Trabalho Prático 5 - Recuperação da Informação

Matheus Aquino Motta[1]

[1]Bacharelado em Matemática Computacional
DCC - Universidade Federal de Minas Gerais

matheusaquino199@gmail.com.br 2018046513

Abstract. *In this report we will briefly discuss the implementation of assignment 5 of the subject Information Retrieval. The problem consisted in a follow-up to the assignment 4, where from the constructed inverted list, given a query of terms we need retrieve information about the links with the query terms, ranking the results with the vector model. Furthermore were provided some sample queries and analysis about the memory and execution time cost for querying each vocabulary term.*

Resumo. *Nesse relatório iremos discutir brevemente a implementação do trabalho prático 5 da disciplina Recuperação da Informação. O problema consistia em uma extensão do trabalho prático 4, onde a partir da lista invertida construída, dada uma requisição de termos precisamos recuperar informação a respeito dos documentos HTML em que os termos da requisição ocorrem, ranqueando-os a partir do modelo vetorial. Além disso, foram disponibilizados exemplos de requisições e uma análise relativa ao tempo de execução da requisição e custo de memória para a de cada termo do vocabulário.*

1. Introdução

No trabalho prático 4 fizemos a construção de uma lista invertida para os termos de uma coleção de 1000068 documentos HTML. A lista foi construída de modo que todos os termos encontrados foram armazenados em um arquivo único ordenado respectivamente de acordo com o identificador de cada termo, documento em que ocorre e sua respectiva posição de ocorrência. Assim, temos ao final uma lista dividida em blocos que armazenam informações relativas as ocorrências de cada termo.

Além disso, foram construídos outros documentos auxiliares, para armazenar informações relevantes a respeito dos termos individuais que compoem o vocabulário e URLs dos documentos indexados, assim como um dicionário de acesso aos blocos de informação específicos de cada termo da coleção. Isto é, para cada termo da lista invertida foi associado um identificador de acesso ao início e fim das linhas relativas a suas ocorrências na lista invertida, de modo que dado um termo podemos acessar informações relativas a ele de forma direta sem a necessidade de percorrer demais partes do arquivo.

Assim, foi construído um sistema simples de queries, cujo dado um termo único t somos capazes de recuperar em tempo quasi-linear (em função do número de ocorrências na lista invertida) todas as URLs em que t ocorre, assim como demais informações relevantes como o *TF-IDF*.

Agora, no trabalho prático 5 realizamos uma pequena alteração no sistema de requisições, de modo que agora recebemos uma requisição qualquer de termos únicos

ou múltiplos e realizamos a busca de modo a recuperar os documentos em que todos os termos dados na requisição ocorram. E por fim, após obter os documentos que satisfazem essa condição, ranqueá-los de acordo com o modelo vetorial, associando a cada documento uma pontuação de similaridade com a requisição dada.

2. Implementação

De um modo geral a implementação do algoritmo consiste no carregamento dos arquivos de dicionário, vocabulário e URLs construídos no trabalho prático anterior e busca nos blocos de informação relativos a cada termo no texto da requisição.

Os processos de construção dos arquivos foram discutidos detalhadamente no trabalho anterior. Entretanto vamos retomá-los de sucintamente para um entendimento mais detalhado do sistema de requisições.

2.1. Construção dos arquivos

Inicialmente, contamos com um conjunto de 1000068 documentos HTML que serão analisados por um *Parser* que irá extrair os termos do texto do código, assim como caracteres não relevantes. Assim, devido à larga quantidade de documentos que não cabem em memória principal particionamo nossa coleção em subconjuntos e realizamos a indexação em arquivos separadamente, atualizando o vocabulária de maneira apropriada e aferindo a cada termo e URL identificadores apropriados. Desse modo, para cada subconjunto ordenamos em memória principal cada sub-lista invertida respectivamente em função do identificador de cada termo, identificador documento e a posição em que o termo ocorre, mantendo assim termos iguais juntos em blocos na lista. Com isso, ao fim, com o uso de técnicas de ordenação em memória externa realizamos a união das sub-lista em um único arquivo mantendo a ordenação dos itens de maneira apropriada de acordo com os critérios supracitados.

Dessa forma, ao fim obtemos uma lista invertida única com informações acerca de todos os termos da coleção, um arquivo vocabulário armazenando os termos únicos dos documentos HTML analisados com seus respectivos identificadores e número de documentos em que cada termo i ocorre n_i, e um arquivo de URLs com os enlaces de cada documento da coleção e seu respectivo identificador.

Entretanto, embora agora tenhamos informação acerca de todos os termos relevantes dos textos da coleção precisamos de uma maneira eficiente de recuperar a informação relativa a cada um dos termos. Assim, foi construído um dicionário que armazena informação acerca de quais posições do arquivo devem ser acessadas para cada termo único do vocabulário. Onde percorremos a nossa lista final resultante e para cada identificador (termo) salvamos a posição em bytes do arquivo onde o identificador ocorre a primeira vez, e última. Dessa forma, podemos acessar em tempo quasi-linear (em função do número de ocorrências do termo) as informações relativas a todas as ocorrências do termo na coleção na lista invertida, como o *IDF* de cada um dos termos.

2.2. Requisições e o Modelo Vetorial

Com os arquivos devidamente criados podemos realizar requisições de maneira minimamente eficiente para termos únicos ou múltiplos na coleção e obter resultados (minimamente) relevantes a partir de uma simples utilização do modelo vetorial.

Podemos visualizar o funcionamento das requisições a partir do fluxograma abaixo

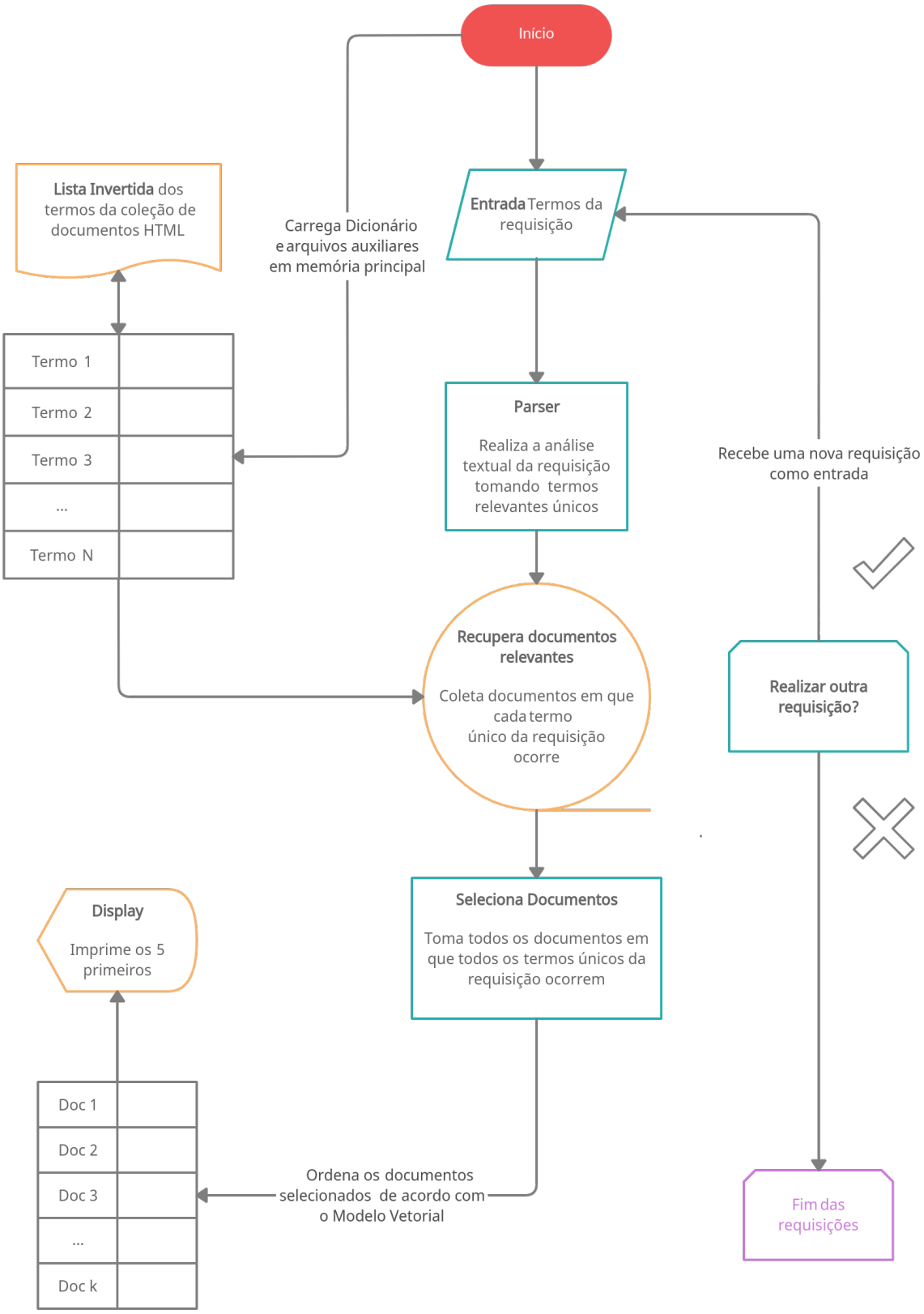


Figure 1. Fluxograma de funcionamento do sistema de requisições.

