

• POD MASKĄ

Jak SumizAI nie pozwala bibliotece zappełnić się prawie-duplikatami?

Zapytaj model o jakość retrievalu cztery razy w ciągu trzech miesięcy, a dostaniesz cztery notatki o jakości retrievalu — każda odrobinę lepsza od poprzedniej, żadna kompletna, wszystkie znajdowalne. **To jest tryb awarii, któremu SumizAI ma zapobiegać**, i kosztuje jeden dodatkowy krok przed każdym zapisem.

Aktualizacja: 2026-08-14 8 pytań 3 min czytania

Q. Co biegnie, zanim nowa notatka zostanie zapisana?

Dwuetapowe sprawdzenie: najpierw tanie, potem kosztowne.

Etap pierwszy to wyszukiwanie: pełnotekstowe i trigramowe w obrębie vaulta dają krótką listę kandydatów. To czysta praca bazy danych i zajmuje milisekundy nawet przy tysiącach notatek.

Etap drugi to ocena: model pomocniczy czyta nowy materiał wobec kandydata i zwraca **ocenę od 0 do 1 z pisemnym uzasadnieniem**. Ten podział ma znaczenie — proszenie modelu o porównanie z każdą notatką byłoby nie do udźwignięcia, a proszenie bazy danych o zrozumienie, że dwa akapity mówią to samo, po prostu nie działa.

Q. Dlaczego dwie różne metody wyszukiwania kandydatów?

Bo zawodzą w przeciwnych kierunkach i nawzajem się kryją.

Wyszukiwanie pełnotekstowe jest precyzyjne co do słów i ślepe na te, których nie wpisałeś. Podobieństwo trigramowe dopasowuje wzorce znaków, więc łapie literówki, formy odmienione i bliskie chybienia, które dokładne dopasowanie gubi.

Jeden szczegół istotny dla języków fleksyjnych: tekst jest normalizowany na potrzeby indeksu — ogonki zdejmowane — a zapytanie używa dokładnie tej samej normalizacji. Pisanie po polsku bez ogonków jest regułą, a nie literówką, a gdyby obie strony się rozjechały, baza po cichu przestałaby używać indeksu.

Q. Co właściwie znaczy ta ocena?

Jest osądem, czy te dwie rzeczy są *tym samym tematem*, a nie czy dzielą słownictwo. To jest zadanie, którego nie da się oddać bazie danych.

Pisemne uzasadnienie jest równie użyteczne jak liczba. „Obie opisują kompromis przy dzieleniu tekstu, nowa dokłada liczby kosztowe” mówi, co zrobiłoby scalenie; „obie wspominają dzielenie tekstu, ale jedna jest o narzędziach” mówi, dlaczego nie powinno do niego dojść.

Ocena jest porównywana z progiem vaulta. Powyżej progu o tym, co dzieje się dalej, decyduje tryb.

Q. Jakie są trzy tryby i kiedy który jest właściwy?

Pytaj jest domyślny i pokazuje kandydata razem z oceną i uzasadnieniem. Decyzję można odłożyć — nie blokuje dalszej pracy. Używaj go, dopóki uczysz się, co aplikacja uznaje za ten sam temat.

Zawsze scalaj automatycznie włącza nowy materiał. Pasuje do vaulta, który jest długim studium kilku pojęć i w którym chcesz garść głębokich notatek zamiast setki płytkich.

Zawsze twórz trzyma wszystko osobno. Pasuje do vaulta-dziennika, w którym dwa wpisy o tym samym w różnych dniach to naprawdę dwa wpisy.

Próg obowiązuje w każdym trybie, bo jest definicją „tego samego tematu”, na której tryb potem działa.

Q. Co dokładnie robi scalanie?

Dopisuje. Nowy materiał dołącza do istniejącej notatki, a nowe pytanie i odpowiedź trafiają do jej materiału źródłowego obok tych, które już tam były.

Nigdy nie edytuje ani nie usuwa tego, co było. To gwarancja strukturalna, a nie obietnica: serwis i repozytorium w ogóle nie mają operacji aktualizacji ani usuwania materiału źródłowego, więc żadna ścieżka kodu nie zrobi tego przypadkiem.

Efektom jest notatka, która rośnie wgłąb i zachowuje kwity. Zawsze możesz odczytać, co każda z rozmów faktycznie powiedziała, jej własnymi słowami, zamiast brać syntezę na wiarę.

Q. Dlaczego zasada „tylko dopisywanie” zasługuje na tyle uwagi?

Bo scalona notatka to twierdzenie złożone z kilku rozmów, a złożone twierdzenie jest warte tyle, ile zapis pod nim.

Bez oryginałów scalanie jest stratne w sposób, którego nie da się później wykryć. Podsumowanie czyta się dobrze, a niuans, że dany wniosek obowiązywał tylko przy jednej konkretnej konfiguracji, zniknął — i nie masz jak się dowiedzieć, że kiedyś tam był.

Z nimi notatka ma dwie warstwy: aktualne najlepsze rozumienie na wierzchu i dowód pod spodem. To sprawia, że można scalać bezpiecznie przez lata.

Q. Co dzieje się z kandydatami poniżej progu?

Nie są wyrzucane. Notatka powiązana, ale nie ta sama, to dokładnie to, do czego służy odnośnik — i bliskie chybenia z tego etapu zasilają krok linkowania.

Jedna praca daje więc dwa użyteczne wyniki: mocne dopasowania stają się decyzjami o scaleniu, a słabsze — połączeniami między notatkami.

Dlatego próg warto stroić, a nie maksymalizować. Ustaw bardzo wysoko, a dostaniesz duplikaty; ustaw bardzo nisko, a scalisz rzeczy, które powinny zostać osobno, i stracisz odnośniki, które by je połączyły.

Q. Co, jeśli scalę dwie notatki i będę tego żałował?

Zmień tytuł scalonej notatki, a jeśli chcesz je z powrotem rozdzielić — podziel ją ręcznie.

Wszystko, czego potrzebujesz, nadal jest w pliku, bo materiał źródłowy obu rozmów zachował się w całości.

To praktyczna korzyść z zasady wyłącznego dopisywania: scalenie nie jest destrukcyjną edycją, więc najgorszy przypadek to notatka, w której jest więcej, niż chciałeś, a nie taka, z której coś zniknęło.

A ponieważ notatki to zwykłe pliki Markdown w katalogu, podział jest operacją w edytorze tekstu. Nie walczysz ze schematem bazy danych, żeby cofnąć decyzję.