



Chatdocument

Andrea Silingardi & Federico Boni
5F, ITIS E.Fermi, 21/05/2025

OBBIETTIVI

- Aiutare l'azienda a ritrovare documenti importanti sparsi in tante cartelle.
- Usare l'intelligenza artificiale per cercare i documenti giusti a partire da una domanda scritta in modo naturale.
- Creare un sito semplice da usare dove fare queste ricerche.
- Offrire un servizio utile per conservare e usare meglio la conoscenza interna all'azienda.



PROBLEMA



- Le aziende producono tantissimi documenti ogni giorno.
- Spesso non vengono organizzati bene e si perdono nei meandri dei server.
- Quando servono, nessuno sa più dove sono o come trovarli.
- Questo fa perdere tempo, soldi e conoscenze preziose.





SOLUZIONE PROPOSTA

TRASFORMAZIONE TESTO

Si prelevano tutti i PDF sparsi e li si trasforma in testo.

ALGORITMO DI RICERCA

un algoritmo scritto da noi trova le parole importanti

RICERCA

Ricerca fatta sul testo dei documenti basata sull'inserimento dell'utente

FUNZIONALITA' IMPLEMENTATE

1. Conversione Automatica di Documenti

- Estrazione del testo da PDF.
- Salvataggio dei testi estratti in formato .txt per l'elaborazione successiva.

2. Indicizzazione con TF-IDF

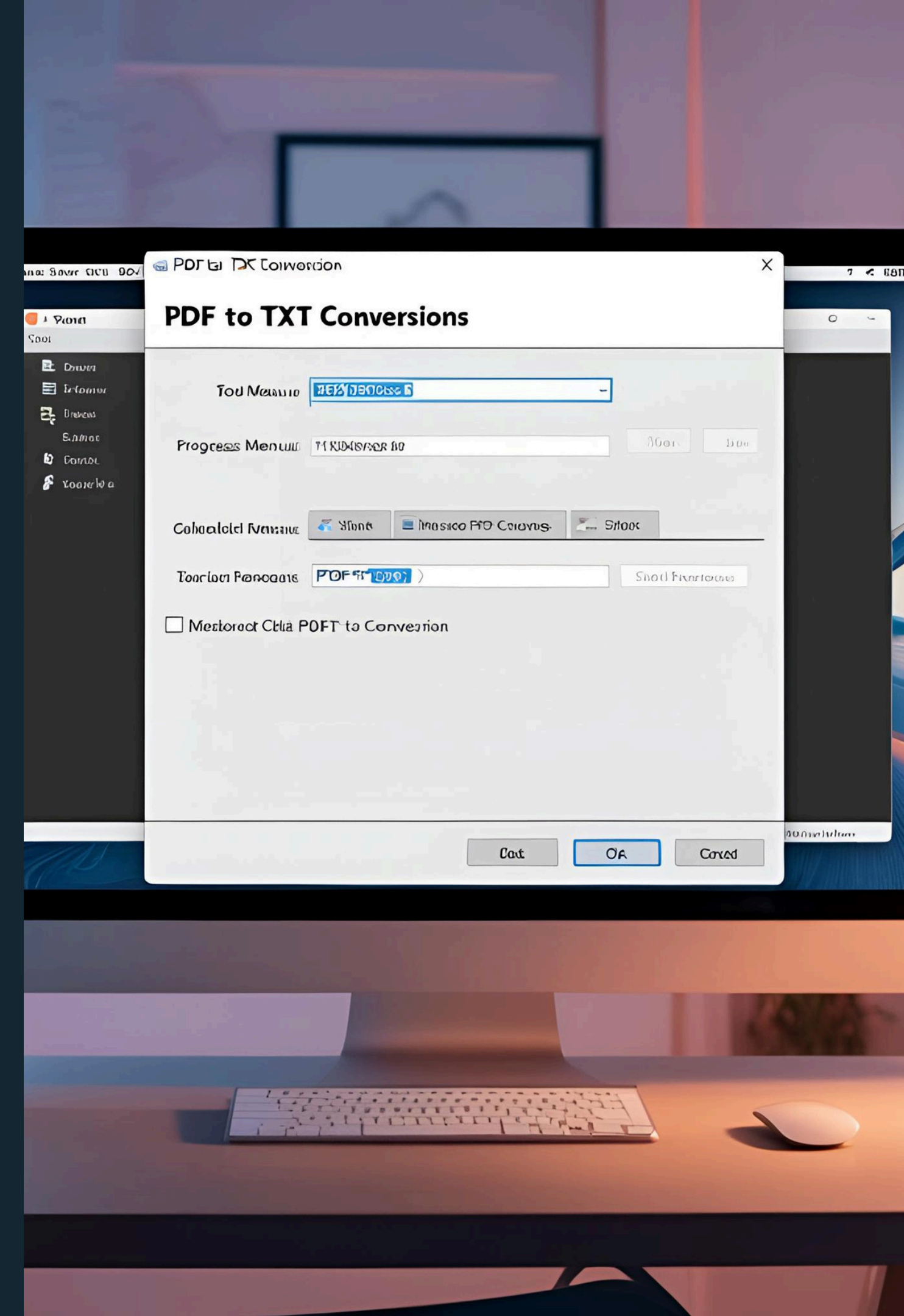
- Utilizzo di scikit-learn per analizzare i testi e calcolare la rilevanza dei termini (TF-IDF).
- Classifica dei documenti in base alla pertinenza rispetto alle query degli utenti.

