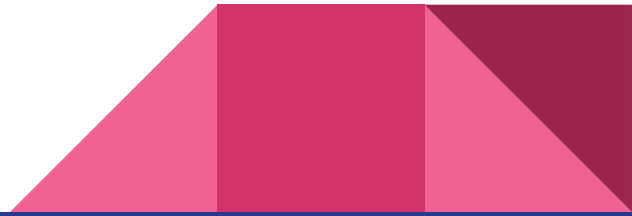
Klasyfikacja list na stronach internetowych z uwzględnieniem atrybutów wizualnych

Marcin Maliszkiewicz

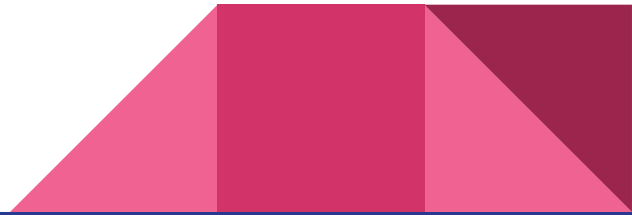
Cel pracy

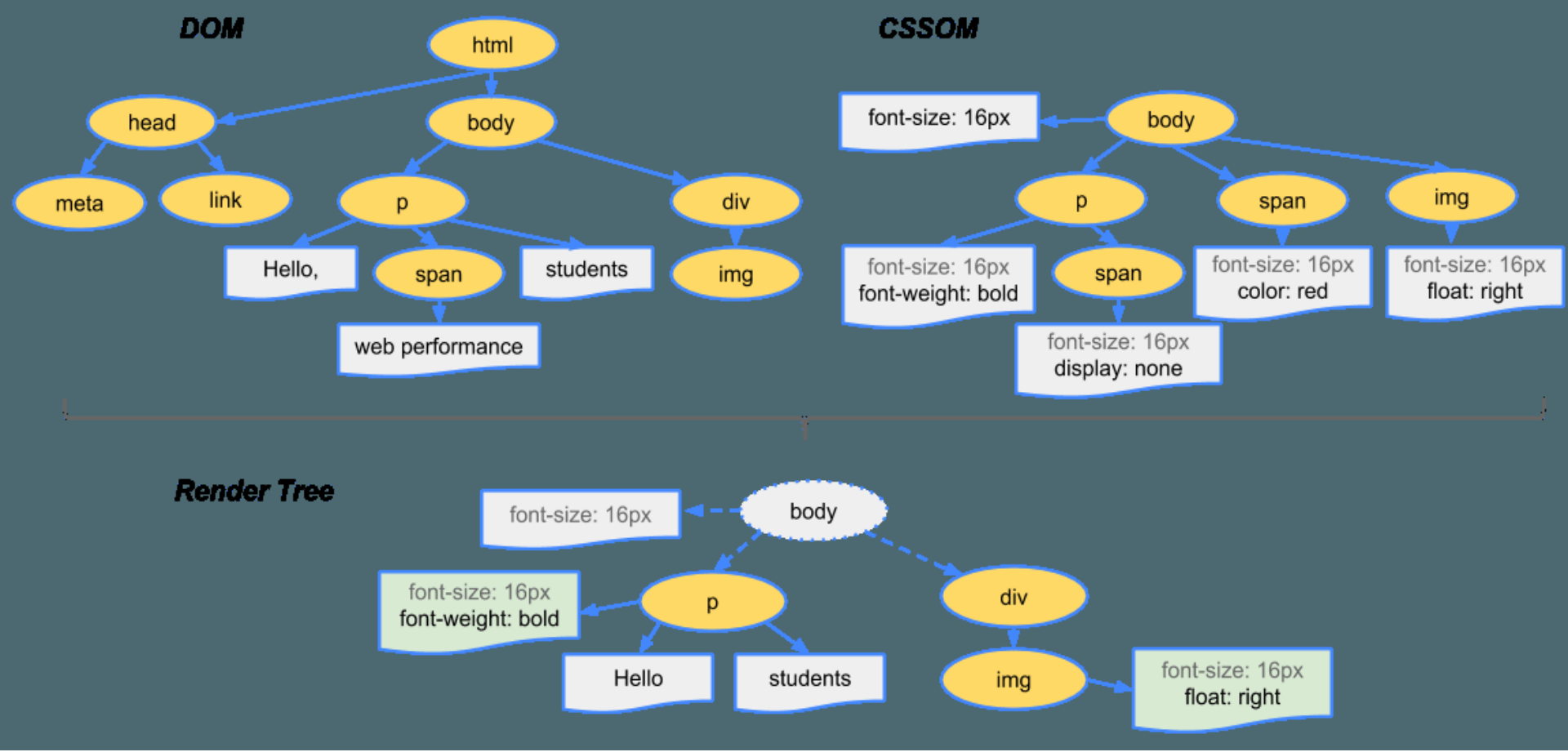
Celem pracy było stworzenie i przetestowanie systemu potrafiącego znaleźć listy na stronach internetowych.



