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

Poniższa praca ma na celu zapewnienie kompleksowego przeglądu projektu aplikacji, która umożliwi osobom fizycznym lepsze zarządzanie i wykorzystywanie własnych zasobów wiedzy. Przedstawiając tym zasady projektowania, szczegóły implementacji i potencjalne rozszerzenia aplikacji.

Aplikacja ułatwia płynne gromadzenie danych, umożliwiając użytkownikom importowanie i przechowywanie plików zawierających cenne informacje. Pliki te, określane zbiorczo jako zbiór danych, obejmują szeroki zakres typów treści, z których każdy reprezentuje wiedzę możliwą do interpretacji przez człowieka. Od prac naukowych po osobiste notatki i pliki multimedialne, projekt systemu uwzględnia obsługę różnych formatów danych, aby zaspokoić indywidualne preferencje i wymagania użytkowników.

Choć początkowo została zaprojektowana jako proste rozwiązanie programowe dla użytkowników prywatnych, aplikacja może być wyposażona w rozszerzenia, które wykorzystują techniki sztucznej inteligencji (AI).

W niniejszym dokumencie zostanie zagłębiony koncept i możliwy rozwój tej aplikacji, podkreślając kluczowe komponenty, takie jak moduł gromadzenia danych, mechanizm kategoryzacji czy funkcjonalność wyszukiwania. Ponadto zostanie zbadana integrację rozszerzeń opartych na sztucznej inteligencji w celu zautomatyzowania takich elementów jak proces klasyfikacji, oferując wgląd w podstawowe algorytmy i strategie wdrażania.

1 Zagadnienia dotyczące projektu

Zaprojektowany system ma uprościć proces gromadzenia informacji dla użytkownika prywatnego. Zgromadzone dane mogą być przechowywane w pliku w systemie plików lub przy użyciu takich narzędzi jak bazy danych. Rozważając rodzaj przechowywania danych należy również wziąć pod uwagę aspekt fizyczny. Rozróżnienie to jest opisane w rozdziale 1.1. Wraz z danymi będą przechowywanie metainformacje, opisujące te dane.

Pozyskiwanie informacji z plików różni się zależnie od ich rodzaju. Aplikacja umożliwia przetworzenie plików tabelarycznych czy tekstowych, poprzez wykorzystanie ich zawartości. Szczególnym przypadkiem są pliki PDF, gdzie proces ekstrakcji zawartości jest utrudniony w związku z kodowaniem i formatem pliku, co zostanie omówione w dziale 1.2.

Projekt aplikacji wykorzystuje architekturę typu klient-serwer. Określenie architektura aplikacji odnosi się do wysokopoziomowej struktury i organizacji systemów oprogramowania. W rozdziale 1.4 została opisana wyżej wymieniona architektura pod kątem charakterystyki i przypadków użycia.

1.1 Sposoby przechowywania informacji

Istnieje wiele sposobów magazynowania danych. Wybór odpowiedniego sposobu zależy od indywidualnych potrzeb użytkownika oraz od ilości informacji. Na poziomie organizacji dzielą się one na serwery lokalne (*ang. on-premises servers*) oraz przechowywanie w chmurze (*ang. cloud storage*). Poniżej zostały opisane obie metody.

Pojęcie **serwerów on-premises** odnosi się do fizycznego sprzętu i infrastruktury, które znajdują się pod kontrolą organizacji lub osoby fizycznej. Podejście to obejmuje konfigurowanie i utrzymywanie serwerów na miejscu w celu przechowywania danych i zarządzania nimi, bez angażowania usługodawców zewnętrznych. Oto przegląd zalet i wad korzystania z serwerów lokalnych do przechowywania danych [1]:

- zalety:

    1. Bezpośrednia kontrola - Organizacje mają pełną kontrolę nad swoim sprzętem i oprogramowaniem, co pozwala na większą elastyczność za-

rządzania w celu spełnienia określonych potrzeb biznesowych i wymogów bezpieczeństwa.

2. Bezpieczeństwo - Serwery lokalne zapewniają wyższy poziom bezpieczeństwa i prywatności danych, ponieważ mogą one wdrażać własne środki bezpieczeństwa i protokoły dostosowane do ich konkretnych potrzeb. Może to obejmować fizyczne środki bezpieczeństwa, takie jak ograniczony dostęp do serwerowni i budynków.
3. Personalizacja - Oferują one elastyczność w zakresie dostosowywania konfiguracji sprzętowych lub konfiguracji sieci w celu optymalizacji wydajności i spełnienia określonych wymagań dotyczących obciążenia.

- wady:

1. Koszty początkowe oraz utrzymania - Konfiguracja serwerów lokalnych wymaga znacznych inwestycji w sprzęt, licencje, konfigurację sieciową i strukturalną. Może to obejmować koszty związane z zakupem serwerów czy systemów chłodzenia. Wymusza to również odpowiedzialność za bieżącą konserwację, zarządzanie oprogramowaniem serwerowym, co może wymagać dedykowanego personelu i zasobów IT.
2. Ograniczenia skalowalności - Należy zakupić dodatkowy sprzęt i rozszerzyć infrastrukturę, aby sprostać rosnącym potrzebom w zakresie przechowywania i przetwarzania danych.

Usługi **przechowywania w chmurze**, takie jak Amazon Web Services (AWS) czy Azure Cloud, zapewniają możliwość przechowywania danych i zarządzania nimi w zdalnych centrach danych utrzymywanych przez zewnętrznych dostawców usług. Poniżej przedstawiono przegląd wad i zalet korzystania z usług przechowywania danych w chmurze:

- zalety:

1. Skalowalność: Usługi przechowywania danych w chmurze oferują praktycznie nieograniczoną skalowalność, umożliwiając organizacjom łatwe

zwiększanie lub zmniejszanie pojemności pamięci masowej w zależności od zapotrzebowania.

2. Opłacalność: Pamięć w chmurze w większości przypadków działa w modelu cenowym pay-as-you-go, w którym organizacje płacą tylko za przestrzeń dyskową i zasoby, które wykorzystują. Może to prowadzić do oszczędności kosztów w porównaniu z tradycyjnymi lokalnymi rozwiązaniami.

3. Dostępność: Zapewniają wszechobecny dostęp do danych z dowolnego miejsca z połączeniem internetowym. Umożliwia to zdalny dostęp, współpracę i udostępnianie danych między rozproszonymi zespołami, poprawiając produktywność oraz rozszerzając działalność firmy globalnie.

4. Wysoka dostępność i niezawodność: Usługi przechowywania danych w chmurze zazwyczaj oferują wbudowane funkcje redundancji i wysokiej dostępności, zapewniając replikację danych w wielu centrach danych i ochronę przed awariami sprzętu lub innymi zakłóceniami. Skutkuje to zwiększoną niezawodnością danych i minimalnymi przestojami.

