spoon

A RESTAURANT REVIEW SEARCH ENGINE

MASSI RICCARDO
CZUBA FILIP

INDICE

KAGGLE

SELEZIONE E CREAZIONE DEL DATASET

KAGGLE

Kaggle è una piattaforma online per data science e machine learning, di proprietà di Google. È nota per la sua vasta raccolta di dataset accessibili gratuitamente a chiunque.

Questi dataset spaziano in diverse aree, dalla finanza alla salute, dal marketing all'ambiente, e sono utilizzati da data scientist e analisti per esplorare, analizzare e sviluppare modelli di machine learning.

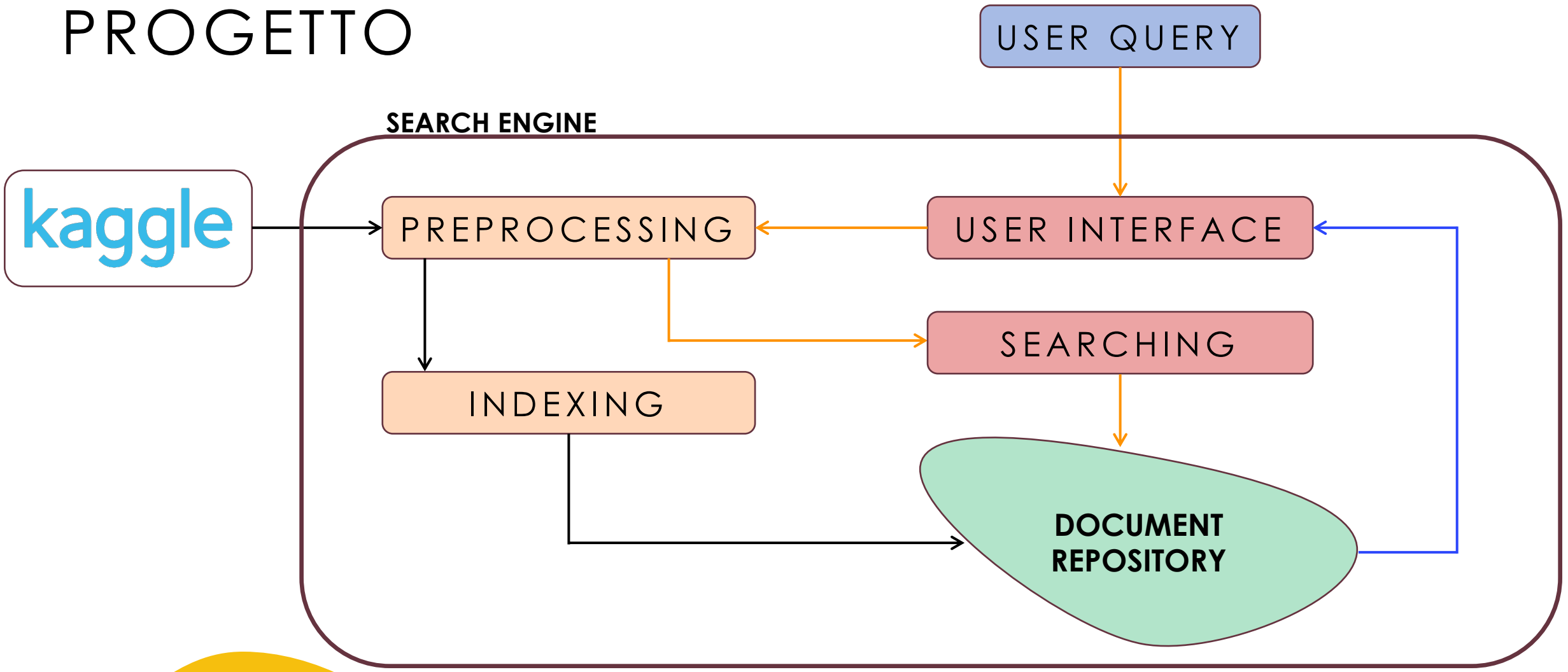
Per The Spoon è stato scelto il dataset proveniente da Yelp contenente recensioni di attività commerciali. Ai fini del progetto tale dataset è stato circoscritto ai soli ristoranti.

Sono state selezionate 21.000 recensioni appartenenti a 3.000 locali distinti.

