

***Magdalena Czechura***

***106425***

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

Projekt z przedmiotu Wybrane zagadnienia współczesnej informatyki pt.:

***Aplikacja webowa do akwizycji i analizy danych z kanału RSS portalu*** [***gofin.pl***](http://gofin.pl/) ***(składki, zasiłki, emerytury)***

**Rzeszów, 2021 r.**

*Link do repozytorium: https://github.com/magda9906/WZWI2020*

**1. Zakres projektu.**

Aplikacja webowa z zaawansowanym GUI (możliwość ustawiania parametrów) zrealizowanym z wykorzystaniem pakietu Shiny, możliwością pobierania wiadomości i magazynowania ich w Apache Solr, analizy częstości występowania słów i wizualizacji chmury słów w wiadomościach zmagazynowanych w Apache Solr, klasteryzacji wiadomości zmagazynowanych w Apache Solr oraz wizualizacji klastrów.

**2. Charakterystyka analizowanych danych.**

Dane użyte w projekcie to treść aktualności ze strony <https://www.gofin.pl/skladki-zasilki-emerytury/>. Dane pobierane są za pomocą kanału RSS w formacie XML. Z pobranego dokumentu wyodrębniana jest kolumna o nazwie ‘entry\_content’ zawierająca treść aktualności. Dane przesyłane są do SOLR, a następnie pobierane. Usuwane są znaki interpunkcyjne oraz tzw. “stop words”. Następnie dane poddawane są lematyzacji. Do klasteryzacji dane są przekształcane za pomocą funkcji DocumentTermMatrix().

**3. Wykorzystane narzędzia.**

Projekt został stworzony w RStudio z wykorzystaniem języka R i pakietu Shiny. Wiadomości magazynowane są za pomocą Apache Solr. Do wygenerowania chmury słów posłużył pakiet ‘wordcloud2’, natomiast do klasteryzacji pakiety ‘dbscan’ i ‘clusters’.

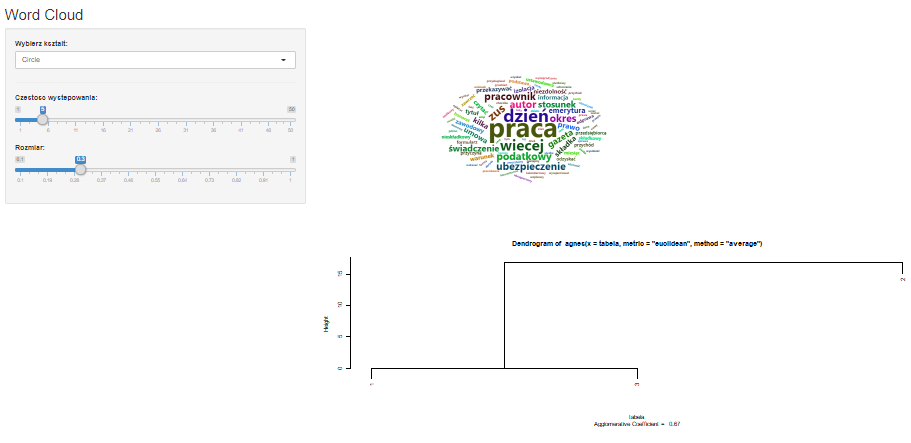
**4. Architektura i implementacja aplikacji.**

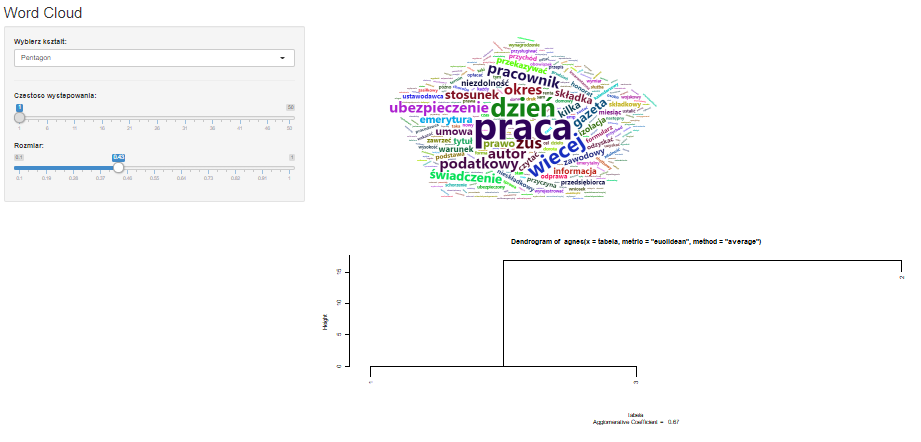
Projekt pobiera dane z kanału RSS portalu gofin w formacie XML za pomocą funkcji tidyfeed() i wyodrębnia kolumnę z treścią aktualności, a następnie zapisuje ją do odpowiednie z pobranych z SOLR danych usuwane są znaki interpunkcyjne oraz stop words. Tak przygotowane dane są lematyzowane odpowiednią funkcją z wykorzystaniem http://ws.clarin-pl.eu/nlprest2/base/process. Za pomocą funkcji TermDocumentMatrix() tworzona jest macierz ze słowami, z których generowana będzie chmura.

Za pomocą funkcji DocumentTermMatrix() tworzona jest tabela, której dane poddawane są klasteryzacji.

**5. Interfejs użytkownika.**

Użytkownik może ustawiać takie wartości jak kształt, rozmiar chmury oraz minimalną częstotliwość występowania słowa.





**6. Podsumowanie.**

RStudio umożliwia wykonanie aplikacji webowej Shiny generującej chmurę słów za pomocą pakietu ‘wordcloud2’. Aplikacja umożliwia użytkownikowi ustawianie parametrów takich jak kształt lub wielkość chmury, a także ograniczenie występowania słów w chmurze. Zebrane dane poddawane są klasteryzacji, czego efekt można zobaczyć na wykresie.