Gestione dell'Informazione
Confronto di sistemi di full text indexing e retrieval su un caso di studio

Presentazione del progetto AA 2024/2025

Il progetto in breve

OBIETTIVO: Progettazione, sviluppo e confronto di prestazioni di 3 sistemi di full text search su un benchmark a scelta.

COLLEZIONE DI DOCUMENTI: La collezione di documenti dovrà riguardare un argomento a piacere e potrà provenire da uno o più dataset accessibili on-line o ottenuto tramite lo scraping di siti web. I documenti scelti dovranno contenere almeno 2 campi (field).

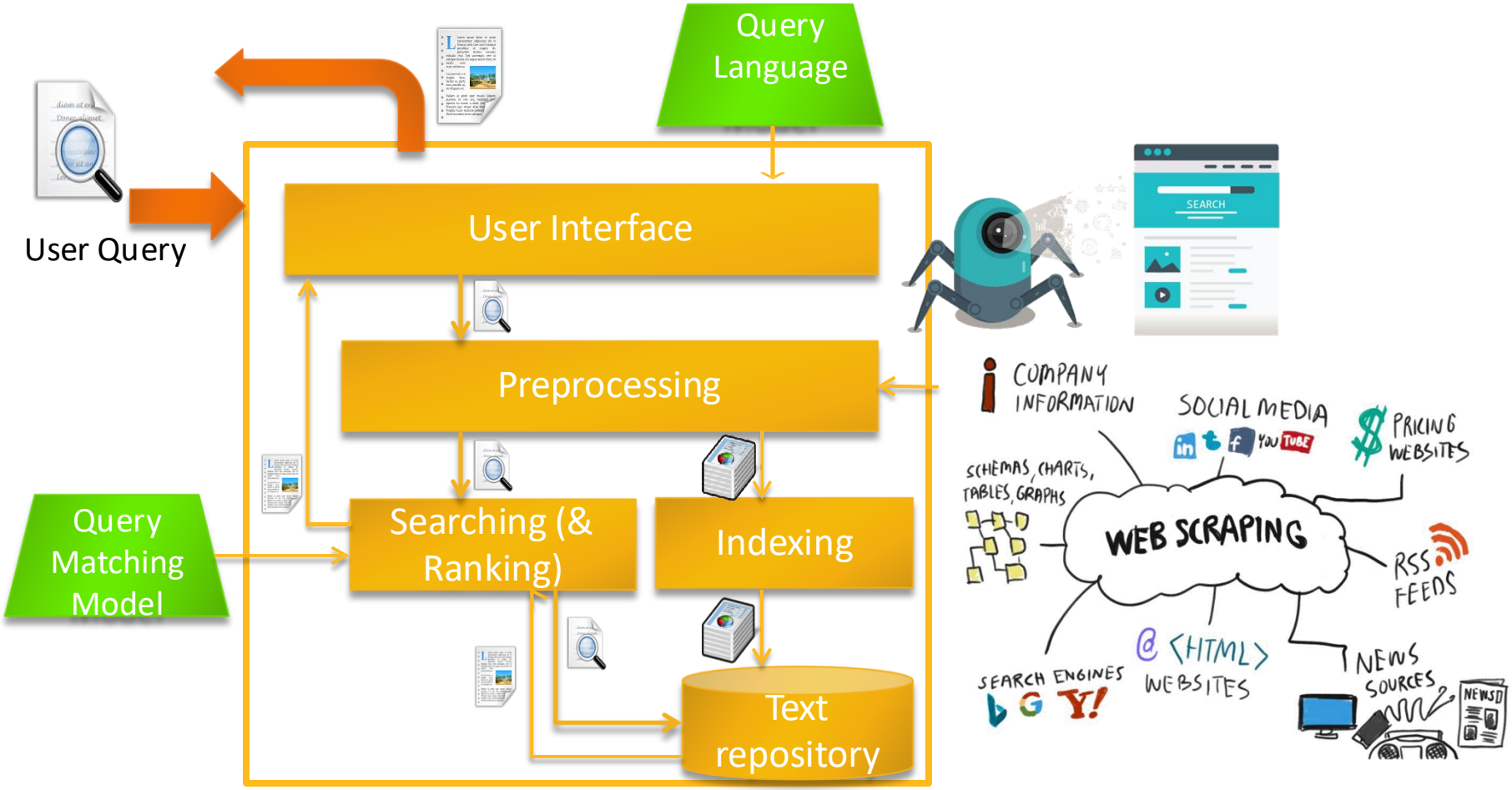
SEARCH ENGINE: Sarà necessario progettare e sviluppare un search engine per la collezione di documenti scelta usando ognuno dei tre sistemi visti a lezione:

- PostgreSQL
- PyLucene
- Whosh

Per ognuno dei search engine sviluppati dovranno essere prodotte almeno 2 varianti che usino altrettanti modelli di ranking.