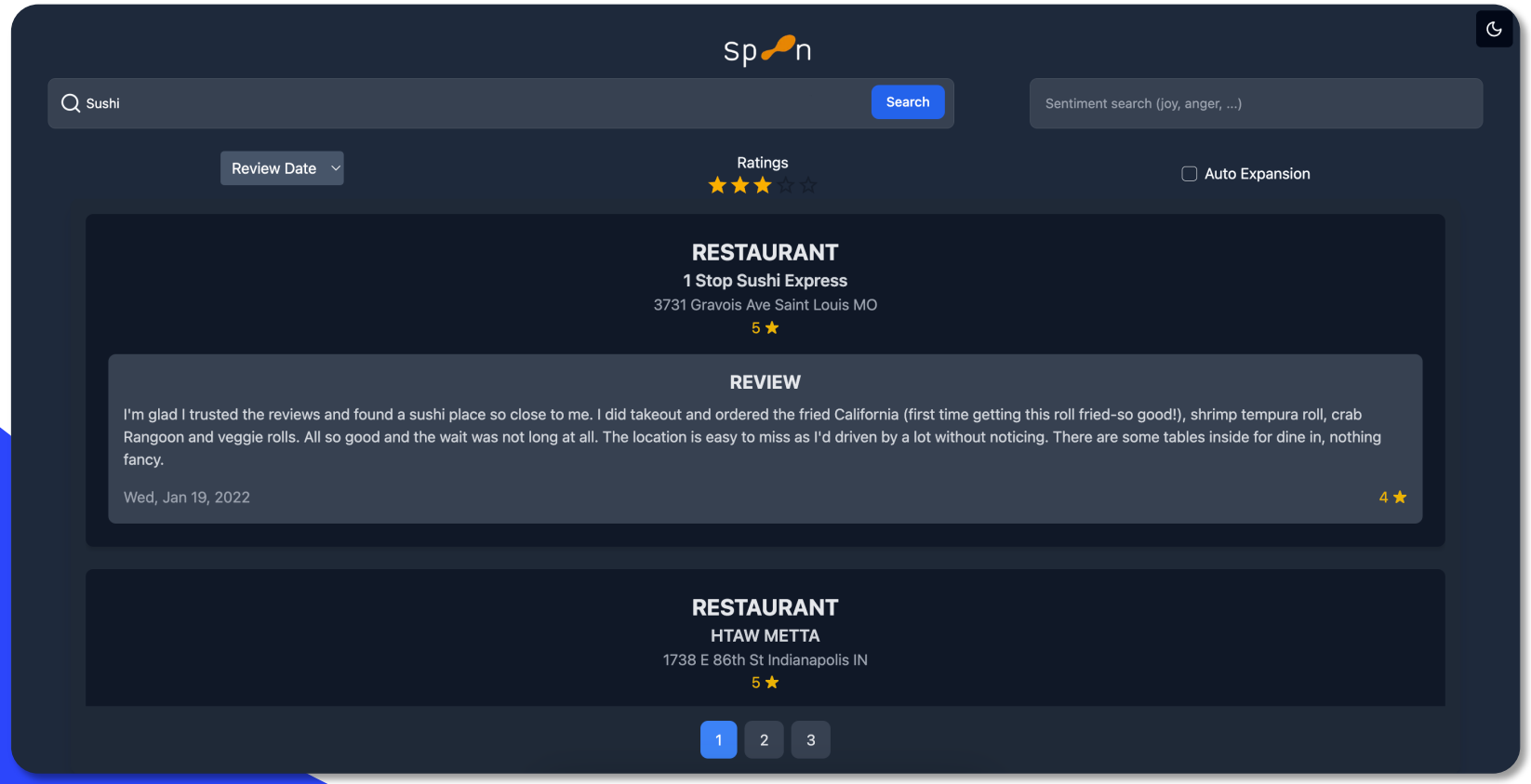
ARCHITETTURA DEL PROGETTO

TECNOLOGIE IMPIEGATE NEL FRONT-END E BACK-END.

ARCHITETTURA DEL PROGETTO



TECNOLOGIE IMPIEGATE



FRONT-END

- **NEXT.JS:** Gestisce la parte grafica dell'applicazione per il rendering dinamico dei contenuti.
- **VERCEL:** Distribuisce e ospita l'applicazione web con aggiornamenti automatici e scalabilità.

