Ingegneria dei dati 2025/2026

Homework 5

(da svolgere in gruppo)

Paolo Merialdo

Homework 5

L'obiettivo del progetto è sviluppare un sistema avanzato di *search* su articoli scientifici, in cui le tabelle siano trattate come oggetti di prima classe e completamente indicizzabili.

1. **Creazione del corpus di documenti**. Scrivere uno script per recuperare da https://arxiv.org tutti gli articoli disponibili in formato HTML il cui titolo o abstract contiene la stringa:
    1. Gruppi A: *"Entity resolution" oppure "Entity matching".*
    2. Gruppi B: *"text-to-sql"* oppure "Natural language to sql"
    3. Gruppi C: "*automatic speech recognition" oppure "speech to text"*
    4. Gruppi C: *"text to speech"*
    5. *Studenti lavoratori: " Query processing" oppure "Query optimization"*
2. **Indicizzazione dei documenti**. Scrivere il codice per indicizzare gli articoli utilizzando Elasticsearch (o Lucene), includendo i seguenti campi: titolo, autori, data, abstract, testo completo.
3. **Funzionalità di ricerca base**. Il sistema deve permettere interrogazioni su uno o più campi, ricerca per parole chiave, combinazioni di query (es. ricerca booleana, full-text). Le funzionalità di ricerca devono essere disponibili tramite una semplice shell su riga di comando e tramite una semplice interfaccia web.
4. **Estrazione delle tabelle**. Scrivere il codice per estrarre dagli articoli del corpus tutte le tabelle con associati dati di contesto. In particolare, per ogni tabella, estrare il corpo, la caption, i paragrafi che la citano, i paragrafi che contengono termini presenti nella tabella o nella caption (evitando di considerare termini non informativi).
5. **Estrazione delle figure**. Scrivere il codice per estrarre dagli articoli del corpus tutte le figure con associati dati di contesto. In particolare, per ogni figura, estrare l'url, la caption, i paragrafi che la citano, i paragrafi che citano termini presenti nella caption (evitando di considerare termini non informativi).
6. **Indicizzazione delle tabelle**. Scrivere il codice per indicizzare le tabelle utilizzando Elasticsearch (o Lucene). Ogni tabella viene indicizzata come un documento, con i seguenti campi: paper_id (ID dell'articolo), table_id (ID della tabella all'interno dell'articolo), caption (testo della caption della tabella), body (contenuto della tabella come testo), mentions (lista dei paragrafi del paper che citano la tabella), context_paragraphs (lista dei paragrafi del paper che contengono termini presenti nella tabella o nella caption).
7. **Indicizzazione delle figure**. Scrivere il codice per indicizzare le figure utilizzando Elasticsearch (o Lucene). Ogni figura viene indicizzata come un documento, con i seguenti campi: url (url della figura), paper_id (ID dell'articolo), table_id (ID della figura all'interno dell'articolo), caption (testo della caption della figura), mentions (lista dei paragrafi del paper che citano la figura).
8. **Funzionalità di ricerca avanzate**. Il sistema deve permettere interrogazioni per tabelle, figure, documenti su uno o più campi, ricerca per parole chiave, combinazioni di query (es. ricerca booleana, full-text). Le funzionalità di ricerca devono essere disponibili tramite una semplice shell su riga di comando e tramite una semplice interfaccia web.

Homework 5

- Ripetere le stesse operazioni, per almeno 500 articoli open access presenti su PubMed che contengono le keyword
  - Gruppi 1: *"cancer risk AND coffee comsumption"*
  - Gruppi 2: "glyphosate AND cancer risk"
  - Gruppi 3: "*air pollution AND cognitive decline"*
  - Gruppi 4: *"ultra-processed foods AND cardiovascular risk"*
- Si possono ottenee da questo url: https://pmc.ncbi.nlm.nih.gov/search/?filter=collections.open_access

Homework 5

- Preparare una relazione di circa 10 pagine che descrive la soluzione tecnica implementata. Descrivere l'architettura della soluzione, e gli sperimenti per valutare le prestazioni (in termini qualitativi e quantitativi) del sistema
- Preparare una presentazione di 15' (che descrive architettura e valutazione sperimentale della soluzione)

Termini di consegna: ore 12:00 il giorno dell'esame

Caricare la relazione e la presentazione attraverso il seguente modulo:

https://forms.office.com/e/EbqbR7dvK4