Ingegneria dei dati 2024/2025

Homework 3

(da svolgere in gruppo)

Paolo Merialdo

Homework 3

Dobbiamo realizzare un sistema di data discovery che permetta di cercare risultati scientifici di interesse in un corpus di documenti.

Nel nostro sistema una query può essere composta da un insieme di termini che descrivono metriche e proprietà di un risultato scientifico (ad es.: record linkage f1 sul dataset wdc).

Poiché i risultati scientifici sono tipicamente rappresentati in tabelle, il risultato di una query dovrà essere un insieme (possibilmente ordinato per pertinenza) di tabelle.

Per rendere più efficiente il sistema, le tabelle presenti negli articoli del nostro corpus sono già state estratte, e organizzate in un repository. Ad ogni tabella sono associate informazioni di interesse, quali didacalia, paragrafi che citano la tabella, note a piè di pagina.

Come caso di studio usiamo i dati disponibili a questo indirizzo: all_htmls

Homework 3

- Progettare e implementare il sistema descritto nella slide precedente
- Definire una metodologia di valutazione del sistema. Per valutare la qualità dei risultati, valutare l'adozione delle metriche: Mean Reciprocal Rank (MRR) e Normalized Discounted cumulative gain (NDCG)

Homework 3

- Preparare una presentazione di 10' (.pdf or .pptx) ed una relazione di max 8 pagine (word o pdf) che descrivano gli aspetti rilevanti del vostro lavoro:
  - Definizione del problema
  - Descrizione della soluzione adottata
  - Valutazione sperimentale
    - Metodo di valutazione
    - Risultati
  - Limiti della soluzioni e possibili sviluppi futuri
- La relazione deve contenere anche un link alla repo del codice

Homework 3

Termini di consegna: inviare la relazione entro
le ore 20:00 del 29 novembre 2024 attraverso il seguente modulo:

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