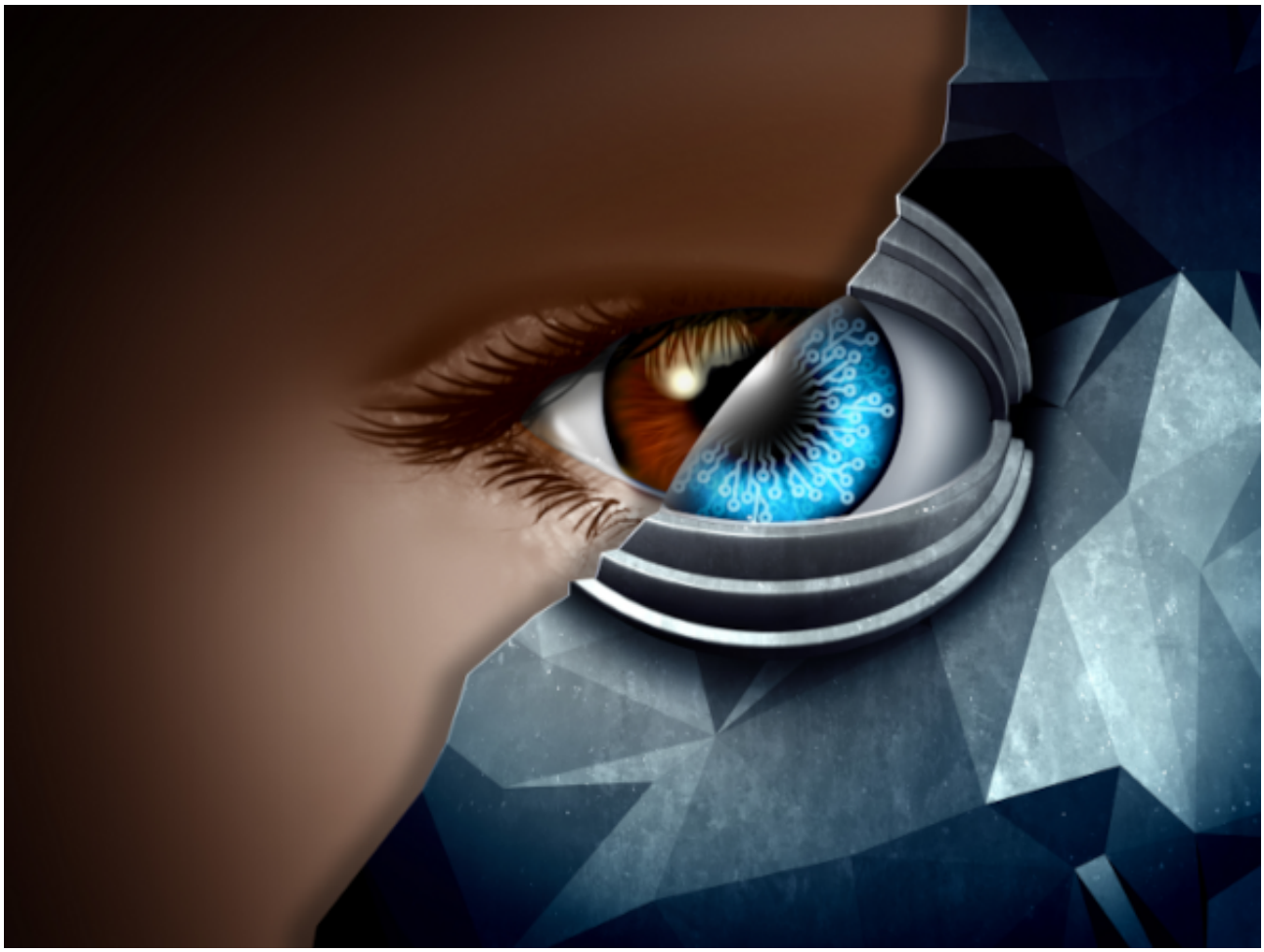
Cassandra Crossing 663- Archivismi- una sessione con Claude



image

(663) - Usare gli LLM non è il male assoluto; se dovete manipolare linguaggio sono tool adatti. L'ultima sessione di Archivismi li ha utilizzati, e questa è la storia.

