Proposte per la creazione di dataset per l’addestramento e il finetuning di LLM

Versione 0.1 ***Bozza*** *– Matteo Rinaldi – 3 Marzo 2024 – CC BY*

Indice generale

Premessa

Quello che segue è un breve documento dove ho raccolto delle idee in merito alla creazione di dataset per l'addestramento e il finetuning di Large Language Models. Va considerato esclusivamente come bozza e accenno a potenziali progetti da discutere ed eventualmente realizzare.
Tutta la parte introduttiva si può saltare e andare direttamente all'elenco schematico delle proposte
Non è stato riguardato più di tanto, non tutti i punti sono stati chiariti a sufficenza, e soprattutto ho ancora molte altre idee di cui vorrei discutere.
Questo documento non è in uno stato che lo rende adatto a nessun tipo di pubblicazione o divulgazione: è fondamentalmente un abbozzo, scritto in un giorno solo, spesso in linguaggio colloquiale e dove sono state solo appuntate idee; lo condivido solo a causa dei rapidi mutamenti delle situazioni che potrebbero renderlo un pochino utile anche in questo stato scheletrico e perché ho promesso da un paio di settimane di scrivere un documento del genere ad alcuni membri della community Discord di Mii-LLM. Andando ad approfondire ogni punto e scrivendolo in un linguaggio accademico probabilmente il numero delle pagine andrebbe a triplicarsi.
Vorrei rilasciarlo innanzitutto con grande umiltà: riconosco come non vi siano "grandi idee" al suo interno ma solo una schematizzazione di alcuni temi ricorrenti per quanto riguarda i dati per l'addestramento dei modelli. L'idea è che possa fungere da riferimento per altre idee e proposte; vuole essere un documento *collaborativo*, sarebbe bello lavorarci a più mani per arrivare a qualcosa di ancora più utile. Non sono neanche sicuro al momento della sua effettiva utilità, ripeto, lo rilascio con umiltà sperando che qualcuno lo apprezzi e lo trovi utile. Spero non sembri pretenzioso, non credo di avere in mano nessuna "ricetta segreta" per la risoluzione di un compito così complesso né penso di voler insegnare niente a nessuno.
Lascio pertanto aperta un'istanza Etherpad: https://pad.disroot.org/p/Proposte_per_LLM dove poter scrivere qualsiasi commento, critica, aggiunta. Firmatevi. E si può pensare a una piattaforma collaborativa migliore.
Consiglio anche la lettura del manifesto *LLMentor* che ho scritto quasi un anno fa come proposta di un progetto di crowdsourcing di dataset di instruction fine tuning rivolto a docenti universitari. Il link è qui: https://github.com/manalog97/LLMentor
Manca inoltre la risposta alla domanda principale: *perché* vogliamo sviluppare un LLM? Quali fini ci stiamo proponendo?

Una prolissa introduzione sull'importanza di modelli multiculturali e di qualità… si può saltare

I due cardini fondamentali di questi dataset sono la qualità e l'essere multilingua. Per quanto riguarda il secondo punto, ciò significa sviluppare dataset in lingue diverse dall'inglese, nello specifico in italiano ma potenzialmente queste linee guida potrebbero applicarsi a progetti analoghi da svolgersi in luoghi diversi, in modo da giungere a una raccolta di dataset utili per l'addestramento di un vero modello multilingua, requisito che i modelli attuali non soddisfano se non parzialmente. [Vedi:]
La situazione attuale, che vede una prevalenza eccessiva dell'inglese nei dati di addestramento, è particolarmente dannosa e critica negli aspetti non solo dell'accessibilità del modello a utenti che non parlano in inglese o che preferiscono utilizzare la loro lingua per interagire con le risorse basate sui LLM, ma anche per quanto riguarda questioni meno dibattute e che non sono risolvibili limitandosi a meccanismi basati sulla *traduzione*. L'appiattimento dei modelli sull'inglese significa anche un appiattimento dei modelli sulla cultura angloamericana, con il rischio di andare a far perdere d'importanza la pluralità di visioni del mondo a vantaggio di una visione anglocentrica assolutamente parziale e incapace di rispecchiare l'umanità nel suo complesso. La lingua non è un

mero mezzo di codifica di informazioni, ma ha la capacità di dare forma al discorso e ritagliare i concetti in tanti modi quante sono le lingue esistenti (non ci si limita a parlare la lingua ma si è anche *parlati* da questa). Anche andando oltre le questioni di differenze linguistiche e semantiche, la lingua si fa anche portatrice di un certo contesto culturale, ed è nocivo che i modelli vengano allenati a considerare il resto del mondo in rapporto al mondo angloamericano; tale nocività non si ferma all'aspetto tecnico di corretto funzionamento dei modelli e usabilità da parte della popolazione globale, ma si estende fino a diventare un potenziale problema sociale ed etico non appena tali modelli nelle loro varie e ancora non ben definite declinazioni entreranno nella vita quotidiana e pubblica della popolazione. Ai fini di tale bozza, possiamo lasciare l'approfondimento dell'argomento a lavori successivi e proseguire con l'altro aspetto, quello relativo alla qualità.
Tralasciando per ora discussioni sull'importanza del migliorare la qualità dei modelli per vari fini (sociali, accademici, di utilità contingente…), possiamo spostarci direttamente *in medias res* constatando in primo luogo la pressocchè totale assenza di dataset per il finetuning in italiano e in secondo luogo le criticità esistenti negli attuali dataset di finetuning pensati per l'inglese.
I dataset italiani per il finetuning al momento reperibili e disponibili con licenze aperte sono per il momento mere traduzione di dataset inglesi; tale approccio, sebbene possa funzionare per effettuare qualche sperimentazione, non è adatto per lo sviluppo di modelli che siano autenticamente multilingua. Dal punto di vista lessicale e sintattico, le traduzioni potrebbero conservare uno stile troppo aderente a quello della lingua inglese, specialmente in considerazione del fatto che tali traduzioni vengono svolte non da traduttori professionisti ma in modo automatico. Il risultato potrebbe essere quello di un modello che, a una analisi più approfondita, non comunica effettivamente in italiano ma al contrario continuerebbe a parlare in inglese tradotto in italiano. Dal punto di vista semantico e dei contenuti, tradurre, peraltro automaticamente, non contribuisce minimamente a mitigare quell'effetto di accentramento sulla cultura angloamericana di cui accennavamo in introduzione. **Avere un modello che parla un italiano un po' inglesizzato e che continua a riferirsi a situazioni, luoghi, fatti e persone tipiche degli Stati Uniti non è un modello multilinguistico e multiculturare.**
Si rende pertanto necessario raccogliere grandi quantità di dati nelle lingue in cui si desidera che il modello possa operare. Preferiribilmente, la parte maggiore di questi dati dovrebbe riferirsi anche alla cultura del luogo in cui tale lingua è parlata: un libro scritto originalmente in italiano è da considerarsi un dato avente un valore maggiore di un libro scritto in inglese e tradotto in italiano.
In questo documento non andrò a soffermarmi troppo sul dataset di pre-training, un dataset grande e che, per forza di cose, non può essere nella sua interezza considerabile "di qualità". Si rende tuttavia necessario sviluppare anche dataset più piccoli di svariati ordini di grandezza che però si distinguano per la loro "qualità". Uno tra i vari esempi di ricerche a supporto dell'aspetto qualità VS quantità è: [phi-2, textbook are all you need]
[Da continuare, bozza!!!]

Dataset piccoli, specifici, curati

Quelle che seguono sono idee per la creazione di dataset piccoli, specifici, curati e che quindi sono più adatti a fasi di finetuning piuttosto che di addestramento; ciò non toglie che, nel caso in cui durante lo sviluppo di un certo modello non si ritenga valido usarli per il fine-tuning, potrebbero tornare utilissimi come risorse per il pretraining.

Criteri per la qualità [bozza]:
Tematiche
Attendibilità degli autori

Difficoltà di ragionamento => Fondamentale perché è da qui che il modello farà astrazione sui dati facendo emergere le capacità più interessanti.