Budowa stron internetowych

1. Strony są zbudowane ze znaczników HTML, stylów CSS, skryptów JavaScript.
2. Przeglądarka tworzy z nich drzewa DOM i CSSOM, a następnie drzewo renderowania.
3. Współczesne strony składają się z bardzo dużej liczby elementów (setki bądź tysiące).





https://developers.google.com/web/fundamentals/performance/critical-rendering-path/render-tree-construction?hl=pl

China orders evacuation of blast area

The Chinese authorities order the evacuation of residents within a 3km radius of the Tianjin blast site as sodium cyanide is confirmed nearby.

2 hours ago | China

Man found alive near blast site

Latest pictures of aftermath

What do we know about blast?

The potent chemicals involved

Migrants die in boat's hold off Libya

At least 40 migrants are found to have died in the hold of a crowded fishing boat intercepted off the Libyan coast, the Italian navy says.

1 hour ago | Europe

Navy video shows migrant rescue

Why is EU struggling with migrants?

Japan emperor 'remorseful' over WW2

Japanese Emperor Akihito expresses "deep remorse" over Japan's role in World War Two on the 70th anniversary of the end of the conflict.

55 minutes ago | Asia

UK veterans mark VJ Day

Saying sorry in Japanese

Lebanon arrests wanted radical cleric

16 minutes ago | Middle East

SPORT **First openly gay player leaves NFL**

4 hours ago

Ex-army chief shot dead in Burundi

53 minutes ago | Africa

IS leader 'raped US hostage Mueller'

2 hours ago | US & Canada

Apple car clues emerge from letter

15 August 2015 | Technology | 234

Minister resigns over BBC interview

16 minutes ago | Asia

SPORT **McCaw sets record as All Blacks win**

4 hours ago | Rugby Union

Also in the News >

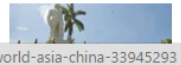
Siri 'saves trapped US teenager'

14 August 2015 | US & Canada

Beavers cut through Siberian railway wires

14 August 2015

The Reporters >



Watch/Listen

LIVE World Service radio

Man found alive near China blast epicentre

4 hours ago | China

Fire helicopter winches driver to safety

1 hour ago

Navy video shows migrant rescue

2 hours ago | Europe

London flypast marks WW2 anniversary

1 hour ago | UK

Anti-Japan protests in South Korea

2 hours ago | Asia

Japan commemorations - in 60 seconds

9 hours ago | Asia

Markets

FTSE 100 >	6,550.74	▼ -0.27%
Dow Jones >	17,477.4	▲ +0.4%
Nasdaq >	5,048.23	▲ +0.29%
Nikkei 225 >	20,519.45	▼ -0.37%

