https://jacek.kwasniewski.org.pl

Jacek Kwaśniewski

Lęk przed śmiercią jako siła kulturotwórcza: pytania badawcze i analiza wspierana przez AI

Budowa agenta AI do analizy literatury (OCR → segmentacja → embeddingi/FAISS → synteza)

2025

Jacek Kwaśniewski

Lęk przed śmiercią jako siła kulturotwórcza: pytania badawcze i analiza wspierana przez AI

Budowa agenta AI do analizy literatury (OCR → segmentacja → embeddingi/FAISS → synteza)

Literatura o lęku przed śmiercią jako zjawisku kulturotwórczym jest bardzo obszerna. Pewne wyobrażenie daje, napisany przez AI, pod moim kierunkiem, tekst ***Lęk przed śmiercią jako siła kulturotwórcza: Wpływ na religię, psychikę, kulturę i cywilizację. Debaty XX wieku*** (2025, https://jacek.kwasniewski.org.pl)

Współpraca z AI nad zagadnieniem lęku przed śmiercią na tym się nie kończy. Opracowałem z AI listę zagadnień/pytań, co konkretnie oznacza, że lęk przed śmiercią jest siłą kulturotwórczą. Zidentyfikowałem, pięćdziesiąt siedem pytań / problemów badawczych zebranych w siedemnaście grup. Przykładowo, lęk przed śmiercią rodzi w perspektywie historycznej coraz to nowe idee, symbole i mity, wytwarza lub modyfikuje rytuały, inspiruje projekty technologiczne i cywilizacyjne.

Chcę zobaczyć, jak rozległa literatura zajmująca się tematem śmierci i lęku przed śmiercią zmierzyła się z tymi pytaniami i problemami. Chcę zobaczyć znalezione przez AI fragmenty tekstów, które traktują dokładnie o sprawach postawionych przeze mnie w kolejnych pytaniach. A oto, jak AI będzie w tym pomocna.

Pytania, jakie stawiam oraz wybrana literatura na ten temat – kilka, kilkanaście albo i kilkadziesiąt tysięcy stron, zostaną umieszczone w wielowymiarowej przestrzeni semantycznej w celu wykrycia fragmentów znaczeniowo najbliższych postawionym pytaniom. Pytania zostały uszczegółowione do postaci lepiej zrozumiałej dla AI dokonującej wyszukiwania. Z pięćdziesięciu siedmiu problemów powstało więc przeszło trzysta szczegółowych pytań. Będę przy tym decydował o parametrach wyszukiwania: liczbie fragmentów literatury (np. 5, 10, 20, 30) przypisanych do każdego pytania, długości tych fragmentów (np. 100, 200, 300 słów) oraz o minimalnym progu podobieństwa semantycznego między pytaniem a przypisanym mu fragmencie literatury.

Celem jest szybkie wsparcie pracy analitycznej. Duże zbiory tekstów mogą zostać w krótkim czasie (praca rzędu kilku, kilkunastu godzin) przyporządkowane do pytań; pod każdym pytaniem wyświetlone będą fragmenty znaczeniowo najbliższe (z danymi bibliograficznymi). Dalsza praca — syntezy, porównania, analizy — będzie prowadzona przeze mnie samodzielnie lub z pomocą innych narzędzi AI.

Do wykonania tych wszystkich zadań zostanie wykorzystany agent AI, który jest przeze mnie aktualnie budowany. Agent AI to złożony system programistyczny, który łączy narzędzia sztucznej inteligencji w sekwencję kroków. Oto one:

- **Konwersja źródeł do tekstu** — zamiana prac w formacie pdf na format txt oraz formaty, z którymi pracuje AI (csv, json);
- **Ekstrakcja treści głównej** — usunięcie elementów niebędących treścią zasadniczą: nagłówków, stopek, przypisów dolnych, list bibliograficznych, stron tytułowych, stron redakcyjnych;
- **Segmentacja** – podział treści głównej na mniejsze fragmenty (chunki) o kontrolowanej długości, którym w kolejnym kroku zostaną przypisane współrzędne (czyli adresy, miejsca) w przestrzeni semantycznej. Ten etap przygotowuje materiał do przetwarzania numerycznego
- **Proces embeddingu** — wysłanie chunków do modelu embeddingowego. Jest to program AI, który przypisuje chunkom współrzędne (miejsca, adresy) w wielowymiarowej przestrzeni semantycznej. Współrzędne chunka lokalizują jego znaczenie / sens dzięki powiązaniom z miliardami innych zestawów słów w tej przestrzeni. Te współrzędne nazywamy embeddingiem danego chunka. Mówiąc nieco bardziej technicznie, embedding to cyfrowa rekonstrukcja znaczenia (sensu) chunka, obliczana przez model na podstawie ogromnych zbiorów tekstów, na których został wytrenowany. Dzięki temu fragmenty, które używają odmiennych słów, ale wyrażają podobne sensy, otrzymują współrzędne położone blisko siebie;
- **Model embeddingowy** — wykorzystuję model ‘text—embedding—3—small’ firmy OpenAI, operujący w przestrzeni 1536-wymiarowej i nadający chunkom adresy (embeddingi) złożone z 1536 liczb;
- **Embedding pytań** — w ten sam sposób zostaną wyznaczone embeddingi pytań, które zamierzam postawić wybranej literaturze;
- **Sparowanie pytań z tekstami** — połączenie pytań z najbliższymi im znaczeniowo chunkami. To zadanie (wektoryzowane wyszukiwanie) realizuje kolejny program AI, tzw. silnik wyszukiwania wektorowego, np. FAISS;
- **Obliczanie podobieństwa pytań z tekstem** — Sparowanie następuje w następujący sposób: każde pytanie—embedding i każdy chunk—embedding to punkty w tej samej przestrzeni (w naszym przypadku 1536-wymiarowej) i zarazem wektory prowadzące od początku układu współrzędnych do danego punktu. Stopień podobieństwa sensu pytania i sensu chunka liczony jest kątem między ich wektorami w przestrzeni semantycznej. Podobieństwo między pytaniem a fragmentem tekstu model oblicza geometrycznie jako miarę kąta między ich wektorami. Im mniejszy kąt, tym większe podobieństwo semantyczne.
- **Efekt końcowy** — to zestaw pytań z fragmentami tekstów, które są najbliższe co do znaczenia i sensu postawionym pytaniom. Z tym materiałem można, wspólnie z AI,

zrobić wiele rzeczy: (1) uporządkować go (usunąć powtórzenia, pogrupować według wątków tematycznych, tradycji intelektualnej itp.); (2) dokonać syntez i porównań (wskazać kontrasty, luki); (3) zbudować mapę tematów (sieci zależności między pytaniami); (4) przeprowadzić krytyczną analizę dyskursu (przesunięcia znaczeń, zmiany chronologiczne) a także zaproponować dalsze pytania badawcze. W opracowaniu „surowych" materiałów dostarczonych przez silnik wektorowy będą pomocne liczne narzędzia, które on oferuje oraz duży model językowy (LLM) ChatGPT-5 Thinking i ChatGPT-5 Pro.

- **Kontrola jakości —** Pomiędzy etapami działają moduły kontrolne, które sprawdzają, czy poprzedni etap został wykonany prawidłowo i dokonują autopoprawek albo wstrzymują cały proces do decyzji użytkownika (wybór jednej z opcji dalszego działania).

Budowa agenta AI jest żmudna, bo dostępne, gotowe moduły (m.in. modele embeddingowe silnik wyszukiwania wektorowego FAISS) przeplatane są w architekcie mojego agenta AI modułami, które muszą być zbudowane od zera. Aktualnie opracowuję w ten sposób heurystykę plus system algorytmiczny do usuwania przypisów dolnych, nagłówków i stopek w tekstach naukowych. Takie gotowe i dobrze działające rozwiązanie „z półki" (off—the—shelf) obecnie nie istnieje (reklamowany GROBID w moich testach miał skuteczność nie przekraczającą 36%, tylko w 4 tekstach na 11 usunął przypisy). Kolejnym modułem, który trzeba stworzyć od początku będzie zamiana tekstów na chunki. Widzę też potrzebę bliższego poznania modelu wektorowego. Po pierwszym sparowaniu pytań z tekstami można go wykorzystać do znacznie bardziej wyrafinowanych operacji.

wrzesień 2025