21 aprile 2026 - Mantenere, migliorare ed aggiornare un'archiviazione periodica, anche piccola come la rubrica di "Cassandra Crossing", è un processo senza fine. Man mano che le cose archiviate aumentano, l'entropia della Rete le complica. Escono nuovi articoli, la formattazione inevitabilmente cambia, alcuni editori iniziano a pubblicarli, altri cessano, i link alle pubblicazioni cambiano o spariscono, qualcuno ripubblica, talvolta cambiando la licenza…

Mentre l'archiviazione di un batch di nuovi articoli può essere molto automatizzata, fino a diventare una faccenda da mezz'ora, gestire il cambiamento e la complessità del passato diventa difficile.

Prendiamo un esempio semplice il sito di Cassandra Crossing.

Generato a mano nel 2009 (all'inizio usando Notepad, oggi con Gedit o Bluefish), minimalista, fatto di solo HTML statico con un unico foglio di stile CSS, si è arricchito di tutto quanto ho salvato in più di venti anni di pubblicazioni di Cassandra Crossing. Gli articoli sono sparsi in Rete in vari posti, e con previdenza, fatica ed autodisciplina la struttura di un articolo è stata mantenuta perlopiù costante; titolo, sinossi e testo con link, formattazione semplice stile markdown, rare illustrazioni.

Tuttavia i link e le informazioni di singoli articoli erano talvolta non standard, incomplete o con errori. Anche la memorizzazione degli articoli come file locali sul mio laptop e sul sito ed i relativi indici aveva questi problemi.

Di recente l'aver fatto, nell'ambito della campagna Archivismi, una archiviazione completa su archive.org ha permesso, come sottoprodotto, di rendere più omogeneo e corretto anche il sito.

La settimana scorsa ho deciso una ulteriore "campagna" di miglioramento del sito e di archiviazione degli ultimi articoli, sia sul sito che su Internet Archive (IA); avevo messo giù qualche obbiettivo:

- definire in maniera inequivoca ed esatta la numerazione e la data di pubblicazione degli articoli (sembra impossibile, ma in due decenni si sbagliano anche quelli);
- controllare la validità e l'univocità degli identificativi di IA;
- inserire sul sito tutti i link con cui ogni articolo è stato pubblicato (cosa difficile, lunga e da fare a mano);
- recuperare ed inserire le sinossi (sottotitolo) nell'archiviazione su IA e sul sito;
- creare una nuova pagina contenente le librerie tematiche di articoli, convertendo quelle presenti su Medium.com, e formattandole in maniera "ricca" con sinossi e link di pubblicazione.

Ora, sembrano tutti compiti semplici, ma sono abbastanza sfrangiati da essere di difficile automazione usando lo scripting.

Di questo fatto sulla raccolta degli articoli ad essa dedicati se ne danno ampi esempi; per questo motivo, dovendo accedere a risorse in linguaggio sia formale che naturale, perdipiù sparse nel web, usare un LLM che potesse anche accedere a risorse in Rete era un opzione interessante, visto che non di programmazione pura né di questioni agentiche si trattava, ma di "semplici" manipolazioni di linguaggi.

Per cui, invece di armarsi di infinita pazienza ed olio di gomito, ho per la seconda volta attivato seriamente un account di un LLM, questa volta per par condicio non ChatGPT (che mi aveva molto deluso) ma Claude, ed usando la versione PRO (20 €/mese) per avere un contesto di dimensioni adeguate rispetto al volume di informazioni da trattare.

Da tutta questa attività, che è costata una giornata piena di lavoro, ma che ha anche generato il risultato che volevo, e che potete trovare nelle pagine Liste ed Articoli di cassandracrossing.org, ho ricavato come sottoprodotto anche questa atipica ma spero utile nuova puntata di Archivismi.

Iniziamo da qualche nota tecnica.

Sapere cosa si acquista in termini di token, di dimensione del contesto e di altre risorse, quali l'utilizzo di tool esterni, è praticamente impossibile. Anche un attentissimo esame dei vari tipi di account e delle loro caratteristiche non fornisce nessun dato affidabile. Questo accade per vari motivi, incluso il fatto che i modelli e le risorse che questi utilizzano sono in continua evoluzione, sia tecnica che "commerciale".