Dataset 1: Coppie domande/risposta [Instruction Fine-Tuning]

Accademiche

È il dataset *classico* per il finetuning delle LLM in particolar modo quando finalizzato allo sviluppo di un *assistente* come ChatGPT o Gemini. Senza stare a specificare ora *perché* serve un tale tipo di dataset, vorrei presentare alcune possibili idee e linee guida.
Le coppie D/R sono necessarie per fornire al modello l'astrazione necessaria per rispondere quando viene interrogato; per questo motivo, ritengo importante porre una certa dose di cura nel compilare tali dati. Sarebbe bene puntare a risposte che non siano semplificatorie e che affrontino anche temi molto complessi scendendo nei dettagli.

Possibilità di realizzazione:

- Coinvolgere studenti; tramite piattaforme, anche elementari, di crowdsourcing di cui discuteremo in seguito, studenti volontari potrebbero caricare documenti contenenti coppie di domande e risposte basate sui loro appunti universitari o sulle loro conoscenze. Non si avrebbe la stessa garanzia di accuratezza come se quel materiale provenisse da docenti universitari (vedi progetto originale LLMentor) ma comunque è da aspettarsi un materiale nel complesso più che valido, sicuramente superiore agli attuali dataset per il finetuning inglesi, composti per la maggioranza di contenuti generati automaticamente da altri LLM
- Non tutte le coppie D/R sono uguali, per prima cosa concentriamoci sulle coppie D/R su argomenti specifici, ad esempio accademici. In questo caso, l'ideale sarebbe privilegiare argomenti che non lasciano troppo spazio a opinioni personali e si riferiscono a questioni più o meno assodate, per quanto un certo grado di dissenso potrebbe comunque risultare altamente utile (vedi: prospettivismo in AI). Nelle linee guida da fornire ai collaboratori sarebbe da reiterare più volte l'idea che vengono privilegiati argomenti complessi, anche molto specifici e, idealmente, che coinvolgono una certa dose di ragionamento e che l'eventuale ragionamento necessario sia spiegato per punti. Questo perché argomentazioni generali probabilmente saranno già ampiamente presenti nei dati di addestramento (es Wikipedia); sarebbe interessante mostrare al modello come affrontare domande complesse e rispondere a queste domande in modo approfondito e ragionato.

Questo testo può essere saltato

Modelli come ChatGPT 3 tendono spesso a una spiegazione di tipo enciclopedico, generalista, caratterizzata da una spesso inutilmente prolissa ripetizione di un contesto iniziale che va ad occupare gran parte della risposta per poi confinare la risposta effettiva alla domanda dell'utente in molto meno spazio. Gli argomenti vengono ogni volta "introdotti" spesso con lunghi giri di parole, in una forma che sembra imitare quella delle introduzioni di Wikipedia, anche quando l'utente chiede risposte dirette e specifiche. Sarebbe interessante pertanto sviluppare domande e risposte su più livelli, da una parte domande generali con risposte, giustamente, fornite di una introduzione, ma anche domande più specifiche e che, comprensibilmente, se vengono poste presuppongono che l'utente che le stia ponendo abbia un certo livello di conoscenza pregressa e desideri andare a fondo dell'argomento piuttosto che restarne in superficie. ChatGPT tende spesso a risposte nello

stile “ELI5”, espressione nata su Reddit che significa “Spiegamelo come se stessi parlando con un bambino di cinque anni”. Ora, questo può andar bene per fare scalpore nel pubblico e può aver senso che un modello che aveva un po’ il ruolo di aprire la stagione dell’IA generativa al grande pubblico fosse impostato con questo stile, ma penso che adesso si possa chiedere di più a questi modelli e cercare di privilegiare la profondità alla semplicità. È vero che potrebbe darsi il caso che a porre una domanda complessa sia un utente poco ferrato nella materia che arrivi alla domanda quasi per caso, ma ciò non toglie che in questo caso l’utente stesso potrebbe chiedere una spiegazione più in generale al modello e inoltre non credo si debba dare priorità alle preferenze di utenti che cercano interazioni semplici e superficiali a scapito di utenze più interessate a tematiche complesse e di approfondimento. Il dataset di D/R dovrebbe, implicitamente, inferire il livello dell’interlocutore dalla domanda posta e rispondere di conseguenza. Conoscere l’utente è il modo migliore per soddisfarlo.

Generali

Riuscire ad ottenere questo dataset di D/R è più difficile rispetto a quello basato su argomenti accademici perché, al contrario di quest’ultimo, non segue dei “binari” stabiliti ma al contrario può toccare non solo qualsiasi argomento ma anche qualsiasi uso del linguaggio.
Generare storie, rispondere a domande di senso comune, rispondere a curiosità, impersonare stili di scrittura… sono solo alcuni dei possibili task. È il dataset più difficile tra quelli presentati:

- Si può spaziare su una infinita varietà di task, è difficile anche solo tirare giù una lista di idee e temi (invito a farlo su Etherpad!); Si può prendere come spunto ad esempio l’ottimo (e piccolo) dataset “Norobots” (https://huggingface.co/datasets/HuggingFaceH4/no_robots). Si notano tematiche assolutamente varie:
  - Hobby (ricette di cucina, consigli per il fai da te...)
  - Aneddoti / cultura pop
  - Domande generali
  - Generazione di:
    - Storie
    - Poesie
    - Descrizioni per post social
    - E-mail
    - Slogan pubblicitari
  - Intrattenimento
  - Impersonare chatbot con stili di risposta e caratteristiche particolari
  - Task di NLU
    - Riassunti
    - Modifica e riscrittura del testo seguendo certe caratteristiche, come variazioni nei sentimenti, nello stile
    - Spiegazione di termini ed espressioni
- Le risposte sono molto più arbitrarie rispetto al dataset D/R accademico, potrebbero con facilità contenere:
  - Opinioni personali degli annotatori
  - Bias di vario tipo
  - Considerazioni etiche
  - Qualità discutibile (è un po’ ridicolo da punto di vista letterario pensare che chiedere a un annotatore di “scrivere una poesia su X e Y” sia un modo per ottenere un esempio di una bella poesia)

Al momento i dataset di instruction finetuning generali di questo tipo sono estremamente problematici da usare in un modello multiculturale: sfogliando uno qualsiasi di questi dataset si nota una quantità eccessiva di riferimenti agli Stati Uniti.
È molto complicato sviluppare un dataset di questo tipo senza un significativo investimento economico per assumere degli annotatori professionisti. Fino a dove ci si potrebbe spingere con il crowdsourcing? Che tipologie di volontari cercare? Come verificare la qualità dei dati?
Nonostante queste difficoltà, si potrebbe provare ad individuare dei sotto-task che siano più semplici e soprattutto meno problematici. La generazione di contenuto creativo come poesie e racconti è sicuramente difficile, tuttavia altri task potrebbero essere somministrati anche sottoforma di "gioco" a dei collaboratori volontari. Ad esempio riassumere brevi testi, cambiarne lo stile da formale a informale, risposte semplici e brevi a qualche domanda generale…
Idee?

Dataset 2: Comprensione del testo

Questo sarebbe un bellissimo dataset, per la cui realizzazione si potrebbe chiedere il favore principalmente a studenti di discipline umanistiche.

> Un primo tipo di testi potrebbe ricalcare gli esercizi di comprensione del testo *classici* somministrati ad esempio nelle scuole e per i concorsi pubblici. Tuttavia ho appena scoperto che si può recuperare una grande quantità di materiale già pronto e che sembra tranquillo dal punto di vista delle licenze. Quindi a questo punto ci si potrebbe limitare a una quantità molto piccola di testi da chiedere a studenti di lettere da usare come *golden label*

Sempre rimanendo in tema comprensione del testo, credo che vi possano essere dei compiti che generalmente non vengono fatti svolgere agli esseri umani ma che potrebbero comunque risultare particolarmente utili per i LLM:

- Elencare tutti gli "enti" presenti in un testo, le loro relazioni e i loro aggettivi; un compito più di tipo NLP classico, che potrebbe essere molto utile con testi complessi;
- Le LLM hanno notoriamente difficoltà quanto i contenuti sono *referenziali* ovvero quando avrebbero bisogno dell'apporto di informazioni multimodali per essere compresi. Si potrebbe pensare quindi a compiti di spiegazione delle situazioni:

  Modello del mondo: il modello ha effettivamente compreso *cosa* sta succedendo in un testo? Descrizione oggettiva dei luoghi, personaggi, ipotesi di dialoghi coerenti tra i personaggi che siano fedelmente aderenti cosa si sta raccontando nel testo e allo stile con cui lo si sta facendo.

