

# **Zaawansowane techniki integracji systemów**

## **Projekt**

Temat 8: Środowisko do pobierania i przechowywania informacji  
z portalu blogowego *salon24*

Krzysztof Papciak

### **1. Wprowadzenie**

W ramach zadania eksplorowane były dane z serwisu blogowego *salon24*. Pierwszym krokiem wykonania projektu było przygotowanie aplikacji pobierającej dane ze strony serwisu (tzw. web scraper).

Portal blogowy oferuje możliwość przeglądnięcia wszystkich blogów z odpowiednim sortowaniem. Po otwarciu danego bloga jest możliwość przejrzania wszystkich notek oraz sprawdzenia właściciela bloga, komentarzy oraz ilości udostępnień na popularnych serwisach społecznościowych. Program pobrał pobierał te wszystkie dane i zapisał je do bazy danych MySQL.

#### **1. 1. Cele projektu**

Podczas pracy nad projektem skupiano się na celach opisanych poniżej.

Cele projektu:

- analiza portalu salon24 pod kątem źródeł do pozyskania danych
- zaprojektowanie web scrappera, który wczyta strony i pomoże wydzielić dane
- przygotowanie modelu bazy danych
- zapis danych do bazy danych przy pomocy napisanych programów
- pobieranie danych (ze wszystkich blogów)
- analiza statystyczna pobranych danych i przygotowanie wizualizacji przy użyciu grafów

#### **1. 2. Opis portalu blogowego salon24**

Salon24.pl to polski serwis blogowy o tematyce społeczno-politycznej, na którym pisać i komentować może każdy zarejestrowany użytkownik. Zawiera on blogi poruszające między innymi następującą tematykę: polityka, biznes, ekonomia, media, Internet, kultura, sport, podróże, zdrowie, nauka.

Serwis został uruchomiony 16 października 2006 r. przez Bognę Janke i Igora Janke.

Blogi założyli tu znani publicyści (m.in. Rafał A. Ziemkiewicz, Jan Pospieszalski, Paweł Wroński,

Janusz Rolicki, Sławomir Sierakowski), a także tysiące „zwykłych” ludzi. Założyciele serwisu – dziennikarskie małżeństwo Bogna Janke i Igor Janke w marcu 2007 roku za stworzenie witryny salon24.pl zostali nominowani do Nagrody im. Dariusza Fikusa w kategorii Wydawca.

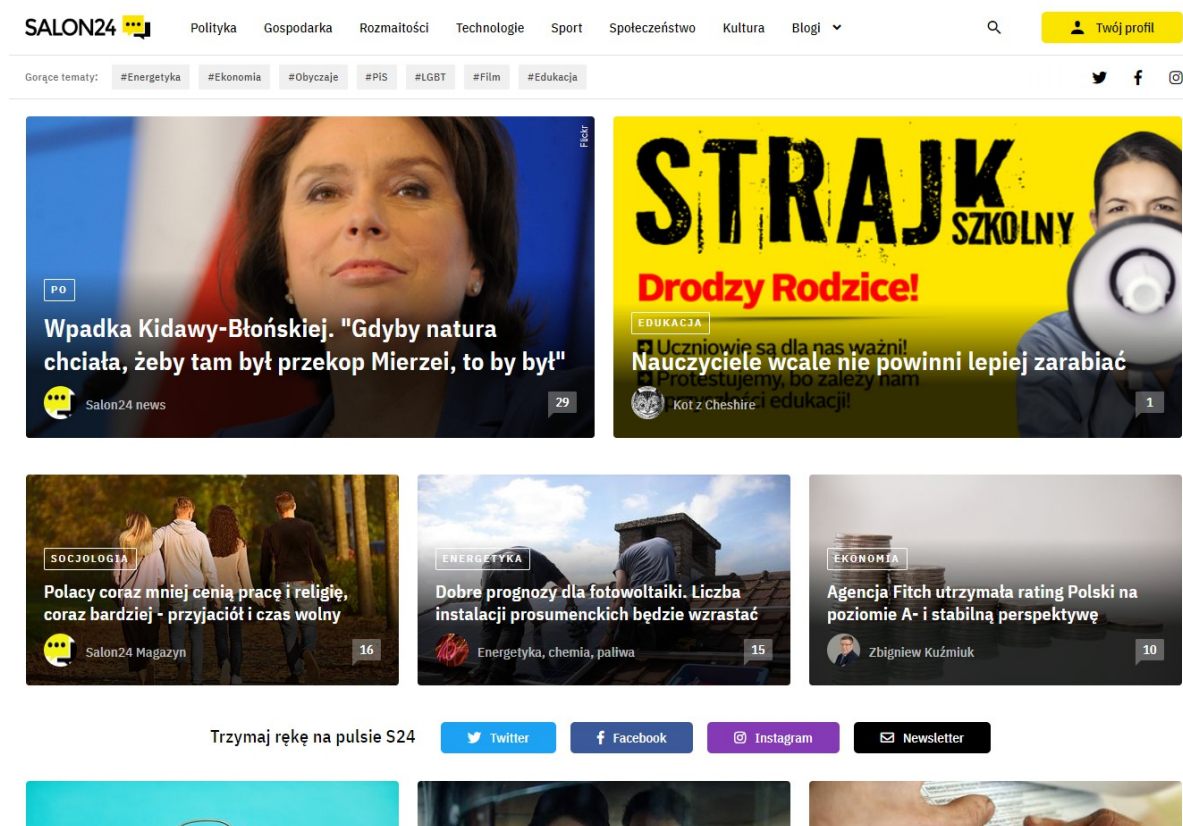
Salon24.pl ma blisko 1 milion unikalnych użytkowników i 8 milionów odsłon miesięcznie (Google Analytics, styczeń 2014). W serwisie jest zarejestrowanych ponad 22 tysiące blogów. Wśród autorów są dziennikarze (m.in. Paweł Lisicki, Marek Magierowski, Piotr Gabryel, Łukasz Warzecha, Mariusz Max Kolonko, Agnieszka Romaszewska, Janina Jankowska, Tomasz Terlikowski, Krzysztof Kłopotowski, Tomasz Rożek), politycy (m.in. Jarosław Gowin, Paweł Kowal, Janusz Wojciechowski, Ryszard Czarnecki, Aleksandra Jakubowska), blogi oficjalne (m.in. Ośrodek Studiów Wschodnich, Instytut Wolności, Instytut Sobieskiego, Forum Rosja-Polska).