Infatti le risorse che si utilizzano lavorando con gli LLM sono fornite, dal punto di vista monetario, gratuitamente od in forte perdita. Anche gli account di livello Enterprise , facendo alcuni conti, fanno si fatturato, ma sono sempre venduti sottocosto, generando non profitti ma perdite.

Questo causa una perenne "vaghezza" delle caratteristiche degli account utilizzabili in un dato momento; ma peggio ancora, le caratteristiche di un account possono cambiare senza preavviso e senza nessuna comunicazione, in pochissimo tempo, addirittura durante una sessione.

In breve, è impossibile sapere cosa si compra.

Si aggiunga a questo il problema di lavorare con un'infrastruttura di cloud computing che deve bilanciare e limitare i carichi in ogni momento. Un problema grosso quando si deve fare un vero lavoro.

Comunque alla fine ho aperto l'account Anthropic Claude PRO, ed ho utilizzato il modello generalista al momento più avanzato, Opus 4.6, senza mai cambiarlo per tutto il lavoro.

Per non finanziare nessun membro dell'"Impero del Male", ho anche utilizzato il Codice del Consumo con il diritto di recesso entro 14 giorni, la lettera di recesso l'ho fatta scrivere da Claude e l'ho passata direttamente al chatbot di Claude; ha funzionato tutto velocemente e molto bene, almeno da questo punto di vista.

Veniamo alla parte più interessante, gli appunti di questo viaggio.

Sono partito da dati già abbastanza strutturati, cioè da un foglio elettronico contenente un elenco semicompleto degli articoli, con numero, titolo e link di prima pubblicazione; per alcuni articoli avevo anche uno o due link alternativi. L'idea era far estrarre dall'LLM la data di pubblicazione e la sinossi, in modo da poter poi generare un pagina di elenco articoli più ricca e con i link sia ai file locali del sito che alle varie pubblicazioni (storicamente sono 7 testate diverse).

Successivamente, sono partito dalle liste tematiche che avevo generato su Medium.com, piattaforma che sto abbandonando per vari ed in parte ovvi motivi, per generare una nuova pagina del sito, contenente le varie liste, con il relativo titolo e la sinossi della lista, e contenente anche, come sezione espandibile, un elenco degli articoli, con numero, link, titolo e sinossi dell'articolo, tutto ben formattato.

Ma per prima cosa dovevo estrarre le informazioni mancanti, prima di tutto data di pubblicazione e sinossi, e poi anche l'identificativo di IA.

Il problema della data era relativamente semplice, perché fin dalla notte dei tempi avevo quasi sempre incluso il numero dell'articolo tra parentesi tonde all'inizio della sinossi, e la data di pubblicazione, in formato con mese in lettere, all'inizio dell'articolo. Infatti la maggior parte di queste informazioni ero già riuscito a recuperarle con degli script bash dai sorgenti degli articoli, come spiegato nelle prime puntate di "Archivismi".

Ho quindi aperto una sessione, fornendo tre file: il foglio elettronico con i dati incompleti, e le due pagine html del sito che contenevano gli elenchi degli articoli.

Con il primo prompt ho cominciato a spiegare dove recuperare le informazioni mancanti, e subito mi sono reso conto che praticamente tutti i siti bannavano le richieste che l'LLM faceva.

Per aggirare questo problema ho individuato fonti più precise per le informazioni: ho fornito i file degli articoli in formato html, e la struttura della collection di Cassandra Crossing, che ho realizzato su archive.org, in formato json, recuperandolo manualmente via API e caricandolo nella sessione dell'LLM.

In questo modo, dopo diversi prompt, sono riuscito ad ottenere una nuova versione del foglio elettronico ben numerato, contenente la maggior parte delle date, delle sinossi e degli identificativi di IA. Ma restavano una trentina (su 650) sinossi mancanti, e malgrado parecchi altri tentativi, costati una strage di token, ne sono rimaste 15 non recuperate.

Pragmaticamente, ho quindi semplicemente completato a mano il foglio elettronico a forza di copia/incolla e l'ho caricato in una nuova sessione, come informazione principale.

Ho poi iniziato a generare la pagina delle liste. Per lo stesso motivo. non potendo far leggere le liste di articoli su Medium.com, le ho prese dal download di un salvataggio dei dati dell'account Medium.com, e li ho caricati come file.