  A tale proposito si possono pensare a delle domande standard da applicare a testi non di facile interpretazione (poesie, canzoni [occhio al copyright!]): alcune bozze di idee:

  - Descrizione dei luoghi e degli avvenimenti **cinematografica**, come se si stesse strutturando la scenografia di un film;
  - Ricostruzione di un possibile dialogo implicito nel testo come se fosse un pezzo teatrale;
  - descrizione grafica delle immagini evocate dal testo (come se si volessero generare prompt per un generatore di immagini IA)
  - Spiegazione delle metafore;
  - Comprensione di riferimenti velati, allusivi, di non univoca interpretazione

I testi di alcuni cantautori classici italiani (penso a De Andrè, Guccini, Battiato, Vecchioni, De Gregori…) sarebbero estremamente adatti a questo tipo di task. In particolare De Andrè fa un utilizzo molto avanzato della lingua e i suoi testi sono spesso di non univoca interpretazione, ricchi di immagini visive ecc… la questione da capire è il copyright. Fair use? Diritto di citazione?

Questi un paio di esempi fatti al volo, con “L’Ultimo Spettacolo” di Vecchioni e “Il Ritorno di Giuseppe” di De Andrè. A mio avviso ChatGPT 3.5 si è comportato malissimo con questo tipo di compito. Non sono esempi esaustivi di quello che ho in mente e che spero di aver fatto capire nei punti precedenti, comunque è un punto di partenza:

https://chat.openai.com/share/6c65122e-5f3b-4417-ab54-7eb4e3f78436
https://chat.openai.com/share/d159037b-d50e-41dc-a1f5-42d423fa3352

Tra l’altro sarebbe anche da affrontare la questione contenuti sessuali/violenti eccetera. Quando si interpreta un’opera artistica, è assurdo fare censura su questo tipo di tematiche… [punto da approfondire]

Vedi in appendice: Task linguistici (ispirati da vecchi siti di materiale didattico trovati online):

In definitiva, i compiti di comprensione del testo sarebbero una risorsa fondamentale da inserire nei dataset considerata la loro grande rilevanza nei compiti di Natural Language Understanding. Si può pensare di andare oltre i tipici esercizi di comprensione del testo (per quanto fondamentali) e pensare a esercizi mirati per le LLM che vadano a lavorare laddove si notano più mancanze. Ci si deve soffermare anche su aspetti particolarmente banali della comprensione, banali per un essere umano ma che potrebbero mettere in luce comportamenti più da “pappagallo stocastico” dei LLM piuttosto che da un modello capace di comprendere. Idee a proposito?

Dataset 3: Ragionamento e Chain of Thought

Rafforzare le abilità di ragionamento è un altro compito fondamentale nell’addestramento di LLM utili e di qualità. “Ragionamento” è un altro termine complesso e ambiguo. Si rimanda ad altro tempo e luogo per una discussione su cosa si intenda per ragionamento, tuttavia, provo ad elencare alcune bozze di idee.

Enigmistica

Il materiale pensato per l’enigmistica può essere a mio avviso estremamente utile per l’addestramento di modelli nella speranza di osservare abilità emergenti relative al ragionamento:

- Enigmi gialli/polizieschi: questo tipo di enigmi racchiudono diverse caratteristiche particolari che possono tornare utili:
  - Richiedono una comprensione avanzata del testo, ovvero 1) attenzione a tutti i dettagli che possono servire a risolvere il caso, anche dettagli piccoli che possono sfuggire a una lettura non accurata; 2) Creazione di un “modello situazionale del mondo” accurato e corrispondente a quanto si vuole esprimere nel testo; 3) Abilità di tenere nella memoria a breve termine una rappresentazione schematica di tutti gli indizi potenzialmente ricavabili dal testo sia le cose ovvie che le nascoste; 4) capacità di filtrare il contenuto non necessario; 5) applicazione di un metodo logico-deduttivo per risolvere caso; 6) applicazione del pensiero laterale per risolvere il caso
  - Si possono scrivere o (meglio) recuperare già fatti, sperando in licenze permissive. L’importante è che vi sia il testo ma anche la soluzione, possibilmente ben argomentata (Chain of Thoughts)
- Cruciverba: meravigliosi esempi di definizioni complesse e ambigue delle parole italiane;
- Indovinelli

- Eccetera eccetera. Praticamente ogni materiale enigmistico potrebbe risultare utile. Inutile dire che un dump della *Settimana Enigmistica* sarebbe una risorsa fantastica, ma impossibile per chiare ragioni di copyright. In teoria tutti i numeri dal 1932 al 1954 sono ormai di pubblico dominio, ma come recuperarli e digitalizzarli? Comunque, online si trova del materiale e anche qui si potrebbe generare oppure scrivere a enigmisti (come Giorgio Dendi, noto per gli enigmi che stimolano il **pensiero lateriale**) sperando che abbiano voglia di donare del vecchio materiale

Ragionamento filosofico

Fonte dal grandissimo potenziale.

Paragrafo da approfondire moltissimo, intanto, appunti (bozza!!!) :

- Logica: dataset di fallacie, dataset di ragionamenti [studenti di filosofia?]
- Argomentazioni filosofiche, classiche e non. Esposizione, commento, critica, controargomentazioni…
- Estrazione dell'argomentazione da un testo
- Dialoghi, discorsi… Questa potrebbe essere una risorsa fondamentale per le questioni legate all'etica: proporre un'argomentazione, confutarla, controargomentarla eccetera eccetra
- Sia materiale di recupera ma sarebbe stupendo stendere testi, anche brevi, che però siano ottimi dal punto di vista dell'argomentazione filosofica

Ragionamento scientifico

(bozza!!!) Problemi scientifici di qualsiasi tipo che necessitano di un metodo di risoluzione logico rigoroso. Inutile fare esempi, se ne possono fare a centinaia e tutti molto banali. Praticamente tutte gli argomenti scientifici (matematica, logica, chimica…) offrono infinite possibilità in questo campo. Sarebbe da capire come recuperarne una grande quantità. Altrettanto interessante quando questo tipo di ragionamento può estendersi a situazioni che non siano legate alle scienze due o ingegneristiche ad esempio problemi di vita quotidiana,

È interessante adottare metodi di ragionamento chiari ed espliciti: individuare le premesse, mostrare le possibili alternative errate, indicare proposte su come risolvere il problema ed eventualmente falsificarle. Sfruttare Chain of Thought.

Ragionamento "in generale"

(bozza!!!)

Dataset 4: Orientamento spaziale e altri task (piccolo)

Un dataset piccolo, molto più piccolo rispetto agli altri proposti, da fare a mano lavorando per poco tempo anche in un piccolo gruppo. È un dataset un po' atipico che però può avere fondamenti scientifici validi, riassumo superficialmente in poche righe, il tutto si potrebbe argomentare decisamente meglio ma è giusto per dare un'idea:

Tolman negli anni “40 propone il concetto di *mappa cognitiva* per spiegare perché i topi fossero in grado di orientarsi in labirinti utilizzati in laboratorio e in particolare perché fossero in grado di *trovare scorciatoie* per raggiungere gli obiettivi; l’idea è che i mammiferi formino nel cervello una mappa dello spazio in cui si trovano che gli consente di trovare percorsi vantaggiosi. Negli anni “80 i coniugi **Moser** scoprono il sistema di orientamento ippocampale-entorinale: *place cells* nell’ippocampo e in seguito *grid cells* e altri neuroni specifici come *head cells* e altri ancora; molto in breve, esistono neuroni che 1) si “accendono” quando ci si trova in un determinato luogo (place cells) 2) complementari a questi c’è un sistema di mappatura esagonale dello spazio (*grid cells*) che può riprogrammarsi a seconda del compito e che fornisce metriche univoche di distanza e direzione. **Bellmund et al, 2019**: studi relativi alla possibilità che questo modello di mappe cognitive evidenziato nel sistema ippocampale-entorinale sia in funzione anche nel pensiero astratto: **mappe contettuali** analoghe a quelle spaziali. L’idea è che “orientarsi” nei concetti sfrutti gli stessi meccanismi neuronali utilizzati per orientarsi nello spazio: le nozioni di distanza concettuale, iperonimia e iponimia, generalizzazione, clustering di argomenti simili eccetera (tutte tematiche fondamentali per il NLU e che potrebbero portare a fruttuose innovazioni *architetturali* dei modelli oltre il Transformer vanilla) sono in questo paper collegate ai meccanismi di orientamento spaziale. Anche il fatto che le metafore concettuale si riferiscono spesso a dinamiche spaziali è a supporto di questa tesi.

