POLITECHNIKA KRAKOWSKA

Narzędzie do generacji tytułów naukowych

Czaja Emilia

Godfryd Jakub

Krasowska Anna

1. Abstrakt

Repozytorium csinva/gpt-paper-title-generator to projekt, który jest związany z generowaniem tytułów do artykułów naukowych. Wykorzystano w nim GPT-3, czyli jeden z najnowszych modeli stworzony przez OpenAI. W repozytorium znajdują się dane, modele oraz kod źródłowy, co pozwala nam na generowanie tytułów naukowych z różnych dziedzin. W projekcie znajduje się również interaktywny notebook, który pozwala na wygenerowanie tytułów na podstawie zadanych tematów i autorów. W celu wygenerowania tytułu, model korzysta z pięciu ostatnich tytułów autorów, którzy napisali co najmniej trzy artykuły związane z uczeniem maszynowym (cs.ML, cs.LG, stat.ML) w serwisie arXiv. Zdarza się, że wyniki są niepoprawne dla danego autora, tzn. nie odpowiadają rzeczywistym zainteresowaniom danego autora i są związane z popularnymi tematami (np. uczenie głębokie) [1].

2. Wstęp

W dynamicznie rozwijającym się świecie nauki, badacze muszą śledzić coraz większą ilość publikacji naukowych. Utrudnia to utrzymanie się z niedawnymi osiągnięciami w dziedzinie nauki. Istnieje potrzeba stworzenia tytułów publikacji naukowych, które przyciągną uwagę i pomogą zrozumieć temat pracy i wyróżnić publikację od innych. Projekt z repozytorium csinva/gpt-paper-title-generator oparty na architekturze GPT, wykorzystując sztuczną inteligencję pozwala na generowanie tytułów artykułów naukowych bazując na słowach kluczowych. Celem narzędzia jest pomoc naukowcom odkrywania istotnych elementów prac naukowych oraz inspirowania się do tworzenia własnych publikacji.

Istnieje wiele stron internetowych, które oferują generowanie tytułów do różnych tekstów, np. do artykułów naukowych, blogów lub opowiadań. Niektóre ze stron oferują tę usługę za darmo, a inne wymagają opłaty za pełne funkcjonalności. Kilka przykładów stron, na których można wygenerować tytuł do tekstu: Portent's Content Idea Generator, Title Generator by Tweak Your Biz, Content Row Link Bait Title Generator oraz SEOPressor Blog Title Generator. Strony, na których można wygenerować tytuł zazwyczaj korzystają z algorytmów sztucznej inteligencji. Tytuły są tworzone na podstawie zadanych słów kluczowych lub po danych tematach. Po wprowadzeniu słów kluczowych lub tematu tekstu, wybrana strona analizuje go i generuje tytuł, który jest podobny już do istniejących tytułów, lub zostaje utworzony nowy tytuł, który opisuje wskazany temat [2].

3. Rozwinięcie

Kod został zmodyfikowany w porównaniu do oryginalnego kodu z repozytorium csinva/gpt-paper-title-generator, by zapewnić jego lepszą czytelność i łatwą możliwość modyfikacji parametrów czy nawet rozwój. Z tego względu zmieniano nazwę zmiennych, parametrów, wydzielano kawałki kodu do odseparowanych metod czy dodawano nowe funkcjonalności pozwalające na ponowne użycie niektórych akcji.

Kluczowe fragmenty kodu:

1. compute_author_stats

Ma na celu obliczenie statystyk dotyczących autorów na podstawie metadanych artykułów naukowych pobranych z bazy danych arXiv. Funkcja przyjmuje jako argument ramkę danych (**dataframe_from_arxiv_metadata**), która zawiera metadane artykułów z arXiv. Funkcja zwraca krotkę, która zawiera słownik z informacją o autorach i indeksach ich artykułów (**authors_dict**) oraz posortowaną listę autorów (**pruned_authors_list**). Wewnątrz funkcji zdefiniowano funkcję pomocniczą **get_authors_list**, która przyjmuje ciąg znaków (**s**) i zwraca listę autorów. Lista autorów jest tworzona poprzez rozdzielenie ciągu znaków na podstawie wyrażeń takich jak "and", "," lub nowa linia. Następnie każdy autor zostaje zamieniony na małe litery i oczyszczony z białych znaków.