- wady:

1. Zależność: Zależność od dostawcy usług w zakresie dostępności, bezpieczeństwa i wydajności danych. Wszelkie zakłócenia lub przerwy w świadczeniu usług po stronie dostawcy mogą mieć wpływ na dostęp do danych i operacje biznesowe.

2. Koszty transferu danych: Podczas gdy przechowywanie danych w chmurze może być opłacalne, organizacje mogą ponieść dodatkowe koszty transferu danych, zwłaszcza w przypadku przenoszenia dużych ilości danych do i z chmury. Koszty mogą znacznie wzrosnąć w przypadku obciążeń wymagających dużej przepustowości.

3. Problemy ze zmianą miejsca przechowywania: Migracja danych między dostawcami usług w chmurze lub powrót do infrastruktury lokalnej mogą być złożone i kosztowne. Organizacje ryzykują uzależnienie od dostawcy, gdy mocno inwestują w ekosystem określonego dystrybutora rozwiązań, co w dłuższej perspektywie ogranicza ich elastyczność.

Wybór miejsca przechowywania danych zależy od takich czynników, jak wymagania bezpieczeństwa, potrzeby skalowalności czy ograniczenia budżetowe. Często korzystne dla organizacji jest przyjęcie podejścia hybrydowego lub wielochmurowego, wykorzystującego mocne strony różnych opcji przechowywania danych. Umożliwia to zaspokojenie różnorodnych potrzeb biznesowych przy jednoczesnym złagodzeniu wpływu niedoskonałości różnych rozwiązań [2].

1.2 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 [3].

Format PDF separuje informacje dotyczące samego znaku a jego wizualną reprezentacją. Jest to rozróżnienie na znak pisarski (grafem) 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 wydobycia elementów PDF z strumienia pliku lub z użyciem analizy obrazów, na przykład technologi optycznego rozpoznawania znaków OCR (*ang. Optical Character Recogintion*). 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 [4]. 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 koordynaty elementów.

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.

1.3 Problem tworzenia streszczeń

Streszczenie artykułu naukowego jest jego kluczowym elementem. Ilość informacji dostępnych dla każdego, kto szuka wiedzy na dany temat, może być przytłaczająca. Jego celem jest przekazanie najważniejszych cech czytanej treści, aby czytelnik mógł określić, czy informacje są dla niego istotne.

W kontekście uczenia maszynowego, kondensacja treści w celu stworzenia abstraktu, najczęściej opiera się na obliczaniu poziomu istotności dla każdego zdania [5]. Skracając czas potrzebny na napisanie abstraktu, automatyzacja ma na celu ułatwienie autorom pisania artykułów. Systemy ATS (ang. *Automatic Text Summarization*) są jednym z cięższych wyzwań sztucznej inteligencji, dotyczących przetwarzania języka naturalnego [6]. Podejścia do tworzenia tych systemów, możemy podzielić na ekstraktywane, abstrakcyjne i hybrydowe.

1.3.1 Metody ekstraktywne a abstrakcyjne

Większość badań nad systemami ATS skupia się na użyciu metod ekstraktywnych, starając się przy tym uzyskać zwięzłe i kompletne streszczenia. Podejście ekstraktywne polega na wybraniu najważniejszych zdań z całego dokumentu, a długość uzyskanego wyniku zależy od wartości stopnia kompresji [7].

Streszczenia stworzone z użyciem metod abstrakcyjnych wymagają głębszej analizy tekstu wejściowego. Generują one podsumowanie poprzez "zrozumienie" głównych pojęć w dokumencie wejściowym. Dzieje się to poprzez implementacje zło-

żonych algorytmów przetwarzania języka naturalnego (*ang. NLP, Natural Language Processing*). Następnie dokonywane jest parafrazowanie tekstu w celu wyrażenia tych pojęć z użyciem słów, które nie należą do oryginalnego tekstu. W praktyce rozwój wspomnianych systemów wymaga kompleksowych zbiorów danych.

Istnieje również podział technik tworzenia streszczeń na nienadzorowane i nadzorowanie. Nienadzorowane tworzą streszczenia tylko na podstawie danych wejściowych, czyli jedynie zawartości wprowadanego dokumentu. Próbują one odkryć ukrytą strukturę w nieoznakowanych danych. Techniki te są zatem odpowiednie dla wszelkich nowo zaobserwowanych danych nie wymagających modyfikacji.

Sposób nadzorowany wymaga trenowania modelu, jak i 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 [6].

1.3.2 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 podzdania (*ang. longest common subsequence*). Te elementy zostały później skondensowane w zbiór kryteriów opisanych akronimem *ROUGE* (*ang. Recall-Oriented Understudy of Gisting Evaluation*) [8]. 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 kożystają z unigramów i bigramów.

- *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.

1.4 Architektura klient-server

Architektura klient-serwer to model projektowania systemów rozproszonych, w którym zadania lub procesy obliczeniowe są podzielone między klientów i serwery. W tej architekturze klientami są zazwyczaj urządzenia użytkowników końcowych (takie jak komputery, smartfony lub tablety), które żądają usług lub zasobów, podczas gdy serwery są scentralizowanymi systemami, które je zapewniają. Architektura najczęściej wykorzystywana w wytwarzaniu oprogramowania internetowego z wykorzystaniem takich narzędzi jak języki opisowe HTML i CSS, czy języka Javascript. Takie aplikacje działają w przeglądarkach internetowych, lecz dzięki takim szkieletom aplikacji jak Electron istnieje możliwość konwersji na program z własnym interfejsem graficznym, oddzielonym od przeglądarki. Opis tego projektu znajduje się w dalszej części rozdziału.

Transfer informacji między elementami aplikacji może odbywa się z pomocą protokołu komunikacyjnego jak HTTP/HTTPS

HTTP (*ang. Hypertext transfer protocol*) jest to bezstanowy protokół, co oznacza, że każde zapytanie jest przetwarzane niezależne od poprzedniego. Jest on zgodny z modelem zapytanie-odpowiedź (*ang. request-response*), gdzie klient wysyła prośbę a serwer odpowiada. Odpowiedź zawiera w sobie kod statusu, wymagane zasoby lub treść błędu.

1.4.1 Pakiet Electron

Electron jest oparty na Chromium, projekcie stojącym za przeglądarką Google Chrome. Oznacza to, że aplikacje wykorzystują ten sam silnik renderujący i obsługę standardów internetowych, co Chrome. Electron osadza instancję Chromium do rende-

rowania treści internetowych w oknach aplikacji.

Standard tworzenia aplikacji z pomocą tej biblioteki jest oparty o komunikację międzyprocesową (*ang. IPC, Inter-process communication*), która jest wynikiem implementacji izolacji wątku.

Izolacja wątku

Separacja wątków odnosi się do możliwości izolacji różnych części aplikacji w celu zapobiegania interferencji oraz zapewnienia stabilności aplikacji. Separacja ta jest osiągana poprzez wykorzystanie wielu procesów, w tym procesu głównego i procesów renderujących.