BACK-END

- **WHOOSH:** Fornisce un motore di ricerca full-text per indicizzare e cercare contenuti nel dataset.
- **NLTK:** Fornisce strumenti per l'elaborazione e l'analisi del linguaggio naturale.
- **PYTORCH (TRANSFORMERS):** Utilizza modelli di Hugging Face per la sentiment analysis.

FRONTEND

NEXT.JS E VERCEL

NEXT.JS E VERCEL

NEXT.JS

Next.js fornisce un framework React estensibile, facile da usare e a prova di produzione.

Next.js è dotato di funzioni che permetteranno di portare un'applicazione da zero alla produzione in pochissimo tempo, offrendo una curva di apprendimento poco ripida, semplicità e strumenti potenti a disposizione.

VERCEL

Vercel porta l'approccio zero-configuration di Next.js nel cloud, in modo da consentire agli utenti di distribuire la propria app in pochi minuti.

La piattaforma Vercel è ottimizzata per l'edge, e consente di fare preview, test e deployment della propria web app senza doversi preoccupare dell'infrastruttura.

SEARCH ENGINES

WHOOSH E ULTERIORI ELEMENTI DEL MOTORE DI RICERCA

WHOOSH E RANKING MODELS

WHOOSH

Whoosh è una libreria open-source scritta in Python e Java per l'indicizzazione e la ricerca full-text. Ideale per applicazioni che richiedono capacità di ricerca testuale, Whoosh permette di creare motori di ricerca personalizzati che possono indicizzare e cercare attraverso grandi quantità di testo.

TF-IDF

È una tecnica di valutazione dell'importanza di una parola all'interno di un documento rispetto a un'intera collezione di documenti.

BM25F

Modello di ranking di default di Whoosh e variante del modello di ranking BM25. BM25F estende BM25 considerando diversi campi di un documento (come titolo, corpo, metadati), assegnando pesi differenti a ciascun campo per migliorare la precisione del ranking.

QUERY SYNTAX E ULTERIORI ELEMENTI

DATA INDEXING

Affinché i risultati restituiti siano più accurati è stato creato un Whoosh Schema, dove a diversi campi sono stati associati pesi differenti, in modo tale da favorirne alcuni durante il parsing multi-field.

QUERY EXPANSION

Nel caso in cui la query presentata dall'utente non restituisca alcun risultato è possibile abilitare la *query expansion*, cioè una query OR-associative a cui vengono aggiunti alcuni sinonimi dei termini della query originale provenienti dal synset fornito da NLTK.

QUERY LANGUAGE

Il query language utilizzato è quello di default fornito da Whoosh. È possibile svolgere:

- Field e range queries;
- Boolean queries;
- Exact phrase queries;
- Wildcard queries.

Inoltre è stata introdotta la possibilità di ordinare i risultati per data o punteggio della recensione.

QUERY SPELLCHECKING

Nel caso in cui la query non produca alcun risultato viene svolto un rudimentale spellcheck sui termini della query, offrendo all'utente la possibilità di correggere la propria query (*forse stavi cercando…*).

SENTIMENT ANALYSIS

Hugging Face

Hugging Face è una piattaforma leader nel machine learning e NLP, nota per la sua libreria Transformers e modelli pre-addestrati, che facilita lo sviluppo di applicazioni avanzate come traduzione e sentiment analysis.

SENTIMENT ANALYSIS

Utilizzando il modulo *Transformers con* il classificatore testuale pre-trained *Roberta* di *Sam Lowe*, si è riuscito ad estrapolare il responso emotivo, categorizzato tramite label differenti, insito nel corpo delle recensioni. Per rappresentare l'umore generale del recensente si è scelta la label con la percentuale di accuratezza maggiore.

La classificazione emotiva del testo permette all'utente finale di filtrare le recensioni selezionando una o più label associate a specifici stati d'animo.

BENCHMARKING

EVALUAZIONE DELLA PERFORMANCE DEI MOTORI DI RICERCA

APPROCCIO AL BENCHMARKING

NDCG

Per ciascuna coppia composta da query e motore di ricerca vengono restituiti i primi 10 risultati ottenuti, i quali vengono valutati con un punteggio da 0 a 3 relativo alla loro rilevanza alla query.