Whittington, Behrens, 2022: RELATING TRANSFORMERS TO MODELS AND NEURAL REPRESENTATIONS OF THE HIPPOCAMPAL FORMATION

Many deep neural network architectures loosely based on brain networks have recently been shown to replicate neural firing patterns observed in the brain. One of the most exciting and promising novel architectures, the Transformer neural network, was developed without the brain in mind. In this work, we show that transformers, when equipped with recurrent position encodings, replicate the precisely tuned spatial representations of the hippocampal formation; most notably place and grid cells. Furthermore, we show that this result is no surprise since it is closely related to current hippocampal models from neuroscience. We additionally show the transformer version offers dramatic performance gains over the neuroscience version. This work continues to bind computations of artificial and brain networks, offers a novel understanding of the hippocampal-cortical interaction, and suggests how wider cortical areas may perform complex tasks beyond current neuroscience models such as language comprehension.

Yamada et al, 2023: “Evaluating Spatial Understanding of Large Language Models”

Mi scuso per la natura estremamente abbozzata di questo paragrafo, che vorrei approfondire accuratamente e ritengo estremamente interessate. Serve tuttavia a giustificare il tipo di task che vorrei proporre ovvero **orientamento, con identificazione di percorsi ottimali, in ambienti spaziali descritti in linguaggio naturale**.

Si possono prendere mappe di città, descriverle a diversi livelli di granularità (zone con singole strade, intera città con quartieri e punti cardinali) e: ad esempio: 1) Descrivere percorsi ottimi 2) Giungere a conclusioni del tipo “se vado verso est *allora* mi trovo davanti a X” e task di questo tipo.

Oltre a città, si può pensare anche a problemi di orientamento in luoghi generici, come case, ambienti ristretti come ciò che si può vedere da una finestra, luoghi immaginari (magari fare un

disegno su carta per evitare errori) eccetera. È un task divertente su cui si può usare molta fantasia, l'unico requisito è rimanere coerenti e possibilmente scrivere task difficili.

Task di questo tipo potrebbero essere interessanti sia per l'addestramento che per il benchmark perché presuppongono abilità complesse di generalizzazione e astrazione; potrebbero anche servire a guidare eventuali innovazioni architetturali specialmente nel caso in cui l'ipotesi esposta brevemente sopra del collegamento tra mappe concettuali, analogie tra il sistema neuronale di orientamento spaziale e concettuale e utilizzabilità di questi concetti da parte di modelli ANN fosse rafforzata.

Dataset 5: Brevi testi estremamente curati

Coinvolgere i docenti universitari in un lavoro tipo LLMentor? https://github.com/manalog97/LLMentor

Vedi anche: L'ipotetica piattaforma di crowdsourcing in particolare "Donazione di tesi e appunti"

Dataset 5...∞: Idee?

I dataset "grandi" di pretraining – Lo scraping

I "grandi classici"

Inutile ora dilungarsi troppo: Wikipedia, Wikisource, Wikitionary eccetera eccetera. Da prendere così come sono. Ovviamente non sono perfetti, possiamo stare a trovare infinite criticità ma credo siano un punto di partenza perfetto.

Libri di pubblico dominio

Inutile stare ad argomentare ora perché sono importanti. **Liber Liber** è una fantastica risorsa con circa 4500 libri di pubblico dominio pronti per essere inseriti nel dataset. Il lavoro è praticamente completato: un annetto fa, più per esercizio personale che altro, avevo rifatto il sito di liber liber passando dal loro sistema basato su campi di testo a un più efficiente DB relazionale. Il lavoro non è più andato avanti perché non c'è stato un grande interesse da parte della comunità di Liber Liber; in compenso, avendo già questo DB con tutti i libri di Liber Liber fino a Maggio 2023 circa, è stato semplice scaricare tutti i Link. Ringrazio Ruggero per la deduplicazione. Il dataset è già su HF ma purtroppo per uno stupido errore circa 1000 libri sono mancanti. Risolverò la cosa il prima possibile, è molto facile recuperare gli altri.

Vanno recuperati anche da altre risorse! Project Gutenberg in Italiano è un punto di partenza ma poi, andando a cercare nei cataloghi delle biblioteche online, si possono trovare tante altre risorse! Il problema più grave è quello relativo a scansioni e OCR.

Usenet e forum

Effettuare lo scraping di fora online è una pratica ampiamente utilizzata nella creazione di dataset per l'addestramento di modelli linguistici. Nei dataset in inglese attualmente esistenti, si nota come Reddit, Quora e StackExchange (piattaforme "moderne") siano spesso fonti di testo ampiamente presenti nei dati di addestramento.

I fora [approfondimenti su cosa sono, storia eccetera rimandati a eventuali documenti successivi] sono una risorsa interessante da includere in un dataset di pretraining. Ci sono degli aspetti problematici come la non verificabilità delle informazioni presenti al loro interno, la possibile presenza di linguaggio tossico e litigi, spesso offensivi (*flame*) tra utenti; nel caso di Usenet c'è anche un certo quantitativo di spam, fortunatamente facilmente identificabile, e la presenza di un linguaggio in certi casi per nulla moderato. Nonostante questi problemi, tuttavia sono anche tanti i punti a sostegno dell'inclusione di questo materiale nei dataset: i fora sono spesso una miniera di informazioni, dettagliate, precise, fornite negli anni da gruppi di utenti particolarmente appassionati ed esperti su specifici argomenti. È possibile trovare al loro interno informazioni non ottenibili altrimenti, frutto spesso di pratica e di esperienze personali. Le informazioni sono organizzate in *discussioni*, quindi seguendo un modello dialogico, il che è importante nell'addestramento di modelli linguistici specie nell'ottica di sviluppare assistenti virtuali. Ci sono fora riguardanti specifici argomenti e, in particolare quando tale argomento è molto delineato, si possono trovare discussioni tecniche di altissimo livello e, spesso, corrette anche perché sottoposte al vaglio di numerosi altri utenti che eventualmente possono dibattere e commentare. Questo dibattito e commento è certamente istruttivo per i modelli nel riuscire ad esporre i concetti argomentandoli e sottoponendoli a critiche. La discussione sui vantaggi dei forum per l'addestramento può andare avanti, ma intato, ricapitoliamo il lavoro fatto e da fare in merito:

Usenet

Scriverò una bella introduzione su Usenet, per ora, basta ricordare che si tratta di una piattaforma distribuita (non centralizzata) facente parte della primissima generazione di Internet, essendo stata sviluppata negli anni Settanta, ben prima della nascita del web. Per quanto riguarda l'Italia, siamo riusciti a raccogliere contenuti dal 1994, anno di nascita della gerarchia "*.it". L'archivio più grande di discussioni Usenet è presente sulla piattaforma "Google Groups" di Google, piattaforma tra l'altro che proprio il 22 Febbraio di quest'anno ha cessato di raccogliere nuovi contenuti.

Un lavoro molto dispendioso in termini di tempo e di risorse computazionali per effettuare lo scraping delle gerarchie "*.it" e "*.italia" dalle pagine di Google Groups è stato effettuato dal sottoscritto a Febbraio 2024. Lo scraping è stato effettuato con degli script Python basati sulla libreria Selenium che verranno resi disponibili su GitHub. Il risultato è un archivio di circa 75GB contenente XXX discussioni divise in XXX newsgroup tematici. In appendice la gerarchia dei newsgroup scaricati. Seguiranno a breve delle statistiche sul materiale scaricato come: quantità di conversazioni, quantità di messaggi, quantità di messaggi per newsgroup, quantità di messaggi per anno, quantità di messaggi per newsgroup e anno e così via.

Come si evince osservando la gerarchia, le tematiche trattate sono svariate e abbracciano un arco diacronico particolarmente lungo (circa trent'anni). I newsgroup hanno conosciuto un calo di

popolarità dopo il 2012 circa a causa del diffondersi dei social network (centralizzati e proprietari) ma comunque è presente un significativo numero di messaggi anche relativi all'ultimo decennio.