Teksty z salonu24.pl były wykorzystywane w radiowych przeglądach prasy oraz przedrukowywane w mediach elektronicznych (m.in. Onet.pl) i w prasie drukowanej (m.in. w „Dzienniku Polskim”, „Rzeczpospolitej”). Blogerzy z Salonu24 występowali także w programach telewizyjnych (np. Warto rozmawiać).

Tekst Igora Janke Nieznośna szybkość bloga o salonie24.pl (opublikowany w „Rzeczpospolitej”, „Plus Minus”, 10–12.11.2006 r.) był jednym z tematów egzaminu maturalnego z języka polskiego 5 maja 2008 r.

*Na podstawie: wikipedia.pl*

### 1. 3. Analiza serwisu *salon24*

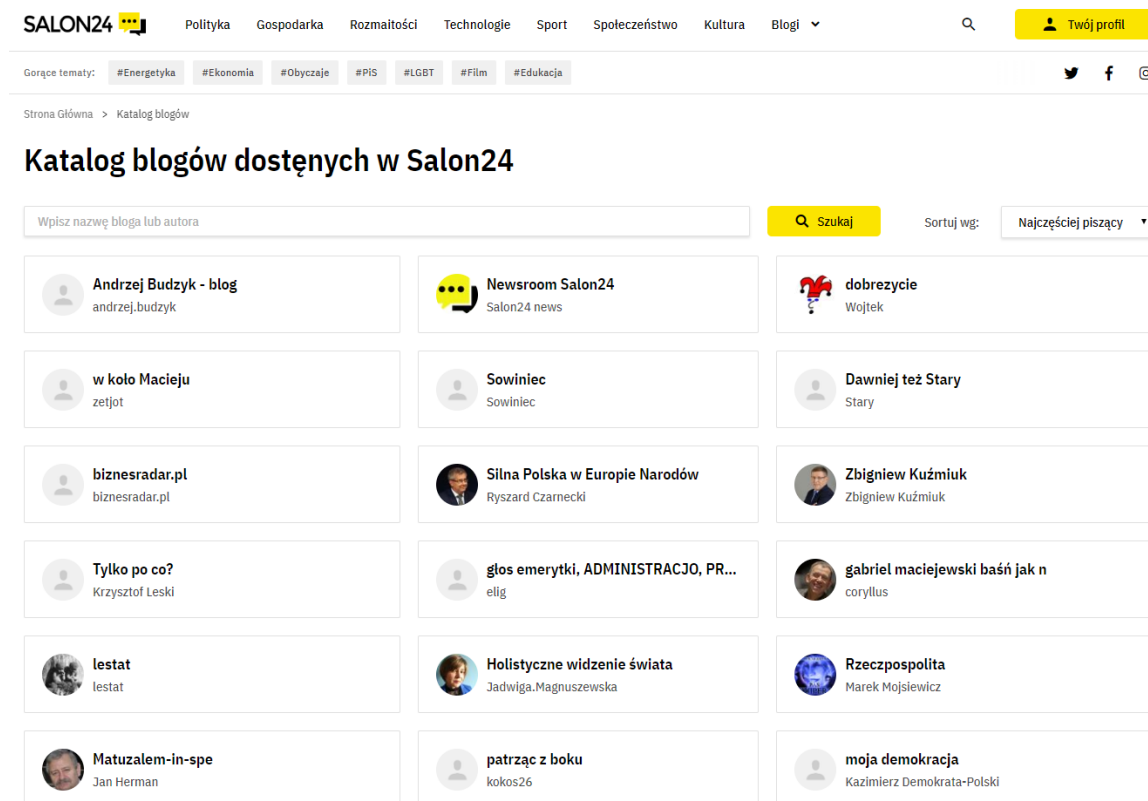


Rys 1. Strona główna serwisu blogowego *salon24*

Pierwszym krokiem wykonania zadania była analiza serwisu *salon24* pod kątem możliwości

automatycznego pobierania danych. Na rysunku 1 pokazano stronę główną serwisu.

Portal blogowy *salon24* posiada możliwość przeglądania blogów stosując różnego rodzaju sortowanie, między innymi: po dacie, alfabetycznie, po najczęściej piszących użytkownikach (rys. 2).



