



AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

WYDZIAŁ GEOLOGII, GEOFIZYKI I OCHRONY ŚRODOWISKA

KATEDRA GEOINFORMATYKI I INFORMATYKI STOSOWANEJ

Projekt dyplomowy

Aplikacja do zarządzania zbiorami danych
Dataset management application

Autor: Monika Hertel
Kierunek studiów: Inżynieria i Analiza Danych
Opiekun pracy: dr Paweł Oleksik

Kraków, 2024

Spis treści

Wstęp	3
1 Zagadnienia teoretyczne	4
1.1 Problem tworzenia streszczeń	4
1.1.1 Metody ekstraktywne a abstrakcyjne	4
1.1.2 Ocena poprawności abstraktu	4
1.2 Tworzenie aplikacji webowych	5
1.3 Ekstrakcja tekstu plików PDF	5
2 Implementacja	7
2.1 Projekt systemu	7
2.2 Architektura systemu	7
2.2.1 Javascript React	7
2.2.2 Python Flask	7
2.2.3 Electron	8
2.3 Struktura aplikacji	8
2.4 Obsługiwane rodzaje plików	8
2.5 Schematy poszczególnych funkcjonalności	10
2.5.1 Dodawanie pliku PDF	11
2.5.2 Wyszukiwanie i wyświetlanie plików	12
2.5.3 Asocjacja plików o podobnej tematyce	13
2.5.4 Obsługa plików csv	14
2.6 Testy funkcjonalne	14
3 Możliwości rozwoju i wykorzystania aplikacji	14
Podsumowanie i wnioski	14
Spis Rysunków	15
Bibliografia	15

Wstęp

(Tutaj niedługo powstanie wstęp)

Cel i zakres projektu

Projekt ma na celu przedstawienie możliwości systemu ułatwiającego użytkownikowi proces zbierania i segregowania informacji. Poniższa praca skupi się na ogólnym koncepcie aplikacji oraz szczegółowym opisie implementacji modułu tworzącego streszczenia.

1 Zagadnienia teoretyczne

1.1 Problem tworzenia streszczeń

Streszczenie w każdej pracy naukowej jest jej ważną częścią. Osoby poszukujące informacji na dany temat nie są w stanie przetworzyć ilości dostępnych informacji. Ma ono na celu przekazać kluczowe informacje o czytany dokumencie, aby czytelnik mógł ocenić czy dany artykuł zawiera informacje mu przydatne.

W kontekście uczenia maszynowego kondensowanie treści, w celu stworzenia kontekstu, opiera się o obliczanie poziomu istotności dla każdego zdania. [1]. Automatyzacja tego procesu dąży do ułatwienia autorom tworzenia prac, poprzez skrócenie czasu, którego wymaga kreacja abstraktu. Systemy ATS (ang. *Automatic Text Summarization*) są jednym z cięższych wyzwań sztucznej inteligencji, dotyczących przetwarzania języka naturalnego [2]. Podejścia do tworzenia tych systemów, możemy podzielić na ekstraktywne, abstrakcyjne i hybrydowe.

1.1.1 Metody ekstraktywne a abstrakcyjne

Większość badań nad systemami ATS skupia się na użyciu metod ekstraktywnych, starając się przy tym uzyskać zwarte i kompletne streszczenia. Podejście ekstraktywne polega na wybraniu najważniejszych zdań z całego dokumentu, a długość uzyskanego wyniku zależy od wartości stopnia kompresji [3].

Streszczenia stworzone z użyciem metod abstraktywnych zawierają zdania, które nie znajdują się w oryginalnym tekście. System musi w pewnym stopniu “rozumieć” tekst, aby móc poprawnie zinterpretować dokument. Abstraktywność wymaga implementacji bardziej złożonych algorytmów przetwarzania języka naturalnego.

Istnieje również podział zadania podsumowania na nienadzorowane i nadzorowane. Nienadzorowane tworzą streszczenia tylko na podstawie danych wejściowych, czyli jedynie zawartości wprowadzanego dokumentu. Do optymalnego wyboru zdań, takie systemy implementują metody natury heurystycznej. Z tego powodu są one odpowiednie do przetwarzania danych na bieżąco.

