogarszać i ludzie coraz gorzej słyszą wysokie dźwięki). Istnieją zwierzęta, które do komunikacji używają dźwięków niższych (stonek komunikują się z tw. infradźwiękami o częstotliwości ponizej 16 Hz) lub wyższych (ultradźwięków, czyli dźwięków bardzo wysokich, o częstotliwościach przekraczających 20 000 Hz używają na przykład nietoperze i delfiny, ale słyszą je również psy). Dźwięk jest organizem częsteczek. Zabój dźwięk się przemieszcza, potrzebny jest ośrodek, czyli materia zbudowana z częsteczek, w której rozejda się drgania, np. powietrze (używasz tego zjawiska codziennie, kiedy rozmawiasz z drugą osobą, wasz dźwięk rozchodzi się właśnie przez powietrze), woda (możesz zanurkować na basenie lub w jeziorze i sprawdzić, czy da się tam porozmawiać z drugim człowiekiem) lub ciało stałe, np. rura (jeśli masz w domu wannę i mieszkanie w bloku, to jak zanurkujesz pod wodę, to czasem możesz usłyszeć sąsiadów, możesz też zbudować prosty model telefonu, jak jest to opisane poniżej). W przestrzeni kosmicznej nie ma częsteczek gazu (a tak naprawdę jest ich bardzo, bardzo mało), nie ma tam ośrodków, w których mógłby się rozchodzić dźwięk, więc panuje tam absolutna cisza. Pytania Czy dźwięki były wyższe, czy niższe wraz z rosnącą długością rurki? Czy możesz znaleźć zależność między długością rurki a częstotliwością dźwięku? Udało Ci się zagrać jakąś melodię? Udało Ci się zagrać duet na tym instrumencie? Słowa kluczowe Drgania, muzyka, ksylofon

Rysunek 4.2: Analizowany tekst wraz z zaznaczonymi słowami potencjalnie trudnymi



	Wartość	Centyl	Inne
Śr. długość zdania (słowa)	17.5	100	100
Indeks FOG	11.57	97	100
Odsetek rzeczowników zakończonych na '-ość'	0.02	97	100
Odsetek rzeczowników długich	0.04	84	100
Odsetek czasowników długich	0.03	71	100
Odsetek czasowników	0.2	70	32
Trudność tekstu	1.25	69	100
Odsetek słów długich	0.11	67	79
Średnia długość słowa (sylaby)	2.12	63	83
Odsetek rzeczowników	0.4	62	100
Odsetek przymiotników	0.11	62	49
Stosunek rzeczowników do czasowników	2.01	55	97
Odsetek trudnych słów	0.02	51	66
Odsetek przymiotników długich	0.01	46	76
Liczba zdań	24	28	0
Liczba trudnych słów	8	23	1
Odsetek rzeczowników odczasownikowych	0	1	0

Showing 1 to 17 of 17 entries

Rysunek 4.4: Szczegółowe statystyki dla analizowanego tekstu

Kolejnym przykładem jest krótki fragment tekstu z Wikipedii dotyczący Wikingów ([18]). Ten przykład pokazuje, jak mogą postrzegać tego typu teksty dzieci w wieku szkolnym. O ile dorosły nie powinien mieć problemu ze zrozumieniem poniższego tekstu, o tyle dziecku mogą sprawić trudność na przykład nazwy własne, które pojawiają się dość licznie.

4. TRUDNE SŁÓWKA – OPIS APLIKACJI

Tekst został zaklasyfikowany jako trudny. Rysunek 4.7 pokazuje, iż wszystkie cechy, definiujące trudność tekstu mają bardzo wysokie wartości – zarówno w obrębie typu tekstu jak i w całej bazie.