Inicialmente recebemos uma linha de texto como requisição em forma de *string*, q com k termos únicos, onde queremos recuperar às URLs dos documentos cujo todos os termos da requisição ocorram. Desse modo, assim como fizemos na indexação dos termos das páginas, de modo a garantir consistência entre os termos da requisição e os termos do vocabulário da coleção, realizamos uma análise textual da *string* q, removendo caracteres de maneira apropriada de modo equivalente ao realizado no *parsing* dos documentos HTML.

Isto é, iremos remover todos os caracteres de marcação e pontução, e todos aqueles que não sejam pertencentes ao alfabeto latino, algarismos arábicos ou algumas de suas extensões. Assim, obtendo uma *string* de termos potencialmente relevantes que serão a priori analisados individualmente, considerando apenas os termos únicos e quantas vezes cada termo ocorre na requsição.

Nesse sentido, para cada termo t da requisição iremos armazenar os identificadores dos documentos em que t ocorre. Com isso, ao fim teremos informação acerca de quais documentos possuem todos termos de interesse na requisição em seu texto, que para o escopo desse trabalho constituem a primeira parte da nossa resposta. A coeleção de documentos minimamente relevantes para a requisição.

Tomados os documentos que satisfazem as condições inciais, iremos ranqueá-los a partir do conceito de similaridade por cossenos, ou mais especificamente, o Modelo Vetorial.

Cada requisição q pode ser representada por meio de um vetor

$$\vec{q} = (w_{1q}, w_{2q}, \ldots, w_{kq})$$

para qual os elementos w_{iq} representam o peso do i-ésimo termo na requisição q, que pode ser obtido através da fórmula

$$w_{iq} = (1 + \log_2 f_{i,q}) \times \log_2 \frac{N}{n_i}$$

Onde $f_{i,q}$ representa TF, isto é, a frequência do termo i na requisição q, isto é, o número de vezes que i ocorre na requisição e $log_2 frac N n_i$ representa o IDF do termo i na coleção, onde N é o número total de documentos da coleção e n_i o número de documentos em que i ocorre.

De modo análogo, cada documento d_j será representado por um vetor

$$\vec{d_j} = (w_{1j}, w_{2j}, \ldots, w_{kj})$$

para qual os elementos w_{ij} representam o peso i-ésimo termo no documento j, que pode ser obtido por uma fórmula similar à representada acima

$$w_{ij} = (1 + \log_2 f_{i,j}) \times \log_2 \frac{N}{n_i}$$

Só que agora estamos interessados nas informações relativas ao documento j.

Assim, calculando a similaridade entre o vetor da requisição para todos os vetores de documentos, podemos obter os potencialmente mais relevantes para a nossa requisição a partir de uma pontuação, *score*, dada pela similaridade entre os cossenos dos vetores $\vec{q}$ e $\vec{d_j}$.

Resultado que pode ser obtido a partir da equação abaixo

$$\begin{aligned}\text{sim}(\vec{d_j}, \vec{q}) = \cos(\theta) &= \frac{\langle d_j, q\rangle}{\|d_j\|_2 \|q\|_2} \\ &= \frac{\sum_{i=0}^{k} w_{iq} \times w_{ij}}{\sqrt{\sum_{i=0}^{k} w_{ij}^2} \times \sqrt{\sum_{i=0}^{k} w_{iq}^2}}\end{aligned}$$

Assim, ordenando a lista de documentos em função da sua similaridade com a requisição, obtemos os enlaces potencialmente mais relevantes.

A complexidade de tempo de cada requisição é dada pela ordem de

$$\mathcal{O}(kN + |D| \log |D|)$$

Isto é, para cada um dos k termos percorremos todas as suas N ocorrências na lista invertida de custo $\mathcal{O}(kN)$ e posteriormente para todos aqueles documentos D que possuem todos os termos da requisição realizamos uma ordenação de custo $\mathcal{O}(|D| \log |D|)$

2.3. Bibliotecas e Estruturas de Dados

Para a implementação do trabalho foram utilizadas diversas estruturas da biblioteca STL do C++, como a estrutura *vector* para representação das listas criadas devido a maior praticidade no acesso das informações em $O(1)$ e alocação dinâmica de memória, com complexidade de espaço da ordem de $O(n)$, onde n é o tamanho do vetor.

Além disso, foi feito amplo uso da estrutura *unordered map* para realizar o *hashing* das palavras e compressão de *strings* em identificadores inteiros, para maior dinamicidade no acesso de posições de memória em estruturas e outras informações relevantes, onde cada operação realizada possui custo de tempo médio de $O(1)$ e complexidade de espaço $O(n)$, onde n é o número de elementos inseridos na estrutura.

Outras funcionalidades da STL do C++ que foram sistematicamente utilizadas são as funções de manipulações de arquivos *tellg* e *seekg* para a busca pelos blocos específicos de cada termo na lista invertida de maneira minimamente eficiente a partir das posições em bytes do arquivo.

3. Resultados

O algoritmo foi desenvolvido para a requisição de informações relativas a um conjunto de termos específicos dados como entrada na lista invertida construída no trabalho prático anterior. E foi executado em uma máquina de 64 BITS com processador INTEL(R) CORE(TM) I5-6400 CPU @ 2.70GHZ X 4 e memória RAM de 7.7 GB.

O algoritmo possui quatro formas de execução, sendo elas

- **$ ./main -b** para a construção dos arquivos.
- **$ ./main -q** para a requisição lista invertida.
- **$ ./main -t** para a requisição de um termo único na lista invertida.
- **$ ./main -c** para obter informações individuais acerca das requisições de todos os termos do vocabulário.

Para a requisição de termos únicos na lista invertida, para todos os 1790353 termos do vocabulário, o sistema de requisições implementado leva aproximadamente em média um tempo de $\mu_t = 0.0010236$, com um desvio padrão de aproximadamente $\sigma_t = 0.0487313$. Entretanto, esses números são pequenos devido aos muitos termos que ocorrem apenas uma vez na coleção, de baixa relevância para fins práticos.

Assim, Considerando diferentes conjuntos dos termos que mais ocorrem temos uma média e desvio padrão dadas pela tabela abaixo

Top termos mais frequentes	Tempo de requisição médio μ_t	Desvio Padrão Médio σ_t
10	15.5879680	8.73970647
10^2	3.77439119	4.97782301
10^3	0.81650909	1.87381318
10^4	0.14554005	0.63548213
10^5	0.01759133	0.20548803

Table 1. Tempo de execução médio e desvio padrão para a requisição dos termos mais frequentes do vocabulário

Essas variações na média e desvio padrão elevadas são devido ao fato que muitos termos são extremamente recorrentes na coleção como, os top 10 que são majoritariamente artigos e preposições comuns da Língua Portuguesa.

Podemos observar essa variação através do gráfico abaixo

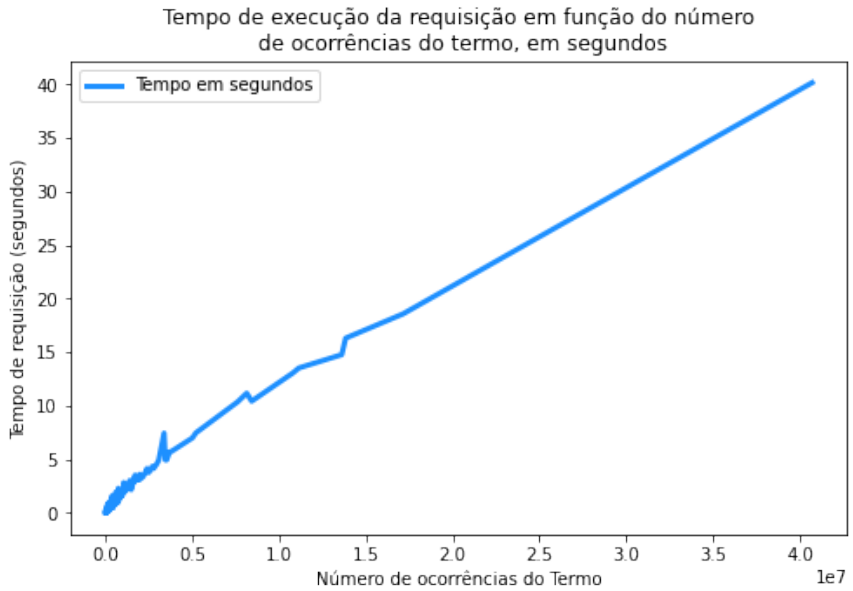


Figure 2. Tempo de requisição de cada termo em função do número de ocorrências.

Uma análise similar pode ser feita para o tamnho dos blocos de dados relativos à cada termo na lista em Megabytes. Onde para todos os termos do vocabulário temos uma média de memória ocupada na lista de aproximadamente $\mu_m = 0.00721086$ e $\sigma_m = 0.59426510$.

