

***Dmytro Polianychko***

***110111***

**Informatyka, studia stacjonarne I stopnia, semestr 5**

Projekt z przedmiotu Wybrane zagadnienia współczesnej informatyki pt.:

**Aplikacja do akwizycji i analizy danych z artykułów z konferencji Petri Nets and Software Engineering 2016**

**Rzeszów, 2021 r.**

**Rozdziały:**

1. **Zakres projektu.**

Zakresem danego projektu była spróba pobrania danych poprzez wybrane technologie ze strony html, umieścienie ich do bazy danych, przetwarzanie z celem podalszego generowania chmury słow i najczęściej wykorzystanych słow, oraz generowania klastrów.

1. **Charakterystyka analizowanych danych.**

Dane do akwizycji były udostępnione na strobie ceur.org, gdzie celem było przejście do spisu artykułow, i pobranie samych treści z plików pdf poprzez automatyzacji ustawiania ścieżki do każdego z nich.

**3.Wykorzystane narzędzia.**

Wykorzystanymi narzędziami są:

-Apache Solr wersji 8.7.0;

-Rstudio oraz R wersji 4.0.3;

- biblioteka w której jest zaimplementowany dany dyzajn: shinydashboard;

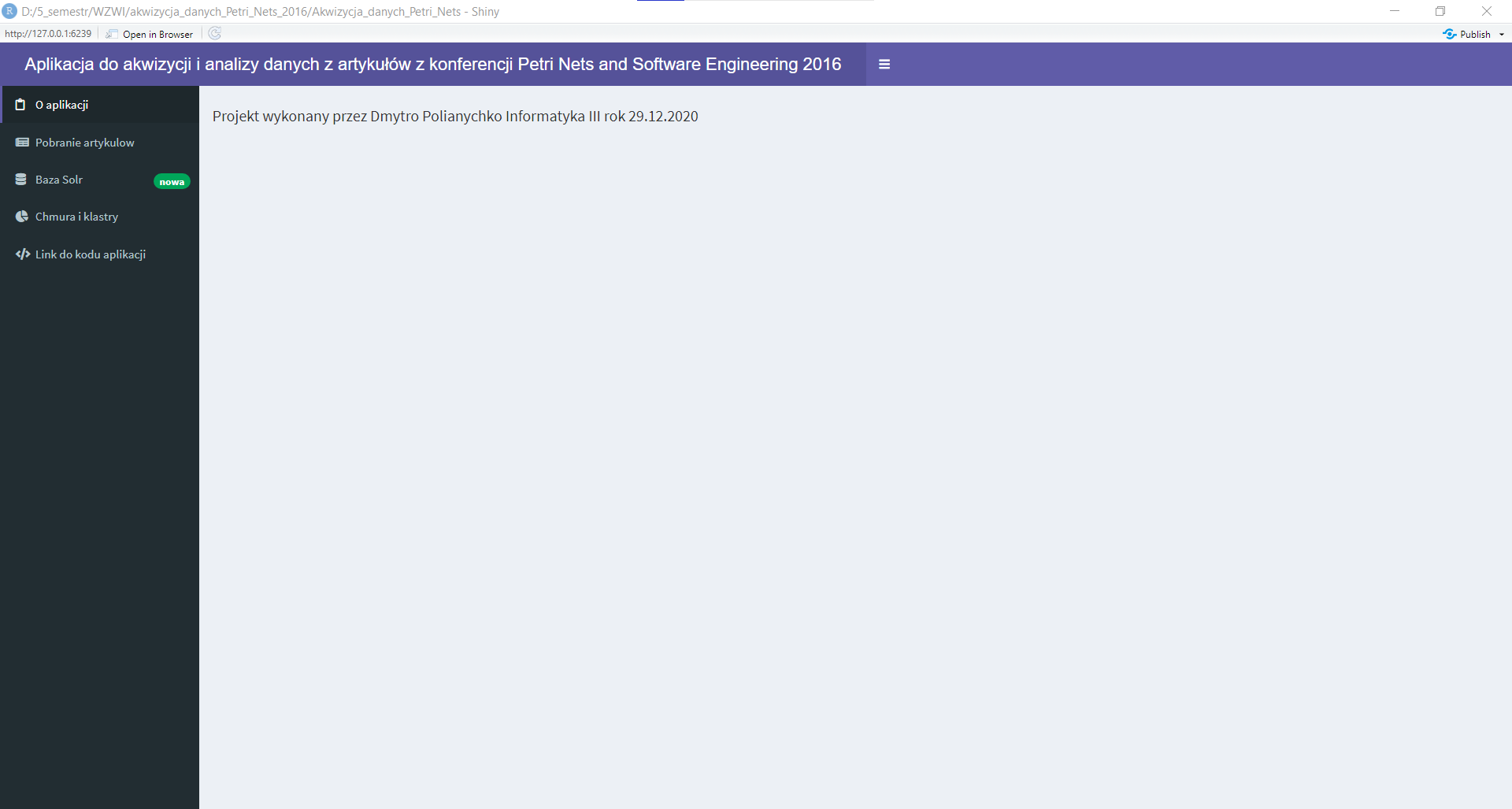
- oraz inne pomocnicze biblioteki.