15 minute delay. Last updated 16:44

Share with BBC News >

Send us an SMS or MMS to +44 7624 800100

Email us at **haveyoursay@bbc.co.uk**

Follow Have Your Say on **Twitter**

Follow Us

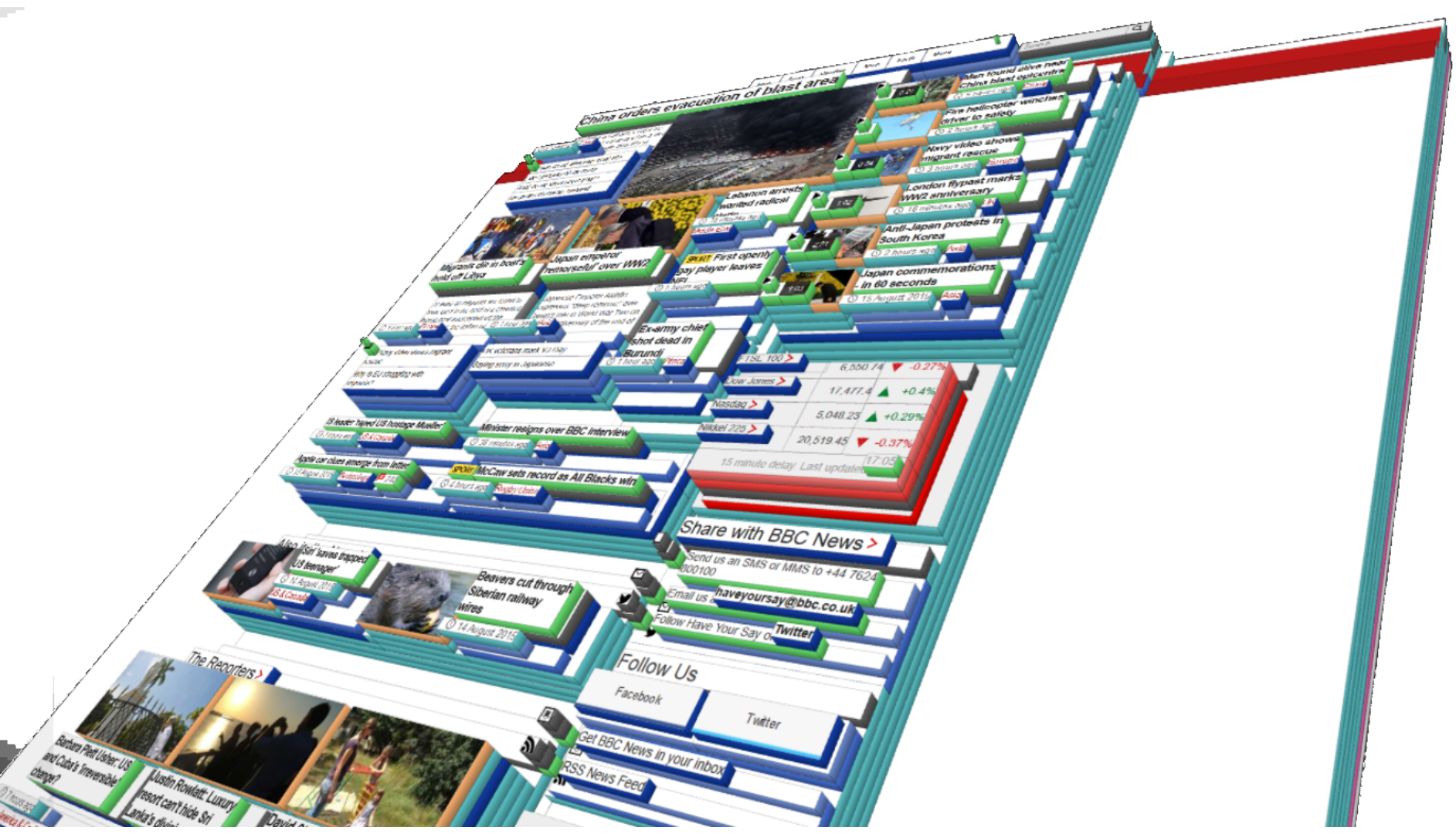
Facebook | Twitter

Get BBC News in your inbox

RSS News Feed

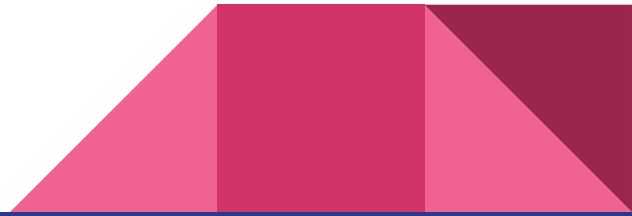
Most Popular

Read | Watched


China orders evacuation of blast area
Man found alive near China blast epicentre
Fire helicopter winches driver to safety
Navy video shows migrant rescue
London flypast marks WW2 anniversary
Anti-Japan protests in South Korea
Japan commemorations in 60 seconds
Lebanon arrests 'wanted radical'
Migrants die in boat's hold off Libya
Japan emperor 'frets herself' over WW2
First openly gay player leaves NFL
Ex-army chief shot dead in Burundi
FTSE 100
6,550.14 -0.27%
Dow Jones
17,477.4 +0.4%
Nasdaq
5,048.23 +0.29%
Nikkel 225
20,519.45 -0.37%
15 minute delay. Last update 17:05
US boxer 'injured US hostage Mueller'
MI5 boss reelgns over BBC interview
Apple car cities emerge from letter
McCaw sets record as All Blacks win
Share with BBC News
Send us an SMS or MMS to +44 7624 800100
Email us haveyoursay@bbc.co.uk
Follow Have Your Say on Twitter
'Fish' saves trapped US teenager
Beavers cut through Siberian railway wires
14 August 2015
The Reporters
Barbara Pert Usher: US and Cuba's 'irreversible change'?