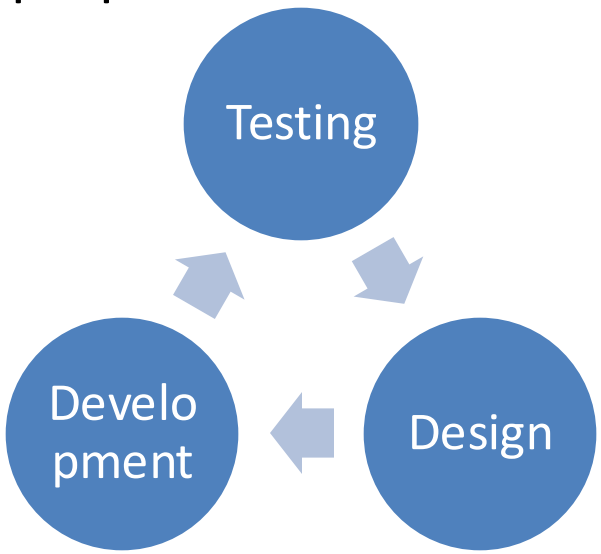
SEARCH ENGINE: Il confronto di prestazioni di efficacia dovrà basarsi su un benchmark di 10 query eseguite sulle varie versione dei search engine prodotte.

Architettura del sistema



I search engine

- Dovrà essere sviluppato un **search engine** per ognuno dei tre sistemi visti a lezione:
  - PostgreSQL
  - PyLucine
  - Whoosh
- Ogni search engine dovrà essere in grado di ricevere richieste keyword based e keyword based su campi specifici
- Per ogni search engine dovranno essere sviluppati almeno due varianti basati su altrettanti modelli di ranking
- L'analisi delle prestazioni dei sistemi sviluppati potrà portare allo sviluppo di ulteriori varianti



Benchmarking

- Gruppo di 10 UIN da eseguire sui search engine
  - Almeno 3 query dovranno usare i field
- Ogni UIN dovrà essere descritta in linguaggio naturale e quindi tradotta nel linguaggio d'interrogazione
- Ogni richiesta avrà una caratteristica particolare in modo da mettere in evidenza le peculiarità del search engine
- Sarà necessario individuare delle **misure di performance adeguate**
- Per ogni query testata e per ogni versione del search engine dovranno essere calcolate le misure di performance
- Il confronto tra le misure di performance ottenute consentiranno di mettere a confronto le versioni del search engine sviluppate

Realizzazione e consegna del progetto

- Il progetto deve esser svolto in gruppi di preferibilmente 2 persone (o anche 3 persone)
- Al termina il gruppo dovrà produrre:
  1. un archivio (ZIP) contenente
     1. il codice realizzato
     2. Una file txt relativo al benchmark contenente una descrizione testuale e la query sottomessa per ogni UIN
     3. README per l'installazione e l'uso dei search engine, per l'esecuzione del benchmark e lettura dei risultati ottenuti eseguendo le query del benchmark
  2. una presentazione

1. Da consegnare una settimana prima dell'appello in cui verrà presentato il progetto
2. Da consegnare il giorno dell'appello

La presentazione

- La presentazione dovrà durare non più di **15 minuti**
- Il numero di slide deve essere commisurato al tempo e comunque non superiore a 20 slide
- La presentazione deve

1. Descrivere brevemente la sorgente dati: caratteristiche e numero di text item
2. Descrivere brevemente l'architettura dei tre search engine
3. Per ogni search engine riportare i modelli di ranking usati
4. Descrivere il benchmark e le sue peculiarità
5. Descrivere e commentare i risultati di confronto tra le varie versioni ottenuti dall'applicazione del benchmark usando
    - Tabelle
    - Grafici

Presentazione del progetto

- Il gruppo presenterà il progetto in occasione di un appello d'esame
  - Tempo 15 minuti per la presentazione (è molto importante rispettare i tempi)
  - Tutti i componenti del gruppo dovranno partecipare alla presentazione
- Il progetto può essere presentato in qualsiasi appello d'esame e il voto avrà validità fino a febbraio 2026
- *Non è necessario* aver superato l'esame propedeutico «Algoritmi e strutture dati» per presentare il progetto mentre è *obbligatorio* per sostenere la prova scritta

Qualità e quantità del lavoro svolto: aspetti valutati

- Caratteristiche del dataset (dimensione e complessità del corpus)
- Qualità dei search engine sviluppati e quantità delle varianti
- Modalità di confronto dei search engine sviluppati
  - Il benchmark è adatto a valutare le performance?
  - Le misure adottate sono adeguate a mettere a confronto le performance?
  - Efficacia delle modalità di presentazione del confronto di performance
    - I grafici mostrati e le misure aggregate usate sono in grado di mettere in evidenza le differenze?
    - I risultati ottenuti sono interpretati adeguatamente?

L’esame…

- 60% del voto finale dipenderà dal voto dello scritto
  - Domande aperte e semplici e brevi esercizi sugli argomenti del corso
- 40% del voto finale dipenderà dal voto del progetto e della presentazione
  - Il voto del progetto sarà personale
  - Il voto dipenderà dalla presentazione
  - Il voto dipenderà dalla qualità e quantità del lavoro svolto