Podemos observar o espaço ocupado em memória em função do número de ocorrências na lista invertida também a partir do gráfico abaixo

Top termos mais frequentes	Memória média ocupada pelo bloco μ_m
10	196.2449376
10^2	37.94120014
10^3	6.73974207
10^4	1.08061915
10^5	0.12673828

Table 2. Espaço médio ocupado pelos blocos de de cada termo na lista invertida para os termos mais ocorrentes

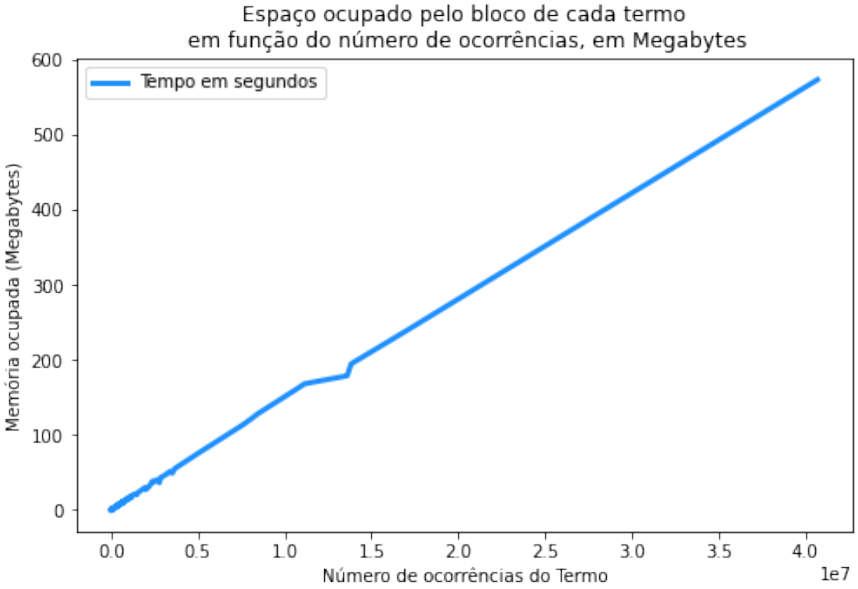


Figure 3. Espaço ocupado por cada bloco na lista invertida, em função do número de ocorrência dos termos.

Além disso, é possível visualizar a disparidade de tempo de requisição e espaço ocupado para cada um dos termos ao analisar os 10 termos mais frequentes individualmente

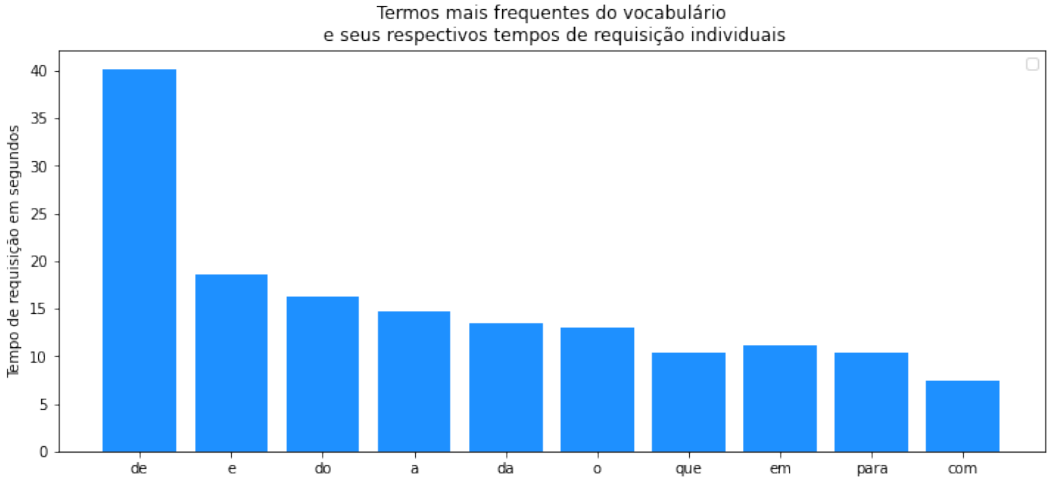


Figure 4. Termos mais frequentes e seus respectivos tempos de requisição

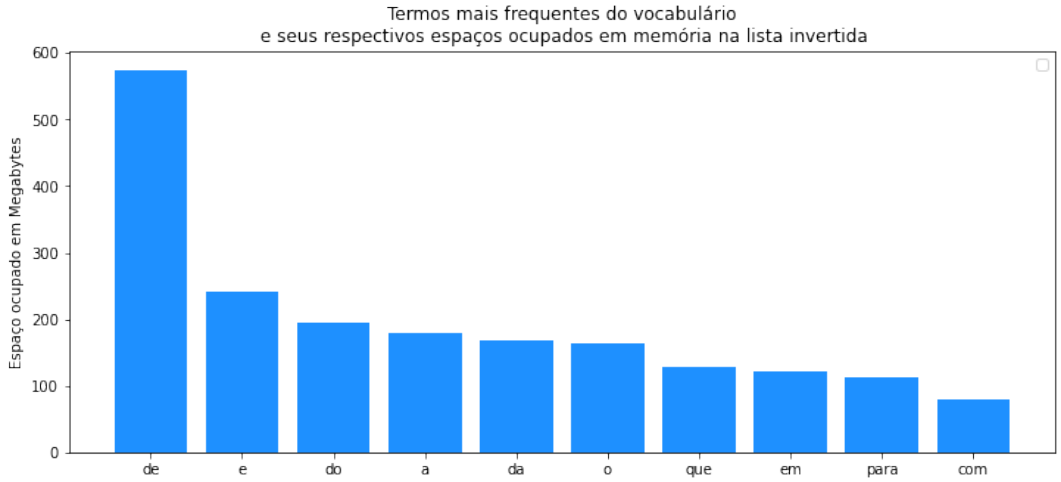


Figure 5. Termos mais frequentes e seus respectivos custos em memória na lista invertida.

A partir dessa análise quantitativa foi possível verificar como o tempo de requisição e espaço ocuapado na lista invertida para cada termo dependem significativamente de sua frequência. Assim, considerando que parte signficativa dessas palavras são *stopwords*, ”palavras vazias”, algumas técnicas poderiam ser utilizadas para otimização de tempo e espaço para alguns casos de busca.

Agora de um ponto de vista qualitativo, 25 exemplos de requisições foram disponibilizados, onde dentre eles 3 sub-conjuntos de consultas podem ser evidenciados. As 9 primeiras são relativas a nomes de Autores conhecidos da literatura internacional, clássicos e contemporâneos: ”Guimarães Rosa”, ”Jane Austen”, ”George Orwell”, ”Tolkien”, ”Muhammad Yunus”, ”Fiódor Dostoiévski”, ”Michael J Sandel”, ”Yuval Noah Harari” e ”Harper Lee”. Essas buscas trouxeram majoritariamente resultados relevantes, recuperando sempre documentos com alguma relação aos autores propriamente ou alguma de suas obras.

Em consonância, outras 9 requisições foram relativas a atualidades recorrentes em notícias do período em que a coleta foi realizada, ”computação quântica”, ”vacinação Brasil”, ”variante brasileira”, ”eleições EUA”, ”uso emergencial vacinas”, ”rover Perseverance”, ”Kamala Harris”, ”coronavírus BH”, ”golpe Myanmar”, apresentando resultados positivos em sua maioria, com notícias, artigos ou reportagens relativas à cada uma das requisições.

Por fim, o último conjunto de 6 requisições foi construido de modo a demonstrar as fraquezas do buscador, haja vista que são requisições que possuem um significado no conjunto de termos como um todo, ou devido à popularidade subjetiva dos termos. As requisições ”O Príncipe”, ”1984”, não tiveram no cabeçalho exemplos de enlaces que diziam respeito às obras literárias de Nicolau Maquiavel e George Orwell. Assim como para as demais requisições como o filme ”irmão urso”, o fenômeno da teoria do caos ”efeito borboleta”, o ”campeonato brasileiro” de futebol, e os renomados autores ”Fernando Pessoa” e ”Victor Hugo”. Não tiveram resultados satisfatórios, haja vista que possuem significados específicos quando consideramos os termos em conjunto, mas todavia como é frequente encontrar tais termos em posições não adjacentes no texto, temos que resultado potencialmente não relevantes também são retornados.