- Proces główny: Zarządza cyklem życia aplikacji i interakcjami z systemem operacyjnym. Działa we własnym, odizolowanym wątku, oddzielnie od procesów renderujących.
- Procesy renderujące: obsługują renderowanie i wyświetlanie treści internetowych w poszczególnych oknach aplikacji. Każdy proces renderujący działa niezależnie, odizolowany od innych procesów renderujących i procesu głównego.

Separacja wątku zapewnia, że zmiany dokonane w jednej części aplikacji nie wpływają na inne części, zwiększając stabilność i bezpieczeństwo. Izolacja umożliwia również efektywne zarządzanie zasobami i skalowalność, ponieważ każdy proces może być zarządzany i optymalizowany niezależnie.

Komunikacja międzyprocesowa (IPC)

Pomimo separacji między procesami, Electron zapewnia mechanizmy komunikacji między nimi poprzez IPC (*ang. Inter-Process Communication*). Pozwala to różnym częściom aplikacji na efektywną wymianę danych, wyzwalanie akcji i synchronizację stanu.

- IPC Main: Electron zapewnia moduł *ipcMain* w głównym procesie, pozwalając mu nasłuchiwać i obsługiwać zdarzenia IPC wysyłane z procesów renderujących.
- IPC Renderer: W procesach renderujących, moduł *ipcRenderer* umożliwia wysyłanie zdarzeń IPC do głównego procesu i odbieranie odpowiedzi.

Dzięki IPC możliwa jest implementacja dwukierunkowej komunikacji między procesem głównym a procesami renderującymi, umożliwiając im koordynację działań, udostępnianie danych i synchronizację stanu w różnych częściach aplikacji.

2 Projekt

Celem pracy było zaprojektowanie i wykonanie programu użytkowego wspomagającego gromadzenie informacji oraz późniejsze przeszukiwanie utworzonych kolekcji. Wśród jego planowanych funkcjonalności należy wymienić:

- możliwość gromadzenia danych różnych rodzajów (dokumenty tekstowe i hipertekstowe, dane tabelaryczne, grafiki i inne),
- automatyczną organizację i kategoryzację plików ze względu na zawartość
- automatyczne generowanie streszczeń dokumentów tekstowych tak, by zapewnić użytkownikowi szybki wgląd w treść bez konieczności zapoznania się z całą zawartością

Docelowym odbiorcą produktu ma być indywidualny użytkownik, korzystający z lokalnych zasobów. Dlatego założeniem projektowym było, aby unikać używania usług zewnętrznych, np. chmurowych.

2.1 Opis projektu

Ze względu na rozbudowany charakter narzędzia, zostało ono zaprojektowane jako system współpracujących elementów. Są to:

- moduł przetwarzania plików - najważniejszy moduł, decydujący jakie akcje zostaną podjęte dla danego pliku,
- moduł organizacji danych - segreguje dokumenty oraz tworzy połączenia między podobnymi plikami,
- wyszukiwarka - pobiera informacje z systemu związane z frazą wejściową i przedstawia użytkownikowi,
- moduł podsumowujący - niezależnie od rodzaju pliku tworzy skrótowy opis zawartości.

2.2 Wymagania funkcjonalne modułów

Moduł organizacji danych automatycznie kategoryzuje i organizuje dane użytkownika, aby ułatwić ich efektywne wyszukiwanie i zarządzanie.

- System tagowania: Użytkownicy mogą samodzielnie przypisywać etykiety do każdego dokumentu lub pozostawić te wygenerowane przez system, umożliwiając im kategoryzowanie i łączenie plików w oparciu o wspólne motywy lub tematy.

- Struktura folderów: Aplikacja zapewnia hierarchiczną strukturę folderów, która pozwala użytkownikom organizować swoje dane w zagnieżdżone kategorie, zapewniając uporządkowany i intuicyjny sposób poruszania się po swoich informacjach.

- Automatyzacja: Wykorzystując zaawansowane algorytmy, aplikacja analizuje zawartość elementów, aby automatycznie kategoryzować je w oparciu o cechy, takie jak słowa kluczowe, tematy lub podobieństwo treści.

Wyszukiwarka umożliwia użytkownikom szybkie i skuteczne znajdowanie określonych informacji w systemie danych.

- Funkcja wyszukiwania pełnotekstowego: Użytkownicy mogą przeszukiwać wszystkie elementy danych w aplikacji za pomocą słów kluczowych lub fraz, pobierając odpowiednie wyniki na podstawie pasujących treści.

- Zaawansowane filtry: Aplikacja oferuje szereg filtrów wyszukiwania, takich jak zakres dat, typ pliku lub tag, umożliwiając użytkownikom zawężenie wyników wyszukiwania i znalezienie dokładnie tego, czego szukają.

- Sugestie i autouzupełnianie: Gdy użytkownicy wpisują swoje zapytania, aplikacja zapewnia sugestie i opcje autouzupełniania, aby przyspieszyć proces wyszukiwania i poprawić dokładność.

Moduł podsumowujący generuje zwięzłe streszczenia długich dokumentów, zapewniając użytkownikom szybki wgląd w główne punkty artykułów.

- Algorytmy podsumowujące tekst: Aplikacja wykorzystuje zaawansowane algorytmy do podsumowywania tekstu, które mogą obejmować metody ekstrakcyjne, które wybierają ważne zdania lub metody abstrakcyjne, które generują nową treść podsumowania.

- Automatyczne podsumowanie: Użytkownicy mogą wybrać opcje dla automatycznego generowania podsumowań dla dokumentów w aplikacji, takie jak ilość zdań lub poziom szczegółowości podsumowań.

Przetwarzanie plików wykrywa rodzaj pliku dodanego przez klienta i wyznacza akcje do przeprowadzenia zgodnie z typem.