Rys. 2. Lista blogów

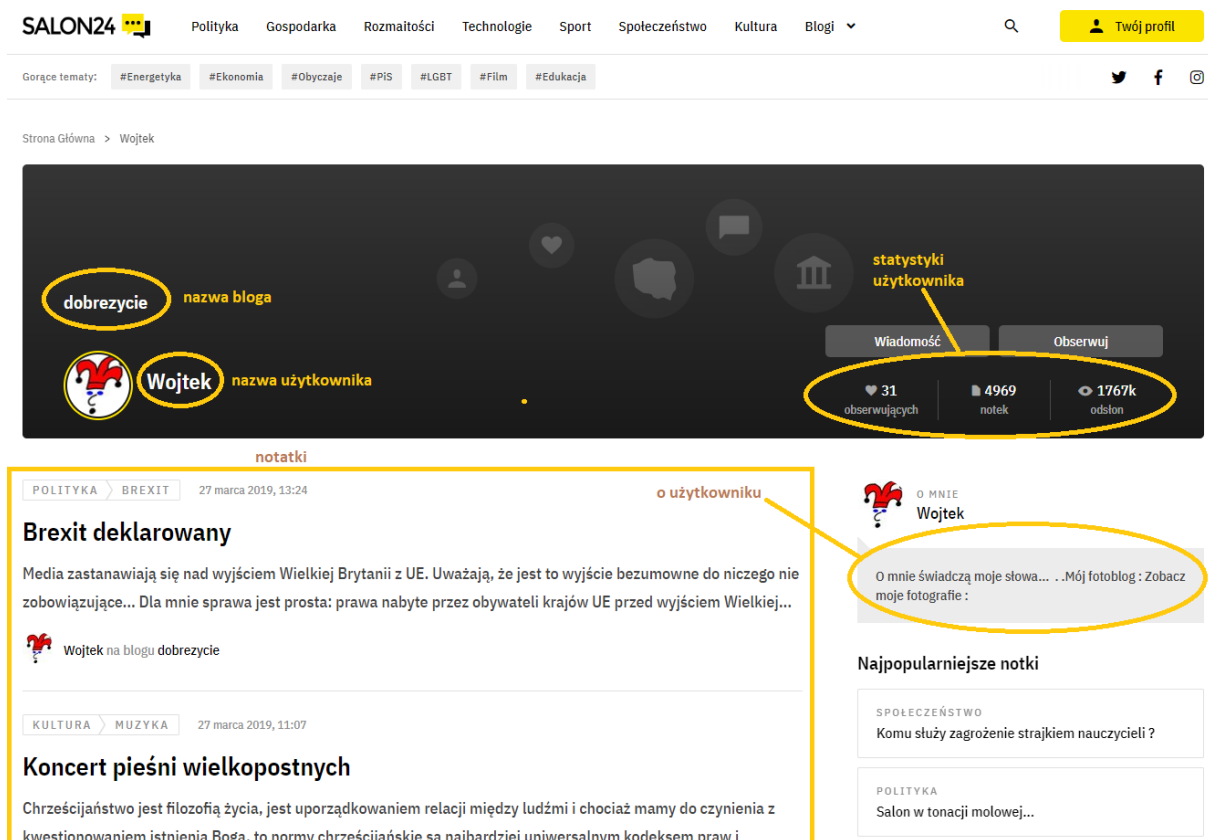
Każdy użytkownik może posiadać jeden blog. Użytkownicy (a zatem także blogi) podzielone są na 1236 stron. Program musi zatem przebiegać po kolejnych stronach i odwiedzać poszczególne blogi.

Każda strona blogu posiada w nagłówku informacje o użytkowniku, a także takie jak ilość wyświetleń, udostępnień na facebooku, liczbę notatek oraz ilość śledzących go użytkowników.

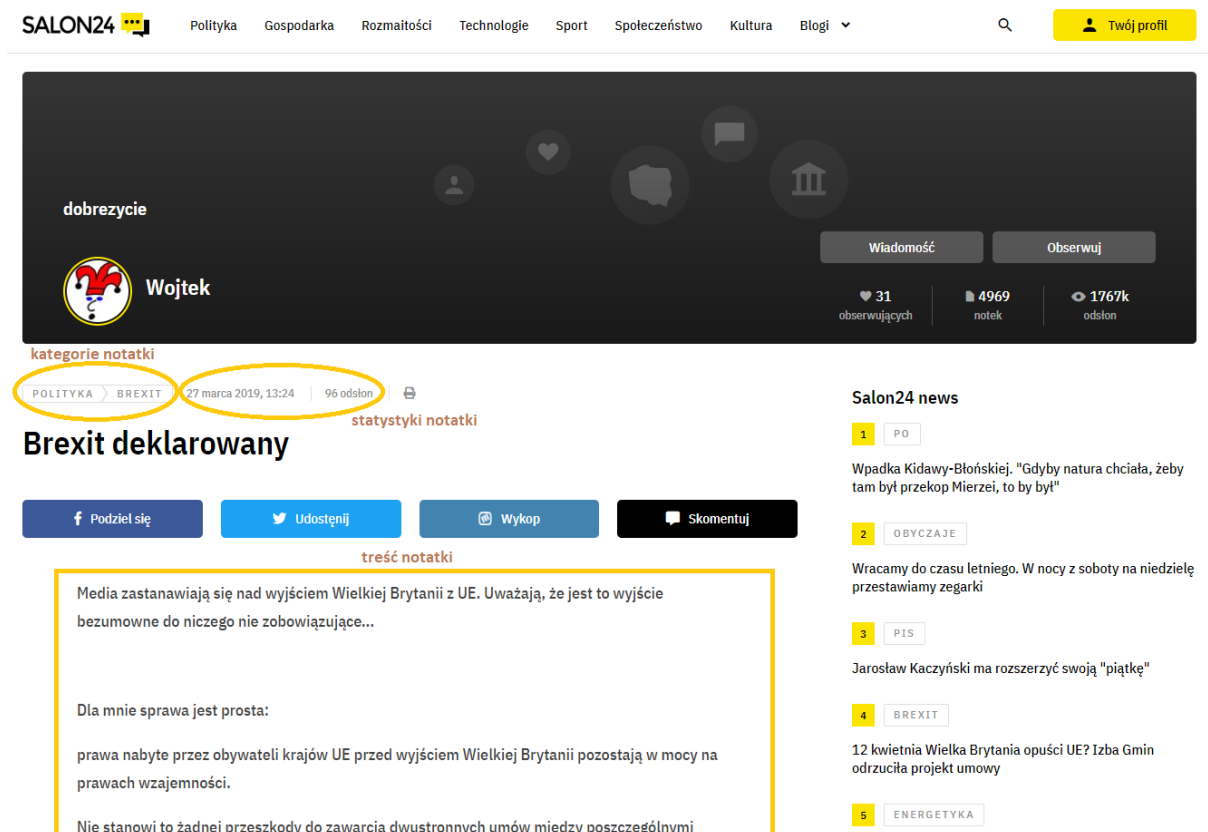
W dalszej części strony znajduje się opis użytkownika (opcjonalny) lista najpopularniejszych notatek, komentarze użytkownika oraz lista notatek (rys 3). Jeżeli notatek jest więcej niż zmieści się na jednej stronie pojawia się przycisk *następna strona*. Web scraper pobiera te wszystkie informacje oraz przechodzi przez kolejne strony notatek i odwiedza poszczególne wpisy.

Każda z notatek zawiera nad samym tekstem kategorie, do których należy oraz datę i liczbę wyświetleń (rys 4).

