



Challenge US - SpazioDati

LLM per l'estrazione di nomi e impieghi

Leonardo Calamita - Paolo Ferragina



Outline

- Introduzione
- Preprocessing
- ChatGPT
- Confronti e analisi
- Conclusioni

Introduzione al problema

Data una pagina web (possibilmente) di contatti, estrarre da essa tutti i nomi di persona e i loro ruoli aziendali.

SpazioDati offre un software aziendale che permette di svolgere questo task.

Obiettivi del ChallengeUS:

- Quanto sono **efficaci** i LLM per questo task rispetto all'approccio di SpazioDati?
- Quanto sono **efficienti** e dunque utilizzabili in pratica i LLM rispetto all'approccio adottato da SpazioDati?

Introduzione al problema

Data una pagina web (possibilmente) di contatti, estrarre da essa tutti i nomi di persona e i loro ruoli aziendali.

SpazioDati offre un software aziendale che permette di svolgere questo task.

Obiettivi del ChallengeUS:

- Quanto sono **efficaci** i LLM per questo task rispetto all'approccio di SpazioDati?
- Quanto sono **efficienti** e dunque utilizzabili in pratica i LLM rispetto all'approccio adottato da SpazioDati?



Preprocessing del dataset fornito

Utilizzate le prime **100 entry**

- meno risorse necessarie
- nessun training da effettuare

Dati considerati:

- codice HTML delle pagine
- persone
 - nome
 - ruolo

Testo estratto con *BeautifulSoup*¹

- tutto il testo presente nella pagina viene estratto, indipendentemente da dove si trovi

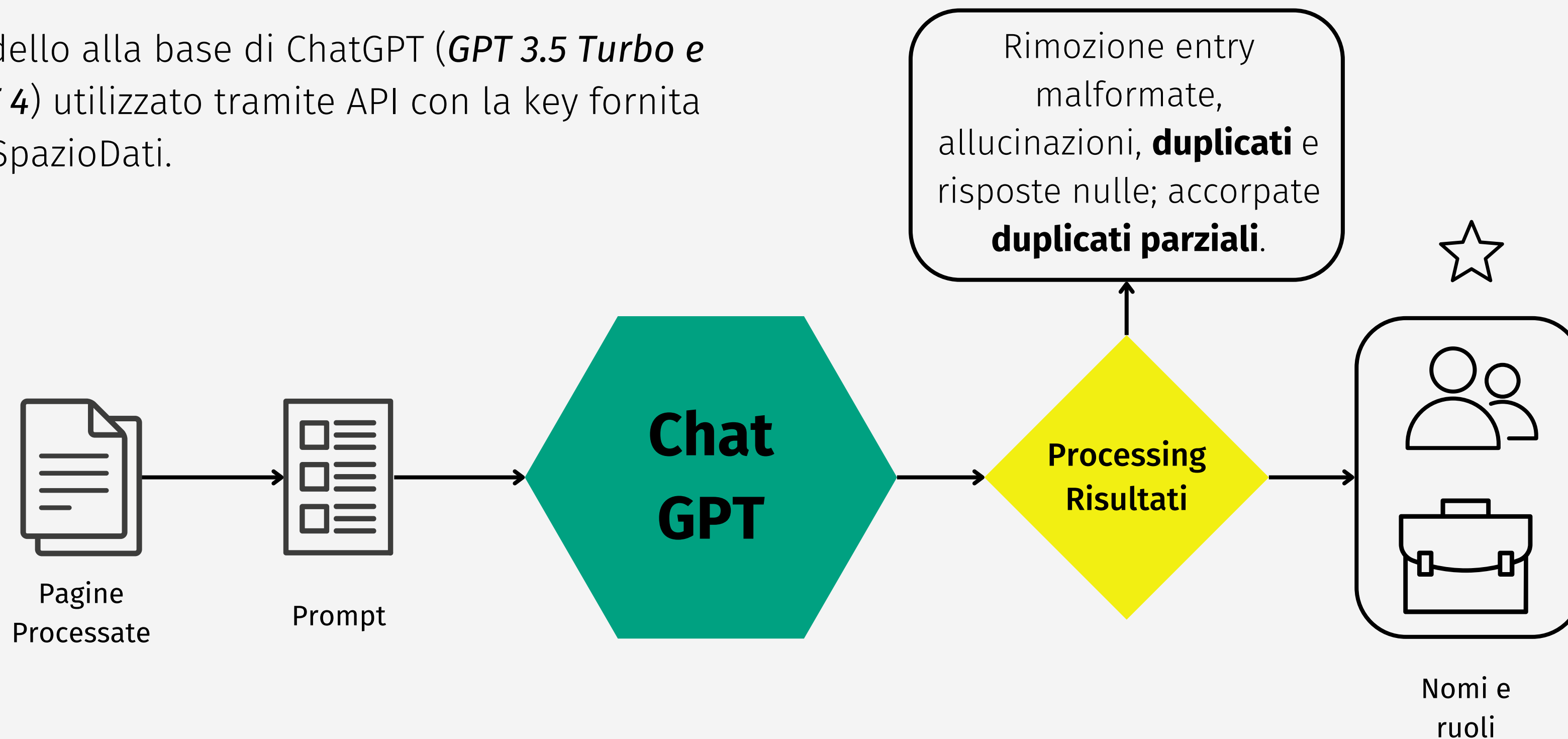
Pulizia generale:

- "\n" -> "."
- rimozione caratteri poco comuni
- PAROLE MAIUSCOLE -> Parole
"Titolate"

1: <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>

ChatGPT

Modello alla base di ChatGPT (*GPT 3.5 Turbo e GPT 4*) utilizzato tramite API con la key fornita da SpazioDati.



Prompt Utilizzato

System: Sei un analizzatore testuale di contenuti di pagine web.

Estrai tutti i nomi propri delle persone dal testo che ti verrà fornito e le rispettive posizioni lavorative.

Il testo in questione è stato estratto da una pagina di contatti del sito web di un'azienda. Le informazioni da estrarre sono necessarie a creare una lista di persone che lavorano per l'azienda.

Se la posizione lavorativa di una persona non è chiara o non è riportata, scrivi solo "?".

La tua risposta deve essere formata unicamente dalla lista dei nomi delle persone, seguita da ":" e poi dalla posizione lavorativa. Ad esempio: "Mario Rossi: CEO" o "Luigi Verdi: ?".

Rispondi solo con la lista di nomi e ruoli lavorativi, non aggiungere commenti o altro.

Di seguito riporto un esempio per aiutarti nell'estrazione di informazioni dal testo reale.

[...]

Analisi quantitativa

	SpazioDati	ChatGPT 3.5*	ChatGPT 4
# persone estratte	534	783	1007
# ruoli estratti	280	543	817
# pagine senza risultati	0	15	5

Maggior numero di nomi e ruoli estratti, ma possibili allucinazioni/errori

Dati basati sui risultati estratti dalle prime 100 pagine del dataset.

**ChatGPT 3.5 utilizzato sui dati estratti con Trafilatura, per confronto.*

Confronto con SpazioDati (1)

	ChatGPT 3.5*	ChatGPT 4
# match esatti nome + ruolo	109 (20,4%)	151 (28,3%)
# match parziali nome + ruolo	162 (30,3%)	236 (44,2%)
# match esatti nome	256 (47,9%)	483 (90,4%)
# match parziali nome	380 (71,1%)	511 (95,7%)

Molti match sui nomi, molti meno sui ruoli; le differenze però non sono necessariamente causate da errori di GPT

Dati basati sui confronti con le 534 persone estratte da SpazioDati, utilizzati come ground truth.

**ChatGPT 3.5 utilizzato sui dati estratti con Trafilatura, per confronto.*

Confronto con SpazioDati (2)

Persone estratte dalle pagine	ChatGPT 3.5*	ChatGPT 4
Più persone estratte rispetto alla baseline	28	51
Meno persone estratte rispetto alla baseline	66	4

Dati basati sui confronti con le 100 pagine prese in esame, rispetto alle persone estratte da SpazioDati.