2.3 Przykładowe schematy funkcjonalności

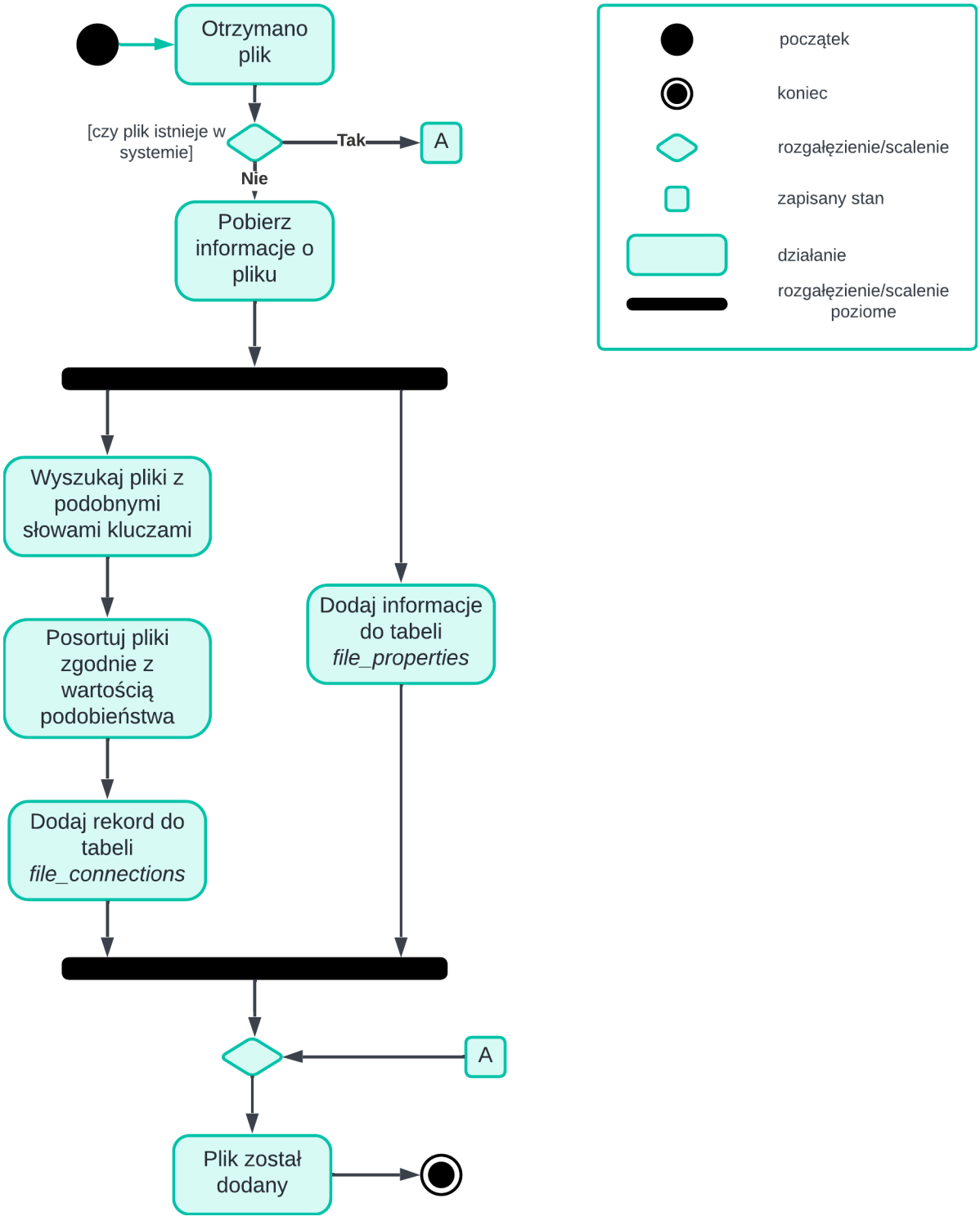
Dodawanie plików PDF

Zachowanie systemu przy dodaniu plików tekstowych.

1. Gdy użytkownik wybierze plik, aplikacja przesyła ścieżkę pliku do serwera,
2. System ekstraktuje z pliku całą zawartość,
3. Z tekstu wytwarzane są słowa klucze,
4. W tym samym czasie jest tworzone streszczenie z użyciem algorytmu *TextRank*,
5. jeżeli plik PDF posiada zakładki (ang. *outlines*) to 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 [9]. Operuje on na zasadach “głosowania” i “rekomendacji”. Jest on używany do tworzenia streszczeń metodą ekstraktywną i nienadzorowaną. Dla przedstawionego systemu, nie jest możliwe wytworzenie odpowiedniego zbioru treningowego, aby móc zastosować metody nadzorowane.

Poniższy diagram przedstawia proces dodawania pliku PDF wraz z uwzględnieniem implementacji asocjacji plików o podobnej tematyce, opisanej w punkcie 2.4.3.

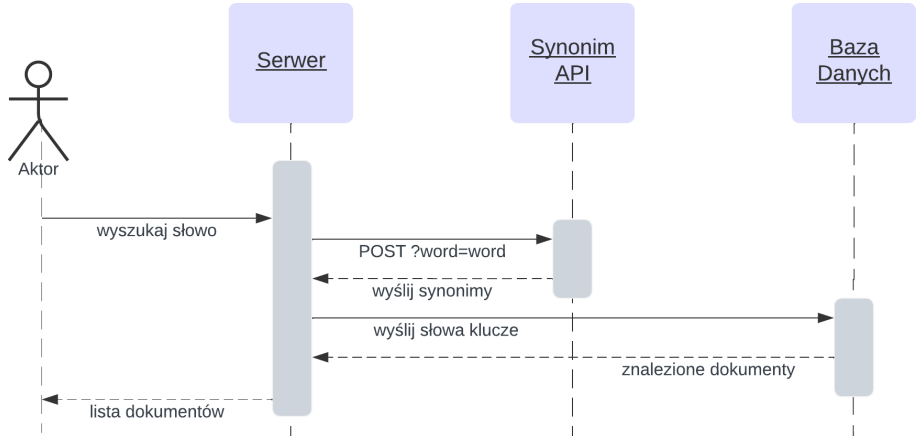


Rysunek 1: Diagram aktywności przesyłania pliku PDF

Wyszukiwanie i wyświetlanie plików

Użytkownik ma możliwość wysukiwania 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,
3. serwer przesyła osobne komunikaty do tabeli w bazie danych, zawierające osobno słowo klucz oraz synonimy
4. baza zwraca informacje dotyczące znalezionych dokumentów oraz słowo klucz,
5. serwer wysyła użytkownikowi posortowaną listę plików.