3. Interfaccia Web Sicura

- Autenticazione utente con crittografia delle password (via bcrypt).
- Ricerca in tempo reale con visualizzazione dei risultati ordinati per punteggio TF-IDF.
- Pagine dedicate a registrazione, login, contatto e gestione errori (es. 404 personalizzato).
- Possibile download dei pdf a seconda della query

4. Gestione di Documenti Non Strutturati

- Supporto per una base documentale disordinata, senza necessità di metadati o struttura predefinita.



ARCHITETTURA DEL SISTEMA

Backend:

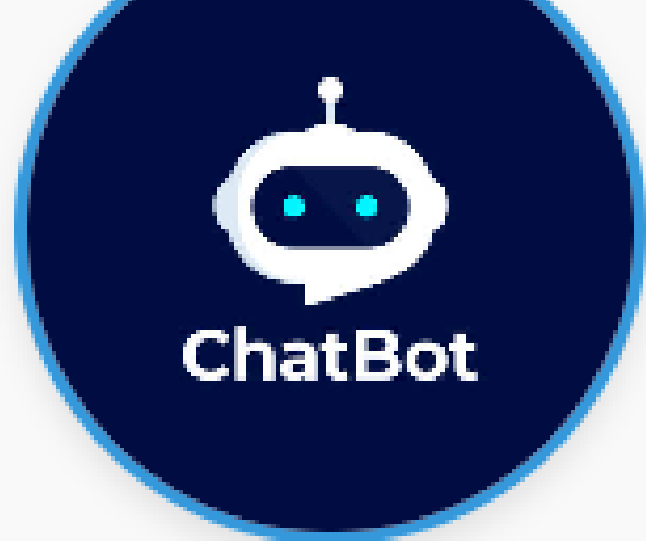
- Framework Tornado per gestire richieste HTTP, sessioni utente e rendering dei template.
- Algoritmo TF-IDF per l'analisi semantica semplificata.

Frontend:

- Template HTML dinamici con integrazione di CSS/JS per un'esperienza utente intuitiva.
- Pagine responsive per ricerca, login, registrazione e contatto.

Persistenza Dati:

- Archiviazione degli utenti in formato JSON.
- Salvataggio dell'indice TF-IDF serializzato (.pkl) per ottimizzare le ricerche successive.



ChatBot

 Username

 Password

 Accedi

ESEMPIO DI UTILIZZO

1. Un utente si registra e accede al sistema.
2. Digita una query (es. "clientex").
3. Il sistema restituisce i documenti più rilevanti, con snippet e punteggio di rilevanza.
4. L'utente può esplorare i risultati e accedere direttamente ai documenti originali.