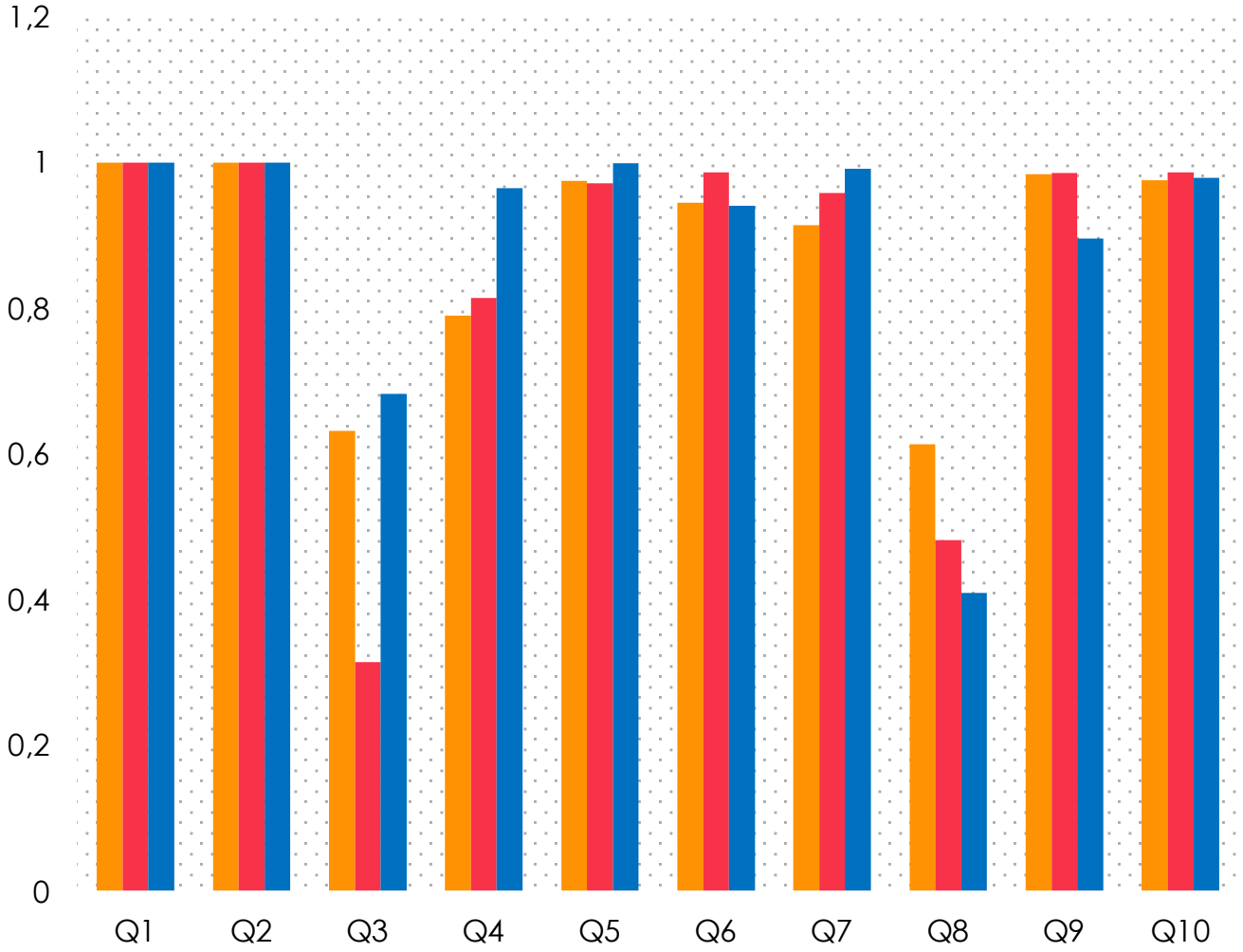
Tali valutazioni sono da svolgere manualmente, compito affidato ad un agente terzo e imparziale.

R-PRECISION

Sono state create manualmente delle collezioni di documenti rilevanti per ciascuna query, con il fine di confrontare i risultati ottenuti dai motori di ricerca con tale insieme per estrapolarne i valori di precisione a valori fissi di recall.