Attualmente per ogni newsgroup esiste un file JSON contentente tutti i dati necessari all'organizzazione dei messaggi nel newsgroup:

{title, original_url, newsgroup, messages: [author, day, month, year, hours, minute, am/pm, text]}

Prima del caricamento su Hugging Face, che avverrà in settimana, i dati saranno riorganizzati in file JSONL aventi come struttura:

{title, author, id, progressive_number, endflag, timestamp, newsgroup, original_url, text}

Dove progressive_number rappresenta l'andamento della discussione (a partire da "0" per il primo messaggio) e il dato booleano endflag vale 1 se si è arrivati all'ultimo messaggio della discussione. Il timestamp sarà in formato ISO-8601-1

Si tratta di una struttura dati altamente inefficiente e ridondante (campi come title, original_url, newsgroup saranno ripetuti milioni di volte) ma al momento è l'unica soluzione pensata per rendere il dataset facilmente fruibile su HuggingFace. La struttura più adatta al dataset sarebbe sicuramente quella di un DB relazionale:

Conversazioni (id,titolo,newsgroup,url)

Messaggi(id,id_conversazione,autore,dataeora,testo)

ma al momento questa strada non sembra essere percorribile sulla piattaforma HuggingFace.

La compressione con buoni algoritmi come LZMA2 o Bzip2 consentono tuttavia di ovviare a questa ridondanza in termini di occupazione di spazio su disco.

Il dataset di Usenet è già completo, al momento conservato su un hard disk esterno e in settimana sarà caricato su HuggingFace e Archive.org. Potrebbe risultare il dataset in Italiano per task di NLP più grande tra quelli liberamente disponibili.

Ringrazio *Ruggero* per il costruttivo confronto durante questo progetto e per avere scaricato gli ultimi due *giga* di materiale quando le mie risorse computazionali non erano più sufficienti, oltre che per aver evidenziato il problema dell'usare il JSON originale come dataset di HuggingFace.

Forum

Ho scritto uno script Python, basato su Beautiful Soup, adatto a scaricare per intero qualsiasi forum. Occorre solo individuare dei campi specifici con un browser internet (div con contenuto, div con titolo, tag che individuano autore e data, meccanismo di paginazione, logica dell'URL, numero massimo di discussioni), inserirli nello script ed eseguirlo.

Ho fatto partire ieri lo scraping di alcuni piccoli forum: matematicamente (matematica), analogica (fotografia) e electroyou (elettronica) per testare lo script e sembra funzionare tutto correttamente. Sto compilando una lista di forum che potrebbe valer la pena scaricare, indicativamente si riusciranno a ottenere in questo modo altri 30GB circa di materiale testuale. A differenza di Usenet, qui potrebbero sorgere questioni legate al copyright pertanto sarebbe bene sentire al più presto un esperto di diritto.

Riviste accademiche delle Università italiane

Punto da approfondire; comunque: ho scoperto che quasi tutte le università italiane hanno portali di questo tipo:

https://ojs.unito.it/

https://rosa.uniroma1.it/

http://www.serena.unina.it/

eccetera eccetera.

Sono riviste di altissima qualità, su tematiche accademiche specifiche e complesse che sarebbero una risorsa magnifica per i dataset di pretraining. Una risorsa così tanto di qualità è probabilmente assente anche in rinomati dataset americati. Bisogna capire la questione copyright: sono tendenzialmente in open access e licenza CC-BY-SA-ND. Possiamo inserirle nei dataset?

Concorsi pubblici

Sito scoperto per caso proprio oggi mentre scrivevo questo documento:

https://www.concorsipubblici.com/quiz/categorie/comprensione-di-testi-1731

Una miniera di risorse perfette per creare dataset di instruction finetuning. Numerosi esempi di comprensione del testo. Che ne pensate? Comincio volentieri a riflettere su come effettuare uno scraping sensato.

Materiale legale di pubblico dominio

Sentenze, atti di processo, codici...

Altro materiale online (senza esagerare… inutile pensare di rifare a mano Common Crawl)

…….Possiamo scrivere una grande lista

L’Ipotetica piattaforma di crowdsourcing

[BOZZA!!!]

Punto importante, ma anche questo verrà lasciato qui solo abbozzato. Però:

1. LLMentor: possibile piattaforma indirizzata a persone del mondo accademico
2. Espandere LLMentor con accesso a studenti universitari che potrebbero:

    1. “Donare” la loro tesi
    2. “Donare” i loro appunti
    3. Scrivere coppie domande e risposte
    4. Valutare le prestazioni dei modelli esistenti
    5. Discutere tra di loro
3. Una sottosezione di LLMentor potrebbe essere aperta al pubblico generale con vari task simil-Amazon Mechanical Turk/LabelStudio come
    1. Proporre domande generaliste
    2. Effettuare valutazioni

Punto da approfondire molto!

Comunque, è naturale che un progetto di crowdsourcing opensource rispetto a un progetto con annotatori pagati fornirà risultati di qualche ordine di grandezza inferiore. È inoltre ancora più difficile a causa del fatto che, al fine di privilegiare la qualità, si stia anche facendo una selezione sugli ipotetici volontari (universitari, dottorandi, docenti…) Tuttavia, credo che possa valere la pena provare a patto di non investire troppe risorse nella piattaforma (La base di LLMentor è quasi pronta, codice semplice scritto a mano ma funzionante) dal momento che:

1. C’è un grande interesse del pubblico generale per questa tecnologia;
2. Si potrebbe contribuire anche con una quantità molto piccola di tempo, come pochi minuti per caricare materiale come tesi e appunti o poche ore per scrivere qualche decina di domande e risposte;
3. Si potrebbero fornire considerazioni etiche che possano far capire a un pubblico di studenti sensibili quanto sarebbe utile contribuire allo sviluppo di modelli aperti, di qualità e multiculturali

Da brevi discussioni informali fatte nell’ambiente universitario, sembra che ci siano studenti che parteciperebbero volentieri a un progetto simile. Sono molti gli esempi di progetti senza fini di lucro nel web che poi portano a risultati considerevoli, come Wikipedia o la piattaforma di calcolo distribuito “Boinc”. Motivare i volontari, ringraziarli e magari farli divertire con idee stimolanti e *gamificare* il tutto con punti e classifiche potrebbe portare a una discreta adozione. Si potrebbero inoltre incollare volantini ben fatti nelle varie Università italiane e magari sperare nella collaborazione delle Università stesse.

Sintetico o naturale?

[BOZZA!!!]

Scriverò qualcosa sulla questione dataset sintetico e naturale, per ora, appunti:

- Ovviamente avere una grande mole di dati in naturale è estremamente complesso;

- I dati naturali sono però ancora al momento insostituibili. Esagerare con i sintetici può portare alla “Mucca Pazza”. Con i sintetici continueremmo inoltre ad avere il problema dell’appiattimento sulla cultura angloamericana (forse Mistral migliora le cose?);
- Si potrebbe però provare ad estendere i dati con metodologie sintetiche, basate però sulla riproduzione dei “nostri” piccoli dataset di finetuning curati, in modo da avere un maggior controllo sulla qualità rispetto che lasciare i modelli “a ruota libera”

Problemi aperti

- Copyright!
- Molto altro...

Appendici:

Un “appoggio” di link con materiale didattico

Scrivendo questa bozza, mi sono imbattuto in alcuni siti con materiale didattico per le scuole che potrebbe essere utile tenere in considerazione. In gran parte si tratta di siti molto vecchi e quindi i problemi di copyright potrebbero essere più gestibili.

Questi link non sono assolutamente esaustivi, sono stati trovati oggi in pochi minuti; invito a cercare più materiale.

