Graboid RFC

Progetto Gestione dell'Informazione
2024-12-08

Dobrovolsky Andrei, Menozzi Matteo,
Turci Gabriele e Turci Sologni Enrico

Indice

1. Dataset

1.1. Scelta del Dataset

Abbiamo dedicato particolare attenzione alla selezione di un dataset adeguato. La scelta è ricaduta sull'insieme di documenti denominati **Request For Comments** (**RFC**), una serie di pubblicazioni tecniche che definiscono standard, metodologie e sviluppi tecnici di Internet.

La selezione di questo dataset si è basata su una serie di requisiti fondamentali e vantaggi distintivi che gli RFC soddisfano pienamente. Di seguito sono riportati i motivi principali e i benefici derivanti dall'uso degli RFC come dataset per il nostro progetto.

1. **Facilità di Recupero**: Gli RFC sono facilmente accessibili grazie alla struttura del loro sistema di identificazione. Ogni documento è identificato da un numero univoco incrementale, il che rende lo **scraping particolarmente semplice**. Per esempio, è possibile recuperare i documenti incrementando sistematicamente un indice numerico. Inoltre, esistono repository ben organizzati che facilitano il download di massa.

2. **Dimensione del Dataset**: Il corpus degli RFC comprende circa 9600 documenti, un numero perfettamente adeguato per il nostro scopo. Questo range consente di analizzare e testare algoritmi di indicizzazione e ricerca su un dataset realistico senza incorrere in problemi di scalabilità o carenza di dati.

3. **Varietà nella Lunghezza e Complessità dei Documenti**: Gli RFC presentano una significativa varianza nella lunghezza dei documenti:

   - Alcuni documenti contano poche decine di righe.
   - Altri arrivano a diverse centinaia di righe.

   Questa varietà consente di testare il motore di ricerca su casi molto diversi in termini di granularità del contenuto.

4. **Ricchezza di Termini Specifici**: Gli RFC sono documenti tecnici caratterizzati dall'uso di terminologia altamente specifica, inclusi acronimi e parole chiave che rappresentano concetti chiave nel dominio informatico. Questa caratteristica li rende ideali per:

   - Creare indici dettagliati.
   - Formulare query tecniche complesse.
   - Effettuare benchmarking di algoritmi di ricerca avanzati.

5. **Presenza di Metadati Strutturati**: Ogni RFC include metadati ben definiti, quali: *Numero identificativo*, *Titolo*, *Autori*, *Data di pubblicazione*, *Status (es. Standard, Informativo, Obsoleto)*, *Abstract e Parole chiave*, *Sezioni strutturate (Introduzione, Specifiche, Conclusione, ecc.)*.

   La presenza di questi metadati consente di implementare funzionalità avanzate, come la ricerca basata su campi specifici o il filtraggio per criteri.

6. **Compatibilità con Query Complesse**: Gli RFC coprono una vasta gamma di argomenti tecnici e standard, il che li rende perfetti per simulare query complesse e specifiche. Ad esempio, è possibile eseguire ricerche che riguardano concetti tecnici, interazioni tra standard o termini legati a specifiche versioni.

1.2. Scraping del Dataset

Abbiamo deciso di costruire un dataset personalizzato effettuando lo scraping manuale dei documenti e dei relativi metadati. Non abbiamo utilizzato dataset precostruti, poiché volevamo garantire che i dati soddisfacessero specificamente i requisiti del nostro motore di ricerca.

1.2.1. Scraping dei Metadati

La prima fase del processo prevede il recupero dei metadati associati a ciascun documento RFC.

- Il codice HTML degli RFC è carente e i metadati inclusi nei documenti stessi risultano spesso inconsistenti e difficili da estrarre in modo affidabile.
- Per ovviare a questa limitazione, abbiamo individuato una fonte alternativa: i metadati sono disponibili in formato tabellare sul sito ufficiale dell'RFC Editor, raggiungibile al seguente link: Fonte dei Metadati.
- I metadati recuperati includono campi come *numero identificativo*, *titolo*, *autori*, *data di pubblicazione*, *stato*, *estratto*, e *parole chiave*. Questi sono stati scaricati e organizzati in modo strutturato per essere utilizzati successivamente.