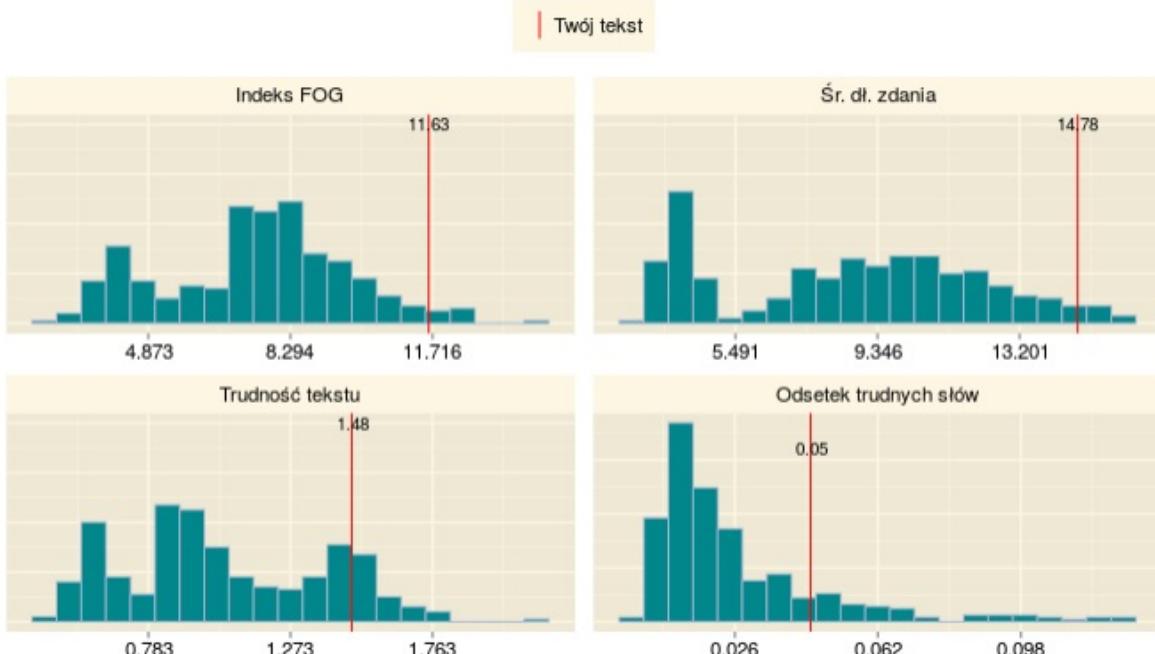
TrudneSłówka - narzędzie do analizy trudności tekstów

The screenshot shows the TrudneSłówka application interface. On the left, there's a text input area with the text: "przepływali aż na Morze Kaspijskie. Wikingowie bogacili się jednak nie tylko ląpieniem ale również handlem. Prowadzili intratny handel z Arabami którzy z srebra z Taszkenatu i Afganistanu dostarczali futra ozdoby z metalu szlachetnych i niewolników z terenów nadbałtyckich. Po fazie najazdów rabunkowych Normanowie zaczęli się osiedlać na zdobytych terenach zwłaszcza na wyspach: Orkachach Szetlandach Islandii a także w Irlandii i Brytanii oraz na półwyspie Cotentin we Francji północnej gdzie założyli księstwo Normandii [potrzebny przypis]. Zamieszkali w nim wikingowie zaczeli używać języka francuskiego a w 919 roku przyjęli chrześcijaństwo. Państwo wikingów w Normandii istniało przez ponad 300 lat stając się w końcu księstwem lennym króla Francji. Dotarli także w głąb dzisiejszej Rosji i Ukrainy (Kijów znany jako Kanugard był ich faktorią handlową) oraz do Wielkiej Brytanii i Europy kontynentalnej." Below this is a button labeled "Analizuj!".

On the right, under the "Wyniki analizy" tab, the results are displayed:

- Oceniony poziom trudności - 2:** bardziej krótkie teksty, zbudowane z długich zdań, trudne, odsetek długich i trudnych słów jest wysoki. Wartość indeksu FOG w obrębie grupy wynosi średnio 10, zatem jest to klasa tekstów trudnych.
- Pierwszy odnotowany napad wikingów miał miejsce w roku 793 na wybrzeżach Anglii (Lindisfarne).** Początkowo ich wyprawy nie wychodziły poza obręb Morza Północnego i wysp szkockich – Orkadów Szetlandów Hebrydów. Około roku 800 po raz pierwszy wylądowali na Wyspach Owczych które szybko zostały przez nich skolonizowane. W roku 874 dotarli również do Islandii. Kolejny wiek przyniósł odkrycie Grenlandii (ok. 982) i Ameryki (późnysyp Labrador i Nowa Fundlandia – ok. 1000[5]) gdzie również założyli swoje osady. Jednak walki z tubylcami i później wewnętrzne rycie spowodowały porzucenie osad w Nowym Świecie. Z biegiem lat bandy wikingów rozrosły się do połkaznych flotylli i zwiększyły się ich zasięg. Niektóre grupy optywując Półwysep Iberyjski dostawały się na Morze Śródziemne i grabyły wybrzeża południowej Galii i Italii. Inne grupy wyprowadzając się na wschód penetrowały szlaki i tereny wzdłuż rzek Dżawiny i Wołchow oraz Dniepu aż do Morza Czarnego gdzie ląpły przybrzeżne miasta Cesarstwa Bizantyńskiego. Wolała przepływać aż na Morze Kaspijskie. Wikingowie bogacili się jednak nie tylko ląpieniem ale również handlem. Prowadzili intratny handel z Arabami którymi za srebro z Taszkenatu i Afganistanu dostarczali futra ozdoby z metalu szlachetnych i niewolników z terenów nadbałtyckich. Po fazie najazdów rabunkowych Normanowie zaczęli się osiedlać na zdobytych terenach zwłaszcza na wyspach: Orkachach Szetlandach Islandii a także w Irlandii i Brytanii oraz na półwyspie Cotentin we Francji północnej gdzie założyli księstwo Normandii [potrzebny przypis]. Zamieszkali w nim wikingowie zaczeli używać języka francuskiego a w 919 roku przyjęli chrześcijaństwo. Państwo wikingów w Normandii istniało przez ponad 300 lat stając się w końcu księstwem lennym króla Francji. Dotarli także w głąb dzisiejszej Rosji i Ukrainy (Kijów znany jako Kanugard był ich faktorią handlową) oraz do Wielkiej Brytanii i Europy kontynentalnej."