Casi con meno estrazioni di SpazioDati

4 casi in cui GPT4 ha estratto meno persone della baseline:

- 3 errori nella baseline
 - a. nome: “Urbano Centro”
 - b. nome: “Maria Elena”
 - i. manca il cognome, nome completo poi presente correttamente
 - c. nome: “Evaristo Dandini”
 - i. nome di una circolo didattico
- 1 pagina senza nessun nome nella pagina html

Confronto con ground truth

Stat	Nomi	Nomi+Ruoli
Ground truth	250	172
SpazioDati	134	37
SpazioDati (parziali inclusi)	136	59
GPT4	247	164

Dati basati su confronto manuale con 35 pagine randomicamente estratte tra quelle fornite.

Analisi qualitativa (1)

Esempio 1: testo molto schematico con pochi nomi

Neox . Trentino Alto Adige . Bolzano: Tecnoassistenza Snc 0471 [Num] . Veneto . Venezia
(Centro storico e isole): C.A.R.E. di Pieretto Fabrizio 041 [Num] ...

Output:

- **risultato corretto:** 5 nomi, nessun ruolo
- **base:** 2 nomi, nessun ruolo
- **ChatGPT 3.5:** 3 nomi, nomi di aziende o città come ruoli
 - 1 allucinazione
- **ChatGPT 4:** 10 nomi, nomi di aziende o città come ruoli
 - 4 allucinazioni (3 nomi di aziende, 1 solo nome senza cognome)
 - 1 nome estratto da una mail

Analisi qualitativa (2)

Esempio 2: testo lungo e con alto numero di nomi, molti dei quali stranieri

La Sig.ra Rossana Carenza insieme al consorte Donato Accettura decisero di avviare un'attività imprenditoriale nel 1968. All'interno di una struttura di proprietà, una splendida villa fine Ottocento, con un parco di 3000 metri ...

Output:

- **risultato corretto:** 62 nomi, ruoli generici o non chiaramente specificati
- **base:** 26 nomi, nessun nome straniero e nessun ruolo
- **ChatGPT 3.5:** 59, ruoli in buona parte assegnati
- **ChatGPT 4:** 53 nomi, tutti i ruoli assegnati
 - alcuni ruoli generici come “ospite” o “artista di fama internazionale”

Analisi qualitativa (3)

Esempio 3: testo semplice, nomi senza ruoli

The Butcher Just Fly Fishing è un negozio specializzato per i pescatori a mosca appassionati di viaggi. Vendita di attrezzatura specifica, ...

Output:


- **risultato corretto:** 3 nomi, nessun ruolo
- **base:** 2 nomi, nessun ruolo
 - un nome non estratto perchè non capitalizzato
- **ChatGPT 3.5:** nessun nome (testo mancante)
- **ChatGPT 4:** 3 nomi, nessun ruolo

Conclusioni

- Unione di tre nuovi fattori per il miglioramento dei risultati:
 - a. Estrazione di tutti i dati testuali dalle pagine
 - b. Utilizzo di GPT4
 - c. Prompt utilizzato con il ruolo *system*
- Rivelati un maggior numero di errori e di problematiche relative alla baseline
- Le allucinazioni in qualche caso sono ancora un problema
 - Spesso individuabili con check su caratteri presenti nei nomi o sul numero di token
- GPT4 non velocissimo (100 pagine -> circa 15 minuti) e costoso (0.06\$ ogni 1000 token in output)

Possibili sviluppi

- sfruttare GPT4 per creare un dataset affidabile su cui poter poi effettuare il fine tuning di un altro LLM open source (ad es. Llama 2)
- approccio ibrido in cui le pagine “semplici” vengono analizzate dal sistema di SpazioDati, mentre le pagine più “complesse” vanno invece analizzate da GPT4
 - possibilità di usare altri modelli più *economici* per classificare la pagina
- riconoscere i nomi tramite un sistema NER ed estrarre solo le parti più rilevanti del testo (ad es. finestra di 20 parole attorno al nome) prima di passarle a GPT4, per diminuire il numero di token da processare



**Grazie per
l'attenzione!**