I Langue Language Mordels quali chatGPT sono una figata, tecnologia allo stato dell'alrte in grado di aumentare la produttività di qualsiasi settore, rivoluzione digitale 2.0

Uno dei punti di forza di chatGPT è la possibilità di ottenere informazioni "chattando", in pratica utilizzare google in forma colloquiale.

Una delle limitazioni di chatGPT è che:
1. non sa tutto quello che sa google (non ha accesso ad internet),
2. non è possibile "insegnarli" qualcosa

Use case: avendo un documento o un ecosistema di documenti (tipo tutta la burocrazia, resoconti fiscali etc. di un'azienda), sarebbe figo dare in pasto tutti questi dati in formato **testuale** a chatGPT e poter chattarci per recuperare informazioni in maniera efficiente (senza leggermi tutto)

ad esempio, gli do 12503 pagine di testo del resoconto dell'andamento dell'azienda negli ultimi 12 anni, anziché leggermeli tutti per trovare l'informazione che cerco, chiedo direttamente a chatGPT qualcosa del tipo: "ascolta ma cos'è cambiato negli ultimi sei anni?"

Sebbene le limitazioni citate prima, ciò è possibile ed è quello che si offre di fare questa applicazione.

Come? Domandando a chatGPT una domanda come quella di prima "ascolta ma cos'è cambiato negli ultimi sei anni?", nella domanda gli si aggiunge anche un contesto da cui cercare e costruire una risposta

In altre parole, è come se nel farti una domanda io ti dessi anche la risposta senza saperlo, ti dico però anche informazioni che potrebbero anche non avere nulla a che fare con la domanda! Sarà chatGPT a capire cosa dare in output creando una risposta usufruendo delle sue incredibili capacità da cervellone digitale

Quindi per ricapitolare, dato in input un pdf del bando di un concorso per l'erasmus, ad ogni domanda dell'utente,

L'applicazione andrà a fare una **ricerca semantica** nel pdf per trovare le porzioni di testo più simili possibili alla domanda in questione, la logica è che più sono simili più è probabile che contengano anche la risposta!

se l'utente chiede
"quando scade l'iscrizione per partecipare al bando?"

a chatGPT arriverà il seguente prompt:

[2] "già fruiti nell'ambito dei precedenti Programmi comunitari. La durata della mobilità dipende non solo dal numero di mesi residui per il proprio ciclo di studio ma anche dalla durata della mobilità prevista dall'accordo inter-istituzionale relativo alla destinazione scelta, dunque se l'accordo è per un semestre la mobilità non può essere di un'annualità. Art. 4 Destinatari della mobilità Possono partecipare al presente bando gli studenti dell'Università degli Studi di Salerno che, nell'anno accademico 2022-2023, siano regolarmente iscritti a corsi di studio di I, II e III ciclo e partecipano al bando relativo allo stesso corso di studio per l'anno 2023/24. Per la partecipazione al bando è essenziale il possesso dei requisiti linguistici richiesti dall'Istituzione ospitante. Coloro che intendono laurearsi nella sessione straordinaria a.a. 2022/2023 dovranno aver concluso la mobilità entro il conseguimento della laurea. In caso di mancato conseguimento del titolo di laurea entro la sessione straordinaria a.a.2022/2023, sarà necessario formalizzare"

[2] "l'iscrizione all'a.a.2023/2024. N.B. È responsabilità di ogni candidato verificare il numero di mensilità Erasmus ancora disponibili per il ciclo di studi di riferimento, ai fini della candidatura al presente bando. In caso di falsa dichiarazione, la mobilità potrà essere annullata prevedendo la restituzione dei contributi erogati. A tal proposito si richiamano inoltre le responsabilità penali legate al rilascio di dichiarazioni mendaci (Codice penale e leggi speciali in materia, ai sensi e per gli effetti dell'art. 76 D.P.R. n. 445/2000). Art. 5 Scegliere l'Istituzione ospitante L'elenco delle destinazioni e dei tutor interni per i diversi accordi di mobilità con le Università straniere è allegato al bando. Le Università straniere di Programme e Partner countries definiscono le SCADENZE per: - Nomination (candidatura, ovvero ricezione dei nominativi) di coloro che effettueranno la mobilità Erasmus+ presso la propria Università. - Application (domanda di ammissione) di coloro che sono stati nominati. Questo processo è curato"