Rysunek 4.5: Analizowany tekst wraz z zaznaczonymi słowami potencjalnie trudnymi



Rysunek 4.6: Histogramy dla wybranych zmiennych

	Wartość	Centyl	Inne
Indeks FOG	11.63	98	100
Śr. długość zdania (słowa)	14.78	97	100
Odsetek czasowników długich	0.05	96	100
Odsetek słów długich	0.14	87	93
Średnia długość słowa (sylaby)	2.28	87	100
Trudność tekstu	1.48	87	100
Odsetek trudnych słów	0.05	86	100
Stosunek rzeczowników do czasowników	3.02	71	100
Odsetek przynimotników długich	0.01	70	84
Odsetek rzeczowników długich	0.03	68	98
Odsetek rzeczowników	0.42	64	100
Odsetek przynimotników	0.1	31	10
Liczba trudnych słów	12	30	2
Liczba zdań	18	18	0
Odsetek czasowników	0.14	12	0
Odsetek rzeczowników odczasownikowych	0	1	0
Odsetek rzeczowników zakończonych na '-ość'	0	1	0

Showing 1 to 17 of 17 entries

Rysunek 4.7: Szczegółowe statystyki dla analizowanego tekstu

Bibliografia

- [1] Banerjee A. & Dave R.N., *Validating clusters using the Hopkins statistic*. IEEE International Conference on Fuzzy Systems: Budapest, Hungary, 2004, 1. 149 - 153 vol.1.
- [2] Bańko M., Górski R., Red. Przepiórkowski A., Red. & Lewandowska-Tomaszczyk B., *Narodowy Korpus Języka Polskiego*. Warszawa : Wydawnictwo Naukowe PWN, 2012.
- [3] Bezdek J., & Hathaway R. *VAT: A tool for visual assessment of (cluster) tendency*. Neural Networks, 2002. IJCNN '02, Proceedings of the 2002 International Joint Conference on 3, 2225-2230.
- [4] Broda B., Ogorodniczuk M., Nitoń B., Gruszczyński W. *Measuring Readability of Polish Texts: Baseline Experiments*. [w:] N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (red.), ELRA, Reykjavik 2014, Materiały konferencji LREC 2014 (9th International Conference on Language Resources and Evaluation, Reykjavik, 26-31 maja 2014), s. 573–580.
- [5] Dale E., Chall J. *A Formula for Predicting Readability*, Educational Research Bulletin, 1948, 27: 11–20+28.
- [6] Datta S., & Datta S., *Comparisons and validation of statistical clustering techniques for microarray gene expression data*. Bioinformatics, 20303, 19(4), 459-466.
- [7] Dębowksi Ł., Nitoń B., Broda B., Charzyńska E. *Jasnopis — A Program to Compute Readability of Texts in Polish Based on Psycholinguistic Research*. [w:] B Sharp, W. Lubaszewski and R. Delmonte (red.), Natural Language Processing and Cognitive Science. Proceedings 2015, s. 51-61. Libreria Editrice Cafoscarina, 2015 (materiały konferencji NLPCS 2015: 12th International Workshop on Natural Language Processing and Cognitive Science, Kraków, 22-24 września 2015)
- [8] Flesch Rudolf. *How to Write Plain English.*, Barnes & Noble, 1981.
- [9] Gruszczyński W., Broda B., Charzyńska E., Dębowksi Ł., Hadryan M., Nitoń B., Ogorodniczuk M. *Measuring Readability of Polish Texts*. [w:] Z. Vetulani, J. Mariani (red.), Wydaw-

nictwo Poznańskie i Fundacja Uniwersytetu im. A. Mickiewicza, Poznań 2015, Materiały konferencji LTC 2015, s 445-449.