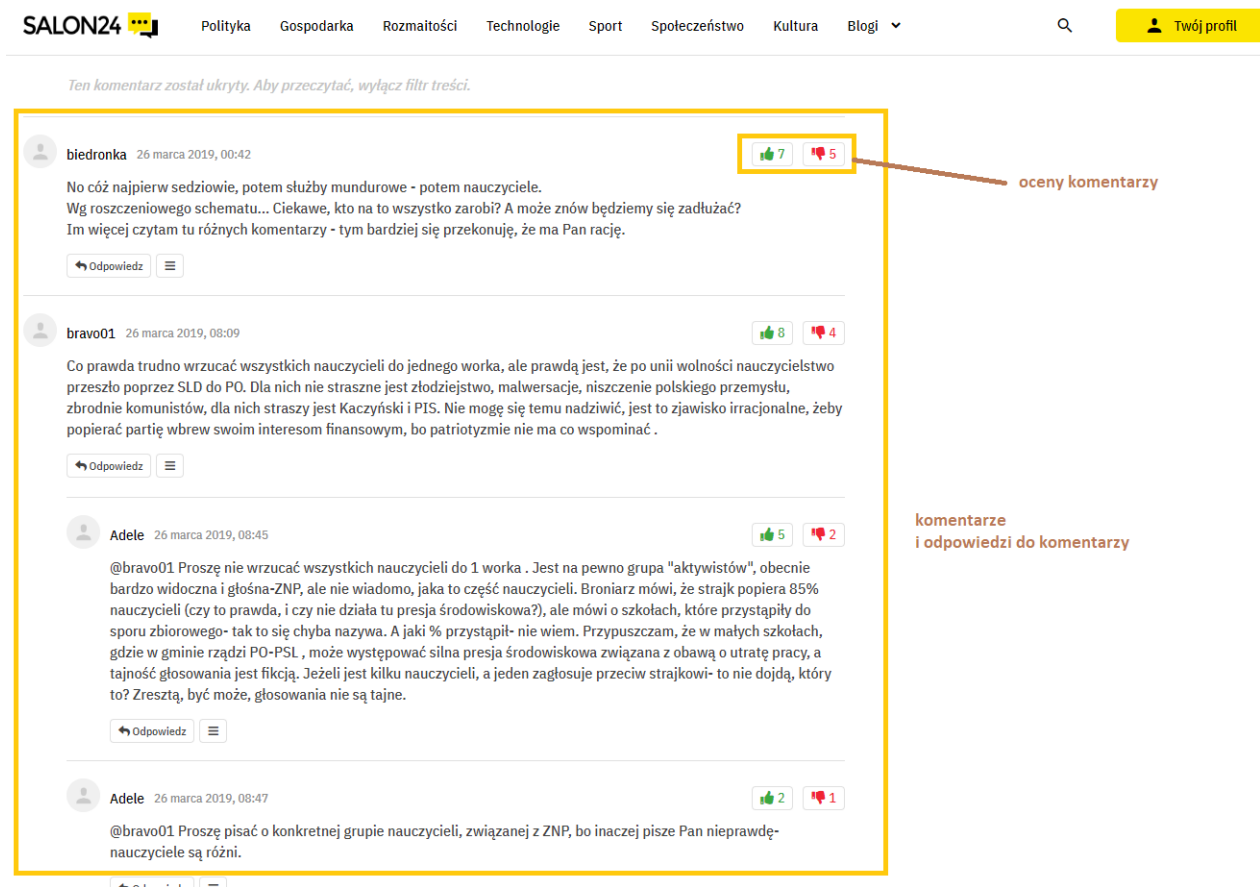
U dołu strony wyświetlane są komentarze. Są one ułożone hierarchicznie (możliwe są odpowiedzi na komentarze). Każdy komentarz posiada możliwość wystawienia pozytywnej lub negatywnej oceny (co wyświetlane jest po prawej stronie każdego wpisu) (rys. 5).



Rys. 3. Widok bloga



Rys. 4. Widok notatki



Rys. 5. Komentarze i oceny

Po przeanalizowaniu źródła danych, przystąpiono do przygotowania modelu bazy danych a następnie do pisania programu pobierającego dane ze strony.

## 2. Model danych

W poniższych punktach opisano model bazy danych oraz ujęto statystyki jej dotyczące.

