**Catalogatore - Guida all’uso**

**Descrizione sintetica**

Il *Catalogatore* è uno strumento software per l’analisi comparativa e la clusterizzazione di insiemi documentali. Sviluppato per applicazioni umanistiche e digitali, consente di gestire in sequenza un processo automatico di estrazione testuale, calcolo di hash, costruzione di matrici binarie e di similarità, e visualizzazione grafica dei risultati. È progettato per facilitare attività di esplorazione, confronto e ricostruzione di relazioni tra file testuali.

**1. Requisiti**

**Software:**

* **Python** 3.9 o superiore
* **LibreOffice** installato localmente (richiesto per l’estrazione del testo)

**Dipendenze Python (in requirements.txt):**

* pandas
* numpy
* matplotlib
* seaborn
* scikit-learn
* scipy
* ttkthemes

**2. Struttura del processo**

Il workflow è suddiviso in sei fasi indipendenti e modulari, attivabili da interfaccia:

| **Fase** | **Descrizione** |
| --- | --- |
| **prepare** | Crea la struttura delle directory di lavoro |
| **extract** | Estrae il testo dai documenti (PDF, DOCX, ODT) tramite LibreOffice |
| **hash** | Calcola gli hash dei contenuti estratti, utili per la comparazione binaria |
| **binary** | Genera una matrice binaria documento × hash |
| **similarity** | Calcola una matrice di similarità tra documenti |
| **visualize** | Visualizza la matrice di similarità tramite heatmap, dendrogrammi, MDS o PCA |

**3. Utilizzo dell’interfaccia**

1. Selezionare la cartella **sorgente** contenente i documenti.
2. Selezionare la cartella **di destinazione** per l’output.
3. Scegliere la **fase da eseguire**.
4. Cliccare su **Esegui**.

L’interfaccia include anche:

* Selettore del **tipo di visualizzazione** (heatmap, dendrogramma, PCA, MDS)
* Opzioni di visualizzazione:
  + Attivazione/disattivazione dei numeri nella heatmap
  + Dimensione del font numerico regolabile
  + Precisione decimale selezionabile
* Menu per il **cambio dinamico del tema grafico**

**4. Output generati**

| **Cartella** | **Contenuto** |
| --- | --- |
| 1-extract/ | Testo estratto dai documenti |
| 2-hash/ | File CSV con hash |
| 3-binary/ | Matrice binaria dei documenti |
| 4-similarity/ | Matrice CSV con indici di similarità tra documenti |
| (visualizzazioni grafiche) | Mostrate a video, non salvate automaticamente |

**5. Note tecniche**

* Il percorso di LibreOffice è configurato per Windows (C:\Program Files\LibreOffice\...) ma può essere modificato nel codice.
* Tutti i passaggi sono modulari e richiamabili singolarmente.
* L’interfaccia è pensata per essere accessibile anche a utenti senza competenze di programmazione.