Rysunek 2: Schemat sekwencyjny wyszukiwania pliku

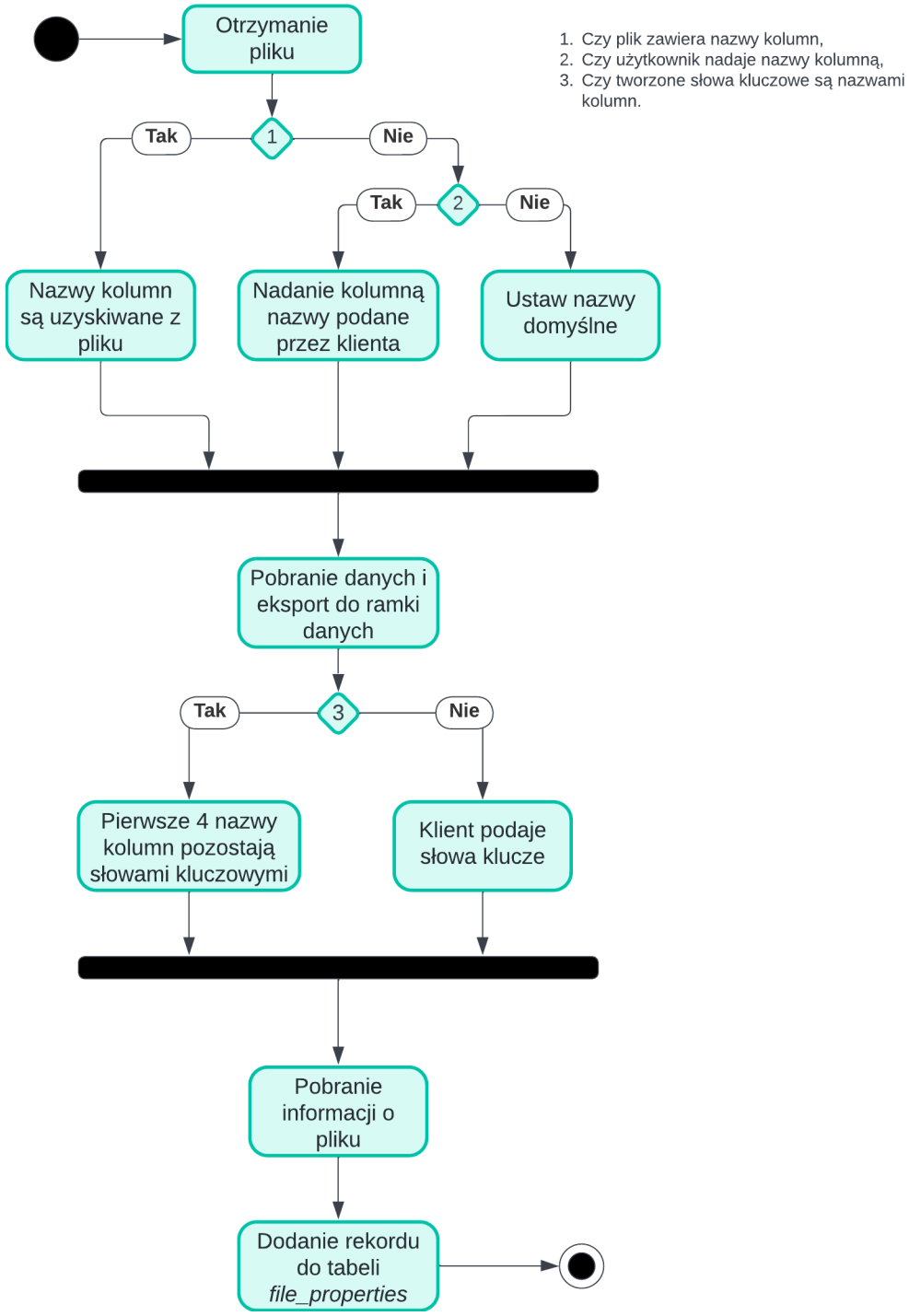
Obsługa plików zawierających dane tabelaryczne

1. użytkownik przy dodaniu pliku podaje następujące informacje:
   - czy plik zawiera informacje o nazwach kolumn, jeżeli nie to czy chce on nadać te nazwy,
   - czy słowa kluczowe będą nadane manualnie lub czy powinny być pobrane z nazw kolumn,

2. dane są odpowiednio wczytywane na serwer do postaci ramki danych,

3. sprawdzany i zapisywany jest typ wartości w każdej kolumnie oraz obliczony procent ilości wartości brakujących,

4. dodanie rekordu do bazy danych.



Rysunek 3: Diagram aktywności przetwarzania plików tabelarycznych

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 zawierającej informacje o zawartości plików, 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 kluczy

   - jeżeli wysłane słowo kluczowe zgadza się fragmentarycznie z otrzymanym tagiem (np. “zaburzenia” a “zaburzenia snu”), to zapisywana jest wartość $\frac{ilosc(dopasowan)}{ilosc\_slow(wejscie)+ilosc\_slow(pobrane)}$,
   - jeżeli słowa są identyczne, zapisywana jest wartość 1,

   (b) wartości dla wszystkich słów kluczy 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 informującej o podobieństwach dokumentów,

   - jeżeli wartość podobieństwa nie przekracza 0.5, to plik nie jest brany pod uwagę.

3 Implementacja

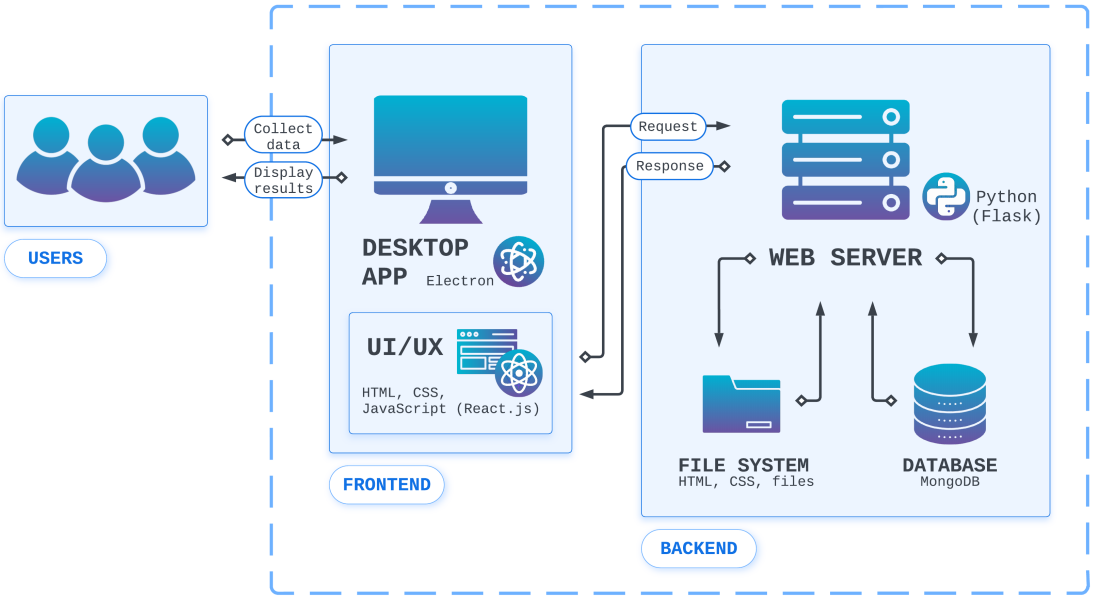
Realizacja miała na celu wykazanie możliwości projektu biorąc pod uwagę następujące aspekty. Po pierwsze, wykonalność aplikacji została oceniona w celu określenia jej praktyczności i potencjału pomyślnego rozwoju. Dodatkowo przeprowadzono analizę złożoności i wymagań dotyczących zasobów związanych z przetwarzaniem różnych typów plików, w tym artykułów, danych tabelarycznych, grafiki i notatek.

Co więcej, wysiłki ukierunkowano na identyfikację wymaganych rozszerzeń lub ulepszeń niezbędnych do optymalizacji funkcjonalności aplikacji, ze szczególnym uwzględnieniem algorytmów klasyfikacji czy modułu tworzenia abstraktów, opartych na sztucznej inteligencji. Wreszcie, podjęto planowanie w celu określenia zakresu wdrożenia w przydzielonych ramach czasowych, zapewniając zgodność z celami projektu i wykonalność w ramach ograniczeń dostępnych zasobów.