- https://digilander.libero.it/italianonelbiennio/
- https://digilander.libero.it/uraniaceleste/
- http://www.lineadidattica.altervista.org/materiali_didattici.html
- https://rossanaweb.altervista.org/blog/area-studenti/esercizi-online/esercizi-online-di-italiano/
- https://www.profgiuseppebettati.it/
- http://www.apprendendo.altervista.org/italiano.html
- https://italianoperstranieri.altervista.org/
- https://www.profwaltergalli.it/per-i-docenti/mediatori-didattici-sottosezioni-da-1-a-7/1-siti-didattici-per-tutte-le-discipline/
- http://www.apprendendo.altervista.org/didattica%20online.html
- https://library.weschool.com/ **(!!!)**
- https://italianoinlinea.com/cruciverba-per-imparare-litaliano/
- https://digilander.libero.it/sussidi.didattici/

- http://www.bibliolab.it/lessico/lessico_index.htm
- https://www.guamodiscuola.it/2014/01/testo-argomentativo-materiali-didattici.html
- http://www.storiadellaletteratura.it/ **(!!!)**
- https://www.dropbox.com/s/tcy2j8nma5hmr1j/LAVORO%20SUL%20TESTO%20-%20Percorso%20didattico%20su%20Pinocchio.pdf?e=1&dl=0
- https://www.fabrizioaltieri.it/wordpress/tutti-i-miei-libri-per-ragazzi/battello-a-vapore/serie-azzurra/comprensione-del-testo-scuola-primaria/6-brani-con-verifica-sulla-comprensione-del-testo/

Appoggio di link enigmistica:

- https://cruciverba.io/soluzioni-recenti
- https://www.nostrofiglio.it/famiglia/indovinelli-difficili-per-adulti
- https://www.enigmatopia.it/category/giochi-a-enigmi-online/la-strada-per-enigmatopia/
- https://yesnogame.net/it
- https://utenti.quipo.it/base5/penslate/latmate.htm **(!)**
- https://utenti.quipo.it/base5/penslate/latclass.htm

Task linguistici (ispirati da vecchi siti di materiale didattico trovati online):

Tipologia	Attività	Obiettivi	Proposte operative
Cancellazione	Riconoscere le parole incluse arbitrariamente in un testo	Sviluppare la riflessione sul lessico, in base al criterio della pertinenza del singolo elemento linguistico, rispetto all'insieme.	P1
Cloze	Inserire in un testo le parole mancanti, fornite in sequenza casuale.	Promuovere la competenza testuale e la capacità inferenziale, mediante il preventivo riconoscimento della categoria grammaticale da inserire.	P2

Decontaminazione	**Distinguere gli elementi testuali che appartengono a due differenti testi.**	**Rafforzare la competenza testuale, sulla base del riconoscimento dei fattori di coerenza e coesione.**	**P3**
Esplicitazione	**Collegare ogni pronome, presente nel testo, al proprio referente.**	**Promuovere il riconoscimento dei fattori che determinano la coesione testuale.**	**P4**
Griglia	**Individuare l'intersezione delle variabili (riconducibili a un testo dato) rappresentate sugli assi di una matrice.**	**Potenziare la comprensione del livello denotativo e connotativo.**	**P5**
Incastro	**Ricostruire l'esatta sequenza delle parole di un testo, presentate in ordine casuale.**	**Sviluppare la competenza morfo-sintattica.**	**P6**
Riassunto	**Ridurre un testo ai nuclei informativi essenziali, da riprodurre secondo una formulazione personale.**	**Incentivare la capacità di riconoscere la gerarchia delle informazioni essenziali, mettendo in atto la globalità dei processi cognitivo-linguistici.**	**P7**
Ricostruzione	**Riprodurre la corretta sequenza dei paragrafi di un testo, proposti in ordine casuale.**	**Potenziare le strategie del processo di comprensione, mediante il riconoscimento dei fattori che determinano la coerenza testuale.**	**P8**
Scelta multipla	**Individuare la risposta corretta, selezionandola tra le varie opzioni date.**	**Guidare il percorso di comprensione**	**P9**
Suddivisione	**Dividere un testo in sequenze e assegnare una titolazione pertinente.**	**Potenziare le abilità di lettura e comprensione.**	**P10**

I giochi linguistici

Proponiamo una serie di possibili esercizi, tutti sul medesimo testo di riferimento, che è il seguente:

Un topolino correva avanti e indietro sopra il corpo di un leone addormentato. Quello si svegliò e afferratolo stava per mangiarselo. Ma il topolino lo scongiurò di lasciarlo libero, dicendogli che se lo avesse salvato gli avrebbe ricambiato il favore; il leone sorrise e lo lasciò andare. Non molto tempo dopo il leone fu catturato da alcuni cacciatori che lo legarono con una corda ad un albero. Il topolino, che aveva sentito i suoi lamenti, rosicchiò la corda e lo liberò, dicendogli: "Un giorno tu sorridesti di me, pensando che io non fossi in grado di ricambiare il favore. D'ora innanzi, invece, sarai convinto che esiste la gratitudine anche presso i topi".

Cloze

- Descrizione: individuazione e inserimento dei termini adatti a riempire gli spazi vuoti di un testo. (cloze classico, cloze mirato, cloze facilitato)
- Obiettivi: comprensione e lettura approfondita, sviluppo delle competenze lessicale o linguistiche

Esempi di cloze

- Cloze classico

  Un topolino (…) avanti e indietro sopra il (…) di un leone (…). Quello si svegliò e afferratolo stava per (…). Ma il topolino lo (…) di lasciarlo libero, dicendogli che se lo avesse (…) gli avrebbe ricambiato il favore; il (…) sorrise e lo lasciò (…).
- Cloze mirato (inserimento dei pronomi personali)

  Un topolino correva avanti e indietro sopra il corpo di un leone addormentato. (…) si svegliò e afferrato (…) stava per mangiarsi (…). Ma il topolino (…) scongiurò di lasciare (…) libero, dicendo (…) che se (…) avesse salvato (…) avrebbe ricambiato il favore; il leone sorrise e (…) lasciò andare.

Puzzle

- Descrizione: ricostruzione della forma originaria di un testo di cui preventivamente sono state divise e mescolate le sequenze o i paragrafi. Si può aumentare la difficoltà dell'esercizio usando due testi

- Obiettivi: lettura approfondita ed analitica, riflessione sui meccanismi di coerenza e coesione testuali

Esempio di puzzle

D'ora innanzi, invece, sarai convinto che esiste la gratitudine anche presso i topi". Quello si svegliò e afferratolo stava per mangiarselo. il leone sorrise e lo lasciò andare. Non molto tempo dopo il leone fu catturato da alcuni cacciatori che lo legarono con una corda ad un albero. "Un giorno tu sorridesti di me, pensando che io non fossi in grado di ricambiare il favore. Un topolino correva avanti e indietro sopra il corpo di un leone addormentato. Ma il topolino lo scongiurò di lasciarlo libero, dicendogli che se lo avesse salvato gli avrebbe ricambiato il favore; Il topolino, che aveva sentito i suoi lamenti, rosicchiò la corda e lo liberò, dicendogli:

Logorally

- Descrizione: revisione e trasformazione di un brano in cui sono stati inseriti, casualmente o meno, parole, sintagmi o proposizioni. Il brano deve conservare coerenza e coesione
- Obiettivi: riflessione sulle diverse situazioni linguistiche (in particolare grammaticali e sintattiche

Esempio di logorally

Prova:

Un topolino foresta correva avanti e indietro sopra il corpo di un leone addormentato. Quello si svegliò e afferratolo sbadiglio stava per mangiarselo. Ma il topolino paura lo scongiurò di lasciarlo libero, dicendogli che se lo avesse salvato vita gli avrebbe ricambiato il favore;

Soluzione:

Un topolino, **appena giunto in una foresta**, correva avanti e indietro sopra il corpo di un leone addormentato. Quello si svegliò e afferratolo, **dopo aver emesso un rumoroso sbadiglio**, stava per mangiarselo. Ma il topolino, **che per la paura non riusciva quasi a parlare**, lo scongiurò di lasciarlo libero, dicendogli che se lo avesse salvato, **anche a costo della vita**, gli avrebbe ricambiato il favore;

Alfabeto, tautogramma, acrostico

- Descrizione dell'esercizio: modificazione di un testo in modo che ogni periodo, seguendo l'ordine narrativo del

testo-base, inizi con la lettera successiva dell'alfabeto (Al). o con la stessa lettera (T), o in una successione di lettere tale da determinare un nome (Ac)
- Obiettivi: abilità lessicali e sintattiche

Esempi di alfabeto, tautogramma, acrostico

- **A**nni fa un topolino correva avanti e indietro sopra il corpo di un leone addormentato. **B**atti e ribatti quello si svegliò e afferratolo stava per mangiarselo. **C**on il muso rigato di lacrime, il topolino lo scongiurò di lasciarlo. **D**isse che …..
- **A**lberto, un simpatico topolino correva avanti e indietro sopra il corpo di un leone addormentato. **A**lla fine quello si svegliò e afferratolo stava per mangiarselo. **A**ffranto dal dolore, il topolino lo scongiurò di lasciarlo libero. "**A**vrò pace soltanto quando ….
- **M**olti anni fa….**A** forza di salti e capriole…"**R**idammi la libertà" scongiurò il topolino. **I**nsistette, dicendo al re della foresta che…**O**ffeso, ma non troppo, il leone...