**4. Architektura i implementacja aplikacji.**

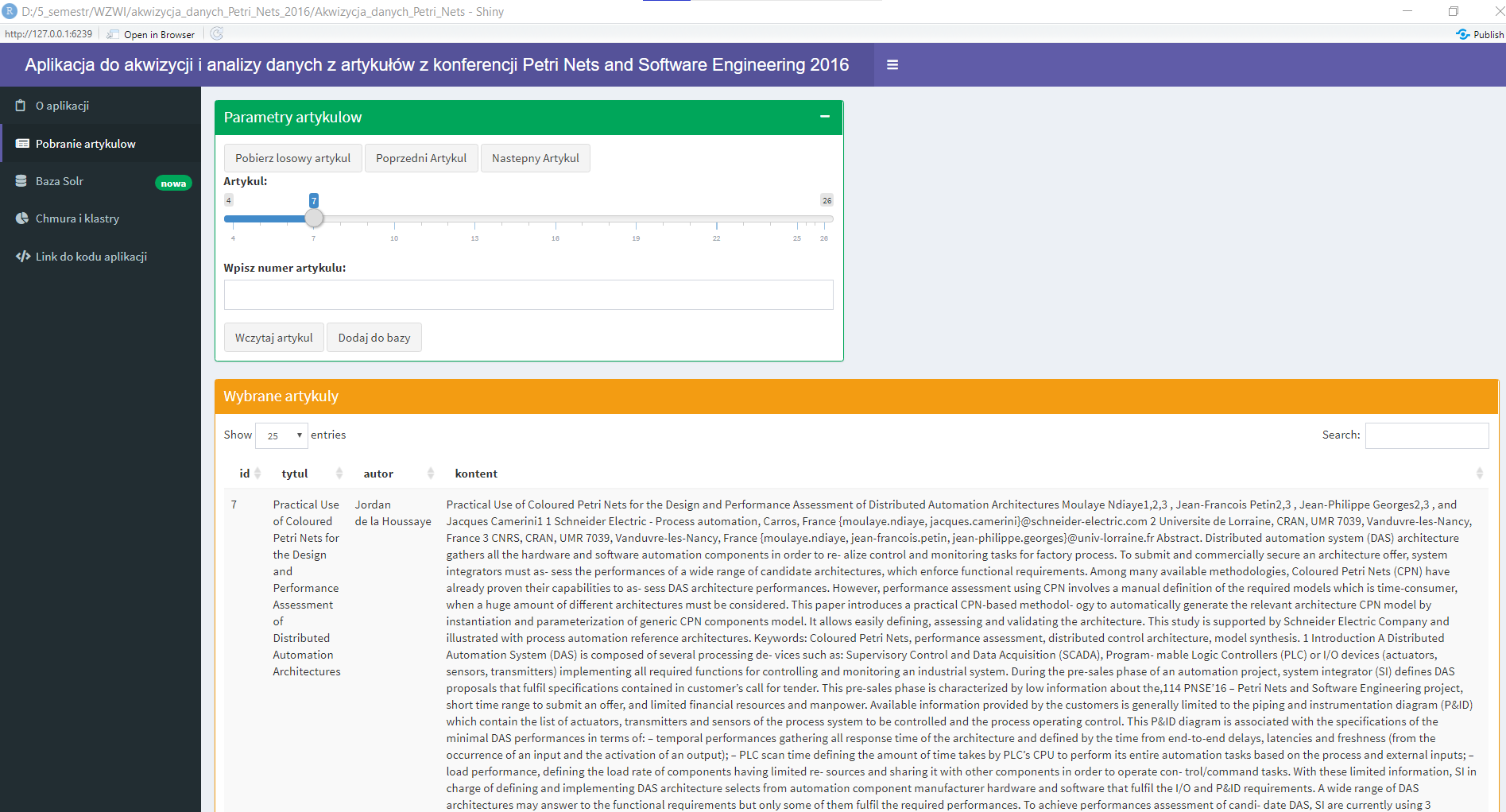
W implementacji danego projektu dane były pobierane ze strony html przy pomocy funkcji read\_html, jako dane występowaly tytuł atrykułu, autor oraz sam artykuł który był pobierany przez funkcje pdf\_text. Dalej w tych artykułach były usuwane wszystkie nadmiarowe spacje (powierzchne przetwarzanie tekstu) i zapisywane do pliku pomocniczego za nazwą dane.txt, ponieważ w każdym pobranym artykurze jest różna liczba części, i dla tego było ciężko wyznaczyć ilość rządów w tablice, do której one były zapisywane. Obsługa bazy Apache Solr była wykonana poprzez zapytania. Generowanie chmury słów wykonane przy pomocy biblioteki wordcloud2, co pozwala na ustawianie różnych dodatkowych dosyć ciekawych parametrów w postaci ksztaltu samej chmury, kolorze zapisanych w niej słów oraz innych. Dla dyzajnu została wybrana biblioteka shinydashboard, która rozszerza możliwości samej shiny (bardziej wygodne i ciekawsze czcionki, kolory, ukrywane i rozwijane box, możliwość stylizacji tablic, i berdzo wiele innych ciekawych rzeczy).