Funkcja **compute_author_stats** następnie stosuje funkcję pomocniczą **get_authors_list** do kolumny authors w ramce danych, tworząc tym samym listę list autorów (**authors_list_arr**). Kolejnym krokiem jest utworzenie słownika **authors_dict** przy użyciu **defaultdict(list)**. W pętli for, funkcja iteruje po indeksie i liście autorów, a następnie dodaje indeks artykułu do listy indeksów dla każdego autora. Po utworzeniu słownika autorów, funkcja tworzy oczyszczoną i posortowaną listę autorów (**pruned_authors_list**). Lista ta jest tworzona poprzez łączenie wszystkich list autorów w jedną listę za pomocą **itertools.chain.from_iterable**, a następnie pozbycie się duplikatów przez zastosowanie set. Na końcu lista jest sortowana alfabetycznie.

Funkcja **compute_author_stats** zwraca krotkę zawierającą słownik z informacją o autorach i indeksach ich artykułów (**authors_dict**) oraz posortowaną listę autorów (**pruned_authors_list**).

2. sanitize_author_collections

Ma na celu oczyszczenie danych autorów, usuwając niepoprawne dane oraz informacje, które nie są związane z autorami. Funkcja ta przyjmuje dwa argumenty: **authors_dict**, który jest słownikiem autorów oraz ich ostatnich pięciu tytułów, i **authors_list**, który jest listą autorów. Zadaniem funkcji jest przetworzenie i oczyszczenie obu struktur danych, aby zawierały jedynie prawidłowe informacje o autorach. Funkcja zwraca krotkę zawierającą oczyszczone dane słownika (**authors_list_clean**) i listy autorów (**authors_list_clean**). Wewnątrz funkcji zdefiniowano kilka funkcji pomocniczych, które są używane do sprawdzania, czy dany autor spełnia określone kryteria:

author_has_more_articles_than - sprawdza, czy autor ma więcej artykułów niż określona liczba (w tym przypadku 3).
does_not_contain_punctuaction - sprawdza, czy nazwa autora nie zawiera żadnych znaków interpunkcyjnych.
author_has_name_and_surname - sprawdza, czy autor ma zarówno imię, jak i nazwisko.
author_does_not_contain_keywords - sprawdza, czy nazwa autora nie zawiera żadnych słów kluczowych, takich jak "universi", "departm", "faculty" itp.

Następnie, funkcja tworzy oczyszczoną listę autorów (**authors_list_clean**) poprzez iterowanie po oryginalnej liście autorów i sprawdzanie, czy dany autor spełnia wszystkie kryteria zdefiniowane przez funkcje pomocnicze. Jeśli autor spełnia te kryteria, zostaje dodany do oczyszczonej listy. Po utworzeniu oczyszczonej listy autorów, funkcja tworzy oczyszczony słownik autorów (**authors_dict_clean**) poprzez iterowanie po oczyszczonej liście autorów i dodawanie ich informacji (ostatnich pięciu tytułów) do nowego słownika. Na koniec, funkcja sprawdza, czy liczba autorów na oczyszczonej liście (**authors_list_clean**) jest równa liczbie autorów w oczyszczonym słowniku (**authors_dict_clean**). Jeśli tak, funkcja zwraca krotkę z oczyszczonymi danymi słownika i listy autorów.