Do implementacji aplikacji wykorzystano kombinację technik typowych dla tworzenia aplikacji internetowych jakie zostały określone w rozdziale 1.4. Interfejs użytkownika (UI) został stworzony przy użyciu języka JavaScript, w szczególności wykorzystując bibliotekę React, wraz z językami opisu HTML i CSS do zdefiniowania struktury i stylizacji witryny. Jednak zamiast polegać na tradycyjnej przeglądarce internetowej, cały interfejs użytkownika został zamknięty w aplikacji natywej, dzięki utylizacji szkieletu aplikacji Electron. Opis tego procesu jest zamieszczony w rozdziale 1.4.

Do przechowywania wyodrębnionych danych wybrano nierelacyjną bazę danych, a jako system zarządzania bazą danych wybrano MongoDB. Decyzja ta była podyktowana zróżnicowanym i niejednolitym charakterem danych, gdzie na strukturę każdego rekordu wpływają takie czynniki jak typ pliku.

Aby ustanowić komunikację między aplikacją a bazą danych, zaimplementowano serwer WWW przy użyciu biblioteki Flask języka Python. Działając jako istotny łącznik, serwer ten zarządza wszystkimi podstawowymi funkcjami aplikacji. Nadzoruje on zadania takie jak przetwarzanie plików, ekstrakcja danych, klasyfikacja i segregacja, wraz z wszelkimi niezbędnymi operacjami manipulacji danymi. Takie podejście zapewnia płynną interakcję między komponentami aplikacji, umożliwiając wydajną obsługę i zarządzanie danymi w całym systemie.



Rysunek 4: Diagram łączenia komponentów

3.1 Opis wykorzystywanych elementów

Javascript to powszechnie używany język programowania do tworzenia stron internetowych, znany ze swojej wszechstronności i kompatybilności z przeglądarkami. Opisywany system wykorzystuje Javasctipt w środowisku *Node.js*, co ułatwia dostęp do różnorodnych pakietów takich jak *React*, których opis jest przedstawiony poniżej.

- *React* - biblioteka pozwalająca na budowanie interaktywnych interfejsów użytkownika. Główną koncepcją React są komponenty, czyli samodzielne, hermetyczne jednostku interfejsu.

- *Axios* - klient HTTP z pomocą którego odbywa się komunikacja z serwerem. Jest to mechanizm oparty o obietnice co pozwala na asynchroniczną relacje z punktem końcowym (*ang. endpoint*). W systemie, po otrzymaniu odpowiedzi od serwera, ustawiany jest stan, czego przykład wykorzystania jest przedstawiony poniżej.

Python to język programowania wysokiego poziomu znany ze swojej prostoty, czytelności i szerokiego wsparcia dla bibliotek, co czyni go odpowiednim wyborem do tworzenia backendu. Stworzenie serwera w tym języku pozwala na korzystanie z dużego zestawu bibliotek dotyczących takich tematów jak przetwarzanie i analiza

```
const [Message, setMessage] = useState();
const handleUpload = () => {
  const fileInput = document.getElementById("fileInput");
  const method = document.getElementById("method").value;
  const filePath = { path: fileInput.files[0]?.path, method: method };
  if (filePath.path) {
    axios
      .post(`${SERVER_URL}/upload`, filePath, {
        headers: { "Content-Type": "application/json" },
      })
      .then((response) => {
        setMessage(response.data.message);
      })
      .catch((err) => {
        console.error(err);
      });
  }
};
```

Rysunek 5: Przykład wykorzystania *Axios*; Implementacja dodania pliku

danych czy szerokiego spektrum bibliotek sztucznej inteligencji. Python ma dużą i aktywną społeczność programistów, którzy przyczyniają się do jego rozległego ekosystemu bibliotek, frameworków i innych narzędzi. Poniżej zostały przedstawione używane w aplikacji elementy.

- Flask - jest to elastyczny framework, pozwalający na 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.
- PdfMiner - biblioteka przeznaczona do wyodrębniania tekstu i metadanych z dokumentów PDF.
- Yake - biblioteka stworzona do identyfikowania i wyodrębniania kluczowych terminów, które reprezentują główne tematy lub wątki dokumentu.
- Sumy - biblioteka umozliwiająca automatyczne tworzenie streszczeń tekstów.

3.2 Architektura

Podstawowe funkcjonalności aplikacji są oparte na architekturze klient-serwer. Interfejs po stronie klienta, opracowany przy użyciu standardowych technologii internetowych przedstawionych w rozdziale 1.4, komunikuje się z zapleczem po stronie

serwera w celu obsługi żądań użytkowników oraz przechowywania i przetwarzania danych.

Nie jest ona jednak aplikacją internetową. Po stronie klienta jest ona zamknięta w *Electronie*, szkielecie (*ang. framework*) umożliwiającym tworzenie wieloplatformowych aplikacji desktopowych przy użyciu technologii internetowych. Osadzając aplikację internetową w Electron, przekształcana jest ona w samodzielną aplikację desktopową, która może być dystrybuowana i uruchamiana natywnie w różnych systemach operacyjnych (Windows, macOS, Linux). Ta integracja oferuje użytkownikom znane doświadczenie interakcji z aplikacją desktopową przy jednoczesnym zachowaniu korzyści płynących z technologii webowej.

3.3 Status realizacji

Ze względu na ograniczenia czasowe, implementacja obejmuje następujące elementy określone w rozdziale 2. Odchylenia od założeń przedstawionych dziale związanym z opisem projektu zostały opisane w rozdziale 3.4, a poniżej wydnieje lista częściowo lub w pełni zawartych elementów w zależności od modułu.

1. Moduł organizacji danych
    - Zrealizowano: Tagowanie plików poprzez generację słów kluczy na podstawie zawartości oraz łączenie plików
    - Odrzucono: Automatyczne wytwarzanie struktury folderów oraz kategoryzacja.
2. Wyszukiwarka
    - Zrealizowano: Wyszukiwanie, prezentacja wyników wyszukiwania oraz możliwość filtrowania wyników,
    - Odrzucono: Sugestie i autouzupełnianie
3. Moduł podsumowujący
    - Zrealizowano: Tworzenie streszczeń metodą ekstraktywną opartą o algorytm TextRank,
    - Odrzucono: Wykorzystanie metod abstrakcyjnych.

4. Przetwarzanie plików

    - Zrealizowano: Przetwarzanie cześciowe dla plików o rozszerzeniu PDF, CSV oraz TXT.

    Pełen opis specyfikacji zależnej od rodzaju pliku jest przedstawiony w rozdziale 3.5.

3.4 Ograniczenia implementacji

Organizacja plików

Podczas tworzenia systemu napotkano problem dotyczący modułu organizacji danych, a dokładniej automatycznego tworzenia struktur katalogów. Proces ten wymaga wzięcia pod uwagę wielu warunków co sprawia, że koszty implementacji są niewspółmierne do oczekiwanych wyników. Przykładowe wymagania zostały wypisanie poniżej.

