

DOCUMENTO DEI REQUISITI

Il sistema permette di eseguire il download di articoli da diversi sorgenti e la loro persistenza su file (il formato di deve essere uguale per tutti gli articoli di tutte le sorgenti), oppure di estrarre i termini con il peso maggiore (ovvero con la presenza in un maggior numero di articoli) a partire dal file in cui vengono memorizzati gli articoli, o entrambe le operazioni. I termini con il peso maggiore vengono memorizzati in un file di testo .txt nel formato "<termine> <peso>".

- Il sistema deve poter supportare nuove sorgenti, in particolare è richiesto di considerare due sorgenti da cui estrarre/scaricare gli articoli: file CSV e The Guardian API.
- Il sistema deve poter supportare nuove modalità di memorizzazione ed accesso agli articoli.
- Il sistema deve poter supportare nuove strutture per memorizzare ed avere accesso ai termini più importanti.

Use Case 1. L'utente sceglie di recuperare un determinato numero di articoli dal file .csv in cui sono stati salvati, per poi aggiungerli a un altro file locale in formato .csv (vedi nota finale). Il file sorgente e quello di destinazione vengono specificati come argomenti di comando dall'utente o vengono recuperati dal file delle proprietà. Il sistema stampa un messaggio di errore nel caso di problemi in apertura/lettura/chiusura del file sorgente o apertura/scrittura/chiusura del file destinazione. Il formato del file sorgente deve essere uguale a quello fornito nel file .csv presente nella cartella example_assets.

Use Case 2. L'utente sceglie di recuperare un determinato numero di articoli tramite le API di The Guardian, per poi salvarli in un file locale in formato .csv (vedi nota finale). La query e il file di destinazione vengono specificati come argomenti di comando dall'utente o vengono recuperati dal file delle proprietà, mentre l'API key va aggiunta necessariamente nel file delle proprietà. Il sistema stampa un messaggio di errore nel caso in caso di errori durante la chiamata alle API e durante il recupero degli articoli.

Use Case 3. L'utente sceglie di leggere il file in formato .csv su cui sono stati salvati degli articoli per poi estrarne una lista dei termini con "più peso" (numero di documenti in cui appaiono) e salvarla in un file .txt. Il sistema carica una stoplist da file locale contenente termini generici da trascurare nella fase di estrazione dei termini più pesanti. Il sistema stampa un messaggio di errore nel caso di problemi in apertura/lettura/chiusura del file sorgente o apertura/scrittura/chiusura del file destinazione, oppure nel caso in cui il file contenente la stoplist non venga trovato. I formati del file sorgente e del file contenente la stoplist devono essere uguali a quelli forniti nel file .csv e .txt presente nella cartella example_assets.

Use Case 4. (Combinazione dei precedenti) L'utente sceglie di recuperare un determinato numero di articoli tramite le API di The Guardian e/o dal file .csv in cui sono salvati, per poi aggiungerli in un file locale in formato .csv. Da quest'ultimo viene infine estratta una lista dei termini più pesanti, che viene salvata in un file .txt.

- NOTA: gli articoli vengono salvati su di un file, saved_articles.csv, che si trova nella cartella principale della repository, la scrittura è impostata su append. Se si vuole eliminare gli articoli salvati in precedenza sarà necessario eliminare tale file. Il programma non impedisce il salvataggio di articoli in multipla copia.