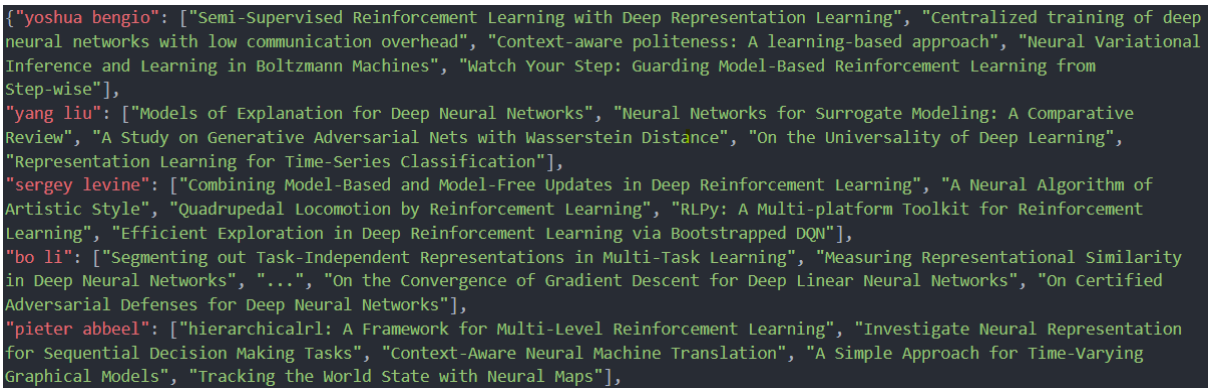
3. generate_authors_titles

Funkcja **generate_authors_titles** jest odpowiedzialna za generowanie tytułów artykułów naukowych dla autorów na podstawie wcześniejszych tytułów autorów, wykorzystując OpenAI API. Jako parametry przyjmuje słownik z informacjami o autorach i ich wcześniejszych tytułach artykułów, funkcję filtrującą autorów na podstawie określonego kryterium, opcjonalnie maksymalną liczbę autorów, dla których zostaną wygenerowane tytuły, oraz opcjonalnie ścieżkę, do której zostaną zapisane wyniki. Na początku funkcja weryfikuje, czy przekazano funkcję filtrującą autorów. Następnie tworzy pusty słownik, który będzie przechowywał wyniki generowania tytułów, oraz listę autorów na podstawie kluczy w przekazanym słowniku. Kolejno filtruje autorów za pomocą funkcji filtrującej, po czym inicjalizuje OpenAI API za pomocą przekazanych ustawień.

Funkcja iteruje przez przefiltrowanych autorów, generując tytuły artykułów za pomocą OpenAI API. Dla każdego autora sprawdza, czy przekroczono limit iteracji (jeśli został ustawiony), a następnie tworzy odpowiedni prompt na podstawie wcześniejszych tytułów autora, korzystając z funkcji **create_prompt**. W dalszej kolejności funkcja tworzy uzupełnienie za pomocą OpenAI API, wykorzystując funkcję **create_openai_completion**, i zapisuje wygenerowane tytuły w słowniku przechowującym wyniki. Ograniczenie liczby iteracji zostało wprowadzone, by ograniczyć koszta korzystania z OpenAI API.

Po przetworzeniu wszystkich autorów, jeśli została podana ścieżka do zapisu wyników, funkcja zapisuje wyniki do pliku. Na koniec funkcja zwraca słownik z wygenerowanymi tytułami artykułów dla każdego autora. W procesie generowania tytułów funkcja korzysta z różnych pomocniczych funkcji, takich jak **configure_openai_api**, **create_openai_completion**, **create_prompt**, **filter_authors** i **save_results**. Te funkcje służą do konfiguracji API, tworzenia promptów, filtrowania autorów i zapisywania wyników, co pozwala na elastyczne i efektywne generowanie tytułów artykułów naukowych na podstawie wcześniejszych publikacji autorów.

Na poniższym rysunku przedstawiono przykładowy output programu, gdzie znajdują się imię i nazwisko autora oraz pięć wygenerowanych tytułów dla niego.



Rys.1. Zapisane tytuły artykułów naukowych, które zostały wygenerowane dla danego autora

4. Podsumowanie

Projekt csinva/gpt-paper-title-generator wykorzystuje architekturę GPT-3 w celu generowania tytułów artykułów naukowych na podstawie wcześniejszych tytułów autorów. Projekt ma na celu pomóc naukowcom w szybszym śledzeniu i zrozumieniu najnowszych publikacji naukowych oraz inspiracji do tworzenia własnych publikacji. Projekt został zrealizowany i dokonano refaktoryzacji kodu. Funkcje zostały zaprojektowane tak, aby zapewnić efektywność generowania tytułów.

Jednym z ograniczeń projektu jest możliwość generowania niepoprawnych tytułów dla niektórych autorów, co wynika z ograniczeń modelu GPT-3 i istniejących danych. W trakcie analizy wyników zauważono, że choć w większości przypadków generowane tytuły są odpowiednie dla danego autora, czasami mogą być niepoprawne lub związane z popularnymi tematami, takimi jak uczenie głębokie, co nie zawsze odpowiada rzeczywistym zainteresowaniom danego autora. W celu dalszego rozwoju projektu, warto by rozważyć wprowadzenie dodatkowych kryteriów filtrowania autorów oraz bardziej zaawansowanych metod oczyszczania danych, aby zapewnić lepszą jakość wyników. Można również rozważyć wykorzystanie innych modeli językowych, takich jak GPT-4, aby dodatkowo poprawić jakość generowanych tytułów.

Podsumowując, projekt csinva/gpt-paper-title-generator jest skutecznym narzędziem dla naukowców, którzy chcą generować inspirujące tytuły artykułów naukowych na podstawie wcześniejszych publikacji. Projekt wykorzystuje potencjał sztucznej inteligencji i może być dalej rozwijany, aby zwiększyć skuteczność i wydajność generowania tytułów.

Bibliografia

[1] https://github.com/csinva/gpt-paper-title-generator
[2] https://seopressor.com/blog-title-generator/