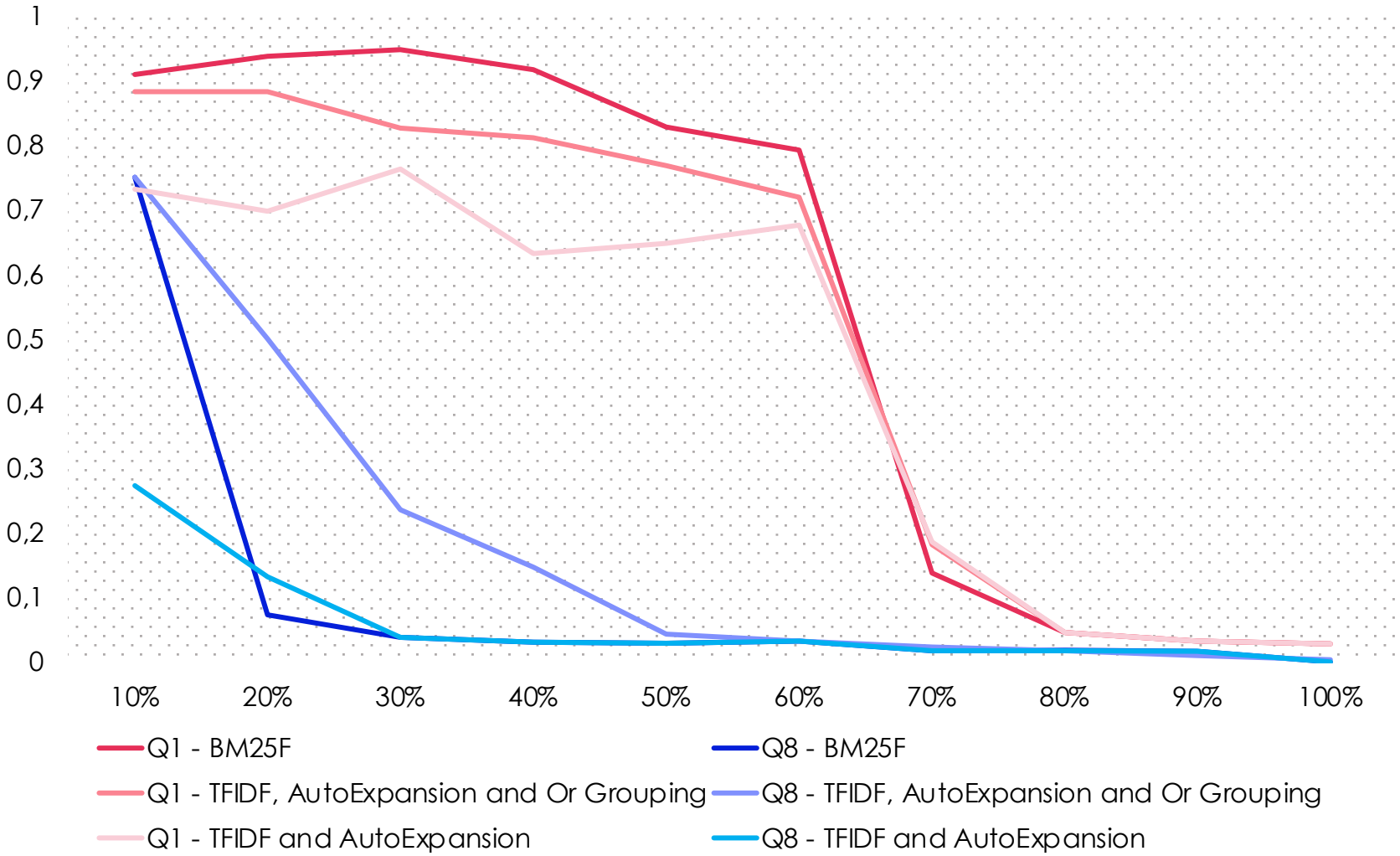
Questa parte del benchmarking viene svolta in autonomia dall'algoritmo.

NDCG



I valori di NDCG ci permettono di confrontare tra loro query e motori di ricerca differenti, permettendo di valutare le loro abilità di ranking e recall.
Si notino le query Q3 e Q8, più articolate delle altre, le quali restituiscono risultati talvolta irrilevanti a causa dell'ambiguità di linguaggio.

R-PRECISION



Si è scelto di mostrare due query piuttosto differenti: una in cui i search engine hanno performato molto bene, un'altra, invece, in cui i search engine hanno faticato maggiormente. Si noti come la combinazione di Auto Expansion e AND Grouping abbassi in entrambe le query il livello di precisione.

GRAZIE PER L'ATTENZIONE!

MASSI RICCARDO

CZUBA FILIP