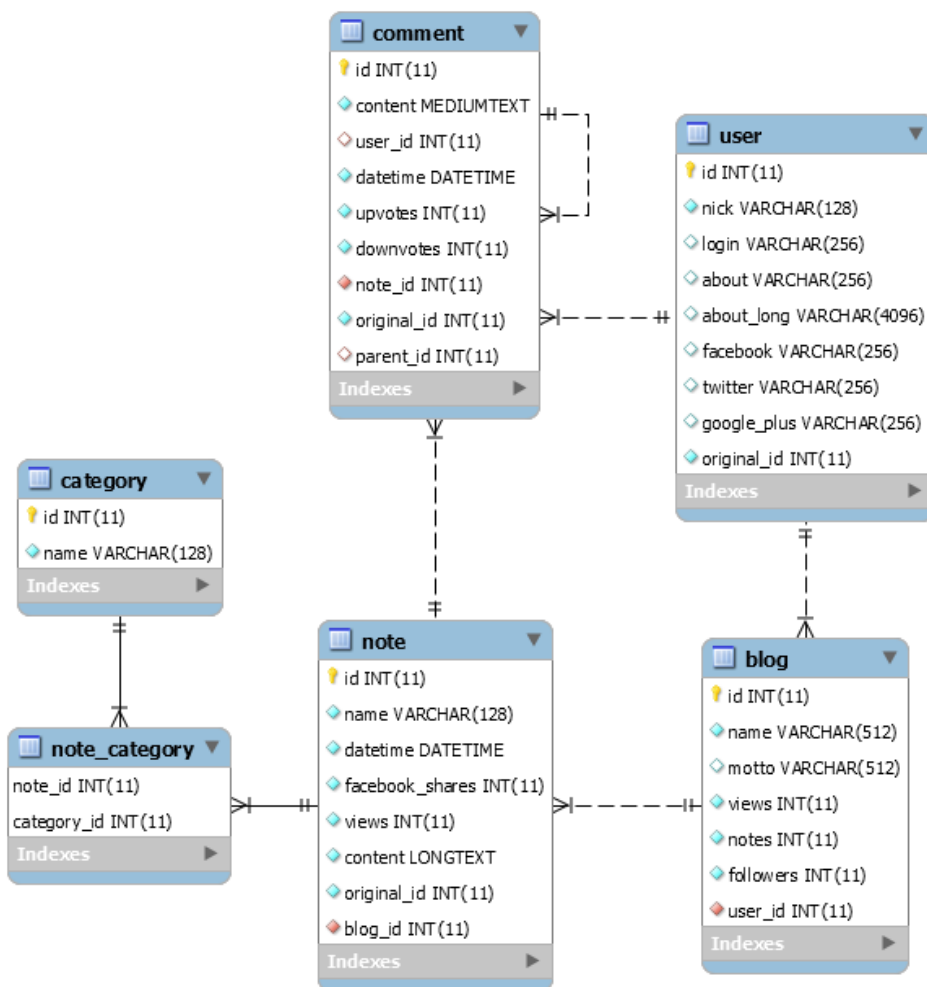
### 2. 1. Baza danych

Do przechowywania pobranych danych wykorzystano lokalną bazę danych SQL (konkretnie *MySQL*). Przygotowano model uwzględniający wszystkie dostępne na stronie informacje.

Portal *salon24* posiada następującą strukturę organizacji danych:

- użytkownik
- blog
  - notatki
  - komentarze

Rysunek 6 przedstawia model bazy danych.



Rys. 6. Model bazy danych

Opis poszczególnych tabel w bazie danych:

- **user** – użytkownik portalu
  - id – klucz główny
  - nick – pseudonim użytkownika
  - login – login użytkownika
  - about – krótki opis użytkownika
  - about\_long – długi opis użytkownika
  - facebook – adres do strony użytkownika na facebooku
  - twitter – adres do strony użytkownika na twitterze
  - google\_plus – adres do strony użytkownika na google plus
  - original\_id – id użytkownika w bazie danych serwisu *salon24*
- **blog**
  - id – klucz główny

- name – nazwa bloga
- motto – motto bloga
- views – liczba wyświetleń
- notes – liczba notatek
- followers – liczba śledzących użytkowników
- user\_id – id użytkownika (właściciela bloga)
- **note** – notatka w blogu
  - id – klucz główny
  - name – tytuł notatki
  - datetime – data i czas dodania notatki
  - facebook\_shares – liczba udostępnień na facebooku
  - views – liczba wyświetleń
  - content – treść notatki
  - original\_id – id notatki w bazie danych serwisu *salon24*
  - blog\_id – id bloga, do którego należy notatka
- **category** – kategoria notatki
  - id – klucz główny
  - name – nazwa kategorii
- **note\_category** – tabela łącznikowa notatki i kategorii  
notatka może posiadać wiele kategorii
- **comment** – komentarz  
notatka może posiadać wiele komentarzy, a komentarze mają strukturę hierarchiczną
  - id – klucz główny
  - content – treść komentarza
  - user\_id – użytkownik komentujący
  - datetime – data i czas dodania komentarza
  - upvotes – liczba pozytywnych ocen
  - downvotes – liczba negatywnych ocen
  - note\_id – id notatki, do której jest komentarz
  - original\_id – id komentarza w bazie danych serwisu *salon24*
  - parent\_id – id nadrzędnego komentarza (jeżeli istnieje)

## 2. 2. Statystyki

Dane były pobierane przez ponad tydzień. Przez ten czas pobrano następujące ilości instancji.  
 Użytkowników: 52 144

Blogów: 18 968

Notatek: 614 725

Komentarzy: 9 915 885

Liczba pobranych blogów nie zgadza się z liczbą figurującą w notatce z wikipedii z powodu usunięcia niektórych blogów.

Rozmiar wyeksportowanego pliku bazy danych to ok. 6.5GB

### **3. Algorytmy**

W podpunkcie opisano zastosowane algorytmy w programie do pobierania danych, a także te zastosowane do przygotowania statystyk i wizualizacji.

#### **3.1. Użyte technologie**

Do napisania programu pobierającego dane użyto języka Python.

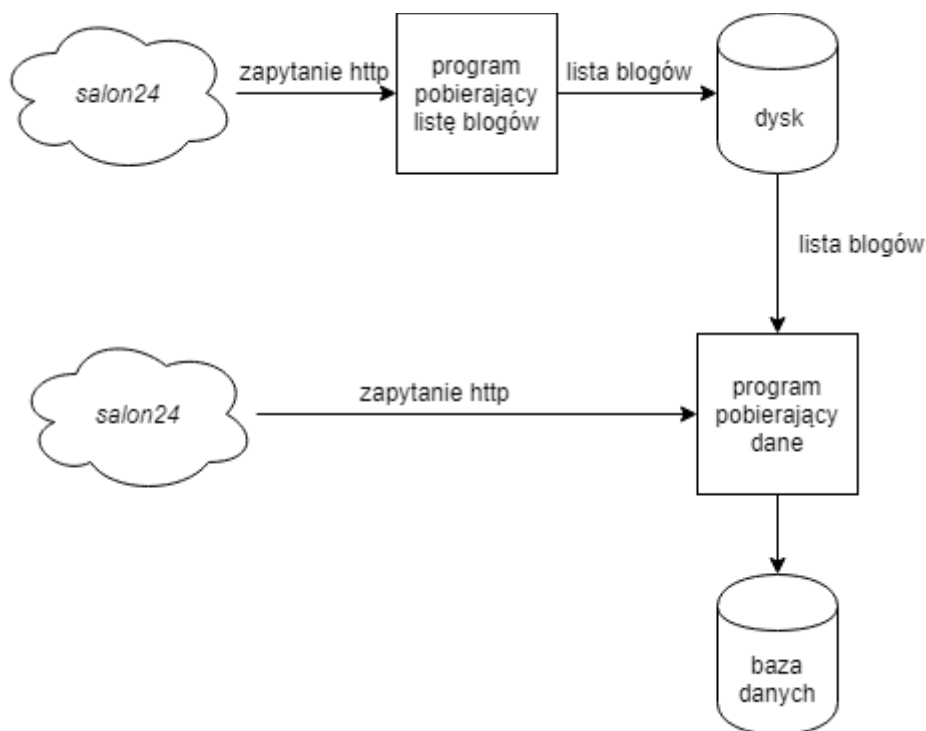
Zastosowane biblioteki:

- requests – wysyłanie zapytań http i pobieranie stron oraz plików JSON
- MySQLdb – obsługa komunikacji z bazą danych MySQL
- BeautifulSoup – parsowanie html strony
- html2text – przerabianie tekstu notatek i komentarzy z html na niesformatowany tekst
- datetime – obsługa konwersji daty i godziny
- pickle – zapis i odczyt danych na dysk do dalszego przetwarzania przez skrypty
- multiprocessing.pool – obsługa wielozadaniowości
- json – parsowanie plików JSON z danymi komentarzy

#### **3.2. Opis programu**

Rysunek 7 przedstawia diagram komponentów i przepływu danych w programach napisanych do pobierania i przetwarzania danych z serwisu *salon24*.





