Gestione dell'Informazione
Opinion rank:
Un search engine per Opinion Corpora

Presentazione del progetto AA 2022/2023

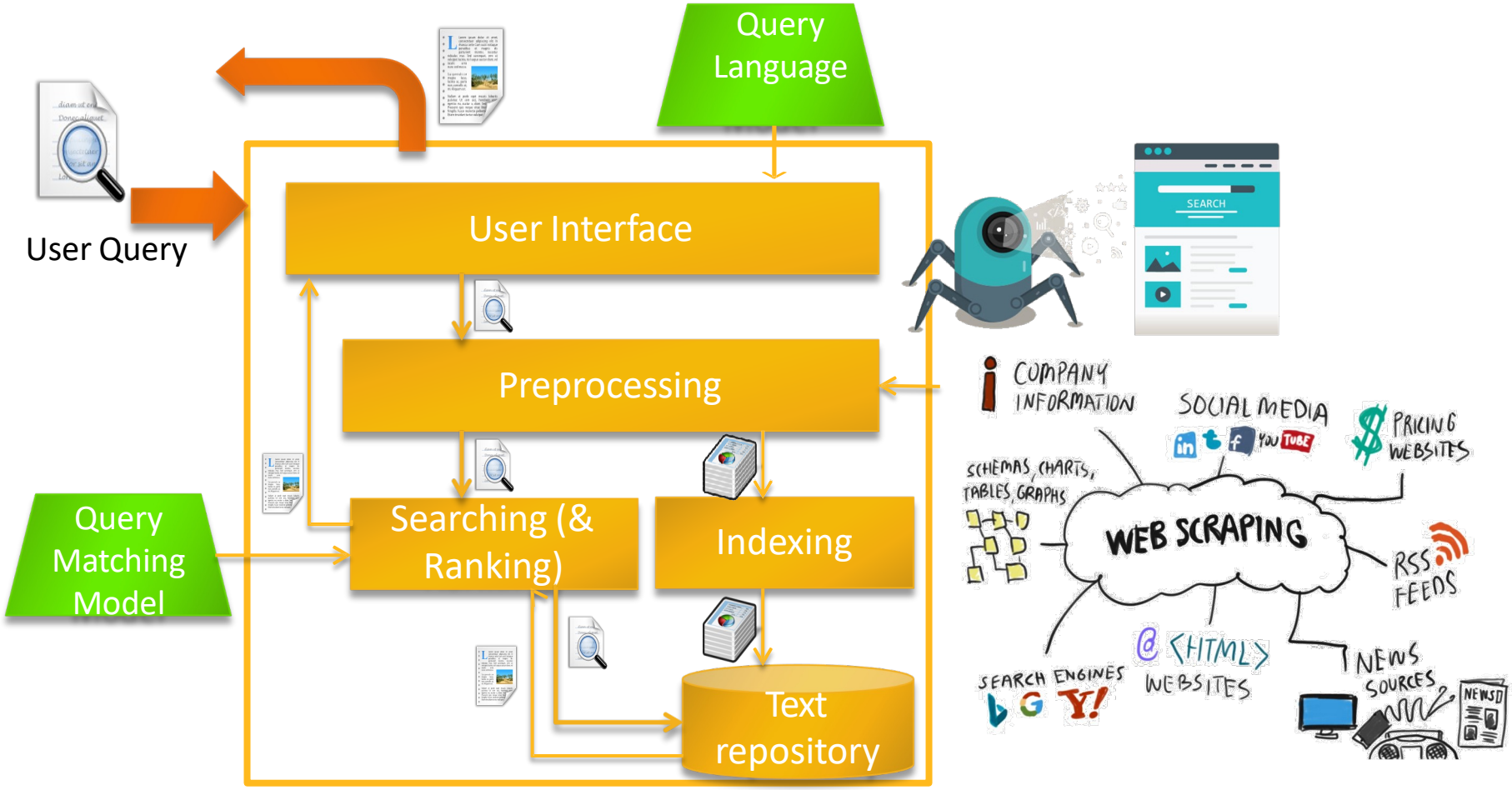
Il progetto in breve

OBIETTIVO: Sviluppare un search engine su uno o più corpus di text item che contengono opinioni che supporti richieste che riguardano sia il contenuto testuale sia **l'opinione espressa**

COLLEZIONE DI DOCUMENTI: Costruita a partire da **un corpus di text item che esprimono opinioni (Opinion Corpora)** come ad esempio review di prodotti/film/servizi, forum o blog. Le sorgenti dati potranno essere siti web o collezioni in vari formati disponibili su siti quali Kaggle.

