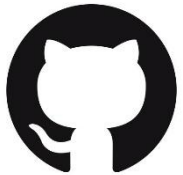
Homework 3

Alessio Marinucci

Riccardo Felici



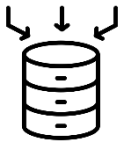
https://github.com/alemari7/Hw3_IDD

Obiettivo

- Implementazione di un sistema di Data Discovery.
- Estrazione di risultati a partire da un corpus di documenti in formato JSON, contenuti all'indirizzo *"all_tables“.*
- Definizione di query sul motore di ricerca per valutare il funzionamento.

Tecnologie Utilizzate

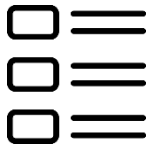
Pipeline



Acquisizione Dati

Pulizia dei Dati

Parsing dei Dati



Indicizzazione

Querying

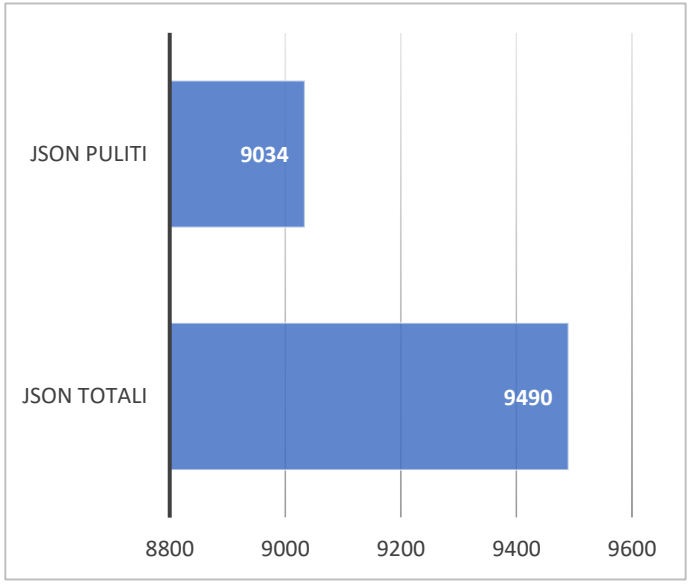
Restituzione dei Risultati

Valutazione e Test

Data Cleaning

Eliminazione delle tabelle non rilevanti (e.g. formule matematiche)

Eliminazione file JSON vuoti o senza formato del documento



Metriche di valutazione

MRR: Valuta il numero medio di volte che l’elemento più rilevante si trova in prima posizione.

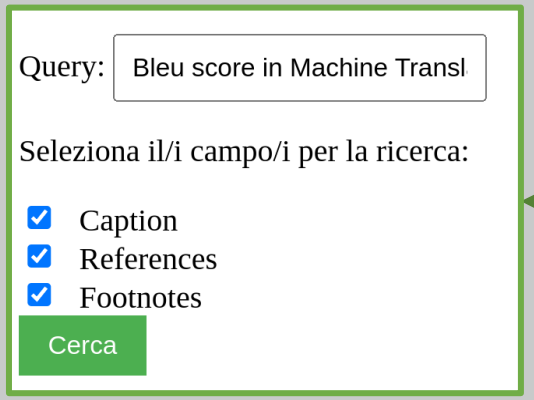
NDCG: Valuta la qualità dell’intero ordinamento dei risultati rispetto alla loro rilevanza, assegnando un peso maggiore ai documenti rilevanti presenti nelle prime posizioni della lista.