Rys 7. Diagram komponentów i przepływu danych

Podstawowym elementem programu do pobierania danych jest web scraper. W początkowej fazie działania uruchomiono skrypt odpowiedzialny za pobieranie listy blogów. Algorytm przechodził po stronach z listą blogów (posortowaną alfabetycznie) i pobierał strony html. Następnie przy pomocy parsera html wyłuskano linki do poszczególnych blogów. Po skończeniu działania lista blogów została zapisana na dysk.

Drugi program wczytywał listę blogów i uruchamiał wątki (w liczbie 10-20), które odpowiedzialne były za pobieranie blogów (każdy wątek pobierał w jednym czasie jeden blog). Wątek uruchamiał kod do pobierania html strony bloga, a następnie za pomocą parsera html wyciągał potrzebne dane. Po pobraniu informacji o blogu, wątek po kolei pobierał adresy do notatek i przechodził do pobierania danych w podobny sposób. Ostatnim etapem dla każdej notatki było pobranie listy komentarzy poprzez pobranie pliku JSON, do którego link był dostępny w kodzie html notatki. Sposób pobierania i parsowania danych komentarzy opisano poniżej.

### 3.3. Zapis do bazy danych

Pobrane i sparsowane dane zapisywano do bazy danych MySQL. Wykorzystywano ułatwiony dostęp do bazy, który umożliwiała biblioteka MySQLdb. Przy pobieraniu danych sprawdzano czy nie ma powtórzeń. W takim przypadku odrzucano dane. Datę i czas przekonwertowano do odpowiedniego formatu. Problemem były brakujące dane w niektórych blogach, notatkach i komentarzach, a także usunięte instancje. Potrzebne było przygotowanie odpowiednich zabezpieczeń.

Sprawdzanie listy użytkowników wykonywane było wiele razy podczas pracy programu, dlatego, aby nie obciążać dysku, pobrano listę loginów użytkowników będących już w bazie i aktualizowano ją w miarę pobierania danych.

### 3.4. Pobieranie komentarzy

Sposób działania strony *salon24* uniemożliwiał bezpośrednie pobranie komentarzy przy pomocy web scrapera i biblioteki requests, ponieważ zastosowano tam technologię z dynamicznym ładowaniem danych.

W kodzie html notatek zawarty jest adres do pliku JSON zawierającego strukturę i dane komentarzy (listing 1.). Wykorzystano to do pobierania odpowiednich danych.

```
{
  "error": null,
  "error_desc": null,
  "data": {
    "sources": [],
    "comments": {
      "sourceId": "Post-944711",
      "sort": "NEWEST",
      "last": "",
      "nextUrl": "",
      "selected": 0,
      "data": [
        {
          "id": "16097205",
          "userId": "7923",
          "created": "1553680597308",
          "content": "KO\u0144sTYtucJA\nczy jako\u015b tak\n:D",
          "format": "text",
          "replies": 0,
          "likes": 39,
          "dislikes": 2,
          "votes": 41,
          "hidden": false,
          "deleted": false
        },
        {
          "id": "16097228",
          "userId": "82359",
          "created": "1553680815918",
          "content": "Ten to dopiero ma leb nie od parady...",
          "format": "text",
          "replies": 3,
          "likes": 38,
          "dislikes": 2,
          "votes": 40,
          "hidden": false,
          "deleted": false,
          "comments": {
            "sourceId": "Post-944711",
            "sort": "NEWEST",
            "last": "",
            "nextUrl": "",
            "selected": 0,
            "data": [
              {
                "id": "16097244",
                "userId": "7923",
                "parentId": "16097228",
                "created": "1553680958497",
                "content": "@DotCa\u00a0\nba\nsam wymy\u015bli\u0142 jak obali\u0107 komunizm w Europie\ntaki
```

```
skromy gieniusz",
  "format": "text",
  "replies": 0,
  "likes": 34,
  "dislikes": 2,
  "votes": 36,
  "hidden": false,
  "deleted": false
},
(...)
```

Listing 1. Fragment przykładowego pliku JSON z danymi komentarzy

Plik JSON parsowano przy użyciu odpowiedniej biblioteki. Wyłuskano między innymi treść komentarza, negatywne i pozytywne oceny oraz użytkownika piszącego.

### 3.5. Zabezpieczenie przed utratą danych

Jako zabezpieczenie przed utratą danych, zapisywano do pomocniczej tabeli w bazie danych informacje o ostatnio pobranym url bloga oraz, która notatka była ostatnio pobrana. Zapobiegało to powtórnemu pobieraniu tych samych danych.