1.2.2. Scraping dei Corpi dei Documenti

Una volta ottenuti i metadati, il passo successivo è stato lo scraping dei corpi dei documenti.

- Gli RFC sono accessibili tramite URL che seguono una struttura numerica incrementale, come: **`https://www.rfc-editor.org/rfc/rfcX.txt`**, dove **`X`** è l'identificatore numerico del documento (es. 1, 2, 3, ...).
- Uno script iterativo automatizza il download di ogni documento a partire dal numero 1 fino all'ultimo RFC disponibile. Per velocizzare il processo, viene impiegato un sistema basato su threadpool, che consente il download parallelo di più documenti, migliorando le prestazioni.

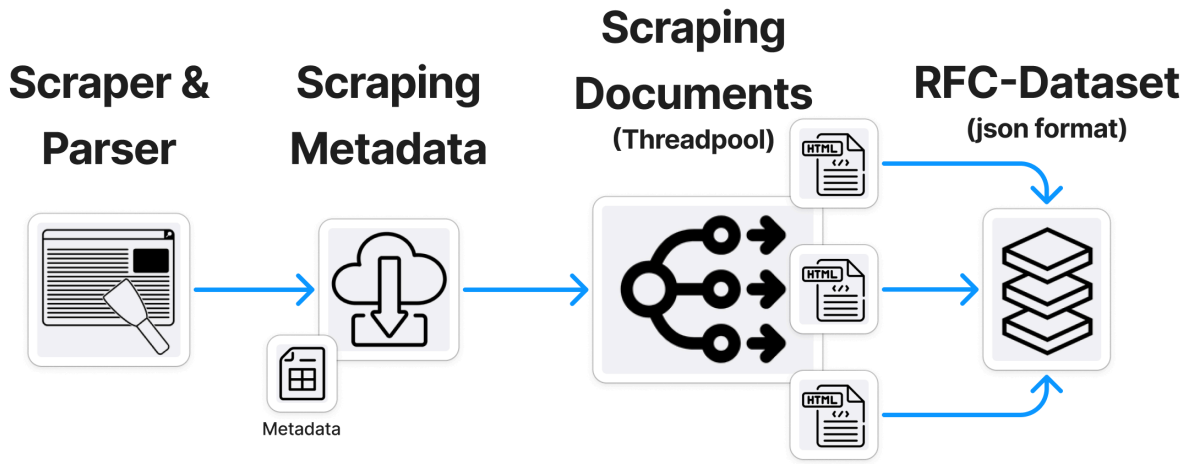

Figura 1: Il diagramma illustra il flusso completo del processo di scraping e costruzione del dataset.

1.2.3. Formattazione del Dataset

Una volta completato il recupero dei metadati e dei corpi dei documenti. Tutte le informazioni sono state consolidate in un unico file in formato **JSON**, dove ogni documento rappresenta un record strutturato contenente sia i metadati sia il testo completo del documento.

Esempio:


```
[
    {
        "Number": "9009",
        "Date": "2021-04",
        "Status": "Proposed Standard",
        "More Info": "Does not obsolete any RFC.",
        "Title": "Efficient Route Invalidation",
        "Authors": [
            "R.A. Jadhav", "P. Thubert", "R.N. Sahoo", "Z. Cao"
        ],
        "Files": [
            "HTML", "TEXT", "PDF", "XML"
        ],
        "Keywords": [
            "NPDAO", "DCO", "no-path", "route", "cleanup"
        ],
        "Abstract": "This document explains ... optimized route invalidation messaging.",
        "Content": "A directed graph having the property ... timer value of 1 second."
    }
]
```


1.3. Locazione dei File Interessati

Script di Parsing:

Lo script responsabile per lo scraping dei metadati e dei corpi dei documenti si trova al seguente percorso:

`".../gestione-info/project/searchengine/myParser/"`

- Il file di interesse è: **`myParser.py`**

Dataset Generato:

Il dataset prodotto dallo script, contenente i metadati e i corpi dei documenti RFC, è situato nella directory:

`".../gestione-info/project/dataset/"`

- Il file interessati sono: **`dataset.json`**, **`example.json`**

2. Tipologia di Utente e Query Language

2.1. Definizione della Tipologia di Utente

Il nostro motore di ricerca è progettato per soddisfare le esigenze di utenti con poca esperienza o conoscenza del dominio tecnico degli RFC, nello specifico:

- **Studenti universitari**: Studenti che potrebbero necessitare di informazioni specifiche sugli standard e sulle tecnologie di Internet per i loro studi o progetti.
- **Non esperti del settore**: Utenti che non hanno familiarità con gli RFC o con il loro contenuto tecnico ma che hanno necessità di cercare informazioni specifiche.
- **Utenti generici**: Chiunque voglia ottenere informazioni dettagliate presenti negli RFC ma non sappia come localizzarle o in quali documenti siano contenute.

