|  |
| --- |
| UNIWERSYTET WARMIŃSKO-MAZURSKI Wydział Matematyki i Informatyki |
| Dokumentacja |
| Praca inżynierska |

|  |
| --- |
| Sebastian Malon  6-4-2023 |

Wykorzystanie SumMedia do obróbki i modyfikacji tekstu

Stowrzenie paczki (package) zintegrowanego z API opublikowanej na pypi.

W dobie cyfrowej transformacji, gdzie ogromne ilości informacji są dostępne w zasięgu ręki, istnieje rosnące zapotrzebowanie na zaawansowane narzędzia do przetwarzania i analizy treści. Niniejsza praca inżynierska prezentuje projekt i implementację pakietu Pythona, "SumMedia", który stanowi krok naprzód w obszarze inteligentnego przetwarzania tekstu. Pakiet ten wykorzystuje zaawansowane algorytmy i techniki uczenia maszynowego, integrując możliwości dostarczane przez pakiet OpenAI oraz Newspaper3k, aby dostarczyć użytkownikom kompleksowe narzędzie do analizy i przetwarzania artykułów internetowych.

SumMedia został zaprojektowany z myślą o efektywności i wszechstronności. Umożliwia on nie tylko wyodrębnienie tekstu z artykułów, ale również jego skrócenie, co jest szczególnie przydatne w kontekście szybkiego przyswajania informacji. Funkcjonalność pakietu obejmuje również pobieranie tagów wraz z obrazkami, oszacowanie czasu potrzebnego na przeczytanie artykułu, a także możliwość transformacji tekstu artykułu w treść odpowiednią dla mediów społecznościowych, takich jak Twitter. Dodatkowo, SumMedia oferuje analizę sentymentu, co umożliwia głębsze zrozumienie tonu i charakteru przetwarzanych treści.

Kluczowym aspektem niniejszej pracy jest nie tylko prezentacja technicznej struktury i funkcjonalności pakietu SumMedia, ale także demonstracja jego praktycznego zastosowania w kontekście współczesnych wyzwań związanych z przetwarzaniem informacji cyfrowych. Praca ta koncentruje się na integracji nowoczesnych technologii przetwarzania języka naturalnego oraz analizy danych, stanowiąc tym samym wkład w rozwój narzędzi wspierających efektywne i skuteczne zarządzanie informacją w erze cyfrowej.

W kolejnych rozdziałach szczegółowo omówione zostaną poszczególne komponenty pakietu SumMedia, wraz z przykładami ich praktycznego wykorzystania. Przedstawione zostaną także studia przypadków, demonstrujące efektywność pakietu w różnorodnych scenariuszach zastosowań, co podkreśla jego wartość i uniwersalność w zakresie przetwarzania treści cyfrowych.

Paczka SumMedia umożliwia wykonywanie następujących operacji na tekstach artykułów:

- Wyodrębnienie tekstu z artykułu

- Skrócenie tekstu artykułu

- Pobranie wszystkich tagów z obrazkami

- Oszacowanie ile czasu zajmie czytanie artykułu

- Zmiana tekstu artykułu w tweeta

- Przeanalizowanie sentymentu artykułu

Opis wykorzystanych technologii:

Projekt SumMedia został zrealizowany przy użyciu stacku technologicznego, który obejmuje zarówno narzędzia do testowania, jak i do zarządzania kodem źródłowym oraz monitorowania go.

**GitHub**

GitHub pełni kluczową rolę w zarządzaniu kodem źródłowym projektu SumMedia. Jest to platforma, która nie tylko umożliwia przechowywanie kodu, ale także wspiera współpracę i przepływ pracy w projektach oprogramowania. Użycie GitHuba w projekcie SumMedia umożliwia śledzenie zmian, zarządzanie zadaniami oraz recenzję kodu, co jest nieocenione w procesie ciągłego rozwoju i utrzymania pakietu. Dzięki integracji z systemami CI/CD (Continuous Integration/Continuous Delivery), GitHub zapewnia również automatyzację procesów testowania i wdrażania, co znacznie przyspiesza cykl życia oprogramowania i podnosi jego jakość. GitHub jest nie tylko repozytorium kodu, ale stanowi także platformę do komunikacji i współpracy wewnątrz zespołu deweloperskiego, co jest szczególnie ważne w projektach open-source takich jak SumMedia

**Pytest**

Jednym z kluczowych elementów jest pytest – zaawansowane narzędzie do testowania w Pythonie. Pytest jest wykorzystywany do zapewnienia jakości i niezawodności pakietu SumMedia poprzez intensywne testowanie jednostkowe. Jest to szczególnie istotne w kontekście przetwarzania i analizy danych, gdzie dokładność i stabilność są kluczowe. Pytest oferuje prostotę w pisaniu testów, a jednocześnie umożliwia tworzenie zaawansowanych scenariuszy testowych, co jest niezbędne w przypadku testowania skomplikowanych funkcji przetwarzania tekstu. Dzięki pytest, projekt SumMedia zapewnia wysoką jakość i niezawodność, a także łatwość w utrzymaniu i rozwijaniu oprogramowania.

**CircleCI**

CircleCI to narzędzie Continuous Integration/Continuous Deployment (CI/CD), które odgrywa kluczową rolę w procesie tworzenia i wdrażania oprogramowania. W projekcie SumMedia , CircleCI jest używane do automatyzacji procesów testowania, budowania i wdrażania aplikacji. Dzięki temu narzędziu, każda zmiana kodu może być automatycznie testowana i weryfikowana, co znacząco zwiększa szybkość rozwoju i zapewnia wysoką jakość ostatecznego produktu. CircleCI umożliwia również łatwe wdrożenie kodu w różnych środowiskach, co jest niezbędne w cyklu życia nowoczesnego oprogramowania.

**Python**

Python, jako język programowania, stanowi podstawę projektu SumMedia. Jest to język szeroko znany ze swojej czytelności, wszechstronności i bogatego ekosystemu bibliotek. W projekcie SumMedia, Python jest używany ze względu na jego mocne wsparcie dla przetwarzania danych i uczenia maszynowego, a także dla łatwości, z jaką można integrować różne API i biblioteki zewnętrzne, takie jak wykorzystywane w tym projekcie OpenAI API i Newspaper3k.

**OpenAI API**

OpenAI API dostarcza zaawansowane algorytmy sztucznej inteligencji i uczenia maszynowego, które są wykorzystywane w SumMedia do analizy i przetwarzania treści tekstowych. Dzięki integracji z OpenAI API, SumMedia zyskuje możliwość wykonywania zaawansowanych operacji na tekście, takich jak generowanie streszczeń, analiza sentymentu i przekształcanie treści artykułów w tweety(X). Wykorzystanie tej technologii znacznie podnosi wartość i efektywność narzędzia, umożliwiając obsługę złożonych zadań związanych z przetwarzaniem języka naturalnego używając do tego LLM(Large language model)

**Newspaper3k**

Newspaper3k to biblioteka Pythona, używana w projekcie SumMedia do ekstrakcji i kuracji treści z artykułów internetowych. Umożliwia ona wyodrębnianie tekstu, obrazów oraz metadanych z różnych źródeł informacji online. W połączeniu z innymi komponentami projektu, Newspaper3k stanowi kluczowy element w procesie analizy i przetwarzania artykułów internetowych, dostarczając niezbędnych danych wejściowych dla dalszych operacji.