1. Występowanie niejednoznacznych lub nakładających się słów kluczowych w treści pliku może prowadzić do niepewności w kategoryzacji. Jeśli słowo kluczowe jest powiązane z wieloma kategoriami lub jeśli różne pliki zawierają podobne słowa kluczowe o różnych znaczeniach kontekstowych, może to spowodować błędną klasyfikację plików do nieprawidłowych katalogów.

2. Zmienność w użyciu słów kluczowych, takich jak synonimy, skróty lub różnice w pisowni, może stanowić wyzwanie w dokładnej identyfikacji i wyodrębnianiu odpowiednich słów kluczowych z zawartości pliku.

Pozostawienie okazała się optymalniejsza zważając na czas przeznaczony implementacji. Zważając na ograniczenia czasowe, została podjęta decyzja o ograniczeniu modułu organizacji danych do implementacji systemu generacji słów kluczowych z zawartości plików oraz łączenia plików ze względu na nie. Schemat działania asocjacji plików został przedstawiony w rozdziale 2.3.

Implementacja abstrakcyjnego tworzenia streszczeń

Abstrakcyjna generacja streszczeń jest procesem wymagającym posiadania zestawu treningowego złożonych z streszczeń i przypisanym im tekstów początkowych. Jest

to proces czasochłonny, lecz jakościowo przewyższa streszczenia metodami ekstraktywnymi.

Samodzielna kreacja takiego modelu nie jest optymalna pod względem ilości zasobów potrzebnych do modelowania a zasobów pamięciowych przeznaczonych na działanie całej aplikacji. Istnieje jednak rozwiązanie tego problemu za pomocą wykożystania *API*.

3.5 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 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. Słowo *outlines* jest zwrotem specyficznym dla plików PDF i są potrzebne do ustalenia tytułu danego pliku. Wynika to z faktu, że nie zawsze nazwa pliku koresponduje tytułowi zawartości.

System wykorzystuje potokowanie zawartości w celu ekstrakcji tekstu, obrazów oraz tabel. Oznacza to, że aplikacja nie "widzi" dokumentu, lecz pobiera informacje z kodu źródłowego pliku. Działanie aplikacji jest oparte o sam tekst. Na potrzeby aplikacji, takie fragmenty jak tabele czy grafiki są jedynie przeszkodą, ponieważ nie powinne one wpływać na wynik algorytmu streszczania. Część elementów tabelarycznych jest niepoprawnie podpisywana jako tekst, co może przyczynić się do generacji niezrozumiałych streszczeń.

Pliki tabelaryczne

Podstawowym zachowaniem aplikacji w obliczu plików zawierających dane tabelaryczne jest generacja opisu zgodnie z następującymi krokami. Dla każdej kolumny pobierana jest jej nazwa oraz rodzaj danych zawartych w niej. Ze względu na specyfikę 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 500 słów, lecz może ona być ustawiona manualnie przez klienta. Po przekroczeniu tego progu, system procesuje plik w sposób podobny do obsługi plików *PDF*.

4 Wyniki

4.1 Końcowy diagram aktywności

4.2 Możliwości rozwoju i wykorzystania aplikacji

Wspomniane funkcjonalności to jedynie początek istnienia tego systemu. Stworzenie dodatkowych zdolności programu jest uproszczone, poprzez oparcie systemu o odpowiedzi w postaci komunikatu http zamiast stosowania szablonów, jak to jest zwykle przy standardowym projektowaniu z użyciem Python Flask. Baza danych wykorzystana w projekcie, MongoDB, jest bazą typu NoSQL, która umożliwia przechowywanie informacji w nie ustrukturyzowany sposób. Dzięki czemu tabele można wypełnić dowolnymi wartościami a sama aplikacja wyświetla wyszystkie elementy tabeli dla wyszukiwanego obiektu.

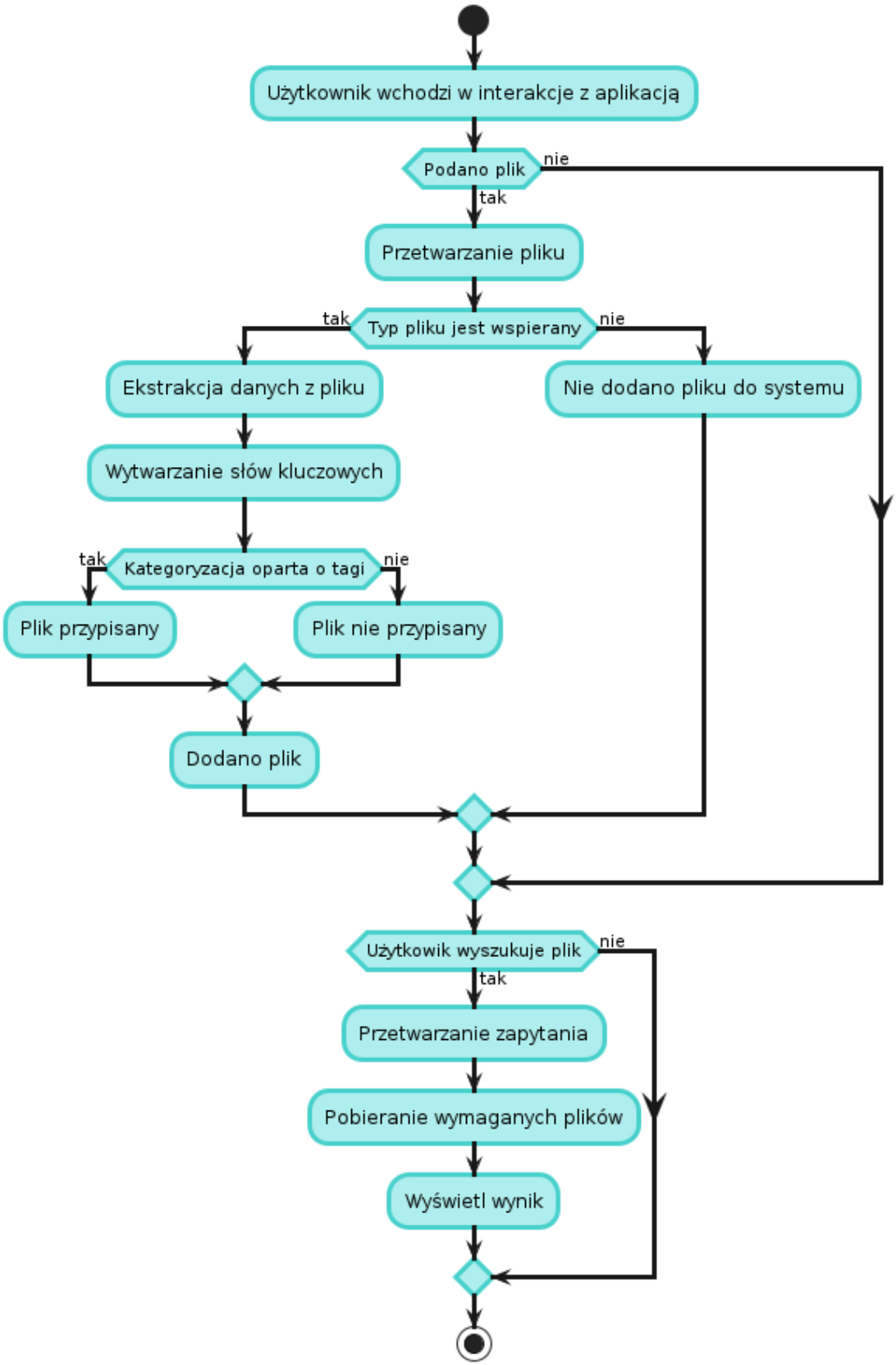
4.2.1 Ekspansja działań dotyczących danych tabelarycznych

