Dokumentacja Specyfikacji Wymagań

System Analizy Skryptów Matematycznych z Wykorzystaniem Metod Text Mining

Data: 2025-06-07
Autorzy: Urszula Urban, Wojciech Nowacki, Jan Pawlikowski

---

1. Wprowadzenie

1.1 Cel dokumentu

W matematyce bardzo ważną rolę odgrywają definicje oraz precyzyjna terminologia. Dowody powinny być ścisłe, a więc nie ma w nich miejsca na wieloznaczne słowa występujące np. w poezji. Unika się również synonimów. Przez wieki rozwoju nauki większość terminów doczekała się jednej precyzyjnej nazwy (oczywiście są wyjątki: niektóre twierdzenia funkcjonują pod kilkoma nazwami, niektóre szkoły matematyki spierają się co do tego co powinno znaleźć się w definicji etc.[1]). Co więcej, skrypty matematyczne pisane są

[1] *Nierówność Cauchy'ego-Schwarza, Schwarza, Buniakowskiego-Schwarza lub Cauchy'ego-Buniakowskiego-Schwarza to nazwy tej samej nierówności.*
W definicji przestrzenii zwartej niektórzy autorzy wymagają, by była ona dodatkowo przestrzenią Hausdorffa, inni pomijają ten warunek. Takich przykładów jest sporo, ale nie deprecjonują one ogólnej zasady: matematyka jest ścisła, również jeśli chodzi o język, którego używa.

zazwyczaj z pełnym rygorem, nawet gdy ich celem jest kształcenie sprawiają wrażenie jakby pisane były jako artykuły naukowe. Wszystko to spotyka się oczywiście z krytyką wielu dydaktyków, którzy uważają, że traci na tym intuicja studentów. Sprzyja to jednak analizie słownikowej.

Teksty bez ironii, niuansów, skomplikowanych metafor etc., zawierające precyzyjne słownictwo nadają się idealnie do zbadania przez pryzmat analizy słownikowej. To właśnie jest celem niniejszego systemu. Zastosowań jest naprawdę wiele, wybrane opisane są w sekcji Scenariusze Użytkowania[2].

Niniejszy dokument stanowi specyfikację wymagań systemu informatycznego przeznaczonego do automatycznej analizy dokumentów tekstowych w formacie PDF z wykorzystaniem metod text mining i data science. System ma na celu wspieranie procesów decyzyjnych oraz procesów badawczych poprzez analizę informacji zawartych w dokumentach.

1.2 Zakres systemu

System Analizy Dokumentów PDF jest narzędziem analitycznym z bazą w środowisku R, umożliwiającym kompleksową analizę treści dokumentów PDF. System wykonuje analizę częstości słów, ocenę sentymentu, badanie podobieństwa dokumentów oraz generuje wizualizację wyników w postaci wykresów i chmur słów.

1.3 Odbiorcy dokumentu

- Wykładowcy matematyki analizujący materiały dydaktyczne.
- Studenci matematyki porównujący skrypty i notatki.
- Autorzy podręczników matematycznych.
- Bibliotekoznawcy klasyfikujący zbiory matematyczne.

[2] Prosimy zajrzeć!

2. Cele systemu

2.1 Cel główny

Stworzenie zautomatyzowanego systemu do kompleksowej analizy danych tekstowych, który umożliwi efektywną eksplorację zawartości dokumentów oraz identyfikację kluczowych wzorców i trendów w analizowanych tekstach.

2.2 Cele szczegółowe

- Porównywalność dokumentów: Umożliwienie analizy podobieństw i różnic między dokumentami.
- Eksploracja częstości słów: Umożliwia zbadanie najczęściej występujących terminów i definicji w danym skrypcie, czy ogólniej- w danym dziale matematyki.
- Przetwarzanie wieloformatowe: Obsługa dokumentów PDF, TXT i CSV z zachowaniem kodowania UTF-8 dla języka polskiego.
- Analiza lingwistyczna: Implementacja zaawansowanych metod przetwarzania języka, dostosowanych do specyfiki języka polskiego.
- Wizualizacja wyników: Generowanie intuicyjnych przedstawień graficznych wyników analizy.
- Reprodukowalność: Zapewnienie pełnej powtarzalności analiz zgodnie ze standardami Reproducible Research.

2.3 Korzyści biznesowe

- Automatyzacja procesu analizy dokumentów tekstowych.
- Oszczędność czasu przy przetwarzaniu dużych zbiorów tekstowych.
- Obiektywna ocena podobieństw między dokumentami.
- Wsparcie dla procesu decyzyjnego opartego na danych tekstowych.

3. Wymagania funkcjonalne

3.1 Wczytywanie dokumentów

Priorytet: Wysoki

Opis: System musi umożliwiać wczytywanie dokumentów w formatach PDF, TXT i CSV z zachowaniem poprawnego kodowania znaków polskich.