L'obiettivo principale è rendere l'esperienza di ricerca intuitiva e accessibile anche per chi non ha conoscenze approfondite del dominio tecnico.

2.2. Scelta del Query Language

Per soddisfare le esigenze di questa tipologia di utenti, abbiamo optato per un approccio di ricerca basato su:

1. **Keyboard-Based Search (Ricerca Libera)**

   Abbiamo scelto di adottare un approccio basato su una ricerca libera (**keyboard-based search**), che consente agli utenti di inserire termini di ricerca nella barra principale in modo diretto e intuitivo. Questo metodo è ideale per chi non ha familiarità con il linguaggio tecnico o con la struttura specifica degli RFC.

2. **Keyboard-Based Search su Campi Specifici**

   Abbiamo inoltre integrato la possibilità di effettuare **ricerche su campi specifici**, come titoli, parole chiave e autori, per consentire una maggiore precisione nella localizzazione delle informazioni.

2.2.1. Funzionalità di Supporto alla Ricerca

A supporto della ricerca libera, sono state implementate funzionalità avanzate per migliorare ulteriormente l'esperienza utente. Tra queste:

1. **Filtri di Stato**: La possibilità di filtrare i risultati in base allo stato degli RFC, come ad esempio Standard Track, Best Current Practice, Informational, Experimental o Historic.
2. **Filtri Temporali**: Gli utenti possono inoltre limitare i risultati a un determinato periodo temporale, specificando un anno preciso o un intervallo di date, rendendo più facile trovare documenti aggiornati o rilevanti per contesti storici.
3. **Funzioni di Correzione e Sinonimi**: Per evitare errori comuni e ampliare i risultati della ricerca, sono state introdotte opzioni di correzione ortografica e sinonimi, che aiutano gli utenti a trovare informazioni pertinenti anche in caso di imprecisioni nei termini inseriti.

La scelta di queste funzionalità e approcci è stata motivata dalla volontà di creare un sistema accessibile, che riduca le barriere tecniche per gli utenti inesperti, senza però sacrificare le capacità avanzate per chi ha esigenze più specifiche.

3. Benchmark

Per valutare le prestazioni del nostro motore di ricerca, abbiamo implementato un sistema di benchmarking basato sui risultati di tre noti motori di ricerca accessibili tramite web: *Google*, *DuckDuckGo* e *Bing*. Questo metodo ci consente di costruire un benchmark costituito da documenti RFC rilevanti per determinate query, associando a ciascun documento un punteggio di rilevanza.

3.1. Obiettivi del Benchmark

L'obiettivo principale è confrontare i risultati dei nostri motori di ricerca con un benchmark che rappresenta uno standard di riferimento. Il benchmark viene costruito raccogliendo i primi risultati restituiti dai tre motori di ricerca per specifiche query e calcolando un punteggio di rilevanza per ciascun documento basato su parametri specifici.

3.2. Costruzione del Banchmark

3.2.1. Fase 1: Recupero dei Risultati

Abbiamo sviluppato uno script, denominato **`autoUrlExtractor.py`**, che utilizza l'API *SerpApi* per effettuare scraping delle pagine di ricerca di *Google*, *DuckDuckGo* e *Bing*. Lo script esegue i seguenti passaggi:

1. Esegue una query predefinita su ciascun motore di ricerca, limitando i risultati al dominio degli RFC (**`site:https://www.rfc-editor.org/rfc/`**).
2. Recupera i primi risultati in formato URL, come ad esempio **`https://www.rfc-editor.org/rfc/rfcNUMBER.html`**, dove **`NUMBER`** rappresenta il numero identificativo dell'RFC.
3. Salva i risultati in un file JSON strutturato, associando ogni query ai relativi URL estratti dai motori di ricerca.