Sposób nadzorowany wymaga przeprowadzenia fazy trenowania modelu, która wymaga wprowadzenia opisanego zbioru treningowego. Takie zbiory powinny posiadać docelowe postacie streszczeń otrzymane z pełnego tekstu dokumentu. Taki proces jest trudny do przeprowadzenia na większej ilości tekstów.

1.1.2 Ocena poprawności abstraktu

Zazwyczaj ingerencja człowieka jest wymagana przy ewaluacji wytworzonego streszczenia. Treść jest sprawdzana pod kątem poprawności gramatycznej, składni czy całościowej spójności. Taka ewaluacja wymaga dużego nakładu pracy, a przy większych projektach jest praktycznie niemożliwe. Dlatego możliwość zautomatyzowania tego procesu jest wręcz wymagana.

Jednymi z pierwszych metod ewaluacji automatycznej były metryki takie jak podobieństwo cosinusowe (*ang. cosine similarity*), *unit overlap* czy miara najdłuższego wspólnego podzdanania (*ang. longest common subsequence*). Te elementy zostały później skondensowane w zbiór kryteriów opisanych akronimem *ROUGE* (*ang. Recall-Oriented Understudy of Gisting Evaluation*) [4]. Poniższy segment przedstawia poszczególne komponenty zestawu *ROUGE*, to jest kryteria *ROUGE-N*, *ROUGE-L* oraz *ROUGE-S*.

ROUGE-N mierzy poziom pokrycia n -gramów, czyli ciągłych sekwencji n słów, pomiędzy wytworzonym tekstem a tekstem źródłowym. Najczęstszym przypadkiem użycia tego kryterium jest ewaluacja poprawności gramatycznej. Miary *ROUGE-1* oraz *ROUGE-2* należą do miar *ROUGE-N* i w kolejności korzystają z unigramów i bigramów. Tymczasem *ROUGE-L* jest wyznaczana na podstawie porównania najdłuższych wspólnych sekwencji słów w zdaniach, występujących w streszczeniu wygenerowanym i wzorcowym. *ROUGE-S* działa na tej samej zasadzie co *ROUGE-2* lecz uwzględnia w swoim działaniu struktury skip-bigram, który jest rozszerzeniem definicji bigramów o możliwość zawierania w sobie maksymalnie jednego przedimka.

1.2 Tworzenie aplikacji webowych

(pomysł)

1.3 Ekstrakcja tekstu plików PDF

Pliki PDF są powszechnie przyjętym standardem przechowywania informacji. Format PDF jest oparty o strukturę binarnego formatu plików, zoptymalizowanego pod kątem wysokiej wydajności odczytu wizualnego. Zawierają w sobie informacje o strukturze dokumentu, takie jak zawartość tekstowa, grafiki czy użyta czcionka. Są one zoptymalizowane pod drukowanie. Niezaszyfrowane PDF mogą być w całości reprezentowane z użyciem wartości bitowych, odpowiadających części zbioru znaków zdefiniowanego w *ANSI X3.4-1986*, symbole kontrolne oraz puste znaki. Wizualnie jednak nie są one ograniczone do zbioru znaków ASCII.

Format PDF separuje informacje dotyczące samego znaku a jego wizualną reprezentacją. Jest to rozróżnienie na znak pisarski i glif, gdzie grafem jest jednostką tekstu a glif, jednostką graficzną. Glif informuje o położeniu znaków na stronie dokumentu, jego czcionce i innych elementach wyglądu.

Otrzymanie zawartości pliku może być wykonane za pomocą wydobywania elementów PDF z strumienia pliku lub z użyciem analizy obrazów, na przykład technologii optycznego rozpoznawania znaków OCR (*ang. Optical Character Recognition*). Podstawowym zadaniem systemu OCR jest konwersja dokumentów w dane możliwe do edytowania czy wyszukiwania. Techniki oparte o analizę obrazów są bezpośrednio zależne od jakości wprowadzonego elementu. Idealną sytuacją dla wykożystania metod OCR jest kiedy posiadany plik zawiera w sobie jedynie tekst i jest obrazem

binarnym [5]. Dodatkowym atutem takich systemów jest możliwość wykrycia pisma i konwersja na tekst.