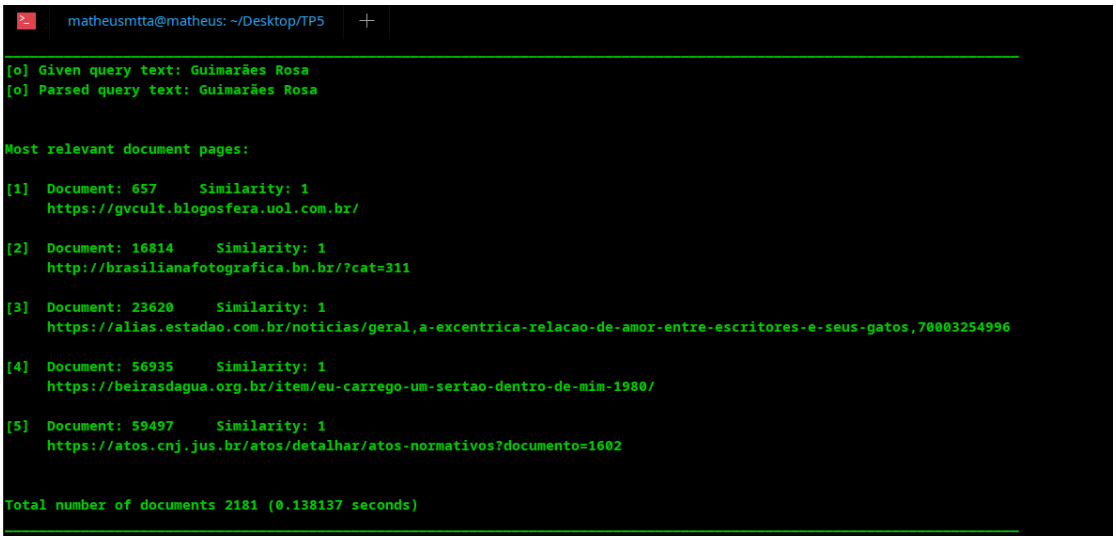
4. Conclusões

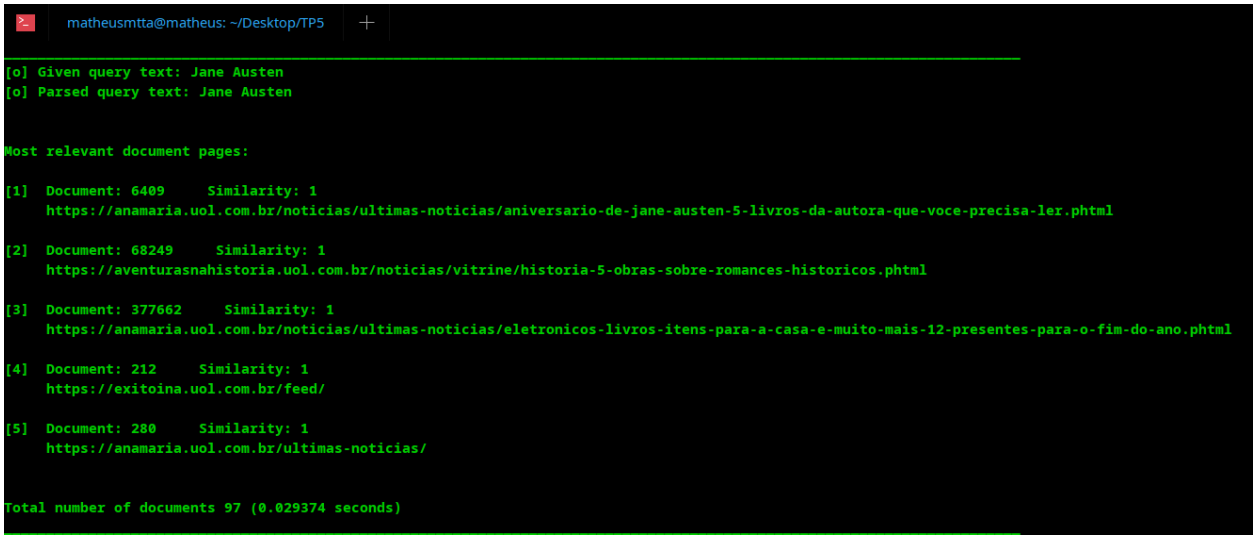
Durante a execução desse trabalho tivemos a oportunidade de utilizar os arquivos construídos no trabalho prático 4 para implementação do modelo vetorial para o ranqueamento de documentos recuperados para uma dada requisição.

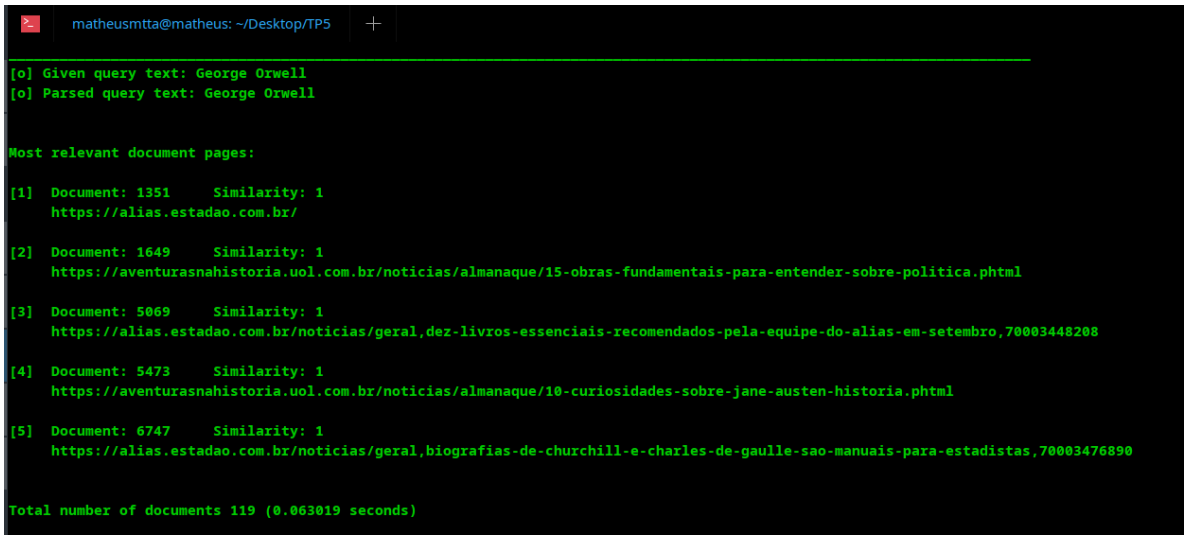
Embora os resultados não sejam significativamente positivos do ponto de vista prático, devido à simplicidade das técnicas utilizadas, foi extremamente positivo e empolgante observar os resultados da implementação de um sistema de requisições minimamente eficientes em uma coleção de 1000068 documentos HTML.

Algumas modificações podem ser feitas para a otimização dos resultados, como o uso do *Page Rank*, e a consideração da posição dos termos no documento para requisições de múltiplos termos. Além disso, uma alteração simples que será realizada para garantir a consistência semântica do buscador é uma modificação na análise textual para colocar todos os caracteres dos textos da coleção em caixa baixa, minúsculos, para garantir que o valor semântico de buscas seja preservado para requisições gramaticalmente distintas, como ”Computação Quântica” e ”computação quântica”.