Justin Rowlatt: Luxury resort can't hide Sri Lanka's divisi
Follow Us
Facebook
Twitter
Get BBC News in your inbox
RSS News Feed

Definicja i przykłady list

Fragment strony będący wyliczeniem obiektów o charakterze informacyjnym.



Przedmioty na aukcji

OBRAZ 30X30 OBRAZY Kuchnia Salon ARTDECO WYPRZEDAŻ

kup teraz **8,00 zł**

z dostawą 17,90 zł

zakończona 23 lipca 2015 o godzinie 17:14

12 osób kupiło

KOMPLET POŚCIEL BAWEŁNA SATYNA 200X220 3D 6-CZ

kup teraz **189,00 zł**

z dostawą 207,00 zł

zakończona 23 lipca 2015 o godzinie 17:15

1 osoba kupiła

62222 AMBITION ZAPARZACZ DO KAWY HERBATY

kup teraz **35,00 zł**

Prognoza pogody

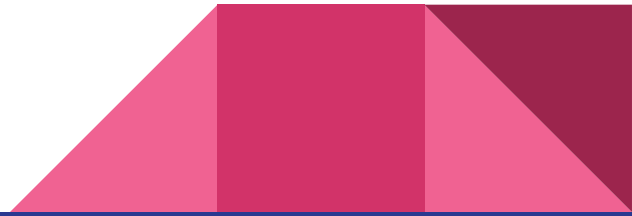
Godzina	Prognoza	Wiatr	Opady	Wilgotność
18 00	25°C Odczuwalna 24°C Częściowo słonecznie	17 km/h WNW Max 17 km/h	Zachm: 58% Deszcz: 0mm	60%
19 00	24°C Odczuwalna	16 km/h	Zachm: 51%	59%

Lista funduszy emerytalnych i ich stopy zysku

Fundusze o najwyzszej stopie zwrotu (12m)	1 mies.	3 mies.	12 mies.
MetLife Specjalistyczny Fundusz...	+1,82 ▲	+9,25 ▲	+32,95 ▲
MetLife Specjalistyczny Fundusz...	+1,37 ▲	+9,67 ▲	+31,84 ▲
NN Specjalistyczny Fundusz Inwestycyjny...	-1,30 ▼	+17,77 ▲	+28,28 ▲
Pioneer Walutowy Fundusz Inwestycyjny...	+1,40 ▲	+2,04 ▲	+26,77 ▲
Investor Parasol Specjalistyczny Fundusz...	-1,75 ▼	+0,99 ▲	+26,71 ▲
MetLife Specjalistyczny Fundusz...	-1,04 ▼	+3,09 ▲	+26,29 ▲