Lipogramma

- Descrizione dell'esercizio: riscrittura di un testo senza mai usare una lettera o un termine precedentemente specificati (lipogramma classico e lipogramma lessicale)
- Obiettivi: sviluppo della competenza lessicale e sintattica

Esempi di lipogramma

- Lipogramma classico (lettera o)

  Un animale, che per abitudine frequenta cantine e ambienti scarsamente puliti, saltava su e giù per le membra del re degli animali che russava in pace. (Un topolino correva avanti e indietro sopra il corpo di un leone addormentato)
- Lipogramma lessicale (parola leone)

  Un topolino correva avanti e indietro sopra il corpo addormentato **di chi di solito ruggisce**….. **Il re della foresta** sorrise e lo lasciò andare. Non molto tempo dopo **il felino** fu catturato….

Statistiche <u>temporanee</u> scraping di Usenet:

Mancano ancora dei newsgroup da sistemare, ma sono già stati scaricati tutti. La gerarchia *italia (discussioni locali) non è al momento presente in statistica. La colonna a destra indica il numero di

conversazioni NON quello dei singoli messaggi (maggiore), che sarà presente in una statistica successiva. Alcuni gruppi purtroppo sono andati perduti perché censurati da Google Groups, forse a causa di troppo spam.

Totale singole conversazioni: 13.098.235

Totale singoli messaggi: *da calcolare*

it.politica	813646
it.sport.calcio.milan	442317
it.economia.borsa	363459
it.discussioni.auto	209997
it.sport.calcio.roma	181430
it.sport.calcio	166006
it.comp.hardware	163391
it.sport.calcio.napoli	160256
it.media.tv	157839
it.media.tv	157839
it.arti.fotografia.digitale	155799
it.comp.macintosh	155526
it.comp.aiuto	152643
it.hobby.motociclismo	146326
it.politica.internazionale	144392
it.hobby.fai-da-te	138488
it.hobby.satellite-tv.digitale	137842
it.comp.os.linux.iniziare	133170
it.tlc.cellulari	127660
it.hobby.viaggi	123933
it.comp.lang.visual-basic	116373
it.arti.cinema	115803
it.arti.fotografia	114277
it.politica.pds	111191
it.arti.musica.strumenti.chitarra	111186
it.comp.console.playstation	110042
it.sport.calcio.inter-f	109206
it.tlc.telefonia.adsl	104615
it.sport.calcio.juventu	99951
it.discussioni.commercialisti	97636
it.hobby.umorismo	93120
it.sport.calcio.inter	90076
it.sesso.discussioni	89924
it.media.video.produzione	87382
it.media.video.produzione	87382
it.hobby.home-cinema	86245
it.politica.polo	85440
it.hobby.cucina	83914
it.discussioni.litigi	83910
it.diritto	83162
it.comp.appl.access	81615
it.lavoro.informatica	80338
it.comp.giochi.action	79174

it.comp.hardware.cd	78026
it.sport.calcio.torino	75571
it.hobby.totoscommesse	73750
it.comp.hardware.palmari	72702
it.hobby.acquari	70399
it.hobby.radioamatori	69671
it.arti.musica.rock	66746
it.discussioni.misteri	65335
it.discussioni.animali.cani	64745
it.discussioni.consumatori.tutela	63063
it.comp.os.win.xp	61986
it.comp.hardware.overclock	61654
it.arti.musica.classica	60054
it.comp.hardware.motherboard	59410
it.cultura.filosofia	59056
it.sport.formula1	58454
it.sport.calcio.fiorentina	57847
it.cultura.religioni	56980
it.arti.cartoni	55926
it.arti.cartoni	55926
it.cultura.single	55807
it.hobby.motociclismo.scooter	55368
it.arti.musica.metal	54985
it.istruzione.scuola	54978
it.lavoro.professioni.webmaster	51924
it.hobby.lotto	51225
it.comp.giochi.annunci	50366
it.comp.os.win.win2000	49655
it.sport.calcio.genoa	49590
it.sport.ciclismo	48954
it.sport.basket	48588
it.comp.java	48529
it.hobby.satellite-tv	48464
it.scienza.matematica	48441
it.comp.grafica	48046
it.comp.grafica	48046
it.aiuto	47043
it.fan.startrek	46051
it.comp.reti.locali	45922
it.cultura.libri	45065
it.lavoro.offerte	44256
it.hobby.scacchi	42853
it.arti.fumetti	42337
it.arti.poesia	42264
it.comp.giochi.sportivi.hattrick	40515
it.hobby.hi-fi	39229
it.comp.hardware.cpu	38584
it.arti.musica	38508
it.discussioni.geometri	38245
it.discussioni.folli	37954

it.comp.www.php	37754
it.comp.os.linux.sys	36822
it.fan.culo	36741
it.arti.musica.rock.progressive	35932
it.fan.studio-vit	35841
it.hobby.elettronica.riparazioni	34833
it.sesso.racconti	34601
it.comp.giochi.simulatori.volo	34370
it.hobby.modellismo	34289
it.comp.lang.javascript	34100
it.comp.hardware.modem	34100
it.politica.ulivo	33218
it.discussioni.varie	33189
it.discussioni.leggende.metropolitane	32435
it.lavoro.consulenti	32354
it.lavoro.consulenti	32354
it.sport.motociclismo	32179
it.discussioni.ufo	32115
it.hobby.scuba	32080
it.sport.calcio.sampdoria	32062
it.sport.calcio.estero	31568
it.diritto.condominio	31507
it.hobby.hi-fi.car	31383
it.lavoro.mlm	31104
it.comp.retrocomputing	30864
it.comp.os.win.win9x	30332
it.sport.windsurf	30327
it.comp.giochi.rpg	30293
it.salute	29591
it.hobby.pescare	29380
it.fan.musica.queen	29297
it.hobby.cicloturismo	29165
it.comp.musica	27983
it.cultura.linguistica.italiano	27052
it.hobby.armi	26886
it.sport.montagna	26538
it.cultura.storia	26527
it.scienza.astronomia	26502
it.discussioni.ingegneria	26406
it.comp.os.linux.software	26333
it.cultura.cattolica	26304
it.comp.lang.delphi	26233
it.comp.software.emulatori	25796
it.arti.musica.strumenti.tastiere	25368
it.hobby.giardinaggio	25263
it.hobby.nautica	24857
it.fan.musica.u2	24640
it.comp.grafica.photoshop	24140
it.arti.musica.jazz	24069
it.tlc.telefonia	23882

it.comp.hardware.storage	23755
it.scienza.medicina	23685
it.annunci.usato	23266
it.comp.lang.c++	23133
it.economia.investire	22883
it.sport.arti-marziali	22422
it.economia	22318
it.fan.tv	22296
it.comp.giochi.sportivi	21990
it.arti.fotografia.segnalazioni	21763
it.arti.ballo.lat-americano	21752
it.comp.os.win.software	21732
it.hobby.elettronica.digitale	21477
it.tlc.gestori.fastweb	20801
it.tlc.cellulari.motorola	20757
it.arti.architettura	20718
it.discussioni.auto.ford	20655
it.sport.americani	20447
it.comp.software.cad	20264
it.cultura.religioni.cristiani	20107
it.sport.tennis	19571
it.sport.tenni	19565
it.sociale.obiezione	19478
it.sociale.obiezione	19478
it.comp.musica.mp3	19370
it.comp.giochi.simulatori	19330
it.cultura	19310
it.arti.fantasy	18856
it.discussioni.motori	18653
it.economia.borsa.estero	18547
it.scienza.fisica	18486
it.sport.formula1.moderato	18463
it.comp.hardware.schede-audio	18444
it.cultura.linguistica.inglese	18266
it.comp.programmare.win32	18116
it.arti.musica.strumenti.basso	17892
it.diritto.assicurazioni	17855
it.tlc.gestori.vodafone	17726
it.cultura.fantascienza	17673
it.arti.hiphop	17465
it.fan.radio.deejay	17463
it.comp.sicurezza.windows	17393
it.fan.starwars	17292
it.fan.musica.lucio-battisti	17218
it.fan.musica.lucio-battisti	17218
it.cultura.horror	17202
it.comp.www	17033
it.comp.reti.wireless	17004
it.discussioni.ristoranti	16865
it.comp.software.newsreader	16808