Kryteria akceptacji:

- Obsługa kodowania UTF-8
- Obsługa wielu plików jednocześnie
- Walidacja poprawności wczytanych danych

3.2 Przetwarzanie tekstu polskiego

Priorytet: Wysoki

Opis: System musi realizować kompleksowe przetwarzanie tekstu w języku polskim, normalizację, usuwanie stop words i podstawowy stemming.

Kryteria akceptacji:

- Konwersja do małych liter.
- Usuwanie znaków interpunkcyjnych i cyfr.
- Eliminacja polskich stop words.
- Implementacja prostego stemmingu dla języka polskiego.
- Usuwanie pojedynczych liter i nadmiarowych spacji.

3.3 Analiza częstości słów

Priorytet: Wysoki

Opis: System generuje statystyki częstości występowania słów w dokumentach.

Kryteria akceptacji:

- Tworzenie macierzy dokument-term.

- Obliczanie częstości słów.
- Filtrowanie słów według minimum występowań.
- Eksport wyników do formatu CSV.

3.4 Wizualizacja danych

Priorytet: Średni

Opis: System musi generować wizualizację wyników analizy w postaci wykresów i chmur słów.

Kryteria akceptacji:

- Generowanie chmury słów z konfigurowalnymi parametrami.
- Tworzenie wykresów słupkowych najczęstszych słów.
- Wykorzystanie pakietu ggplot2 do wizualizacji.
- Możliwość dostosowania kolorystyki i stylu.

3.5 Analiza podobieństwa dokumentów

Priorytet: Średni

Opis: System musi umożliwiać porównanie podobieństwa między dokumentami oraz identyfikację słów wspólnych i unikalnych.

Kryteria akceptacji:

- Identyfikacja wspólnych słów między dokumentami.
- Wyodrębnianie słów unikalnych dla każdego dokumentu.
- Eksport wyników porównania do plików tekstowych.

3.6 Eksport wyników

Priorytet: Średni

Opis: System musi umożliwiać eksport wszystkich wyników analizy do plików zewnętrznych.

Kryteria akceptacji:

- Zapis częstości słów do CSV.
- Eksport macierzy dokument-term do CSV.
- Generowanie plików tekstowych z wynikami porównań.
- Zachowanie kodowania UTF-8 w eksportowanych plikach.

4. Wymagania niefunkcjonalne

4.1 Wydajność

Opis: System musi efektywnie przetwarzać dokumenty tekstowe o różnej wielkości.

Kryteria:

- Czas przetwarzania dokumentu PDF do 10MB nie powinien przekraczać 60 sekund.
- System musi obsługiwać jednoczesną analizę do 10 dokumentów.
- Wykorzystanie pamięci RAM nie powinno przekraczać 2GB dla typowych analiz.

4.2 Niezawodność

Opis: System musi zapewniać stabilne działanie i obsługę błędów.

Kryteria:

- Walidacja danych wejściowych.
- Mechanizmy recovery w przypadku błędów przetwarzania.
- Logowanie operacji i błędów systemowych.

4.3 Używalność

Opis: System musi być łatwy w użytkowaniu dla analityków danych.

Kryteria:

- Przejrzysty i skomentowany kod źródłowy
- Automatyczna instalacja wymaganych pakietów R

- Intuicyjne nazewnictwo zmiennych i funkcji
- Dokumentacja inline w kodzie

4.4 Przenośność

Opis: System musi działać na różnych systemach operacyjnych.
Kryteria:

- Kompatybilność z Windows, Linux i macOS.
- Wykorzystanie standardowych pakietów R.
- Niezależność od specyficznych ścieżek systemowych.
- Obsługa różnych wersji R.

4.5 Skalowalność

Opis: System musi umożliwiać rozszerzenie funkcjonalności.
Kryteria:

- Modularna struktura kodu.
- Możliwość dodawania nowych metod analizy.
- Parametryzowalność algorytmów.

5. Interfejsy użytkownika i wymagania dotyczące danych

5.1 Interfejs użytkownika

System działa w środowisku R/RStudio jako skrypt konsolowy. Interakcja z użytkownikiem odbywa się poprzez:

- Konfigurację wejściową: Modyfikacja parametrów w kodzie źródłowym.
- Wybieranie dokumentów do przeanalizowania: Umieszczenie odpowiednich plików w tym samym folderze co skrypt.

- Wykonanie analizy: Uruchomienie skryptu R.
- Przegląd wyników: Analiza wygenerowanych wykresów i raportów HTML.

5.2 Wymagania dotyczące danych wejściowych

Format plików:

- PDF: Dokumenty tekstowe w formacie PDF.
- TXT: Pliki tekstowe w kodowaniu UTF-8.
- CSV: Pliki CSV z danymi tekstowymi.