### 3.6. Wielowątkowość

Zastosowano wielowątkowość, aby przyspieszyć pobieranie danych poprzez wysyłanie większej ilości zapytań http i równoczesne przetwarzanie pobranych danych. Wykorzystano fakt, że baza danych MySQL synchronizuje wykonywane transakcje.

```

Scraping blog https://www.salon24.pl/u/prostamysl/ (page: 30)
Scraping note 1/2
Scraping note 2/2
Scraping blog https://www.salon24.pl/u/okruchyduszy/ (page: 30)
Scraping note 1/3
Scraping note 2/3
Scraping note 3/3
Scraping blog https://www.salon24.pl/u/adavila/ (page: 30)
Scraping blog https://www.salon24.pl/u/adawu/ (page: 30)
Scraping note 1/5
Scraping note 2/5
Scraping note 3/5
Scraping note 4/5
Scraping note 5/5
Scraping blog https://www.salon24.pl/u/adam/ (page: 30)
Scraping note 1/1
Scraping blog https://www.salon24.pl/u/addendum/ (page: 30)
Scraping blog https://www.salon24.pl/u/addexteram/ (page: 30)
Scraping blog https://www.salon24.pl/u/adek/ (page: 30)
Scraping blog https://www.salon24.pl/u/adele/ (page: 30)
Scraping blog https://www.salon24.pl/u/adelepustakov/ (page: 30)
Scraping blog https://www.salon24.pl/u/teczka/ (page: 30)
Scraping note 1/3
Scraping note 2/3
Scraping note 3/3
Scraping blog https://www.salon24.pl/u/kurwaadelix/ (page: 30)
Scraping blog https://www.salon24.pl/u/adam67/ (page: 30)
Scraping note 1/9
Scraping note 2/9
Scraping note 3/9
Scraping note 4/9
Scraping note 5/9
Scraping note 6/9
Scraping note 7/9
Scraping note 8/9
Scraping note 9/9
Scraping blog https://www.salon24.pl/u/adibrand/ (page: 30)
Scraping note 1/33
Scraping note 2/33
Scraping note 3/33
Scraping note 4/33
Scraping note 5/33
Scraping note 6/33
Scraping note 7/33
Scraping note 8/33
Scraping note 9/33
Scraping note 10/33
Scraping note 11/33
Scraping note 12/33
Scraping note 13/33

```

Rys. 8. Przykład działania programu

## 4. Eksperymenty

Do przetwarzania pobranych danych do eksperymentów napisano skrypt odpowiedzialnych za pobieranie z bazy danych listy par właściciel bloga – użytkownik komentujący.

Napisano także kilka skryptów w języku SQL do tworzenia statystyk.

### 4. 1. Podstawowe statystyki

nazwa bloga	liczba śledzących
kataryna	616
Układ Warszawski - rekonstrukcja	614
Bez dekretu	600

Carthago delenda est	522
Oszolom, jaskiniowy antykomuch	486
Jest super, więc o co mi chodzi?	455
gabriel maciejewski baśń jak n	403
Witold Gadowski	370
Moim interesem jest interes Polaków i Polski	365
Jarosław Kaczyński - Prawo i Sprawiedliwość	331

Tab. 1. TOP10 śledzonych blogów

<b>nazwa bloga</b>	<b>wyświetlenia</b>
Newsroom Salon24	14424000
gabriel maciejewski baśń jak n	12175000
Układ Warszawski - rekonstrukcja	9289000
Tylko po co?	6013000
Sowiniec	5040000
W hołdzie mistrzom ciętej riposty	4991000
Ludzie myślcie, to nie boli	4895000
Krzysztof Osiejuk	4718000
PASAŻER RP	4632000
Moim interesem jest interes Polaków i Polski	4280000

Tab 2. TOP10 wyświetlanych blogów

<b>nazwa bloga</b>	<b>liczba notatek</b>
--------------------	-----------------------

Andrzej Budzyk - blog	6284
Newsroom Salon24	5941
dobrezycie	5003
w koło Macieju	4643
Sowiniec	4046
Dawniej też Stary	3878
Silna Polska w Europie Narodów	3300
biznesradar.pl	3277
Zbigniew Kuźmiuk	3153
Tylko po co?	3086

Tab. 3. 10 blogów z największą liczbą notatek

<b>nazwa kategorii</b>	<b>liczba notatek w kategorii</b>
Polityka	386321
Rozmaitości	89199
Kultura	71814
Społeczeństwo	26745
Gospodarka	24790
Historia	18977
Technologie	12712
Nauka	8977
Religia	5210
Podróże	5124

Tab 4. 10 kategorii z największą ilością notatek

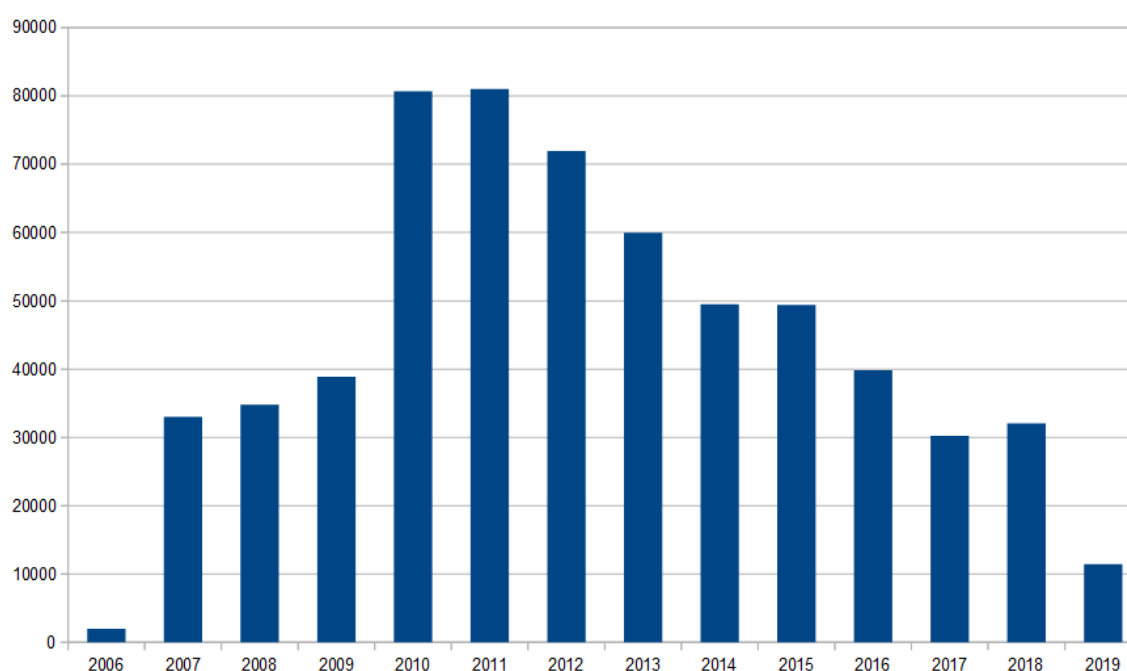
tytuł notatki	data	liczba wyświetleń
Młodzi gniewni, oszukani, wykluczeni, czyli mieszanka wybuchowa	13-09-02 00:13	3200424
Cichy „przewrót” Prawa i Sprawiedliwości	16-05-21 13:30	1755311
Natychmiast aresztować Tuska	12-08-15 18:30	1187974
Wszyscy won!	18-08-09 11:06	651282
Obywatelska, czyli największe kłamstwo PO	15-03-28 16:45	390635
Jak zmienić swój numer IP? Poradnik dla zielonych.	10-09-08 21:56	174055
Sensacyjne odkrycie w Całunie Turyńskim	18-04-24 08:23	172794
Nowoczesny patriotyzm gospodarczy	11-02-18 18:45	163316
Dlaczego NATO, USA oraz Rosja boją się polskiego czołgu ?	17-06-26 08:51	142238
Jedno zdanie Kaczyńskiego	18-07-27 12:21	137446

