Ingegneria dei dati 2024/2025

Homework 2

(da svolgere in gruppo)

Paolo Merialdo

Homework 2

1. scrivere un programma Java che indicizza i file .html disponibili a questo indirizzo: all_htmls

   Considerare almeno due campi (e quindi creare gli indici corrispondenti). Ad esempio, oltre al campo corrispondente all'intero contenuto del file (trascurando i tag html), si può considerare un campo per il titolo e uno per gli autori (o per l'abstract). Per ciascun campo utilizzare un analyzer appropriato

2. scrivere un programma Java che legge una query da console, interroga l'indice e stampa il risultato. Usare una semplice sintassi per la query (ad esempio, una query inizia con la parola chiave nome o contentuto seguita da una sequenza di termini (eventualmente racchiusi tra virgolette per esprimere una phrase query) oppure scrivere una piccola applicazione web in cui le interrogazioni possono essere scritte attraverso una form

3. testare il sistema con una decina di query diverse

Scrivere una relazione che, oltre a riportare l'url del proprio progetto su Github (o analogo) descriva:

- gli analyzer che si è scelto di utilizzare (motivando le scelte)
- il numero di file indicizzati, i tempi di indicizzazione e altri risultati sperimentali ritenuti interessati
- le query usate per testare il sistema

Homework 2

Termini di consegna: inviare la relazione entro
le ore 20:00 del 12 novembre 2024 attraverso il seguente modulo:

https://forms.office.com/e/5nmvtKgY11