Architektura stworzonego systemu

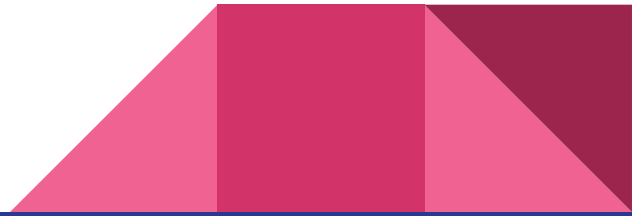
1. serwer
   a. przechowywanie próbek stron
   b. klasyfikatory SVM, Naive Bayes
2. klient
   a. pobieranie elementów i atrybutów
   b. interfejs oznaczania próbek
   c. (przyciski sterujące)



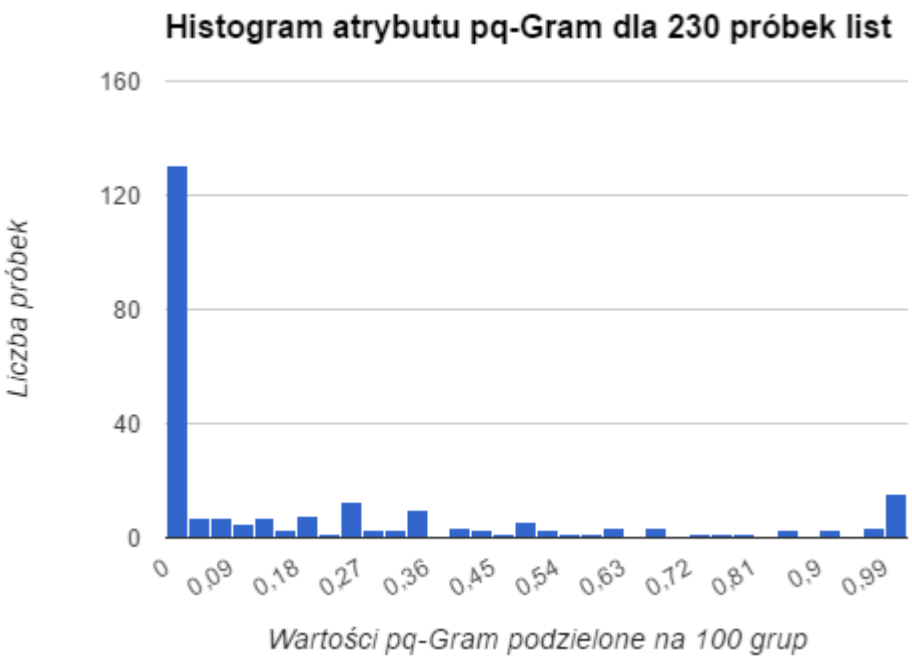
Demonstracje

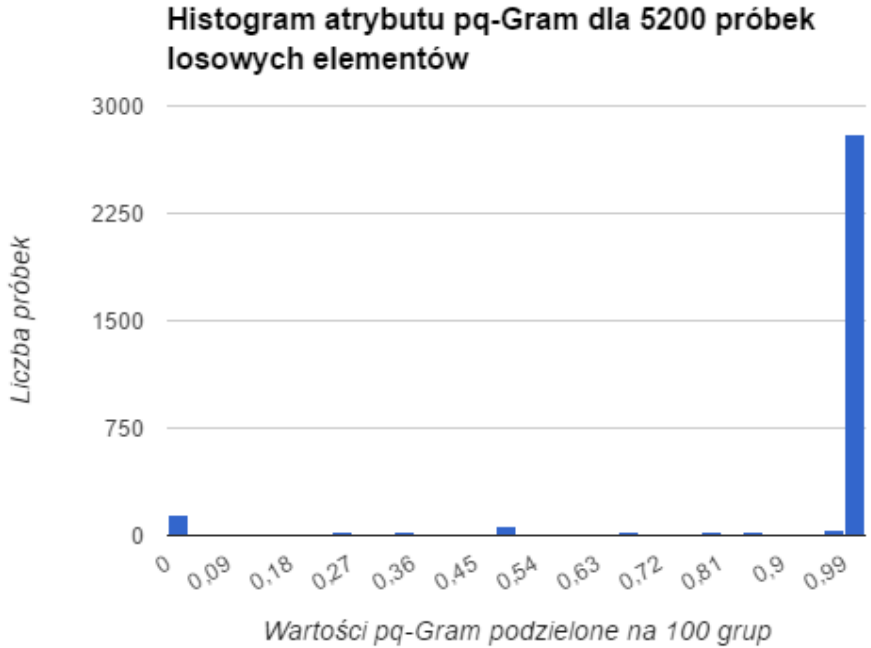
Dobór atrybutów

- rozmiar i położenie,
- rozkład koloru tekstu,
- rozkład wielkości tekstu,
- **wewnętrzne podobieństwo struktury (pq-Gram)**,
- **średnia liczba słów wśród elementów potomnych,**
- **minimalna wariancja rozmiaru i położenia.**