it.hobby.enigmi	16422
it.tlc.gestori.wind	16332
it.sport.nuoto	16326
it.comp.sicurezza.varie	16324
it.sport.sci	16167
it.arti.musica.strumenti	16076
it.arti.cartoni.mercatino	16052
it.hobby.fantasport	15991
it.news.net-abuse	15872
it.news.net-abuse	15872
it.fan.musica.battiato	15711
it.discussioni.sessualita	15706
it.discussioni.sentimenti	15683
it.hobby.giochi.gdr.dnd	15551
it.fan.tv.friends	15421
it.cultura.storia.militare	15157
it.comp.reti.cisco	15121
it.sport	14835
it.tlc.telefonia.voip	14752
it.fan.tv.buffy	14718
it.sociale.scout	14284
it.comp.lang.c	14053
it.hobby.vino	13967
it.cultura.antagonista	13739
it.discussioni.ingegneria.civile	13469
it.fan.musica.baglioni	13404
it.cultura.militare	13391
it.hobby.satellite-tv.digitale.mod	13190
it.diritto.internet	13136
it.cultura.newage	13085
it.hobby.radio-cb	12981
it.istruzione.universita.ingegneria	12943
it.scienza.ambiente	12917
it.discussioni.psicologia	12798
it.news.gruppi	12772
it.tlc.gestori.tim	12735
it.comp.software.database	12364
it.comp.hardware.dvd	12326
it.economia.banche	12283
it.news.aiuto	12280
it.comp.reti.ip-admin	12011
it.comp.software.mailreader	12001
it.politica.destra	11946
it.comp.hardware.scsi	11921
it.scienza.chimica	11799
it.fan.scrittori.tolkien	11513
it.discussioni.droghe	11442
it.discussioni.giustizia	11430
it.discussioni.giustizia	11430
it.fan.stephen-king	11412

it.comp.os.win.nt	11132
it.comp.programmare	10535
it.sport.atletica	10473
it.fan.musica.ligabue	10305
it.hobby.home-cinema.titoli-dvd	10099
it.discussioni.auto.mod	10083
it.istruzione.universita	9870
it.istruzione.universita	9870
it.discussioni.energie-alternative	9817
it.hobby.audiovisivi	9781
it.cultura.ebraica	9775
it.scienza.biologia	9612
it.salute.alimentazione	9202
it.salute.alimentazione	9202
it.fan.musica	9082
it.tlc.gestori.telecom	9012
it.cultura.letteratura.italiana	9006
it.comp.appl.macromedia	9002
it.comp.os.linux.development	8943
it.lavoro.prevenzione	8879
it.scienza.astronomia.amatoriale	8861
it.arti.varie	8842
it.comp.software.shareware	8501
it.comp.os.amiga	8491
it.comp.giochi.avventure.testuali	8469
it.lavoro.richieste	8465
it.hobby.piante.cactus	8273
it.politica.cattolici	8271
it.discussioni.animali	8223
it.sociale.handicap	8120
it.cultura.religioni.bahai	8108
it.discussioni.sogni	8066
it.comp.software.divx	8014
it.hobby.radioascolto	7920
it.comp.software.tex	7810
it.comp.giochi.sviluppo	7722
it.fan.musica.rem	7672
it.associazioni.cri	7652
it.hobby.armi.moderato	7427
it.comp.os.dibattiti	7380
it.comp.os.win.windows7	7295
it.politica.sinistra	7282
it.comp.giochi.rpg.ultimaonline	7227
it.discussioni.agricoltura	7205
it.comp.os.linux.debian	7162
it.hobby.vari	7139
it.comp.lang.perl	7132
it.comp.appl.notes-domino	7058
it.comp.appl.notes-domino	7058
it.faq	7030

it.hobby.aquiloni	7006
it.fan.tv.dawsons-creek	6972
it.sport.calcio.palermo	6903
it.comp.hardware.palmari.gps	6823
it.fan.musica.springsteen	6627
it.tlc.provider	6558
it.news.annunci	6470
it.cultura.cybersocieta	6428
it.discussioni.giallo	6332
it.fan.musica.carmen-consoli	6238
it.comp.software.emulatori.console-recenti	6224
it.discussioni.iso9000	6217
it.sport.rally	6146
it.fan.tv.er	6107
it.comp.os.win.windows10	6106
it.comp.os.win.windows10	6106
it.industria.elettrotecnica.normative	5998
it.fan.musica.elio	5922
it.hobby.viaggi.inter-rail	5905
it.hobby.viaggi.inter-rail	5905
it.cultura.filosofia.moderato	5871
it.fan.musica.de-andre	5714
it.media.tv.fantascienza	5622
it.news.moderazione	5462
it.cultura.linguistica	5389
it.comp.software.browser	5368
it.sociale.adozione	5354
it.scienza.astronomia.seti	5266
it.cultura.linguistica.giapponese	5208
it.comp.lang.vo-clipper	5179
it.sociale.primosoccorso	5155
it.arti.musica.classica.mod	5054
it.arti.fumetti.manga	5054
it.comp.os.os2	5013
it.cultura.cybersocieta.lamer	4810
it.cultura.cybersocieta.lamer	4810
it.arti.animazione	4588
it.comp.lang.pascal	4449
it.istruzione.scuola.informatica	4366
it.fan.marco-ditri	4163
it.comp.os.dos	4156
it.fan.tv.mai-dire-gol	4137
it.comp.os.win.vista	4130
it.arti.musica.polifonia	4092
it.hobby.creativi	3972
it.salute.aids	3865
it.cultura.classica	3661
it.lavoro.sindacato	3562
it.sport.rugby	3561

it.scienza.geologia	3422
it.hobby.volo.ultraleggero	3338
it.hobby.radioamatori.moderato	3326
it.news.gestione	3205
it.comp.sicurezza.pgp	3155
it.cultura.religioni.buddhismo	3069
it.comp.software.irc	2875
it.fan.radio	2867
it.comp.appl.eudora	2865
it.news.votazioni	2859
it.comp.os.linux.mandrake	2759
it.comp.os.linux.ubuntu	2660
it.tlc.gestori	2606
it.fan.nutella	2581
it.sociale.anorexbulimia	2546
it.scienza.informatica	2484
it.fan.tv.babylon5	2484
it.discussioni.energia	2403
it.fan.japan.sailor-moon	2315
it.salute.cefalee	2281
it.lavoro.professioni.pubblicita	2203
it.comp.os.linux.annunci	2072
it.fan.japan.r-takahashi	2000
it.comp.os.unix	1978
it.arti.musica.strumenti.voce	1974
it.comp.sicurezza.crittografia	1950
it.fan.matrix	1947
it.comp.software.libero	1935
it.politica.internazionale.israele	1922
it.discussioni.astrologia	1899
it.hobby.motociclismo.viaggi	1822
it.sport.golf	1700
it.sport.golf	1700
it.comp.sicurezza.unix	1668
it.cultura.storia.moderato	1560
it.cultura.linguistica.francese	1475
it.comp.os.linux.redhat	1438
it.fan.asimov	1343
it.arti.musica.strumenti.chitarra.mod	1180
it.hobby.robotica	923
it.arti.cinema.recensioni	890
it.fan.musica.pearl-jam	870
it.comp.accessibilita	804
it.arti.musica.studio	776
it.sociale.globalizzazione	707
it.sport.calcio.moderato	641
it.hobby.volo	638
it.fan.tv.scrubs	492
it.comp.dotnet	424
it.comp.dotnet	424

it.comp.virtualizzazione	390
it.comp.software.editor	267
it.comp.lang	252
it.scienza.divulgazione	240
it.lavoro.professioni	226
it.comp.os.openbsd	206
it.politica.m5s	152
it.scienza.medicina.tumori	123
it.tlc.provider.disservizi	0
it.sport.volley	0