Ekstrahowanie danych z strumienia pliku, wiąże się z kilkoma problemami. Biorąc pod uwagę możliwość że plik pdf może posiadać różne kodowanie takie jak *UTF-8*, *ASCII* czy *Unicode*, możemy doświadczyć utraty informacji spowodowanej schematem kodowania pliku. Automatyczna ekstrakcja zawartości polega na selekcji znaków znajdujących się pomiędzy zdefiniowanymi słowami kluczowymi. Pliki PDF są przystosowane do drukowania, z tego powodu reprezentacja tekstu w strumieniu może się znacząco różnić od tej na stronie. Ponieważ pozycje znaków na poszczególnych stronach są absolutne, przedstawienie w strumieniu bierze pod uwagę spacje między elementami.

Niezależne od sposobu pobierania informacji, nie jest możliwe zagwarantowanie ich poprawności względem dokumentu wejściowego. Brak formalnej definicji struktury, przynajmniej jeżeli chodzi o artykuły naukowe, uniemożliwia stworzenie uniwersalnego algorytmu ekstrakcji.

2 Implementacja

2.1 Projekt systemu

Aplikacja została zaprojektowana z myślą o użytkownikach będących w posiadaniu dużej ilości plików z danymi. Głównym celem jest automatyzacja procesu segregowania i przetwarzania informacji. System etykietuje otrzymane pliki poprzez ekstrakcje słów kluczowych i daje możliwość wyszukiwania plików w bardziej optymalny sposób. Otrzymane dokumenty są kategoryzowane ze względu na słowa klucze. Po otrzymaniu elementów o podobnej strukturze i zawartości, są one łączone ze sobą. Informacja o takim połączeniu jest umieszczana w bazie danych, tak aby przy odczycie jednego z elementów, drugi był sugerowany jako następny plik do wyświetlenia.

2.2 Architektura systemu

Aplikacja została zbudowana z myślą o zapewnieniu interfejsu użytkownika przy użyciu JavaScript React, serwera HTTP opartego na Flask Python, oraz bazy danych MongoDB, która jest bazą typu NoSQL. Zintegrowana całość jest uruchamiana z wykorzystaniem pakietu Electron, co umożliwia stworzenie aplikacji natywnej, zachowując przy tym funkcje przeglądarki oraz pozwalając na dostęp do zasobów systemowych.

2.2.1 Javascript React

Jest to biblioteka pozwalająca na budowanie interaktywnych interfejsów użytkownika. Główną koncepcją React są komponenty, czyli samodzielne, hermetyczne jednostki interfejsu. Pozwala to na reakcje na zmiany wywołane przez użytkownika lub sam system i przy tym automatyczne aktualizowanie tych stron. React używa składni JSX (*Javascript XML*), który jest rozszerzeniem składni Javascript. Integruje ona kod Javascript z deklaratywnym opisem struktury interfejsu.

2.2.2 Python Flask

Flask to lekki i elastyczny framework, pozwalający na szybkie tworzenie aplikacji internetowych przy minimalnej ilości kodu. Dostarcza on jedynie niezbędne narzędzia, stawiając tym na dużą swobodę implementacji innych funkcji. Funkcja serwera jest uruchamiana w sytuacji gdy na podany adres URL zostanie wysłany komunikat. Informacja o aktywującym adresie jest przekazywana z użyciem dekoratora `.route()`. Rysunek X przedstawia przykład takiego wywołania.

Jedną z możliwości tej biblioteki jest generowanie plików strony jako wynik końcowy działania funkcji. Korzystamy w takiej sytuacji z metody `render_template()`, która otrzymuje na wejściu uprzednio utworzony szablon HTML.

Implementacja serwera z pomocą pakietu Flask umożliwia korzystanie z szerokiego spektrum bibliotek dotyczących sztucznej inteligencji.

```

from flask import Flask

app = Flask(__name__)

@app.route("/")
def hello_world():
    return "<p>Hello, World!</p>"

```

Rysunek 1: Przykładowa implementacja funkcji serwera

2.2.3 Electron

Biblioteka Electron jest projektem open-source umożliwiającym tworzenie aplikacji natywnych używając technologii web. Standard tworzenia aplikacji z pomocą tej biblioteki jest oparty o komunikację międzyprocesową (*ang. IPC, Inter-process communication*), która jest wynikiem implementacji izolacji kontekstu. Izolacja kontekstu odpowiada za separację procesu renderowania strony od głównego procesu aplikacji w środowisku Node.js. Zwiększa to bezpieczeństwo aplikacji, negując możliwość dostępu do użytych API z strony.