- [10] Gruszczyński W.Broda B. Niton B., Ogrodnik M., *W poszukiwaniu metody automatycznego mierzenia zrozumiałości tekstów informacyjnych.*, Poradnik Językowy, 2015, 9-22.
- [11] Gunning R. *The technique of clear writing*, McGraw-Hill, 1952.
- [12] Hastie T., Tibshirani R., & Friedman J., *The Elements of Statistical Learning.*, New York, NY: Springer New York, 2009.
- [13] Kassambara A., *Practical Guide To Cluster Analysis in R, Unsupervised Machine Learning*, STHDA, 2017.
- [14] Koronacki J., & Ćwik J., *Statystyczne systemy uczące się* (Wyd. 2 [rozsz.]. ed.), Warszawa: Akademicka Oficyna Wydawnicza EXIT, 2008.
- [15] Pisarek W., *Jak mierzyć zrozumiałość tekstu.*, Zeszyty Prasoznawcze, Wydawnictwo Uniwersytetu Jagiellońskiego, 1969, 4(42):35–48.
- [16] Tibshirani R., Walther G., & Hastie T., *Estimating the number of clusters in a data set via the gap statistic.*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(2), 2001, 411-423.
- [17] <https://github.com/morfologik/polimorfologik> dostęp na dzień 19.11.2018 r.
- [18] Fragment wpisu z Wikipedii dotyczący Wikingów, https://pl.wikipedia.org/wiki/Wikingowie#Etapy_ekspansji, dostęp na dzień 08.12.2018.
- [19] Zasady dzielenia wyrazów na sylaby w języku polskim, <https://wordlist.eu/strona/18,podzial-na-sylaby>, dostęp na dzień 06.12.2018.

Spis rysunków

1.1	Jasnopolis.pl, ekran wejściowy	15
1.2	Jasnopolis.pl, wyniki dla fragmentu tekstu dotyczącego Wikingów ([18])	16
1.3	Logios.pl, ekran wejściowy	16
1.4	Logios.pl, wynik działania programu	17
2.1	Wykres VAT, zbiór Iris, a). dane oryginalne, b). dane wygenerowane losowo z rozkładu jednostajnego	20
2.2	Wykres usypiskowy dla zbioru Iris, metoda k-średnich	21
3.1	Macierz korelacji zmiennych użytych do budowy modelu	31
3.2	Wykres VAT a). dane oryginalne, b).dane losowe	32
3.3	Ocena optymalnej liczby skupień: a). statystyka odstępu, b). metoda sylwetki, c). metoda wykresu usypiskowego	33
3.4	Rozkład obserwacji w klasach dla wybranych modeli	34
3.5	Wykresy sylwetek dla wybranych modeli	36
3.6	Rozkłady poszczególnych zmiennych w klasach, model k-średnich $K = 5$	39
4.1	TrudneSłówka – ekran startowy	41
4.2	Analizowany tekst wraz z zaznaczonymi słowami potencjalnie trudnymi	42
4.3	Histogramy dla wybranych zmiennych	42
4.4	Szczegółowe statystyki dla analizowanego tekstu	43
4.5	Analizowany tekst wraz z zaznaczonymi słowami potencjalnie trudnymi	44
4.6	Histogramy dla wybranych zmiennych	44
4.7	Szczegółowe statystyki dla analizowanego tekstu	45

Spis tabel

1.1	Przedziały określające klasy trudności ([11])	13
1.2	Wartości indeksu FRES wraz z ich interpretacją ([8])	13
1.3	Wartości indeksu Dall'ego–Chall'a wraz z interpretacją ([5])	14
1.4	Przedziały wartości indeksu Pisarka określające trudność tekstu ([15])	14
3.1	Cechy wyznaczone dla każdego tekstu	29
3.2	Przykładowe rekordy ze słownika Polimorfologik ([17])	30
3.3	Wewnętrzne miary jakości dopasowania	35
3.4	Procent błędnie zaklasyfikowanych obserwacji, wyznaczony na podstawie wartości sylwetek poszczególnych obserwacji	36
3.5	Miary stabilności	37