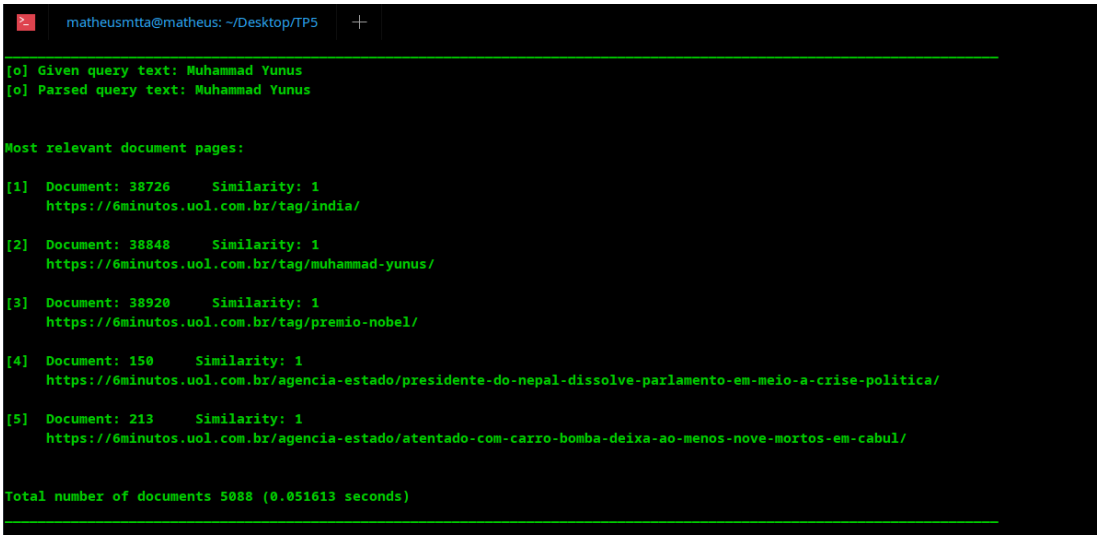
5. Apêndice

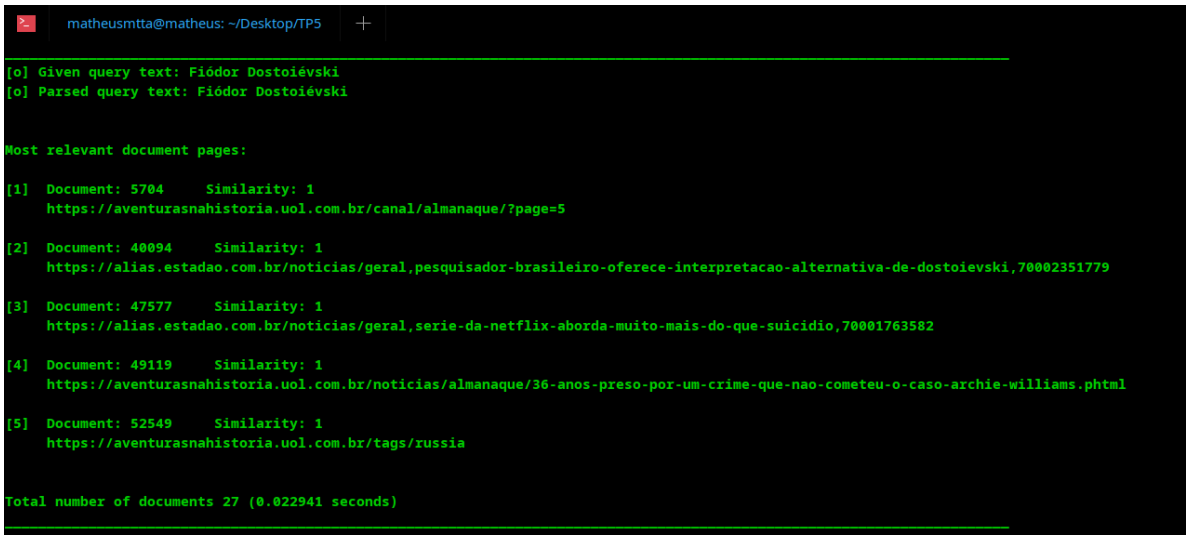

matheusmtta@matheus: ~/Desktop/TP5
[o] Given query text: Guimarães Rosa
[o] Parsed query text: Guimarães Rosa
Most relevant document pages:
[1] Document: 657 Similarity: 1
https://gvcult.blogosfera.uol.com.br/
[2] Document: 16814 Similarity: 1
http://brasilianafotografica.bn.br/?cat=311
[3] Document: 23620 Similarity: 1
https://alias.estadao.com.br/noticias/geral,a-excentrica-relacao-de-amor-entre-escritores-e-seus-gatos,70003254996
[4] Document: 56935 Similarity: 1
https://beirasdagua.org.br/item/eu-carrego-um-sertao-dentro-de-mim-1980/
[5] Document: 59497 Similarity: 1
https://atos.cnj.jus.br/atos/detalhar/atos-normativos?documento=1602
Total number of documents 2181 (0.138137 seconds)

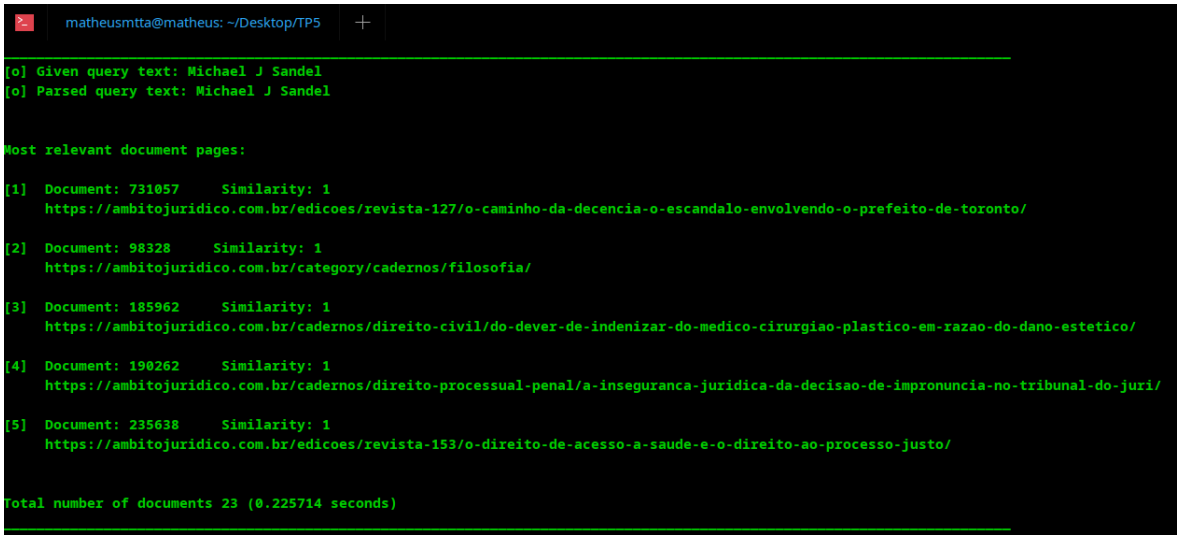

matheusmtta@matheus: ~/Desktop/TP5
[o] Given query text: Jane Austen
[o] Parsed query text: Jane Austen
Most relevant document pages:
[1] Document: 6409 Similarity: 1
https://anamaria.uol.com.br/noticias/ultimas-noticias/aniversario-de-jane-austen-5-livros-da-autora-que-voce-precisa-ler.phtml
[2] Document: 68249 Similarity: 1
https://aventurasnahistoria.uol.com.br/noticias/vitrine/historia-5-obras-sobre-romances-historicos.phtml
[3] Document: 377662 Similarity: 1
https://anamaria.uol.com.br/noticias/ultimas-noticias/eletronicos-livros-itens-para-a-casa-e-muito-mais-12-presentes-para-o-fim-do-ano.phtml
[4] Document: 212 Similarity: 1
https://exitoina.uol.com.br/feed/
[5] Document: 280 Similarity: 1
https://anamaria.uol.com.br/ultimas-noticias/
Total number of documents 97 (0.029374 seconds)

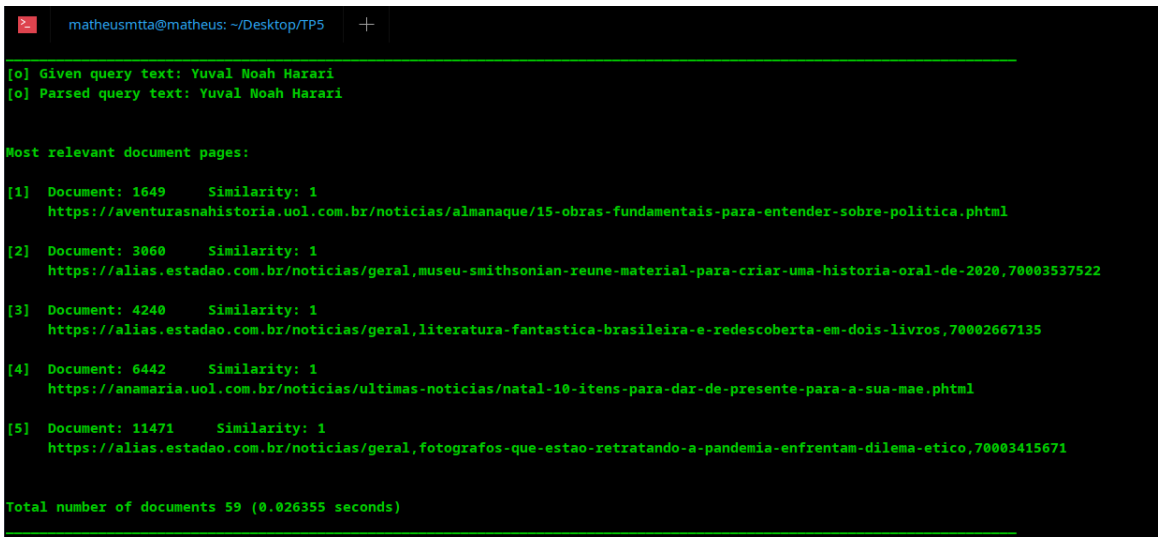

matheusmtta@matheus: ~/Desktop/TP5
[o] Given query text: George Orwell
[o] Parsed query text: George Orwell
Most relevant document pages:
[1] Document: 1351 Similarity: 1
https://alias.estadao.com.br/
[2] Document: 1649 Similarity: 1
https://aventurasnahistoria.uol.com.br/noticias/almanaque/15-obras-fundamentais-para-entender-sobre-politica.phtml
[3] Document: 5069 Similarity: 1
https://alias.estadao.com.br/noticias/geral,dez-livros-essenciais-recomendados-pela-equipe-do-alias-em-setembro,70003448208
[4] Document: 5473 Similarity: 1
https://aventurasnahistoria.uol.com.br/noticias/almanaque/10-curiosidades-sobre-jane-austen-historia.phtml
[5] Document: 6747 Similarity: 1
https://alias.estadao.com.br/noticias/geral,biografias-de-churchill-e-charles-de-gaulle-sao-manuais-para-estadistas,70003476890
Total number of documents 119 (0.063019 seconds)