Esempi di query

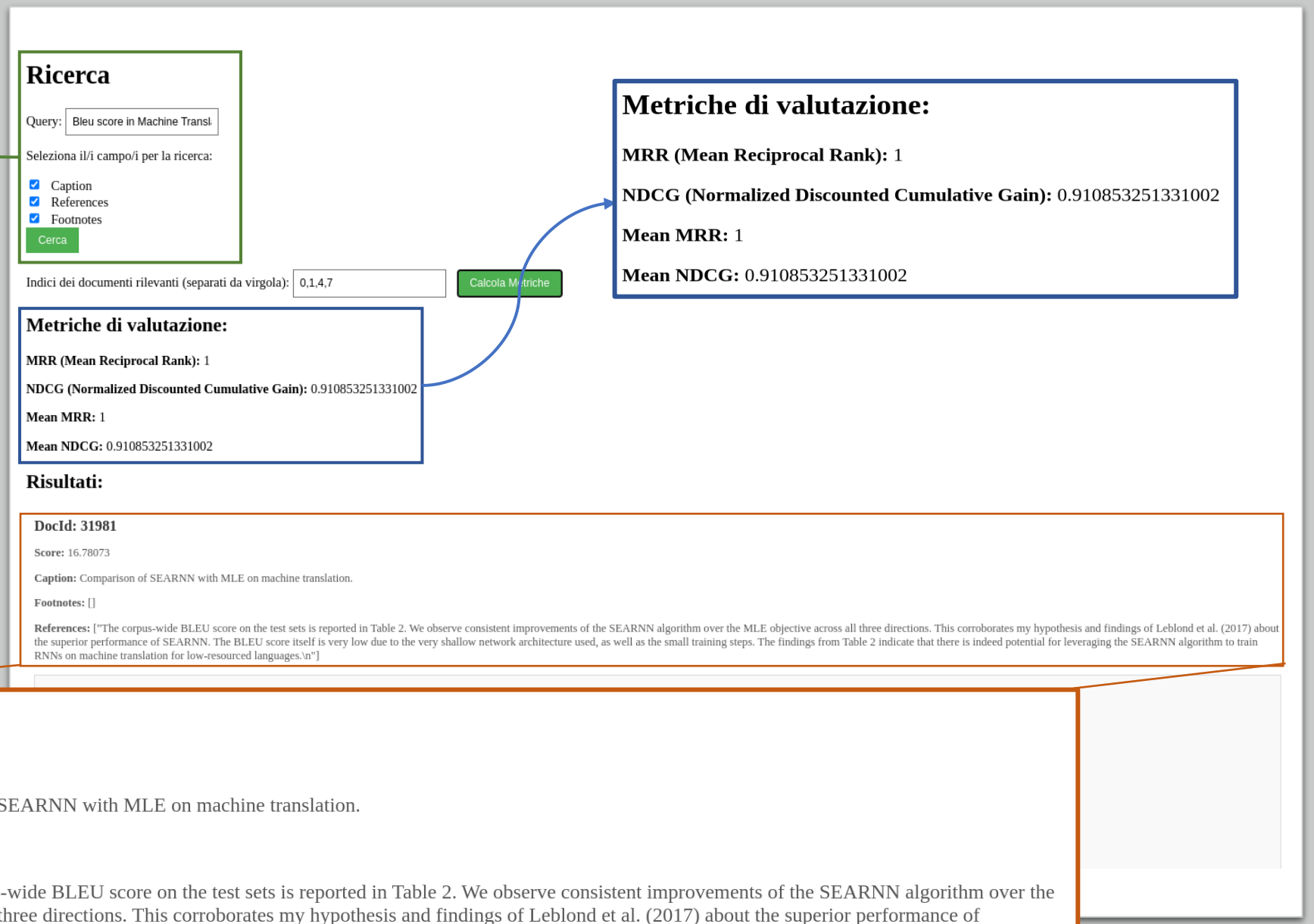
```
1
Inserisci la query di ricerca: data
Trovati 2341+ hits risultati. Tempo di risposta: 71 ms.

Documento ID: 49006 | Score: 1.9990929
Caption: Table 5: Related research studies for Medical data properties in FL for medical applications, consisting of data partitions, data distribution (i.e., non-IID) char
```

```
Scegli il campo su cui fare la ricerca:
1. Caption
2. References
3. Footnotes
4. Ricerca su più campi
1
Inserisci la query di ricerca: data
Trovati 2341+ hits risultati. Tempo di risposta: 71 ms.

Documento ID: 49006 | Score: 1.9990929
Caption: Table 5: Related research studies for Medical data properties in FL for medical applications, consisting of data partitions, data distribution (i.e., non-IID) char
acteristics, possible data privacy attacks, and data privacy protections.
Footnotes:
References: ["Related research studies have explored critical aspects such as data partitions, non-IID characteristics, potential data privacy attacks, and corresponding da
ta privacy protections, as summarized in Table 5."]
Table: Property Study Data partitions [132, 130] Data distribution (non-IID) characteristics [133] Data privacy protections [54]
Fonte: 2405.13832.json

Documento ID: 5596 | Score: 1.9864684
Caption:
Single model performance on MedNLI developlment data. Naiïve means simply integrating all medical-domain data; Ratio means using MedNLI as in-domain data and other medical
domain data as external data; Ratio+MNLI means using medical domain data as in-domain and MNLI as external.

Footnotes: []
References: ["#S3.T5"]
Table: Model Dev Set Test Set WTMed - 98.0 PANLP - 96.6 Ours 91.7 93.8 Sieg - 91.1 SOTA 76.6 -
Fonte: 1906.04382v1.json

Documento ID: 8621 | Score: 1.9775481
Caption: Table 3: The summary of deep learning-based cross-domain data fusion models in urban comput
hical Data ▲; Social Media Data ◆; Demographical Data ❖; Environmental Data 🍀. Notice that method n
otherwise, they are named after the first authors.
Footnotes: ["Zhang et al. [2016]\n\nZhang, J., Zheng, Y., Qi, D., Li, R., Yi, X., 2016.\n\n\nDnn-bas
h ACM SIGSPATIAL international conference on advances in geographic information systems, pp. 1\u2013
```

```
Inserisci gli indici dei documenti rilevanti (separati da virgola): 2,0,1
MRR: 0.5
NDCG: 0.8675034925694373
Media MRR: 0.25
Media NDCG: 0.7479444335353337

Vuoi fare un'altra ricerca? (s/n):
```