**5. Interfejs użytkownika.**

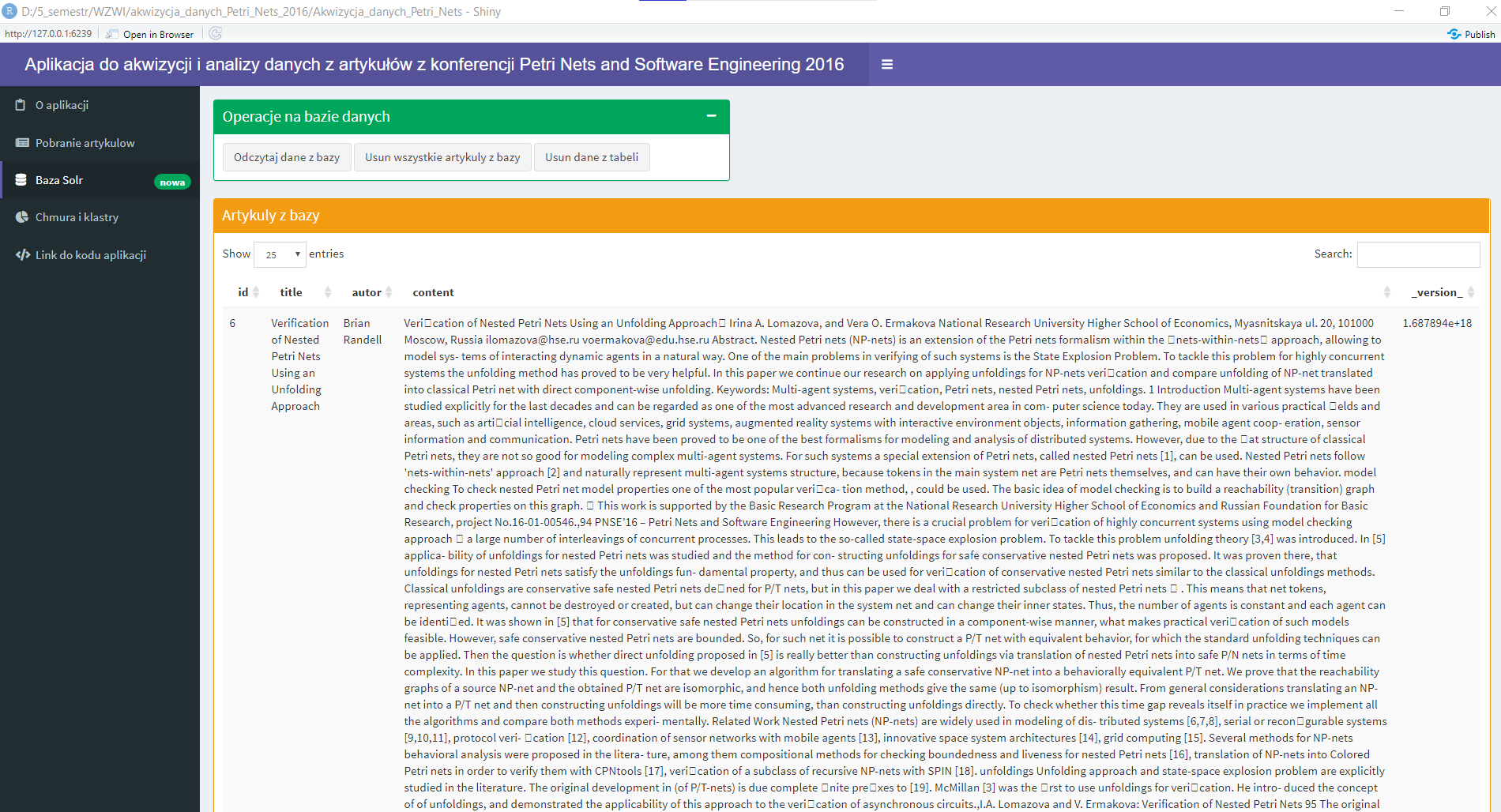
Odrazu przy otwarciu aplikacji widzimy przed sobą zakładke „O aplikacji”, gdzie jest informacja o autorze projektu