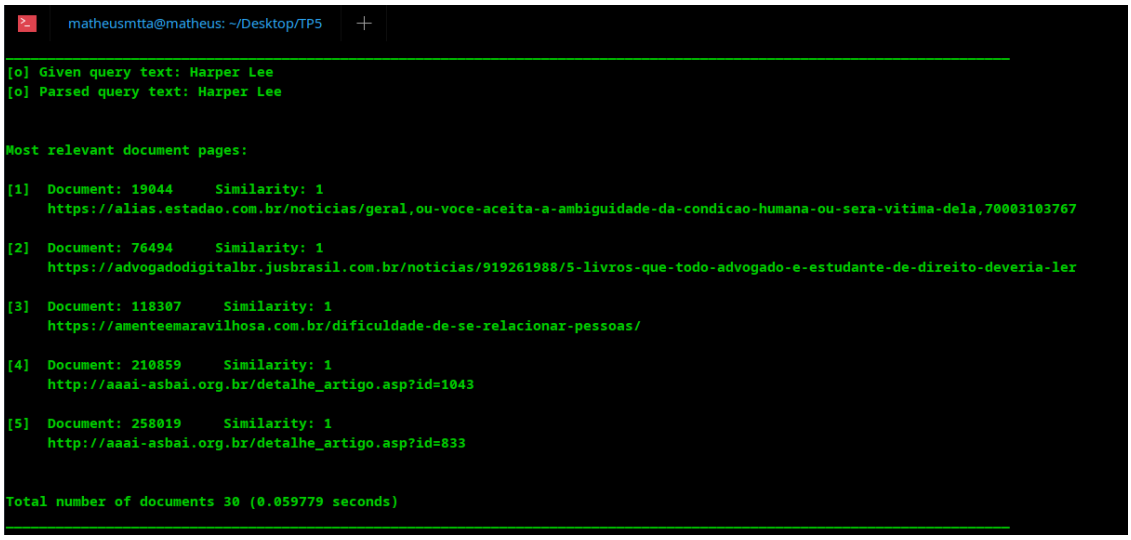

matheusmtta@matheus: ~/Desktop/TP5
[o] Given query text: Tolkien
[o] Parsed query text: Tolkien
Most relevant document pages:
[1] Document: 1461 Similarity: 1
https://cultura.estadao.com.br
[2] Document: 3067 Similarity: 1
https://cultura.estadao.com.br/blogs/babel/um-livro-por-semana-36-tolkien-o-pai-cartas-do-papai-noel-e-sr-boaventura/
[3] Document: 5847 Similarity: 1
https://alias.estadao.com.br/noticias/geral,no-centenario-do-fim-da-1-guerra-mundial-livros-recriam-periodo,70002595710
[4] Document: 6977 Similarity: 1
https://cultura.estadao.com.br/moda
[5] Document: 8832 Similarity: 1
https://alias.estadao.com.br/noticias/geral,em-tempos-sombrios-as-pessoas-olham-para-escritores-afirma-ursula-k-le-guin,70001753157
Total number of documents 74 (0.013597 seconds)

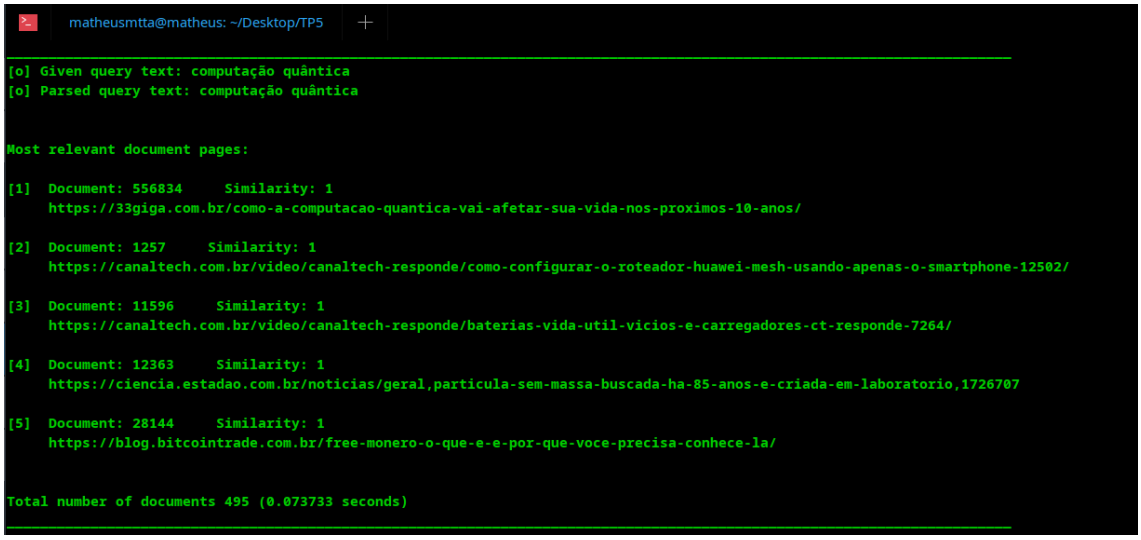

matheusmtta@matheus: ~/Desktop/TP5
[o] Given query text: Muhammad Yunus
[o] Parsed query text: Muhammad Yunus
Most relevant document pages:
[1] Document: 38726 Similarity: 1
https://6minutos.uol.com.br/tag/india/
[2] Document: 38848 Similarity: 1
https://6minutos.uol.com.br/tag/muhammad-yunus/
[3] Document: 38920 Similarity: 1
https://6minutos.uol.com.br/tag/premio-nobel/
[4] Document: 150 Similarity: 1
https://6minutos.uol.com.br/agencia-estado/presidente-do-nepal-dissolve-parlamento-em-meio-a-crise-politica/
[5] Document: 213 Similarity: 1
https://6minutos.uol.com.br/agencia-estado/atentado-com-carro-bomba-deixa-ao-menos-nove-mortos-em-cabul/
Total number of documents 5088 (0.051613 seconds)

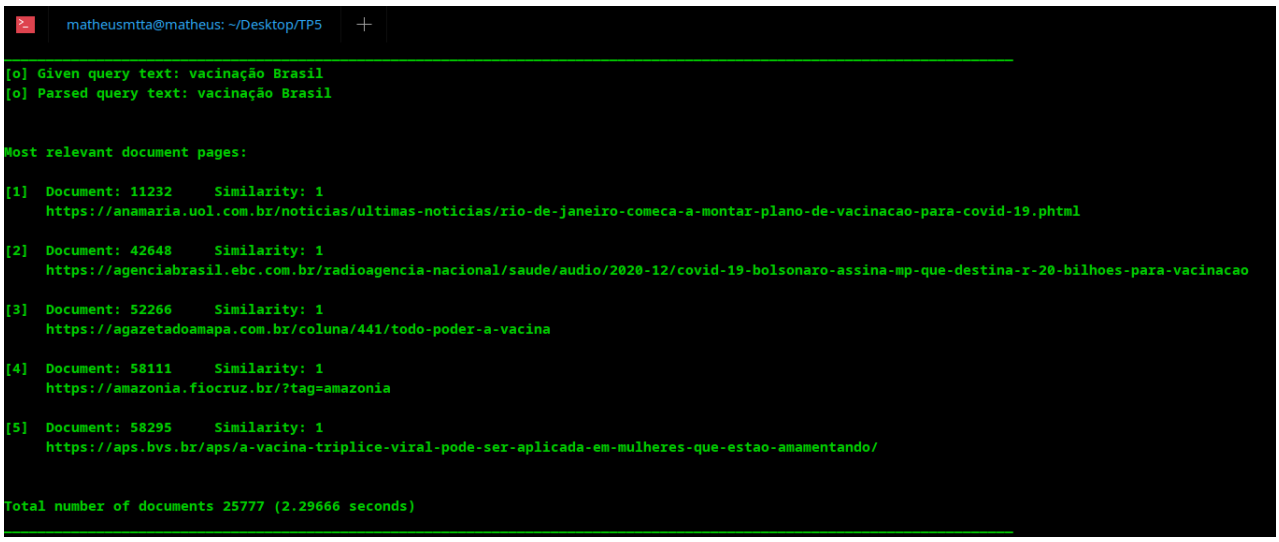

matheusmtta@matheus: ~/Desktop/TP5
[o] Given query text: Fiódor Dostoiévski
[o] Parsed query text: Fiódor Dostoiévski
Most relevant document pages:
[1] Document: 5704 Similarity: 1
https://aventurasnahistoria.uol.com.br/canal/almanaque/?page=5
[2] Document: 40094 Similarity: 1
https://alias.estadao.com.br/noticias/geral,pesquisador-brasileiro-oferece-interpretacao-alternativa-de-dostoievski,70002351779
[3] Document: 47577 Similarity: 1
https://alias.estadao.com.br/noticias/geral,serie-da-netflix-aborda-muito-mais-do-que-suicidio,70001763582
[4] Document: 49119 Similarity: 1
https://aventurasnahistoria.uol.com.br/noticias/almanaque/36-anos-preso-por-um-crime-que-nao-cometeu-o-caso-archie-williams.phtml
[5] Document: 52549 Similarity: 1
https://aventurasnahistoria.uol.com.br/tags/russia
Total number of documents 27 (0.022941 seconds)