Tab 5. 10 najczęściej wyświetlanych notatek

tytuł notatki	data	udostęp- nienia
Słynny „łowca nazistów” powiedział to wprost: Żydzi to ludobójcy ! mordowali Polaków	18-05-13 18:45	35957
Sensacyjne odkrycie w Całunie Turyńskim	18-04-24 08:23	21475
Wanda - sanitariuszka Powstania Warszawskiego, uszanował ją wróg okaleczyli kaci	16-02-08 14:18	13056
Prawdziwą siłą demokratycznej Polski są i pozostaną Narodowcy	18-04-11 12:02	12365
Fabrykę papieru w Kostrzynie sprywatyzowali za 80 zł	18-06-22 18:18	11468
Dlaczego NATO, USA oraz Rosja boją się polskiego czołgu ?	17-06-26 08:51	10331

Żydzi w UB - to kaci i mordercy Polaków. Przeczytajcie, fakty są porażające	18-04-27 08:33	9836
Coraz więcej porzuconych zwierząt. Pies ma świadomość 3-letniego dziecka i cierpi	18-07-10 08:04	9545
800 zł na pracowników z Ukrainy, 500 + i stypendia	18-03-22 16:33	9230
AstroTurfing czyli kto organizuje Majdan w Warszawie	17-07-22 14:57	7992

Tab 6. 10 najczęściej udostępnianych notatek na facebooku



Rys. 9. Liczba notatek w danym roku

#### 4. 2. Graf 500 blogów najczęściej komentujących

Do zwizualizowania połączeń najczęściej komentujących użytkowników wykorzystano program Gephi.

Wielkość wierzchołków:

Ilość wszystkich komentarzy

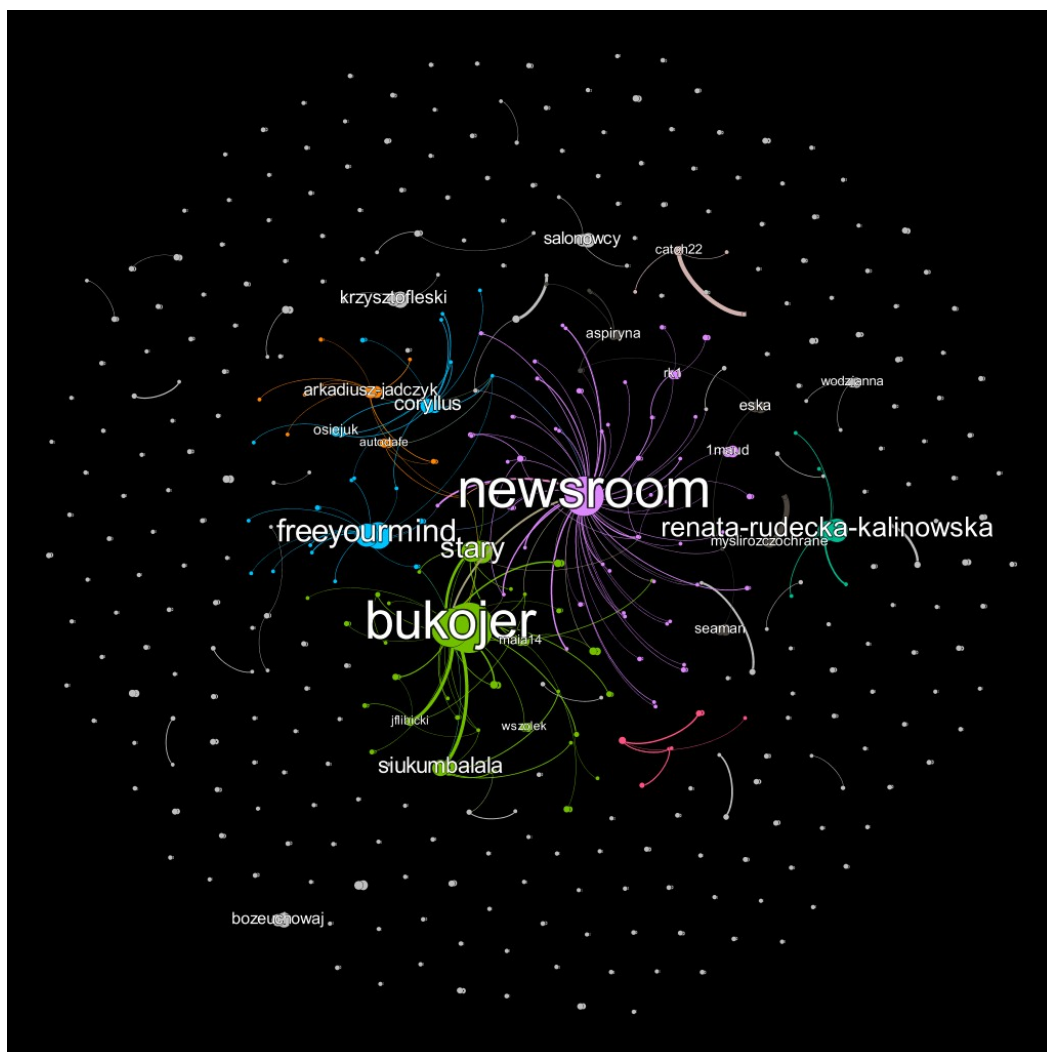
Grubość krawędzi:

Ilość komentarzy między blogami

Kolor:

Modularity class





Rys 10. Graf 500 blogów najczęściej komentujących