3.2.2. Fase 2: Calcolo della Rilevanza

Il secondo script, denominato **`createBenchMark.py`**, legge il file JSON generato nella fase precedente e calcola la Rilevanza di ciascun documento RFC. Questo processo si basa su due parametri principali: la **Frequenza del Documento** e il **Punteggio basato sulla Posizione** nei risultati.

3.2.2.1. Parametri Indiziali

Per ogni documento RFC, calcoliamo i seguenti parametri:

1. **Frequenza del Documento**: La frequenza indica in quanti motori di ricerca il documento compare tra i risultati. È definita come:

$$\text{Frequenza}_{\text{doc}} = \#\ \text{Numero di motori di ricerca in cui doc compare}$$

2. **Punteggio in base alla Posizione**: Per ciascun motore di ricerca, calcoliamo il punteggio di un documento in base alla sua posizione nei risultati. Il punteggio diminuisce con il crescere della posizione, secondo la seguente formula di decremento logaritmico:

$$\text{PunteggioSingolaPosizione}_{\text{doc,engine}} = \frac{1}{\log_2\left(\text{Posizione}_{\text{engine}}(\text{doc}) + 1\right)}$$

   Qui, la posizione $\mathbf{Posizione_{engine}(doc)}$ rappresenta l'indice del documento nei risultati di un motore di ricerca. Il punteggio totale del documento basato sulle posizioni viene calcolato sommando i singoli punteggi di tutti i motori di ricerca in cui il documento compare:

$$\text{PunteggioTotale}_{\text{doc}} = \sum_{i=1}^{E} \text{PunteggioSingolaPosizione}_{\text{doc},i}$$

3.2.2.2. Punteggio Finale

Successivamente calcoliamo per ciascun documento la Rilevanza combinando il Punteggio Totale e la Frequenza ottenute in precedenza, utilizzando la formula:

$$\text{Rilevanza}_{\text{doc}} = \text{PunteggioTotale}_{\text{rfc}} * (1 + \alpha * \text{Frequenza}_{\text{rfc}})$$

Dove α è un fattore di peso configurabile tra **`0`** e **`1`**, che determina l'importanza della frequenza.

3.2.2.3. Normalizzazione

Le Rilevanze vengono normalizzate tra 1 e 3 per rendere i risultati comparabili e più interpretabili:

$$\text{RilevanzaNormalizzata}_{\text{doc}} = \left\lfloor 2 \cdot \frac{\text{Rilevanza}_{\text{doc}} - \text{Min}}{\text{Max} - \text{Min}} \right\rceil + 1$$

Dove **Min** e **Max** sono rispettivamente la minima e la massima rilevanza calcolata per i documenti.

3.2.3. Risultato del Benchmark

Alla fine del processo, otteniamo un benchmark strutturato che associa a ogni query una lista di documenti RFC con i relativi punteggi di rilevanza. Questo benchmark fornisce un riferimento oggettivo per confrontare le prestazioni dei nostri motori di ricerca (*Whoosh*, *PyLucene*, *PostgreSQL*) rispetto ai motori di ricerca di riferimento (*Google*, *Duckduckgo*, *Bing*).

Di seguito un esempio di **output strutturato** dello script:

```
Testo della Query: 'Prompt query 2 site:rfc-editor.org'
Rfc: 9308, Rilevanza: 4.89279, Normalizzata: 3.00000, Arrotondata: 3
Rfc: 9001, Rilevanza: 4.00000, Normalizzata: 2.58616, Arrotondata: 3
...
Rfc: 9287, Rilevanza: 2.79203, Normalizzata: 2.02622, Arrotondata: 2
...
Rfc: 9297, Rilevanza: 0.60206, Normalizzata: 1.01109, Arrotondata: 1
Rfc: 9443, Rilevanza: 0.57813, Normalizzata: 1.00000, Arrotondata: 1
```

3.3. Locazione dei File Interessati

I due Scripts:

Gli script responsabili per lo scraping dei risultati ed il calcolo della rilevanza si trovano al seguente percorso:

`".../gestione-info/project/searchengine/myBanchmark/"`

- I file di interesse sono: **`autoUrlExtractor.py`**, **`createBenchMark.py`**