ACCESSO ALLA COLLEZIONE: L'utente avrà la possibilità formulare richieste sulla base di un linguaggio di interrogazione definito che dovrà integrare un linguaggio standard di IR con un linguaggio che consenta di esprimere richieste sull'opinione espressa. I risultati di ogni richiesta inoltrata al sistema dovranno essere presentati in una lista ordinata in ordine decrescente di rilevanza dove la rilevanza dovrà dipendere dal contenuto e dall'opinione espressa nel text item.

Architettura del sistema



Raccolta dei text item e preprocessing

- Necessario individuare una o più Opinion Corpora su un tema di interessa
- La collezione dei text item sarà costruita a partire da pagine Web o collezioni disponibili in vari formati
  - Ad esempio su Kaggle trovate diversi dataset per Opinion Mining
- In base al formato della sorgente dati, sarà necessario usare tool diversi, ad esempio:
  - parser in grado di processare documenti nei diversi formati
  - Web API
  - Web crawler per effettuare il download di pagine web
  - Web scraper per l'estrazione di contenuti da pagine web
- Ogni text item subirà una fase di preprocessing finalizzata al trattamento della parte testuale e all'estrazione dell'opinione espressa
  - **Sentiment analysis**

Query language e modello di IR

- Il linguaggio d'interrogazione dovrà consentire all'utente di interrogare il dataset esprimendo richieste che coinvolgono:
  - Gli elementi testuali
    - Ad esempio: Apple Watch 8
  - Le opinioni espresse
- I risultati del preprocessing di ogni item dovranno essere memorizzati in un inverted index che memorizzi sia il contenuto sia il sentimento espresso ovvero l'output del processo di sentiment analysis
  - Si potranno *provare più tecniche di sentiment* analysis
- Il modello di IR dovrà supportare una funzione di ranking $\mathbf{R(q_i, dj)}$: $\mathbf{Q \times D \rightarrow R}$ definisca un ordinamento totale basato sia sul contenuto testuale sia sul sentimento espresso
- Il modello dovrà essere integrato nel search engine

Benchmarking

- Gruppo di 10 query da eseguire sul search engine
- Ogni richiesta 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
  - Linguaggio d'interrogazione
  - Query processing
- Per ogni query testata sul sistema si dovranno produrre le misure di performance DCG
- Si potranno mettere a confronto le performance di diverse configurazioni di modelli di IR standard e tecniche di sentiment analysis

Realizzazione e consegna del progetto

- Il progetto deve esser svolto in gruppi di preferibilmente 3 persone (o anche 2 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 dell'applicazione e 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

- Il numero di slide deve essere commisurato al tempo e comunque non superiore a 20 slide
- La presentazione deve

1. Descrivere la sorgente dati
2. Descrivere il linguaggio di interrogazione
3. Descrivere la/le tecniche di sentiment analysis usata/e
4. Descrivere il modello di IR adottato o i vari modelli
5. Descrivere la modalità di integrazione del modello per il query processing
6. Descrivere il benchmark e le sue peculiarità
7. Descrivere e commentare i risultati ottenuti dall'applicazione del benchmark

- Nel mostrare le soluzioni progettate è importante essere chiari su **«come»** il problema è stato risolto ovvero descrivere **«quale» soluzione tecnica** è stata individuata (approccio funzionale, metodologico) mentre non è necessario mostrare il codice, se non dei piccoli frammenti

Presentazione del progetto

- Il gruppo presenterà il progetto in occasione di un appello d'esame
  - Tempo 20 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 2024
- *Non è necessario* aver superato l'esame propedeutico «Algoritmi e strutture dati» per presentare il progetto mentre è *obbligatorio* per sostenere la prova scritta

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 .
    Aspetti valutati:
    - Dimensione e caratteristiche del dataset
    - Tecniche di sentiment analysis usate
    - Modello di IR definito
    - Meccanismo di query processing implementato
    - Approccio alla costruzione del benchmark e discussione dei risultati ottenuti