Le liste erano 35, e stranamente per alcune, come successo per gli articoli, non si riusciva a recuperare la sinossi di lista.

Il problema era che l'LLM subiva una caduta fuori contesto delle prime informazioni che recuperava, prima di essere riuscito a trovare le ultime.

Aprendo ancora una nuova sessione, e fornendo oltre al resto anche la pagina html delle liste già parzialmente generata, riuscivo a farmi completare la pagina HTML.

Bellina, con un html abbastanza ordinato, ma notavo che alcune delle sinossi erano molto banali, non sembravano scritte da me.

Un controllo a campione rivelava che l'**LLM, senza comunicarlo in nessun modo, aveva "generato" le sinossi che non riusciva a recuperare "riassumendo" in qualche modo il testo degli articoli!!**

"Interrogando" il fedifrago, si otteneva una parziale confessione del fattaccio, ma nessuna collaborazione.

Per sistemare le cose si è creata una nuova sessione, fornendo il foglio elettronico e la pagina "parzialmente falsificata". Con qualche insistenza, e fornendo nuovamente i file delle liste, sempre prelevati dal salvataggio precedente, ho infine ottenuto il risultato oggi in linea.

Con operazioni simili, e sempre partendo dall'ormai completissimo foglio elettronico e dalla pagina html preesistente, si è ottenuta la pagina degli articoli, che contiene i link ai file ed a tutte le pubblicazioni di ogni articolo.

Essendo un maniaco della precisione, e volendo continuare in futuro a poter modificare a mano l'HTML, ho esaminato i sorgenti HTML delle due pagine, trovando varie "licenze poetiche" realizzate dall'LLM, nonché parecchie inutili complicazioni.

Ho quindi nuovamente usato l'LLM, a parità di aspetto della pagina, per far modificare e riformattare il sorgente della pagina stessa, eliminando la formattazione inline, come gli operatori "font" sparsi per ogni dove, riconducendo le formattazioni al CSS esistente, e formattando il sorgente stesso in maniera leggibile. Anche qui è stato lungo riuscire a scrivere prompt che mantenessero il contesto, differenziando con precisione la formattazione del sorgente da quella della pagina HTML.

Altra strage di token; Claude ha confuso ripetutamente le due cose, scrivendo arrosti fantasiosi, prima di generare i sorgenti come li volevo.

Infine ho inserito i link degli articoli su archive.org, che sono stati "sintetizzati" dall'LLM partendo nuovamente dal file JSON della collection; è stato però necessario "sistemare" a mano 6 di essi, perché per varie peripezie di archiviazione, i loro ID avevano una formattazione diversa dalla stragrande maggioranza degli altri, poiché su IA un ID creato male non si può né cancellare né riutilizzare..

Riassumendo: in totale una giornata piena di lavoro ed una serie di esperienze interessanti su come lavora il modello che è stato utilizzato.

A consuntivo, il modello scriveva ed eseguiva script in bash e python, anche una dozzina per ogni prompt. Gli script in Python facevano in realtà uso di sofisticati moduli standard Python per l'elaborazione di testi e la codifica in HTML. il LLM si limitava a suddividere in passi il lavoro richiesto, e realizzare gli script per ogni passo, che poi eseguivano il lavoro vero. Interessante il fatto che dopo ogni script, di solito ne veniva girato un altro che doveva verificare il lavoro del primo!

La possibilità che la chat di Claude fornisce di seguire i vari passi di "ragionamento", ed all'interno dei passi vedere le azioni pratiche, ed esaminare (anche a posteriori in una chat archiviata) gli script che vengono generati ed utilizzati, è stato molto interessante ed istruttivo.

Un ultima nota. In almeno tre occasioni durante il lavoro mi è stata chiesta l'autorizzazione di poter leggere e scrivere file sul mio computer e lanciare comandi, inveci di caricare e scaricare i file.

Indovinare la risposta corretta è lasciato agli intuitivi 24 lettori come esercizio finale.

Abbiamo finito per oggi; appuntamento alla prossima puntata di "Archivismi".

Scrivere a Cassandra - Mastodon - Buttondown.com
Videorubrica "Quattro chiacchiere con Cassandra
Lo Slog (Static Blog) di Cassandra
L'archivio di Cassandra: scuola, formazione e pensiero

Cassandra per i posteri: l'archivio su Internet Archive