Risultati per: "prova"

Documento	Dettagli	Scarica
Enrico_Fermi	Termine: 'prova' - Punteggio: 0.0114	 PDF
DIALOGO SOPRA I DUE MASSIMI SISTEMI	Termine: 'prova' - Punteggio: 0.0065	 PDF
MANZONI-PROMESSI SPOSI	Termine: 'prova' - Punteggio: 0.0049	 PDF

TECNOLOGIE UTILIZZATE

- Python: Scelta per la sua versatilità nelle librerie AI e automazione.
- Scikit-learn: Per l'implementazione dell'algoritmo TF-IDF.
- Tornado Web Server: Gestione efficiente delle connessioni e sicurezza integrata (es. cookie sicuri).
- NLTK: Pulizia del testo (rimozione stopword in italiano).
- Pickle: viene usato per salvare e caricare l'indice TF-IDF.





VANTAGGI PER LE AZIENDE

- **Recupero Efficiente:** Trovare informazioni in secondi, anche in basi documentali caotiche.
- **Riduzione dei Costi:** Elimina ore di ricerca manuale.
- **Sicurezza:** Dati utente protetti con crittografia avanzata.
- **Scalabilità:** Aggiunta di nuovi documenti senza riprogettare l'infrastruttura.

SECURITY

Viene utilizzata la libreria bcrypt per gestire in modo sicuro le password degli utenti.

Quando un utente si registra, la sua password non viene salvata in chiaro, ma viene:

1. Convertita in una sequenza di byte.
2. "Hashata" (cioè trasformata) con l'algoritmo bcrypt, che include anche un "sale" (salt) casuale per aumentare la sicurezza.
3. Salvata nel file users.json sotto forma di stringa crittografata (password_hash).



```
    "password_hash": "$2b$12$aw1WPYbT1xeSJUTFyAXR1.hDWjS.2Lf7TI fzJyZKR.m09oF0VvRiG",
    "email": "andreasilingardi@icloud.com",
    "created_at": "2025-04-01T17:09:14.277519"
  },
  "federico": {
    "password_hash": "$2b$12$asQyI/i89eW/NdtL5YxY/.WK.MJxcDx4ppTV/uip2zaIlF5KD0U2a",
    "email": "federico.boni2000@gmail.com",
    "created_at": "2025-04-02T08:17:22.045287"
  },
  "dhd": {
    "password_hash": "$2b$12$3Czz2wdSlfJAeGI2sjoQZ..rMHaPS29S72N/pI7q8xhQuB0YYRjXu",
    "email": "dhd@gmail.com",
    "created_at": "2025-05-13T09:40:49.658203"
  },
  "silli06": {
    "password_hash": "$2b$12$5h.RjL0FlhxBRlbe5T0aCO35dWwCAvWVXLs4ptMJvSpqmP81BxB.",
    "email": "dhds@gmail.com",
    "created_at": "2025-05-13T09:43:23.391285"
  },
  "fede": {
    "password_hash": "$2b$12$GGyHJrSwHYMTzEHBA4D/Fu/I7JWvFHQ2iDrGiAt/vj01DB.NL03ni",
    "email": "fede.boni@gmail.com",
    "created_at": "2025-05-13T14:20:53.365217"
  },
  "toni": {
    "password_hash": "$2b$12$ax7L9g0g.XemjUcYgf0zE.rtQXmiMl18u4IeS35pFRQbZebbFh5F0",
    "email": "toniandreasilingardi2000@gmail.com",
    "created_at": "2025-05-17T09:54:24.055431"
  }
}
```

ESEMPIO LIVELLO DI SICUREZZA

SVILUPPI FUTURI

- Integrazione di modelli NLP avanzati per comprendere query complesse.
- Aggiunta di un motore di ricerca semantico.
- Dashboard amministrativa per monitorare l'utilizzo e gestire i documenti.
- Implementazione di API REST per integrazione con altri applicativi aziendali.



Grazie per l'attenzione