System posiada jeden główny proces *“main”*, którego celem jest tworzenie aplikacji i kierowanie podprocesami, oraz wiele procesów odpowiedzialnych za renderowanie zawartości strony zwanych procesami *“renderer”*. Biblioteka umożliwia również dodanie skryptów, które działają przed renderowaniem strony (*ang. preload scripts*). Otrzymują one dodatkowe możliwości w porównaniu plików *renderer*, poprzez posiadanie uprawnień do korzystania z Node.js APIs. Mogą one bezpiecznie udostępnić części API dla procesów strony.

2.3 Struktura aplikacji

Zakres implementacji obejmuje stworzenie serwera działającego w oparciu o protokół HTTP. Serwer w odpowiedzi na zwraca odpowiedź w postaci pliku JSON, który następnie jest przetwarzany przez stronę React, czego wynik jest wyświetlany klientowi. po stronie klienta.

2.4 Obsługiwane rodzaje plików

System jest przystosowany do obsługi plików o rozszerzeniach takich jak *PDF*, *CSV* i *TXT*. Każdy rodzaj pliku ma osobne funkcjonalności opisane poniżej. Do plików przypisujemy tag opisujący rodzaj pliku, lecz nie jest on odrębną jednostką od słów kluczowych generowanych na podstawie zawartości pliku.

Pliki PDF

Proces obsługi plików pdf jest zależny od wielu czynników. Najprostrzymi plikami do analizy są artykuły posiadające zakładki (*ang. outlines*) o ujednoliconej strukturze,

zawierające jedynie tekst. *Outlines* są potrzebne do ustalenia tytułu danego pliku, ponieważ nie zawsze nazwa samego pliku koresponduje tytułowi.

System wykorzystuje potokowanie zawartości w celu ekstrakcji tekstu, obrazów oraz tabel. Działanie aplikacji jest oparte o sam tekst pliku co Na potrzeby aplikacji, takie fragmenty jak tabele czy grafiki są jedynie przeszkodą , ponieważ nie powinny one wpływać na wynik algorytmu streszczania. Część elementów tabelarycznych jest niepoprawnie podpisywana jako tekst, co może przyczynić się do generacji niezrozumiałych streszczeń.

Pliki tabelaryczne CSV

Podstawowym zachowaniem aplikacji w obliczu plików zawierających dane tabelaryczne w systemie przecinkowym generacja opisu zgodnie z następującymi krokami. Dla każdej kolumny pobierana jest jej nazwa oraz rodzaj danych zawartych w niej (*double*, *string*, itp.).