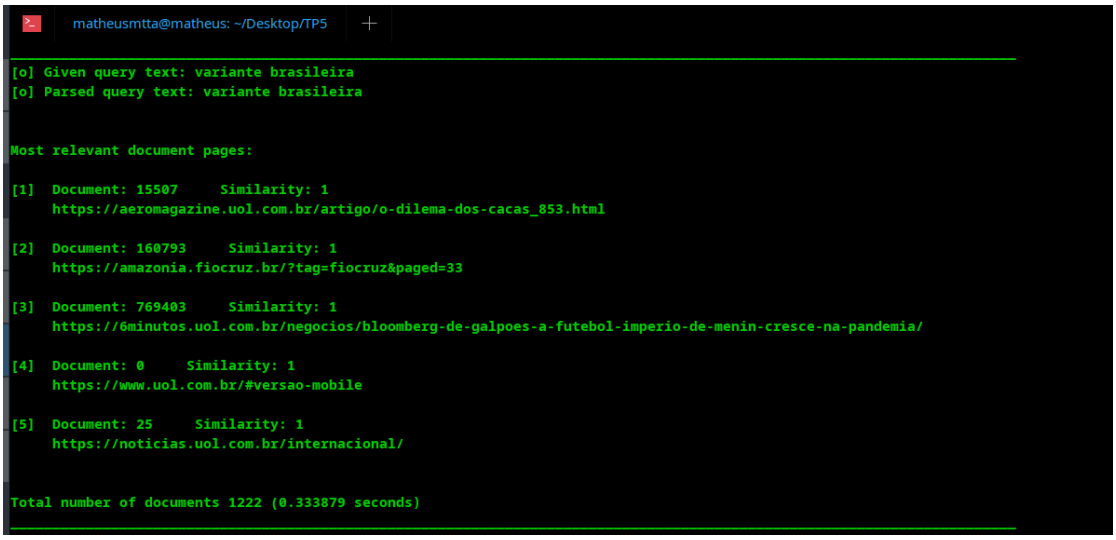

matheusmtta@matheus: ~/Desktop/TP5
[o] Given query text: Michael J Sandel
[o] Parsed query text: Michael J Sandel
Most relevant document pages:
[1] Document: 731057 Similarity: 1
https://ambitojuridico.com.br/edicoes/revista-127/o-caminho-da-decencia-o-escandalo-envolvendo-o-prefeito-de-toronto/
[2] Document: 98328 Similarity: 1
https://ambitojuridico.com.br/category/cadernos/filosofia/
[3] Document: 185962 Similarity: 1
https://ambitojuridico.com.br/cadernos/direito-civil/do-dever-de-indenizar-do-medico-cirurgiao-plastico-em-razao-do-dano-estetico/
[4] Document: 190262 Similarity: 1
https://ambitojuridico.com.br/cadernos/direito-processual-penal/a-inseguranca-juridica-da-decisao-de-impronuncia-no-tribunal-do-juri/
[5] Document: 235638 Similarity: 1
https://ambitojuridico.com.br/edicoes/revista-153/o-direito-de-acesso-a-saude-e-o-direito-ao-processo-justo/
Total number of documents 23 (0.225714 seconds)


matheusmtta@matheus: ~/Desktop/TP5
[o] Given query text: Yuval Noah Harari
[o] Parsed query text: Yuval Noah Harari
Most relevant document pages:
[1] Document: 1649 Similarity: 1
https://aventurasnahistoria.uol.com.br/noticias/almanaque/15-obras-fundamentais-para-entender-sobre-politica.phtml
[2] Document: 3060 Similarity: 1
https://alias.estadao.com.br/noticias/geral,museu-smithsonian-reune-material-para-criar-uma-historia-oral-de-2020,70003537522
[3] Document: 4240 Similarity: 1
https://alias.estadao.com.br/noticias/geral,literatura-fantastica-brasileira-e-redescoberta-em-dois-livros,70002667135
[4] Document: 6442 Similarity: 1
https://anamaria.uol.com.br/noticias/ultimas-noticias/natal-10-itens-para-dar-de-presente-para-a-sua-mae.phtml
[5] Document: 11471 Similarity: 1
https://alias.estadao.com.br/noticias/geral,fotografos-que-estao-retratando-a-pandemia-enfrentam-dilema-etico,70003415671
Total number of documents 59 (0.026355 seconds)


matheusmtta@matheus: ~/Desktop/TP5
[o] Given query text: Harper Lee
[o] Parsed query text: Harper Lee
Most relevant document pages:
[1] Document: 19044 Similarity: 1
https://alias.estadao.com.br/noticias/geral,ou-voce-aceita-a-ambiguidade-da-condicao-humana-ou-sera-vitima-dela,70003103767
[2] Document: 76494 Similarity: 1
https://advogadodigitalbr.jusbrasil.com.br/noticias/919261988/5-livros-que-todo-advogado-e-estudante-de-direito-deveria-ler
[3] Document: 118307 Similarity: 1
https://amenteemaravilhosa.com.br/dificuldade-de-se-relacionar-pessoas/
[4] Document: 210859 Similarity: 1
http://aaai-asbai.org.br/detalhe_artigo.asp?id=1043
[5] Document: 258019 Similarity: 1
http://aaai-asbai.org.br/detalhe_artigo.asp?id=833
Total number of documents 30 (0.059779 seconds)


matheusmtta@matheus: ~/Desktop/TP5
[o] Given query text: computação quântica
[o] Parsed query text: computação quântica
Most relevant document pages:
[1] Document: 556834 Similarity: 1
https://33giga.com.br/como-a-computacao-quantica-vai-afetar-sua-vida-nos-proximos-10-anos/
[2] Document: 1257 Similarity: 1
https://canaltech.com.br/video/canaltech-responde/como-configurar-o-roteador-huawei-mesh-usando-apenas-o-smartphone-12502/
[3] Document: 11596 Similarity: 1
https://canaltech.com.br/video/canaltech-responde/baterias-vida-util-vicios-e-carregadores-ct-responde-7264/
[4] Document: 12363 Similarity: 1
https://ciencia.estadao.com.br/noticias/geral,particula-sem-massa-buscada-ha-85-anos-e-criada-em-laboratorio,1726707
[5] Document: 28144 Similarity: 1
https://blog.bitcointrade.com.br/free-monero-o-que-e-e-por-que-voce-precisa-conhece-la/
Total number of documents 495 (0.073733 seconds)

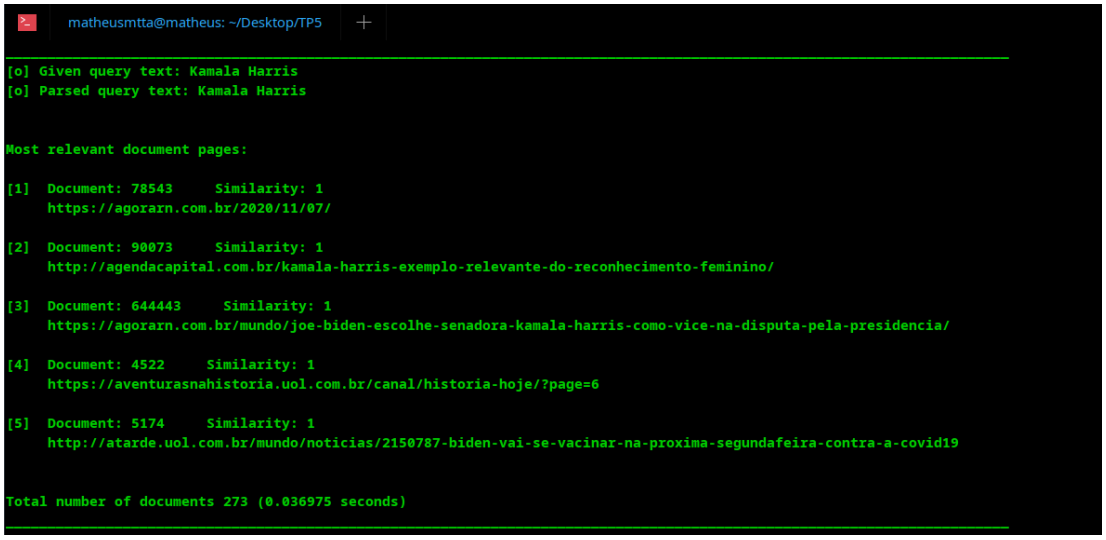

matheusmtta@matheus: ~/Desktop/TP5
[o] Given query text: vacinação Brasil
[o] Parsed query text: vacinação Brasil
Most relevant document pages:
[1] Document: 11232 Similarity: 1
https://anamaria.uol.com.br/noticias/ultimas-noticias/rio-de-janeiro-comeca-a-montar-plano-de-vacinacao-para-covid-19.phtml
[2] Document: 42648 Similarity: 1
https://agenciabrasil.ebc.com.br/radioagencia-nacional/saude/audio/2020-12/covid-19-bolsonaro-assina-mp-que-destina-r-20-bilhoes-para-vacinacao
[3] Document: 52266 Similarity: 1
https://agazetadoamapa.com.br/coluna/441/todo-poder-a-vacina
[4] Document: 58111 Similarity: 1
https://amazonia.fiocruz.br/?tag=amazonia
[5] Document: 58295 Similarity: 1
https://aps.bvs.br/aps/a-vacina-triplice-viral-pode-ser-aplicada-em-mulheres-que-estao-amamentando/
Total number of documents 25777 (2.29666 seconds)