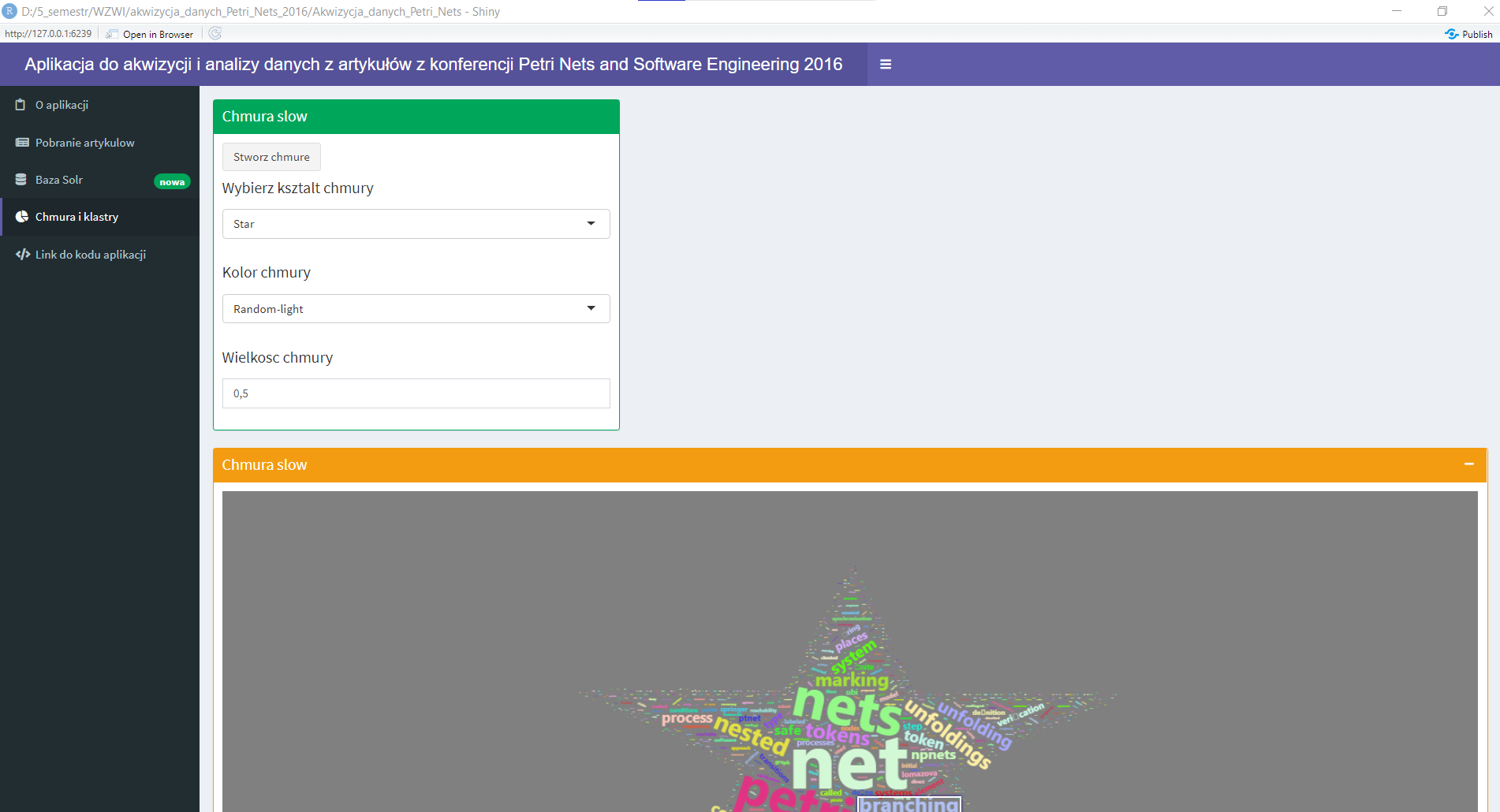
W następnej zakładce „Pobieranie artykulów” pobieramy wybrane przez nas artykuły ze strony internetowej i dodajemy ich do bazy Apache Solr, i mamy takie funkcjonalności: losowe pobieranie artukułu, możliwość zobaczenia poprzedniego i następnego artykułu (dla łatwości preczytania treści), dla tego żeby dodać artykuł do bazy Solr, mamy napisać id artykułu w pole textInput i potem nacisnąć na przycisk „dodaj do bazy”.