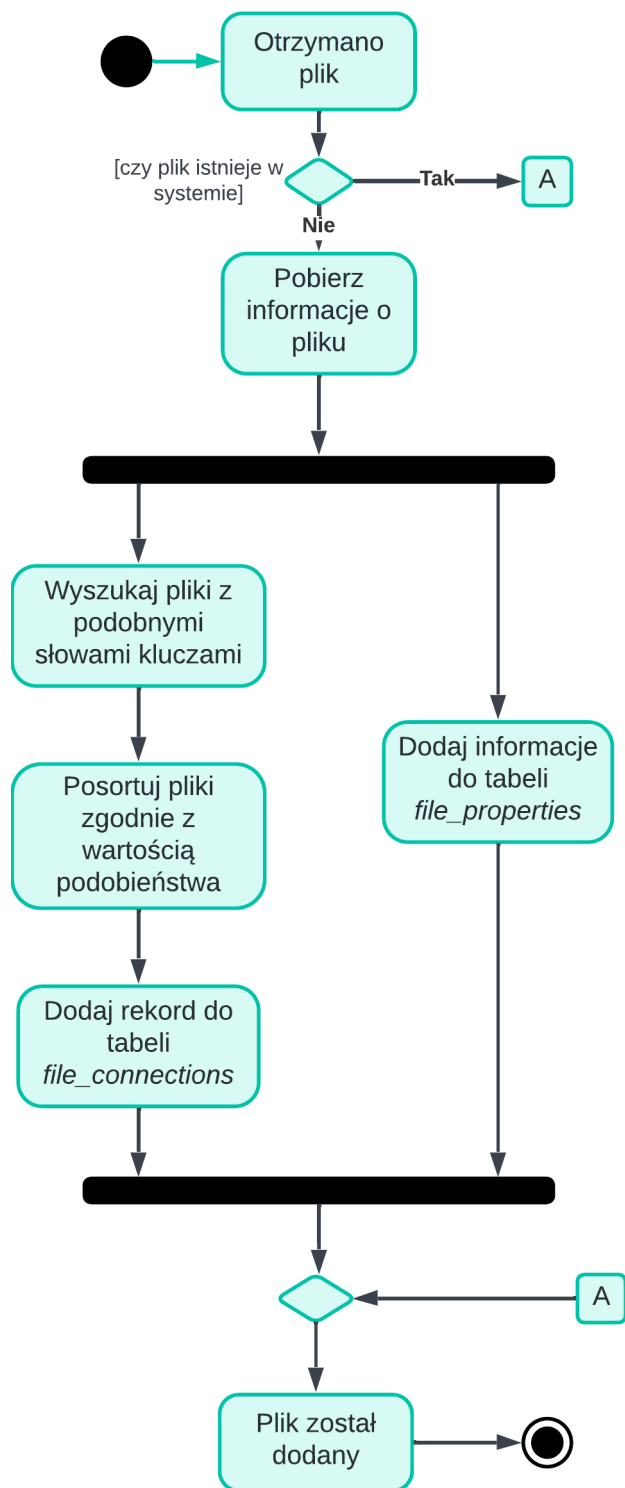
Ze względu na rodzaj pliku generacja słów kluczy może przebiec na dwa sposoby: tagami zostają nazwy kolumn lub użytkownik samodzielnie je dodaje. Oba podejścia mają swoje wady. Nazwy kolumn nie są wymagane podczas kreacji tabel, więc tworzenie tagów w takiej sytuacji nie jest optymalne. Drugie podejście wymaga od użytkownika większego wkładu w ten proces. Potrzebna do tego jest pewna znajomość zestawu danych, który chcemy wprowadzić do systemu aby odpowiednio przypisać słowa klucze do plików.

Pliki tekstowe .txt

Zawartość plików tekstowych nie jest uwarunkowana żadnymi normami, co czyni obsługę tych plików niemożliwą do standaryzacji. Podczas dodawania do aplikacji, wymagają one od użytkownika określenia rodzaju zawartości. Opcjami, które użytkownik ma do wyboru są: tekst lub dane tabelaryczne.

W pierwszym przypadku program sprawdza długość tekstu i na tej podstawie decyduje o następnych krokach. Domyślną długością graniczną jest 200 słów, lecz może ona być ustawiona manualnie przez klienta. Po przekroczeniu tego progu, system procesuje plik w sposób podobny do obsługi plików *PDF*. Tworzone jest streszczenie.

2.5 Schematy poszczególnych funkcjonalności



Rysunek 2: Diagram aktywności przesyłania pliku

2.5.1 Dodawanie pliku PDF

Zachowanie systemu przy dodaniu plików tekstowych.

1. Gdy użytkownik wybierze plik, aplikacja przesyła ścieżkę pliku do serwera,
2. System ekstrahuje z pliku całą zawartość z pomocą biblioteki języka Python *PdfMiner*,
3. Z zawartości zostają wyciągnięte słowa kluczowe. Tworzenie tzw. tagów odbywa się z pomocą biblioteki *yake* oraz metody *KeywordExtractor()*. Przyjmuje ona następujące parametry:
 - *lan* - język danego tekstu,
 - *n* - maksymalna ilość słów w jednym tagu,
 - *deduplim* - szansa na powtórzenie się słów w różnych słowach kluczach,
 - *top* - ilość elementów wyjściowych