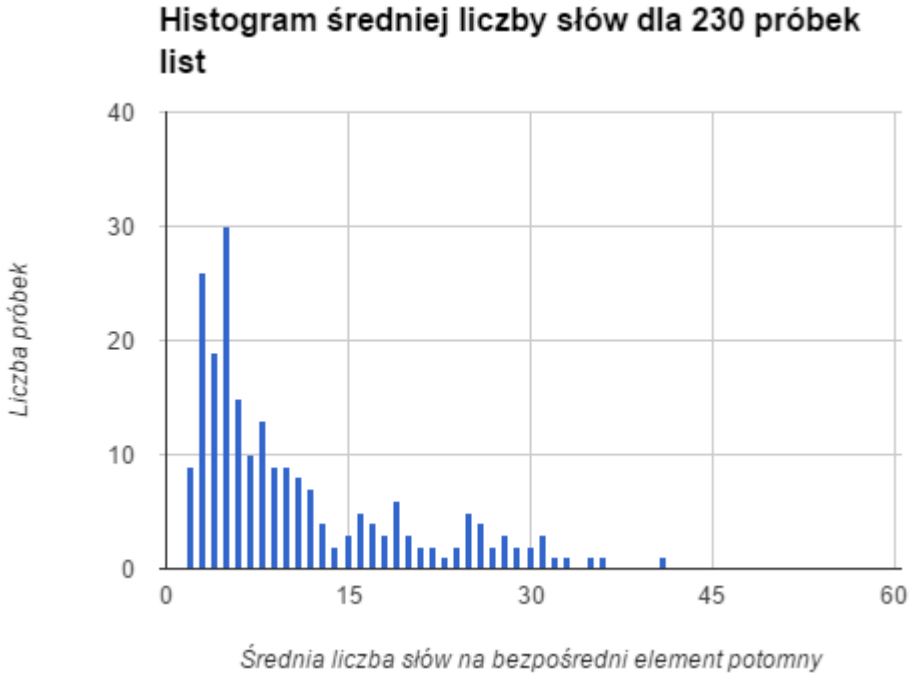


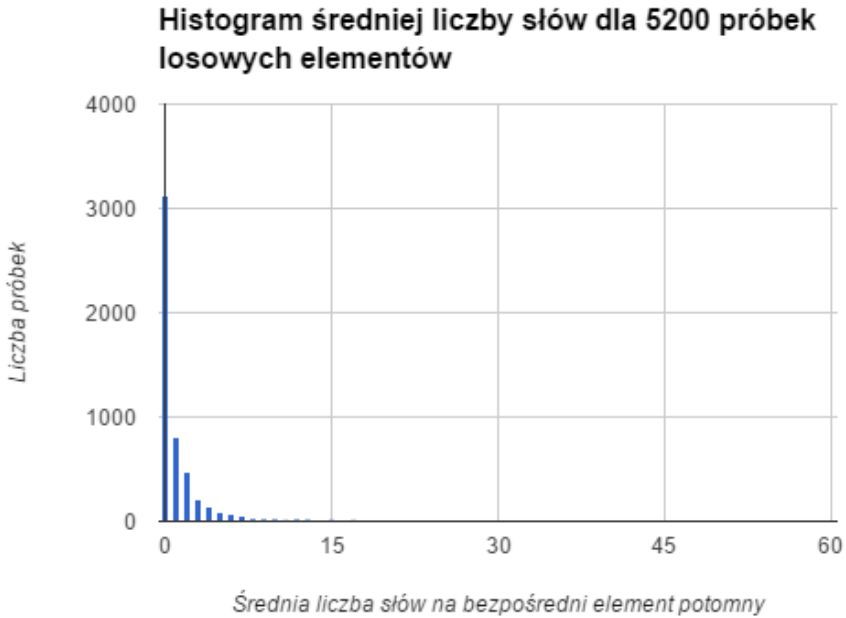
pq-Gram


Histogram atrybutu pq-Gram dla 230 próbek list
160
120
80
40
0
Liczba próbek
0
0,09
0,18
0,27
0,36
0,45
0,54
0,63
0,72
0,81
0,9
0,99
Wartości pq-Gram podzielone na 100 grup

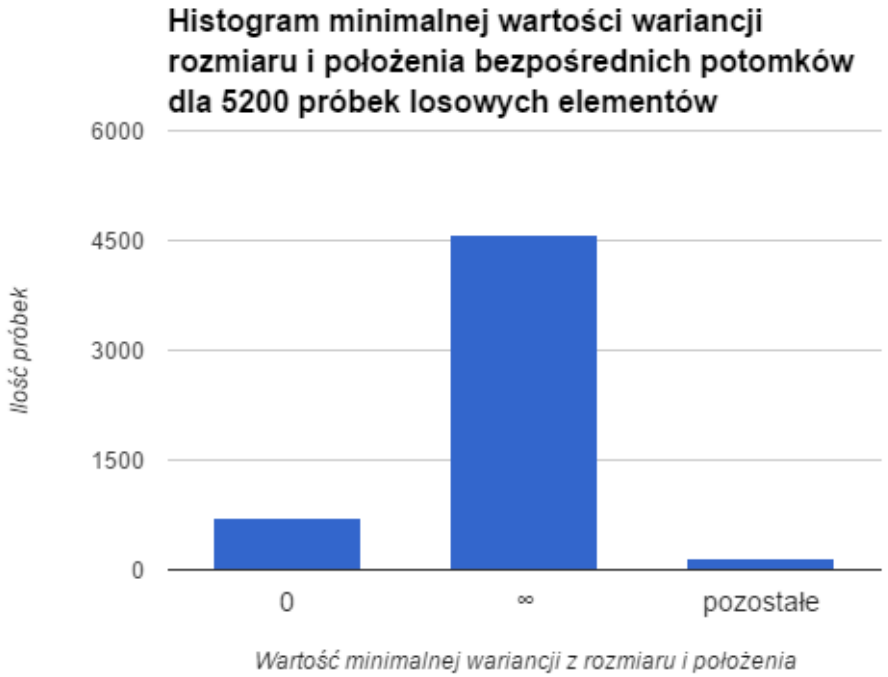

Histogram atrybutu pq-Gram dla 5200 próbek losowych elementów
3000
2250
1500
750
0
Liczba próbek
0
0,09
0,18
0,27
0,36
0,45
0,54
0,63
0,72
0,81
0,9
0,99
Wartości pq-Gram podzielone na 100 grup

Średnia liczba słów





Wariancja rozmiaru i położenia



Rezultaty

Próbki grupy trenującej, X pozytywnych + Y negatywnych	SVM		Naive Bayes	
	Czułość [%]	Precyzja [%]	Czułość [%]	Precyzja [%]
125 + 1716	52	40	85	35
125 + 125	96	20	89	29
125 + 0	100	4	100	4

Tabela 6.2: Ocena klasyfikacji trenowanej z różnym udziałem próbek negatywnych, wyliczona dla 250 próbek pozytywnych i 5200 próbek negatywnych.

Dziękuję za uwagę