https://jacek.kwasniewski.org.pl

Jacek Kwaśniewski

Il timore della morte come forza creatrice di cultura: domande di ricerca e analisi supportata dall'IA

Costruzione di un agente AI per l'analisi della letteratura (OCR → segmentazione → embedding/FAISS → sintesi)

2025

Jacek Kwaśniewski

Il timore della morte come forza creatrice di cultura: domande di ricerca e analisi supportata dall'IA

Costruzione di un agente AI per l'analisi della letteratura

(OCR → segmentazione → embedding/FAISS → sintesi)

La letteratura sul timore della morte come fenomeno fortemente culturogenico è vastissima. Un'idea introduttiva si può trovare in un testo scritto dall'IA sotto la mia supervisione, intitolato Il timore della morte come forza creatrice di cultura: Influenza sulla religione, la psiche, la cultura e la civiltà. Dibattiti del XX secolo (2025, https://jacek.kwasniewski.org.pl).

La collaborazione con l'IA su questo tema non si esaurisce qui. Nelle pagine seguenti presento un elenco di questioni che chiariscono cosa significhi concretamente che il timore della morte sia una forza creatrice di cultura. Insieme all'IA ho identificato cinquantasette problemi di ricerca dettagliati, raccolti in diciassette gruppi. Ad esempio, in prospettiva storica il timore della morte genera costantemente nuove idee, simboli e miti, produce o modifica rituali, ispira progetti tecnologici e civili.

Queste domande intendo porle a un ampio corpus di testi che trattano della morte e del timore della morte — e ricevere risposte sotto forma di frammenti individuati dall'IA, che trattano esattamente le questioni poste. Ecco in che modo l'IA risulterà utile in questo lavoro.

Le domande, insieme alla letteratura selezionata — nell'ordine di decine di migliaia di pagine — verranno collocate in uno spazio semantico multidimensionale (vettoriale), per individuare i frammenti più vicini dal punto di vista del significato. Sarò io a decidere i parametri della ricerca: il numero di frammenti associati a ogni domanda (ad esempio 5, 10, 20, 30), la lunghezza di tali frammenti (100, 200, 300 parole) e la soglia minima di similarità semantica tra la domanda e il frammento assegnato.

L'obiettivo è fornire un supporto rapido al lavoro analitico. Ampi insiemi di testi possono essere in poche ore assegnati alle domande; sotto ciascuna domanda verranno mostrati i frammenti semanticamente più vicini (con i dati bibliografici). Il lavoro successivo — sintesi, confronti, analisi — sarà svolto da me, eventualmente con l'aiuto di altri strumenti di IA.

Per realizzare tutto questo sarà impiegato un agente AI, che sto attualmente costruendo. Un agente AI è un sistema software complesso, che collega diversi strumenti di intelligenza artificiale in una sequenza di fasi. Ecco le fasi:

• Conversione delle fonti in testo — trasformazione di documenti in formato PDF in TXT e in formati utilizzabili dall'IA (CSV, JSON).

• Estrazione del contenuto principale — eliminazione di elementi non essenziali: intestazioni, piè di pagina, note a piè di pagina, bibliografie, pagine di titolo e redazionali.

• Segmentazione — suddivisione del contenuto principale in frammenti più piccoli (chunks) di lunghezza controllata, cui nel passo successivo verranno attribuite coordinate (cioè "indirizzi") nello spazio semantico. Questo prepara il materiale all'elaborazione numerica.

• Embedding — invio dei chunks a un modello di embedding. È un programma AI che assegna a ciascun chunk delle coordinate in uno spazio semantico multidimensionale. Queste coordinate localizzano il significato del frammento grazie alle relazioni con miliardi di altri insiemi di parole in quello spazio. Le chiamiamo embedding del chunk. In termini più tecnici, l'embedding è una ricostruzione digitale del significato del chunk, calcolata dal modello sulla base di grandi corpora testuali. In questo modo, i frammenti che utilizzano parole differenti ma esprimono significati affini ricevono coordinate collocate vicino tra loro nello spazio semantico.

• Modello di embedding — utilizzo del modello text-embedding-3-small, che opera in uno spazio a 1536 dimensioni e assegna a ciascun chunk un embedding composto da 1536 numeri.

• Embedding delle domande — le stesse procedure verranno applicate alle domande di ricerca.

• Accoppiamento domande-testi — collegamento delle domande con i chunks più vicini dal punto di vista semantico. Questa operazione di "ricerca vettoriale" è realizzata da un motore di ricerca vettoriale, ad esempio FAISS (Facebook AI Similarity Search).

• Calcolo della similarità — ogni domanda (embedding) e ogni chunk (embedding) sono punti nello stesso spazio (1536 dimensioni) e al tempo stesso vettori tracciati dall'origine del sistema di coordinate fino al punto corrispondente. Il grado di somiglianza semantica è calcolato in base all'angolo tra i vettori: più piccolo è l'angolo, maggiore è la similarità.

• Risultato finale — un insieme di domande con i frammenti di testo più vicini per significato e senso. Con questo materiale sarà possibile, insieme all'IA: ordinarlo (eliminare duplicati, raggruppare per temi, tradizioni intellettuali, ecc.); realizzare sintesi e confronti; costruire una mappa dei temi; condurre un'analisi critica del discorso; formulare nuove domande di ricerca.

• Controllo di qualità — tra le fasi operano moduli di controllo che verificano se la fase precedente è stata eseguita correttamente, introducono correzioni automatiche oppure sospendono l'intero processo in attesa della decisione dell'utente (scelta dell'opzione successiva).

La costruzione di un agente AI è laboriosa, perché i moduli già disponibili (ad esempio i modelli di embedding e il motore di ricerca vettoriale FAISS) devono essere integrati con componenti da sviluppare ex novo. Attualmente sto elaborando un sistema euristico e algoritmico per eliminare note a piè di pagina, intestazioni e piè di pagina nei testi scientifici. Non esiste un software pronto e affidabile off-the-shelf (già pronto all'uso); nelle mie prove, ad esempio, il molto pubblicizzato GROBID non ha superato il 36% di efficacia, riuscendo a eliminare le note solo in 4 testi su 11. Un altro modulo da sviluppare sarà la suddivisione dei testi in chunks. Vedo inoltre la necessità di approfondire la conoscenza del modello vettoriale: dopo il primo accoppiamento delle domande con i testi, esso può essere sfruttato per operazioni molto più sofisticate.

settembre 2025