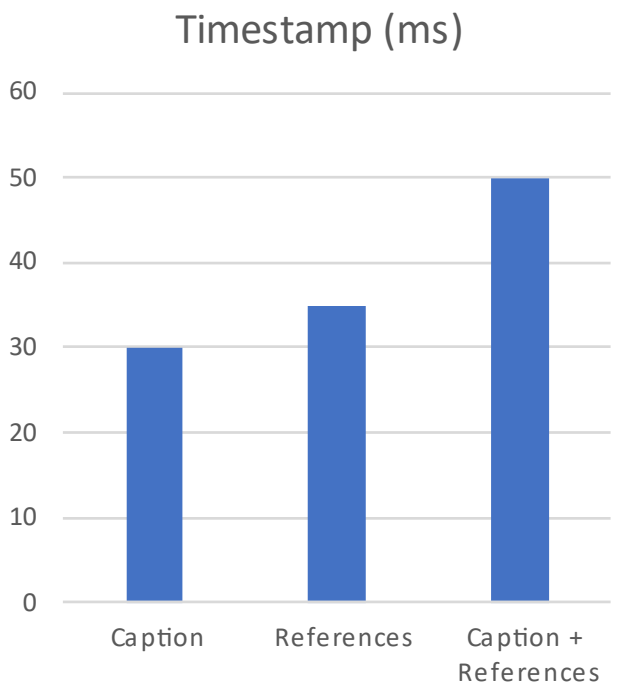


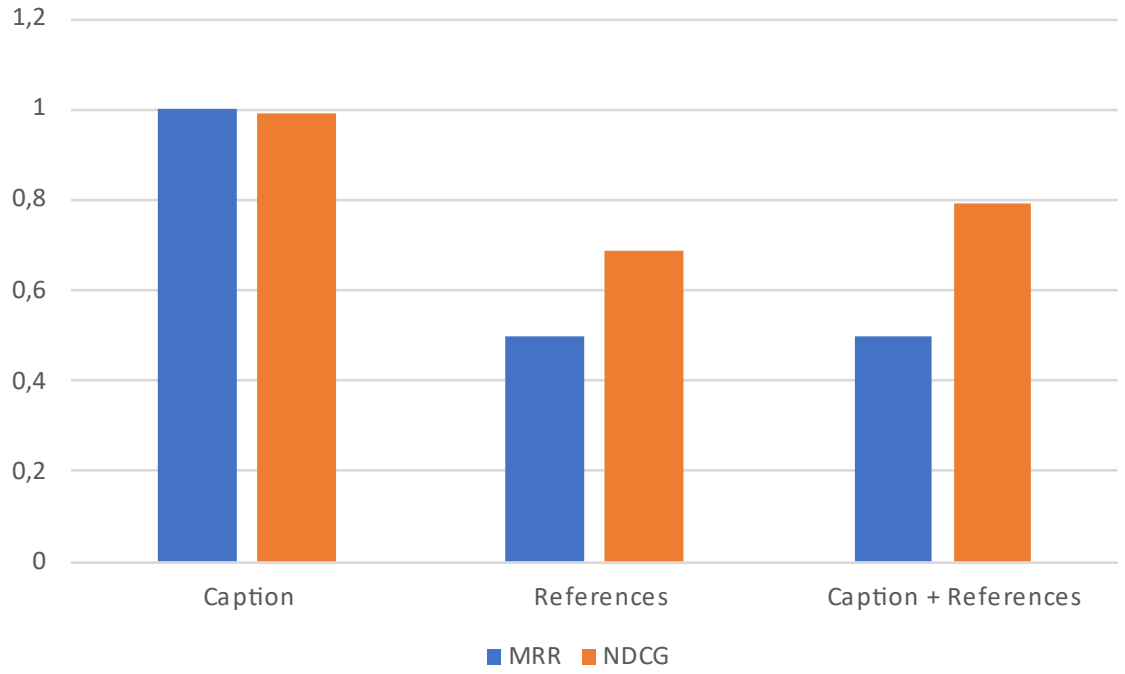
Esempi di query



Risultati ottenuti

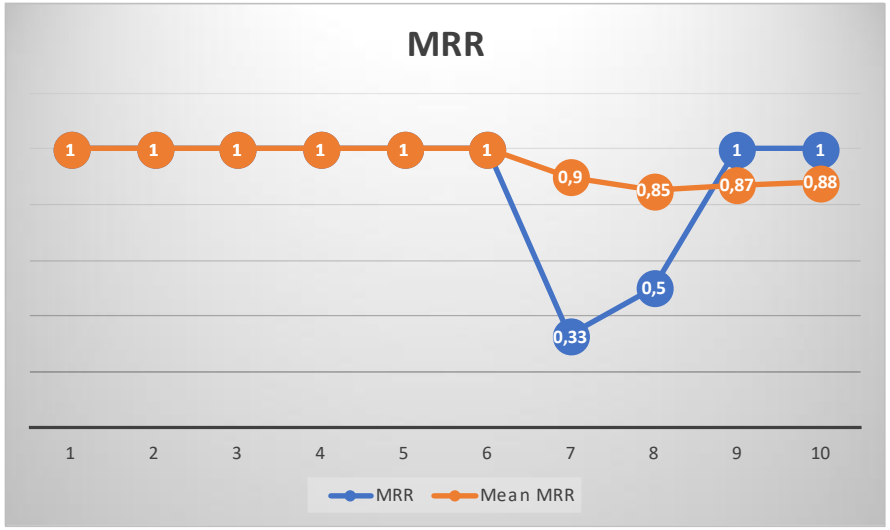
Risultati ottenuti sottoponendo la query «*Bleu score in Machine Translation*» usando rispettivamente una ricerca solo sul campo *caption*, solo sui *paragrafi di riferimento* o su entrambi i campi.

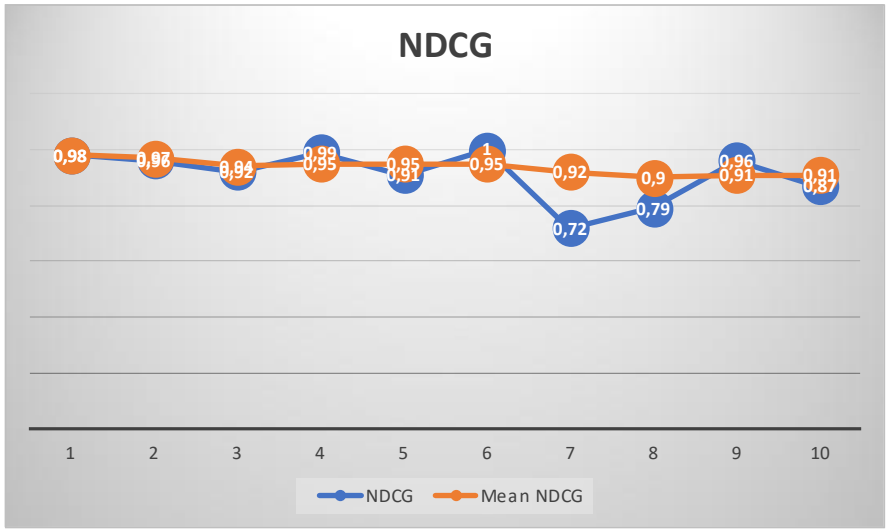




Risultati ottenuti

1. Bleu score ranges for Transformer models in Machine Translation
2. Optimal hyperparameters for GPT-3.5 fine-tuning tasks
3. Accuracy comparison between BERT and RoBERTa on text classification tasks
4. Training time vs performance for large language models
5. ROUGE scores for abstractive summarization models
6. Energy consumption of Transformer models during training
7. Comparison of perplexity scores across GPT variants
8. Impact of data augmentation on Bleu scores in Machine Translation
9. Zero-shot vs few-shot performance for GPT models
10. FLOPs required for training vs evaluation for language models





Conclusioni

Problemi incontrati

- **Calcolo del valore di rilevanza per i documenti:** è risultato complesso identificare un metodo semplice e intuitivo per gestire il calcolo della rilevanza.
- **Creazione query sul campo 'footnotes':** problemi relativi all'indicizzazione in quanto footnotes presenti in differenti formati.

Nozioni imparate

- **Creazione di un'interfaccia grafica**: è stato fatto per rendere più semplice l'interazione dell'utente con il sistema.
- **Costruzione di un motore di ricerca con una gestione di file JSON**: è servito per prendere dimestichezza con la tecnologia Lucene e costruire un motore di ricerca efficiente per rispondere alle query.

Sviluppi Futuri

- **Implementazione modelli di neural networks**: servono a migliorare la rilevanza dei risultati e automatizzare ulteriormente la classificazione dei documenti.
- **Sviluppo di un framework automatico per la valutazione delle metriche MRR e NDCG**: serve a ridurre la necessità di intervento manuale durante i test.
- **Integrare modelli di embeddings**: tecniche più recenti come BERT e Sentence Transformers, per rappresentare in modo semantico i contenuti delle tabelle e migliorare l'accuratezza nel confronto tra query e documenti.

GRAZIE PER L’ATTENZIONE!