Politechnika Krakowska
im. Tadeusza Kościuszki

Informatyka
Wydział Informatyki i Telekomunikacji

Projekt z przedmiotu
Przetwarzanie Języka Naturalnego

Narzędzie do tworzenia podsumowań tekstu

Filip Rydzewski, Marcin Ścieszka, Szymon Śledź
semestr VII, gl03

Kraków, grudzień 2022

Wstęp

Brak czasu i chęć szybkiego poznania informacji to podstawowy problem współczesnego świata. Szukamy sposobów, aby zaspokoić naszą ciekawość bez wysiłku. Jednym z takich sposobów jest czytanie podsumowań zawierających tylko najważniejsze informacje zamiast obszernych artykułów, które są czasochłonne i uciążliwe do czytania.

Cel i zakres projektu

Celem projektu jest zaimplementowanie API, dzięki któremu użytkownik będzie w stanie stworzyć sensowne podsumowanie na podstawie dowolnego dokumentu tekstowego.

Istnieją dwa typy tworzenia podsumowań tekstu:

- Podsumowywanie abstrakcyjne - metody abstrakcyjne wybierają słowa w oparciu o zrozumienie semantyczne, nawet jeśli nie występuje ono w tekście. Ich celem jest stworzenie nowego tekstu na swój sposób. Sprawdzają i interpretują tekst używając zaawansowanych technik języka naturalnego, po to by wygenerować krótszy tekst zawierający kluczowe informację z oryginalnego.

Input document → understand context → semantics → create own summary.

- Podsumowywanie ekstraktywne - metody ekstraktywne tworzą podsumowanie wybierając zbiory najważniejszych słów, które zawierają najwięcej informacji.

Input document → sentences similarity → weight sentences → select sentences with higher rank.

Podsumowywanie abstrakcyjne jest zdecydowanie bardziej zaawansowaną techniką tworzenie podsumowań i wymaga głębszego zrozumienia tekstu w stosunku do metod ekstraktywnych. Z tego powodu zazwyczaj lepszym rozwiązaniem korzystanie jest z podsumowań ekstraktywnych, które są stabilniejsze.

Wykorzystane technologie

Do implementacji projektu został wykorzystany język programowania Python.
Wykorzystane frameworki:

- FastAPI - służący do budowania API z użyciem języka Python.
- NLTK - biblioteka służąca do symbolicznego i statystycznego przetwarzania języka naturalnego
- NumPy - biblioteka dodająca wsparcie do wielowymiarowych tablic i metryk wraz z dużą ilością wysokopoziomowych matematycznych funkcji
- NetworkX - biblioteka do tworzenia, manipulacji oraz uczenia struktur, dynamik i funkcji skomplikowanych sieci.

Dodatkowo do walidacji dokumentów JSON zostały wykorzystane moduły pydantic będące zgodne (complaint) ze standardem JSON Schema
Do wygenerowania dokumentacji został użyty Redoc

Proces tworzenia podsumowań tekstów

Do stworzenia podsumowań zostało wykorzystane podejście uczenia nienadzorowanego, którego zadaniem jest odkrywanie w zbiorze danych wzorców bez wcześniej istniejących etykiet i przy minimalnej ingerencji człowieka. Dodatkowo zakłada ono brak obecności oczekiwanego wyjścia w danych uczących. Do znalezienia podobieństw wśród zdań reprezentowanych przez wektory została wykorzystana odległość Jaccarda.

Przedstawienie procesu generowania podsumowań:

Input article → split into sentences → remove stop words → build a similarity matrix → generate rank based on matrix → pick top N sentences for summary.

API:

API udostępnia 2 endpointy:

- POST /api/summarize - streszcza podany w obiekcie json tekst
- GET /api/articles - zwraca ostatnie 5 streszczonych tekstów dla danego tokenu (token przesyłany w headerze)

1. POST /api/summarize

Request:

Header	
Authorization	string

Body	
text	string

Przykład 1:

Body:

{
"text": "For years, Facebook gave some of the world's largest technology companies more intrusive access to users' personal data than it has disclosed, effectively exempting those business partners from its usual privacy rules, according to internal records and interviews. The special arrangements are detailed in hundreds of pages of Facebook documents obtained by The New York Times. The records, generated in 2017 by the company's internal system for tracking partnerships, provide the most complete picture yet of the social network's data-sharing practices. They also underscore how personal data has become the most prized commodity of the digital age, traded on a vast scale by some of the most powerful companies in Silicon Valley and beyond. The exchange was intended to benefit everyone. Pushing for explosive growth, Facebook got more users, lifting its advertising revenue. Partner companies acquired features to make their products more attractive. Facebook users connected with friends across different devices and websites. But Facebook also assumed extraordinary power over the personal information of its 2 billion users - control it has wielded with little transparency or outside oversight.Facebook allowed Microsoft's Bing search engine to see the names of virtually all Facebook user's friends without consent, the records show, and gave Netflix and Spotify the ability to read Facebook users' private messages."
}

Response:

Body	
summary	List

Przykład 2:

Body:


```
{
   "summary": [
      "for years facebook gave some of the worlds largest technology companies more intrusive access to users personal data than it has disclosed effectively exempting those business partners from its usual privacy rules according to internal records and interviews",
      "facebook users connected with friends across different devices and websites"
   ]
}
```


2. GET /api/articles

Request:

Header	
Authorization	string

Body	

Response:

Body	
articles	List

Przykład:


```
{
   "articles": [
      [
         "for years facebook gave some of the worlds largest technology companies more intrusive access to users personal data than it has disclosed effectively exempting those business partners from its usual privacy rules according to internal records and interviews",
         "facebook users connected with friends across different devices and websites"
      ],
      [
         "for years facebook gave some of the worlds largest technology companies more intrusive access to users personal data than it has disclosed effectively exempting those business partners from its usual privacy rules according to internal records and interviews",
         "facebook users connected with friends across different devices and websites"
      ]
   ]
}
```


Baza danych

W bazie danych przechowujemy 5 ostatnich zapytań dla danego tokenu. Użyta baza danych to mongodb ze względu na format danych json, który używany jest w api.

Wnioski

Stworzone API pozwala w prosty sposób stworzyć podsumowanie dowolnego tekstu. Stworzone podsumowania są sensowne, nie zawierają zdań nie przechowywujących żadnych informacji.

Możliwości rozwoju

Głównym sposobem na rozwój aplikacji będzie dodanie warstwy wizualnej, która zdecydowanie ułatwi korzystanie z API. Dodatkowym pomysłem jest dodanie użytkowników, którzy mogliby przechowywać swoje podsumowania w bazie danych. Razem z dodaniem użytkowników należałoby zrobić ich odpowiednią autentykację oraz nałożyć limit na ilość żądań przez nich wysyłanych.

Źródła:

https://towardsdatascience.com/understand-text-summarization-and-create-your-own-summarizer-in-python-b26a9f09fc70
https://fastapi.tiangolo.com
https://docs.pydantic.dev/usage/validators/
https://www.mongodb.com/docs/
https://docs.docker.com/