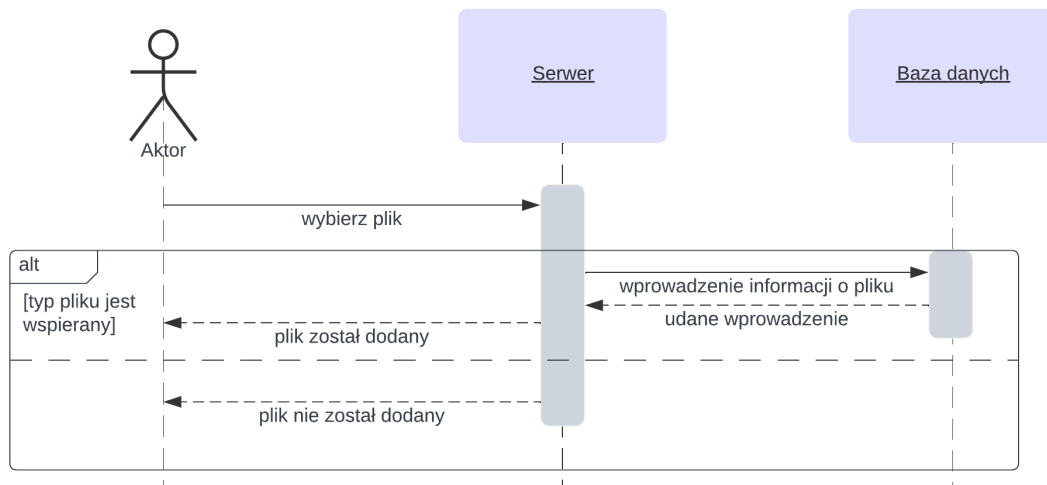
```
lang = "en"
max_ngram_size = 3
deduplication_treshold = 0.5
num_keywords = 6
kw_extract = yake.KeywordExtractor(lan=lang, n=max_ngram_size,
                                   dedupLim=deduplication_treshold,
                                   top=num_keywords, features=None)
keywords = kw_extract.extract_keywords(text)
```

Rysunek 3: Przykład użycia *KeywordExtractor()*

4. w tym samym czasie z pomocą biblioteki *sumy* oraz funkcji *TextRankSummarizer*, jest tworzone streszczenie z użyciem algorytmu *TextRank*,
5. jeżeli plik pdf posiada zakładki (ang. *outlines*) to metoda *.get_outlines()* ekstrahuje je i z wyniku możemy otrzymać tytuł dokumentu i nadać plikowi taką nazwę. a jeżeli ich nie ma to tytułem pliku zostaje bazowa nazwa pliku,
6. ostatecznie wszystkie informacje są zbierane i przesyłane do bazy danych.

Model *TextRank* zastosowany w projekcie, jest algorytmem rankingowym opartym na grafach [6]. Operuje on na zasadach “głosowania” i “rekomendacji”. Jest on używany do tworzenia streszczeń metodą ekstraktywną i nienadzorowaną. Dla naszego systemu, nie jest możliwe wytworzenie odpowiedniego zbioru treningowego, aby móc zastosować metody nadzorowane.

Biblioteka *PdfMiner* języka Python

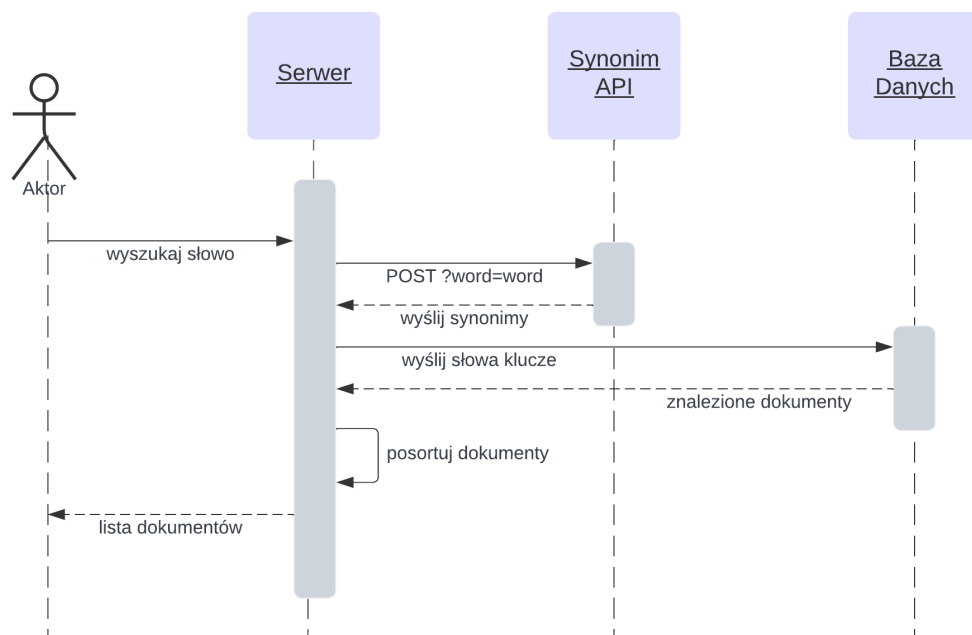


Rysunek 4: Schemat sekwencyjny dodawania pliku

2.5.2 Wyszukiwanie i wyświetlanie plików

Użytkownik ma możliwość wyszukiwania plików po słowach kluczach lub tytule pliku. Tutaj przydatne jest użycie API generującego synonimy dla wyszukiwanego słowa. System zachowuje się w poniżej opisany sposób.