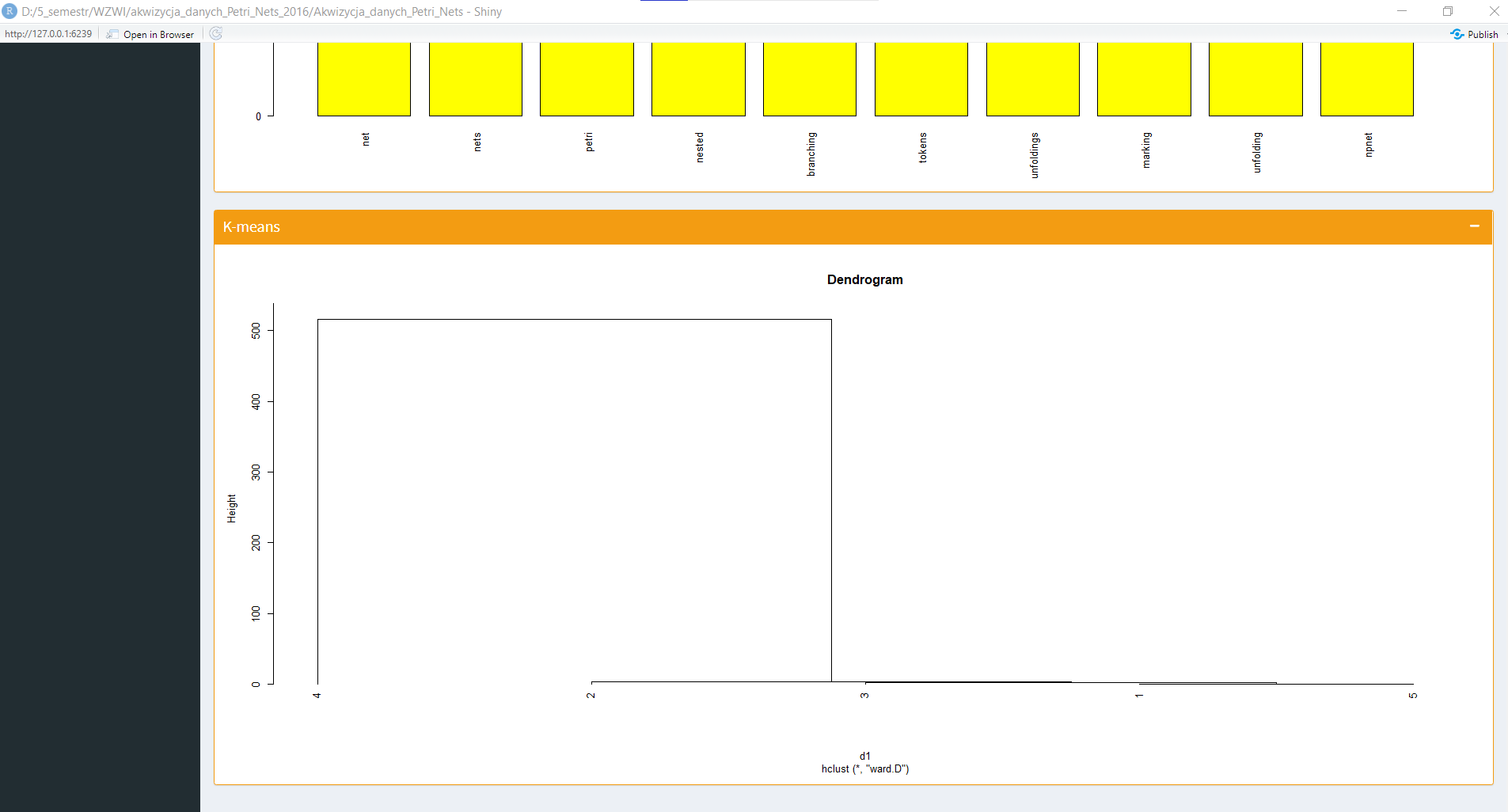
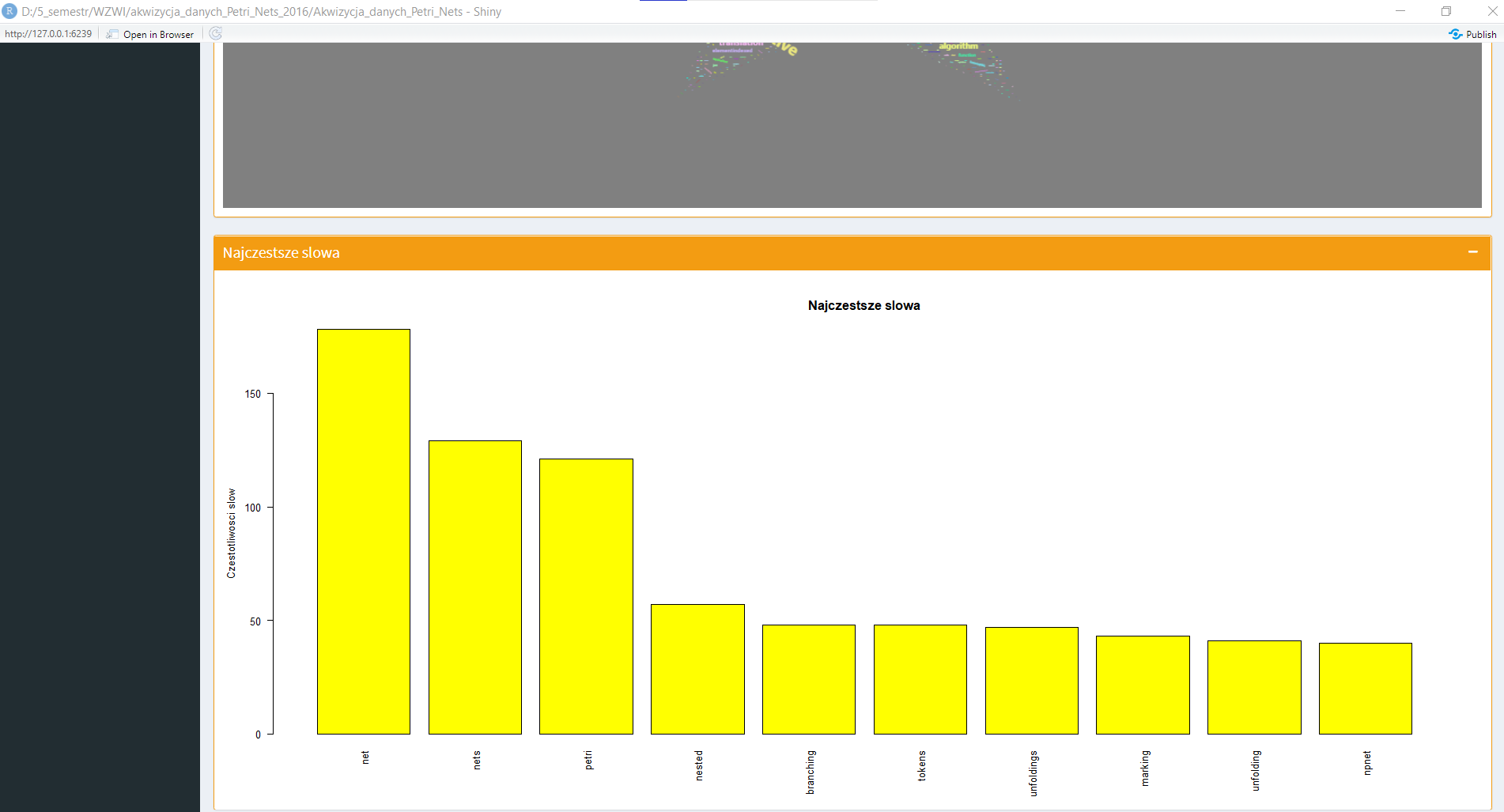


Dalej przechodzimy do zakładki „Baza Solr”, gdzie po naciśnięcu przycisku „Odczytaj dane z bazy” możemy wyświetlić wszystkie artykuły, które są w naszej bazie danych, także są takie funkcjonalności jak usuwanie wszystkich artykulów z bazy i usuwanie wszystkich artykulów z tabeli która jest wyświetlona poniżej.



W następnej zakładce „Chmura i klastry” możemy wygenerować chmure slów z tych artykulów które wsześniej dodaliśmy do Apache Solr, jako parametry do chmury możemy ustawiać ksztalt rysowanej chmury, także jej wielkość i kolor slów, dla chmury jest także wygenerowany wykres o najczęściej używanych słowach i dendrogram.





W ostatniej zakadce “Link do kodu aplikacji” przechodzimy do github, gdzie jest umieszczony kod danej aplikacji.