Wymagania jakościowe:

- Pliki muszą zawierać tekst w języku polskim.
- Minimalna długość dokumentu: 100 słów.
- Maksymalny rozmiar pojedynczego pliku: 50MB.

5.3 Dane wyjściowe

- Pliki CSV: Tabele z częstością słów i macierzą dokument-term.
- Pliki TXT: Listy słów wspólnych i unikalnych.
- Wykresy PNG: Wizualizacje chmur słów i wykresów częstości.
- Raport HTML: Kompletny raport z wynikami analizy.

6. Słownictwo dokumentacji

Termin	Znaczenie
Reproducible Research	Metodologia pozwalająca na powtarzalność badań i analiz.
Tokenizacja	Podział tekstu na mniejsze jednostki – tokeny (np. słowa).
Chmura słów	Graficzna reprezentacja częstości występowania słów.

Sentiment analysis	Analiza nastrojów (pozytywne, negatywne, neutralne emocje w tekście).

7. Przypadki użycia (Use Cases)

- UC1: Użytkownik umieszcza wybrane przez siebie pliki w tym samym folderze co skrypt.
- UC2: Użytkownik uruchamia analizę tekstu z pliku.
- UC3: System wczytuje dane i przetwarza tekst.
- UC4: System generuje wizualizacje danych (np. chmurę słów).
- UC5: System eksportuje wyniki do HTML.

8. Scenariusze użytkownika (User Stories)

- Student drugiego roku bardzo chciałby zapisać się na kurs Ekonometrii w drugim semestrze. Zrobione przez niego kursy to: Rachunek Prawdopodobieństwa, Analiza Matematyczna, Geometria Liniowa, nie zrobił jednak żadnego Kursu ze Statystyki i obawia się, że jego wiedza jest niewystarczająca. Żeby się przekonać, czy tak jest faktycznie używa systemu do Analizy Skryptów Matematycznych i bada podobieństwa między skryptem z Ekonometrii, a skryptami z innych przedmiotów. Wychodzi mu wysoki współczynnik korelacji, sugerujący dużą ilość terminów ze skryptu ze statystyki występujących w skrypcie z ekonometrii, a nie pojawiających się w skrypcie z innych przedmiotów. To sugeruje mu, że nie jest jeszcze gotowy na wzięcie kursu z Ekonometrii. Student decyduje się przyłożyć do nauki statystyki.

- Na wydziale matematyki UW, niektórzy (w tym prowadzący) zastanawiają się czym właściwie jest przedmiot matematyka obliczeniowa. Czy jest to GAL III (na MIMie są dwa kursy z geometrii i algebry liniowej), czy może bardziej analiza. By rozwiać te

wątpliwości badacz używa systemu by zbadać podobieństwo skryptu profesora Krzyżanowskiego (tak naprawdę to nie jest skrypt tylko wręczałki, ale powinno wystarczyć) ze skryptami z Geometrii Liniowej (dr Męcel) i z Analizy Matematycznej (prof Strzelecki). Wychodzi mu, że matematyka obliczeniowa to faktycznie przede wszystkim algebra liniowa.

- Student wybierając z czego się uczyć, chce dowiedzieć się jak trudne w odbiorze są różne skrypty i podręczniki. W tym celu sprawdza w każdym z nich występowanie słów w stylu: proste, trywialne, widać, oczywiście. Zakłada, że skrypty z dużą ilością tych słów są przeznaczona dla ludzi obeznanych już z materiałem i wybiera ten, który takowych zawiera jak najmniej.

- Badacz chce sprawdzić następującą hipotezę: w przedmiotach podobnych do analizy używa się dużo literki “x”. W przedmiotach podobnych do algebry liniowej używa się dużo literki “n”.

- Topologia (szczególnie przedmiot Topologia I) słynie z olbrzymiej ilości definicji, których trzeba się nauczyć na pamięć. Na egzaminie jest również część teoretyczna, w której należy sformułować wybrane twierdzenia, bądź pojęcia; często wraz z dowodem. Student II roku zastanawia się które z tych pojęć są najważniejsze, tj. do czego powinien przyłożyć największą wagę. W tym celu przeprowadza analizę częstości słów, by sprawdzić które definicje są kluczowe, tj. wielokrotnie pojawiają się w skrypcie.

- Istnieją (przynajmniej) dwa sposoby na zapisywanie pochodnych: notacja Lagrange’a i notacja Leibniza[3]. Badacz matematycznych symboli używa systemu by dowiedzieć się którą preferują fizycy, którą biolodzy, a którą ekonomiści i w jakich kontekstach.

[3] Problemy z notacją Leibniza są spore, dla zainteresowanych takimi niuansami - dobra wypowiedź klaryfikująca wiele spraw: https://math.stackexchange.com/questions/3266639/notation-for-partial-derivative-of-functions-of-functions/3270436#3270436