1. aplikacja przesyła komunikat z wyszukiwanym słowem,
2. serwer używając API synonimów pobiera 5 najbliższych słów do słowa szukanego przypisując im ranking,
3. serwer przesyła osobne komunikaty do tabeli *file_properties* w naszej bazie danych, zawierające osobno słowo klucz oraz synonimy
4. baza zwraca informacje dotyczące znalezionych dokumentów oraz słowo klucz,
5. serwer zwraca listę plików użytkownikowi, posortowane zgodnie z rankingiem



Rysunek 5: Schemat sekwencyjny wyszukiwania pliku

2.5.3 Asocjacja plików o podobnej tematyce

Poniższa sekcja przedstawia proces łączenia dokumentów na podstawie wygenerowanych słów kluczowych. System z każdym dodanym plikiem przeszukuje przestrzeń dokumentów i aktualizuje już utworzone rekordy, w sytuacji dopasowania. Schemat przedstawia tworzenie połączeń dla nowo dodanego pliku.

1. dla każdego słowa kluczowego, jest pobrana lista ID plików z tabeli *file_properties*, pod warunkiem dopasowania przynajmniej jednego tagu,
2. następnie obliczany jest współczynnik podobieństwa, którego schemat postępowania jest przedstawiony poniżej:
 - (a) wyznaczana jest wartość dla poszczególnych słów kluczowych
 - jeżeli wysłane słowo kluczowe zgadza się fragmentarycznie z otrzymanym tagiem (np. “zaburzenia” a “zaburzenia odżywiania”), to zapisywana jest wartość $2 \cdot \frac{\text{ilosc}(\text{dopasowan})}{\text{ilosc_slow}(\text{wejście}) + \text{ilosc_slow}(\text{pobrane})}$,
 - jeżeli słowa są identyczne, zapisywana jest wartość 1,
 - (b) wartości dla wszystkich słów kluczowych są sumowane i dzielone przez liczbę tagów.
3. do ID pliku przypisana jest wartość podobieństwa, i po posortowaniu od największej wartości, ID są dodawane do jednego rekordu w tabeli *file_connections*
 - jeżeli wartość podobieństwa nie przekracza 0.3, to plik nie jest brany pod uwagę

2.5.4 Obsługa plików csv

2.6 Testy funkcjonalne

3 Możliwości rozwoju i wykorzystania aplikacji

Podsumowanie i wnioski

Spis rysunków

1	Przykładowa implementacja funkcji serwera	8
2	Diagram aktywności przesyłania pliku	10
3	Przykład użycia <i>KeywordExtractor()</i>	11
4	Schemat sekwencyjny dodawania pliku	12
5	Schemat sekwencyjny wyszukiwania pliku	13

Literatura

- [1] B. Mutlu, E. A. Sezer, and M. A. Akcayol, “Candidate sentence selection for extractive text summarization,” *Information Processing & Management*, vol. 57, no. 6, p. 102359, 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0306457320308542>
- [2] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, “Automatic text summarization: A comprehensive survey,” *Expert Systems with Applications*, vol. 165, p. 113679, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0957417420305030>
- [3] M. Gambhir and V. Gupta, “Recent automatic text summarization techniques: a survey,” *Artificial Intelligence Review*, vol. 47, no. 1, pp. 1–66, Jan 2017. [Online]. Available: <https://doi.org/10.1007/s10462-016-9475-9>
- [4] C.-Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: <https://aclanthology.org/W04-1013>
- [5] R. Mithe, S. Indalkar, and N. Divekar, “Optical character recognition,” *International journal of recent technology and engineering (IJRTE)*, vol. 2, no. 1, pp. 72–75, 2013.
- [6] R. Mihalcea and P. Tarau, “TextRank: Bringing order into text,” in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, D. Lin and D. Wu, Eds. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 404–411. [Online]. Available: <https://aclanthology.org/W04-3252>