[3] "online di posizionamento che si terrà il 17 febbraio alle ore 15. I corsi avranno inizio a partire dal 1° marzo 2023 per gli studenti in partenza nel primo semestre. Gli studenti in partenza nel secondo semestre verranno riconvocati successivamente. Art. 6 Candidatura La candidatura, la cui scadenza è fissata per il giorno 6 febbraio 2023 – ore 23:59, deve essere compilata online entrando con le proprie credenziali nell'area riservata di ESSE3. Nella sezione "Mobilità internazionale" -> "Bandi di mobilità" lo studente accederà al bando inerente il proprio ambito disciplinare e potrà scegliere, in ordine di preferenza e per il ciclo di studi di riferimento 3 istituzioni ospitanti tra i Programme countries e altrettante per i Partner countries. Per tutti i corsi di doppio e triplo titolo la scelta della sede è ovviamente vincolata. Gli studenti che presentano una doppia candidatura (per Programme countries e Partner countries) devono informare la"

[9] "non viene corrisposta per il periodo trascorso in Italia o in un paese diverso da quello di assegnazione della mobilità. Art. 28 Trattamento dei dati Ai sensi del Regolamento UE del Parlamento Europeo e del Consiglio in data 27 aprile 2016 (pubblicato nella Gazzetta Ufficiale dell'Unione Europea n° L. 119/1 del 4 maggio 2016) l'Università degli Studi di Salerno garantisce che il trattamento dei dati personali sarà improntato a principi di liceità, correttezza e trasparenza nei confronti dell'interessato. In particolare, i dati personali saranno raccolti in maniera adeguata, pertinente e limitata alle finalità connesse e strumentali al presente bando di concorso ed all'eventuale gestione del rapporto con l'Ateneo, e successivamente trattati in modo compatibile con tale finalità. Essi saranno conservati in una forma che consenta l'identificazione degli interessati per un arco di tempo non superiore al conseguimento delle suddette finalità, nonché per fini statistici, previa adozione di misure tecniche e"

[8] "contributi sarà comunicata direttamente ai vincitori. Qualora la copertura finanziaria non fosse disponibile del tutto o in parte, sarà comunque possibile svolgere la mobilità assegnata sostenendo autonomamente le spese. Si precisa che, in caso di insufficienza dei fondi, il numero di mensilità coperte dal contributo, previsto per il Paese di destinazione, potrà essere ridotto proporzionalmente. Art. 20 Modalità di erogazione del contributo Il contributo sarà erogato in due rate. La prima rata si compone del 100% di fondi UE e del 50% di fondi MUR previsti per fascia di ISEE. 8 I fondi MUR saranno erogati esclusivamente a favore di coloro che alla data di stipula del contratto avranno presente in Esse3 l'ISEE relativo all'anno di emanazione del bando. Per i contratti del primo semestre o annuali l'ISEE del 2023 dovrà essere disponibile entro il 1 giugno 2023. L'importo della prima rata, moltiplicato per i mesi di permanenza stabiliti"

Instructions: Compose a comprehensive reply to the query using the search results given. Cite each reference using [Page Number] notation (every result has this number at the beginning). Citation should be done at the end of each sentence. If the search results mention multiple subjects with the same name, create separate answers for each. Only include information found in the results and don't add any additional information. Make sure the answer is correct and don't output false content. If the text does not relate to the query, simply state 'Text Not Found in PDF'. Ignore outlier search results which has nothing to do with the question. Only answer what is asked. The answer should be short and concise. Answer step-by-step.

un quantitativo immane di parole che, **se e sottolineo SE** conterrà la risposta alla domanda, chatGPT sarà in grado di trovare:

> "La scadenza per la candidatura è fissata per il giorno 6 febbraio 2023 alle ore 23:59 "

Si può dire quindi che la ciccia della nostra applicazione è trovare le porzioni di testo interessanti, adottando le tecniche di:
ricerca semantica e di **embedding**

Il flusso dell'applcazione lo si può suddividere in 4 fasi:

1. INPUT

L'utente da in input due componenti:

Documento Testuale

Domanda

2. EMBEDDING

per l'embedding dei documenti e della domanda,
utilizzeremo un modello allo stato dell'arte
che si chiama SBERT "all-MiniLM-L6-v2"
https://www.sbert.net/docs/pretrained_models.html

3. RICERCA SEMANTICA

Per trovare le porzioni "chunk" dei documenti
da passare assieme alla domanda allo step successivo
calcolo della cosine similarity tra i vettori

```
# quando la classe viene usata come metodo, calcolo la cosine similarity tra gli emb
def __call__(self, text): # text è la domanda input dell'utente
    top_k = min(5, self.dim_corpus)
    domanda_embeddings = self.encode([text]) # embedding applicato alla domanda
    cos_scores = util.cos_sim(domanda_embeddings, self.corpus_embeddings)[0]
    top_results = torch.topk(cos_scores, k=top_k)
    return [self.data[i] for i in top_results[1]]
```

4. EMBEDDING

Chiamata alle api di chatGPT col prompt appena creato, restituzione dell'output all'utente con la risposta contestualizzata alla sua domanda

EMBEDDING

Nei problemi di NLP, uno dei primi problemi da affrontare è la codifica del testo in un formato che una macchina sia in grado di comprendere. Occorre adottare una strategia di codifica.

Soluzione primitiva:

ONE-HOT ENCODING

Supponiamo un "dataset" da 4 frasi:

- Il cane abbaia
- Il cane morde
- Il cane corre
- Il gatto morde

A valle di un processo di pulizia del testo (rimozione di stop words, maiuscole, plurali etc.) in un primo momento si andrà a creare i "token" del dataset. Il testo lo si può tokenizzare in più modi: prendendo due parole, tre, un singolo carattere, etc.
sia un token = una parola, avremo:

token	frequenza
cane	3
morde	2
gatto	1
abbaia	1
corre	1

Soluzione primitiva:

ONE-HOT ENCODING

Supponiamo un "dataset" da 4 frasi:

- Il cane abbaia
- Il cane morde
- Il cane corre
- Il gatto morde

token	frequenza
cane	3
morde	2
gatto	1
abbaia	1
corre	1

L'HOT ENCODING è molto semplice, se una frase contiene il token, verrà assegnato 1 all'apposita colonna, 0 altrimenti

msg_id	token_count	cane	morde	gatto	abbaia
1	2	1	0	0	1
2	2	1	1	0	0
3	2	1	0	0	0
4	2	0	1	1	0

Il risultato è una matrice sparsa contenente prevalentemente zeri. Siamo riusciti ad ottenere una rappresentazione numerica del nostro dataset, un input adatto da dare in pasto ad un ipotetico modello neurale.

Problemi e limitazioni:

Dimensionalità:

- al modello arriveranno n input dove n sarà il numero di token coinvolti!

Consideriamo le due frasi: "amo il cioccolato ma odio la focaccia" e "amo la focaccia ma odio il cioccolato"
con la rappresentazione dell'hot encoding le due frasi risultano identiche, nonostante però abbiano significato opposto!

Inoltre la tecnica dell'hot encoding non prende in considerazione l'ordine delle parole nella frase
E non viene preso in considerazione neanche la frequenza di utilizzo del singolo token nella frase

EMBEDDING

Nei problemi di NLP, uno dei primi problemi da affrontare è la codifica del testo in un formato che una macchina sia in grado di comprendere. Occorre adottare una strategia di codifica.

word2vec

anziché andare a rappresentare ogni token con un banale flag 0 | 1, nel 2013 è stato creato word2vec un modello in grado di rappresentare ogni parola con un vettore

$$\vec{king} = \begin{bmatrix} 0.125976562 \\ 0.0297851562 \\ 0.00860595703 \\ \dots \\ 0.251953125 \end{bmatrix}$$

Applicare quindi il Coseno di similitudine per determinare la similarità tra le parole

https://arxiv.org/pdf/1301.3781.pdf https://code.google.com/archive/p/word2vec/

EMBEDDING

Nei problemi di NLP, uno dei primi problemi da affrontare è la codifica del testo in un formato che una macchina sia in grado di comprendere. Occorre adottare una strategia di codifica.

SBERT

WIP