Aktualny system pozwala na implementacje metod wstępnego przetwarzania danych (*ang. preprocessing*) z wnętrza aplikacji. Preprocessing oznacza zbiór działań miejących na celu obróbkę i przygotowanie danych do dalszej manipulacji. Do takich działań należą usunięcie wartości brakujących, filtracja czy usuniecie wartości odstajacych.

Celem stworzenia systemu jest automatyzacja procesów związanych z przechowywaniem informacji. Dodanie takich możliwości jak łączenie tabel w pojedyńcze pliki czy w przypadku posiadania przez plik osobnych arkuszy danych pozwolenie na rozdzielenie ich na osobne pliki, powinno być celem kolejnych implementacji.

4.2.2 Identyfikacja obrazów

Identyfikacja obrazów mogą przysłużyć się sukcesowi systemu. Dla człowieka proces identyfikacji elementów obrazu dzieje się podświadomie, a w przypadku programu, jest on w stanie rozpoznać i przypisać grafikę do odpowiedniej kategorii, jedynie w sytuacji gdy jest on zaprogramowany do wykrywania ich. Komputery kategoryzują obrazy, porównując układ pikseli wejściowego obrazu ze schematami zapisanymi w systemie. Z punktu widzenia maszyny, obraz jest ustrukturyzowaną macierzą liczb,



Rysunek 6: Diagram aktywności końcowej aplikacji

której każdy element zawiera informacje o nasyceniu i kolorze piksela. Aby stworzyć system odpowiedzialny za kategoryzację obrazów, należy przejść przez następujące kroki.

1. Gromadzenie danych,
2. Przygotowanie danych,
3. Wybór, trenowanie i testowanie modelu,
4. Wprowadzenie modelu do systemu.

Najważniejszym elementem skutecznego systemu rozpoznawania obrazów jest dobrze skonstruowany zbiór danych. Etap gromadzenia danych obejmuje pozyskiwanie różnorodnych obrazów reprezentujących obiekty, sceny lub wzorce, które system nauczy się rozpoznawać. Kompleksowy zbiór danych przyczynia się do zdolności systemu do uogólniania swojego zrozumienia w różnych scenariuszach.

Przygotowanie danych (*ang. Preprocessing*) to proces przekształcania i udoskonalania zebranych danych, aby były odpowiednie do przeprowadzenia treningu systemu. Może to obejmować takie zadania, jak usuwanie niespójności, zmiana rozmiaru obrazów do spójnego formatu i normalizacja wartości pikseli. Skuteczny preprocessing danych wspomaga zdolność modelu do nauki istotnych wzorców z danych.

Kluczową decyzją jest wybór architektury modelu. Wybór architektury powinien być dostosowany do wymagań zadania i zasobów obliczeniowych. Określa ona sposób, w jaki system będzie przetwarzał i interpretował dane wejściowe. Trening modelu to proces uczenia wybranej architektury rozpoznawania wzorców i cech w danych treningowych. Osiąga się to poprzez iteracyjną optymalizację, w której parametry modelu są dostosowywane w celu zminimalizowania różnicy między przewidywanymi a rzeczywistymi wynikami. Trening trwa do momentu, aż model osiągnie optymalny poziom dokładności i uogólnienia, który sprawi, że będzie on biegły w rozpoznawaniu obiektów na obrazach.

Ostatnim etapem, poprzedzającym implementacje modelu, jest testowanie wytrenowanego modelu. Jest ono niezbędne do oceny jego wydajności dla nowych danych. System jest zasilany obrazami, których nie napotkał podczas szkolenia, umożliwiając ocenę jego zdolności do generalizacji. Metryki takie jak dokładność,

precyzja i czułość zapewniają wgląd w mocne i słabe strony modelu. Rygorystyczne testowanie pomaga w dostrojeniu modelu w celu uzyskania lepszej wydajności.

Powszechnie dostępne, ogromne bazy danych, takie jak *Pascal VOC* czy *ImageNet*, pozwalają na stworzenie i wdrożenie takiego modelu. Zawierają one mnóstwo oznaczonych wzorców opisujących obiekty znajdujące sie na obrazach.

Podsumowanie

W ramach projektu zaprezentowano aplikację umożliwiającą komfortowe przechowywanie oraz zmniejszenie ingerencji użytkownika w proces interpretacji plików. Opisano funkcjonalności aplikacji oraz możliwe rozszerzenia aplikacji w dalszym ciągu rozwoju.

Decyzja o wykorzystaniu MongoDB, nierelacyjnej bazy danych, była kluczowa dla dostosowania się do zróżnicowanego i nieustrukturyzowanego charakteru plików, umożliwiając elastyczne i wydajne rozwiązania w zakresie przechowywania danych. Wybór ten podkreślił zdolność aplikacji do obsługi różnych typów plików i struktur danych, zwiększając tym samym jej użyteczność i zdolność adaptacji.

Badania nad aplikacją wyeksponowały ciekawy koncept jakim jest praca z pilkami o rozszerzeniu PDF, ponieważ proces automatycznego pobierania tekstu jest zależny od wielu czynników. Okazało się, że rozkład tekstu na stronie jak i program, którego użyto do stworzenia pliku może przyczynić się do błędnego eksportu tekstu. Nie jest możliwe uwzględnienie każdego przypadku stuktury pliku aby zapewnić prawidłowy proces ekstrakcji. Należałoby zastanowić się nad poprawą implementacji tego algorytmu lub zastosowanie innych sposobów, takich jak *Optical Character Recognition*.

Spis rysunków

Literatura

[1] D. Fagbuyiro, "On-premise vs cloud solutions: "choosing the best fit for your business"," Feb 2023. [Online]. Available: https://strapi.io/blog/on-premise-vs-cloud-solutions-which-is-the-best-for-your-business

[2] R. Ho, "On premise vs cloud: What's the difference," May 2022. [Online]. Available: https://blog.tessaract.io/on-premise-vs-cloud-whats-the-difference

[3] International Organization for Standardization, "ISO 32000:2008," Geneva, Switzerland, 2008, document management – Portable document format – Part 1: PDF 1.7.

[4] 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.

[5] 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

[6] 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

[7] 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

[8] 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

[9] 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