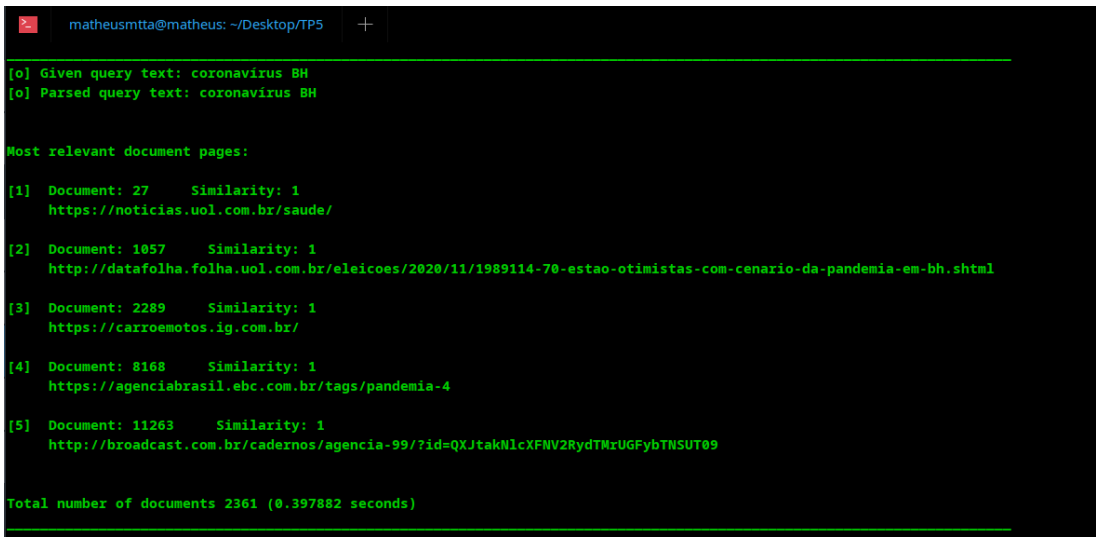

matheusmtta@matheus: ~/Desktop/TP5
[o] Given query text: variante brasileira
[o] Parsed query text: variante brasileira
Most relevant document pages:
[1] Document: 15507 Similarity: 1
https://aeromagazine.uol.com.br/artigo/o-dilema-dos-cacas_853.html
[2] Document: 160793 Similarity: 1
https://amazonia.fiocruz.br/?tag=fiocruz&paged=33
[3] Document: 769403 Similarity: 1
https://6minutos.uol.com.br/negocios/bloomberg-de-galpoes-a-futebol-imperio-de-menin-cresce-na-pandemia/
[4] Document: 0 Similarity: 1
https://www.uol.com.br/#versao-mobile
[5] Document: 25 Similarity: 1
https://noticias.uol.com.br/internacional/
Total number of documents 1222 (0.333879 seconds)

```
[o] Given query text: eleições EUA
[o] Parsed query text: eleições EUA


Most relevant document pages:

[1]  Document: 241189     Similarity: 1
     https://18horas.com.br/mundo/trump-admite-vitoria-de-joe-biden-nas-eleicoes-dos-eua-e-depois-volta-a-acusar-fraudes-no-pleito/

[2]  Document: 643695     Similarity: 1
     https://andrebona.com.br/tag/eleicoes-nos-eua/

[3]  Document: 906396     Similarity: 1
     https://agorarn.com.br/ultimas/com-apuracao-nas-eleicoes-dos-eua-mercados-internacionais-tem-manha-de-instabilidade/

[4]  Document: 919545     Similarity: 1
     https://aprovinciadopara.com.br/eleicoes-nos-eua-urnas-abrem-no-leste-95-milhoes-ja-votaram/

[5]  Document: 29     Similarity: 1
     https://noticias.uol.com.br/confere/


Total number of documents 9778 (0.393188 seconds)
```

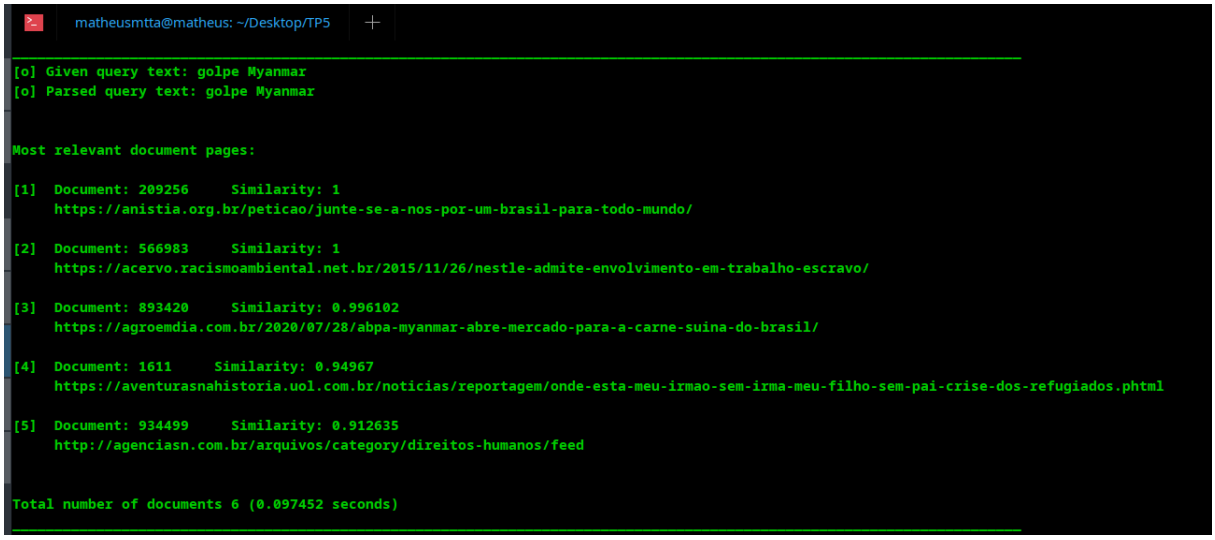
```
matheusmtta@matheus: ~/Desktop/TP5
[o] Given query text: uso emergencial vacinas
[o] Parsed query text: uso emergencial vacinas


Most relevant document pages:

[1]  Document: 87     Similarity: 1
     https://cultura.uol.com.br/

[2]  Document: 1560     Similarity: 1
     https://coronavirus.atarde.com.br/category/brasil/

[3]  Document: 7265     Similarity: 1
     https://bigdata.icict.fiocruz.br/desestimulo-oficial-vacina-prejudica-o-pais

[4]  Document: 13569     Similarity: 1
     https://brpolitico.com.br/tags/anvisa/

[5]  Document: 17328     Similarity: 1
     https://brpolitico.com.br/noticias/leia-o-plano-nacional-de-imunizacao-contra-a-covid-19/


Total number of documents 7712 (0.869078 seconds)
```

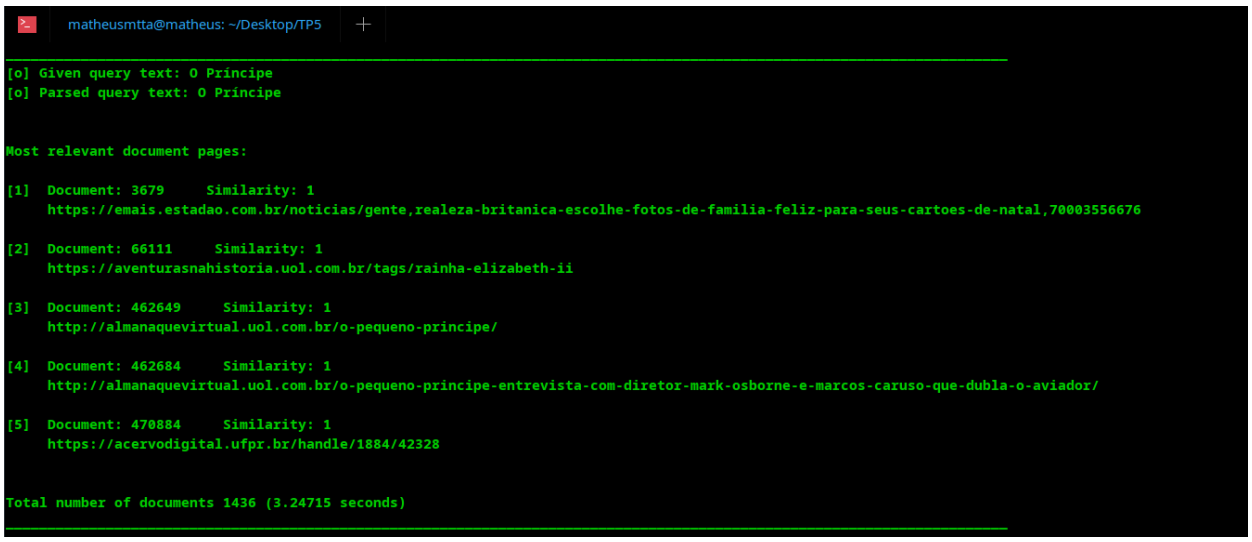
```
matheusmtta@matheus: ~/Desktop/TP5
[o] Given query text: rover Perseverance
[o] Parsed query text: rover Perseverance


Most relevant document pages:

[1]  Document: 819458     Similarity: 1
     http://ansabrasil.com.br/brasil/noticias/brasil/tecnologia/2020/07/30/nasa-lanca-nova-missao-para-marte-em-busca-de-tracos-de-vida_f9ff90bb-8a3d-4141-83
8f93.html

[2]  Document: 997078     Similarity: 1
     http://ansabrasil.com.br/brasil/noticias/brasil/tecnologia/2020/07/13/paises-enviarao-3-missoes-para-marte-a-partir-do-dia-157_3d8a16c5-1b2c-4403-93ba-7
.html

[3]  Document: 6359     Similarity: 0.994651
     https://ciencia.estadao.com.br/noticias/geral,nasa-lanca-nesta-quinta-feira-robo-perseverance-para-a-superficie-de-marte,70003380718

[4]  Document: 6024     Similarity: 0.8972
     https://canaltech.com.br/espaco/cao-robo-au-spot-podera-explorar-a-superficie-e-as-cavernas-de-marte-176534/


Total number of documents 4 (0.024676 seconds)
```

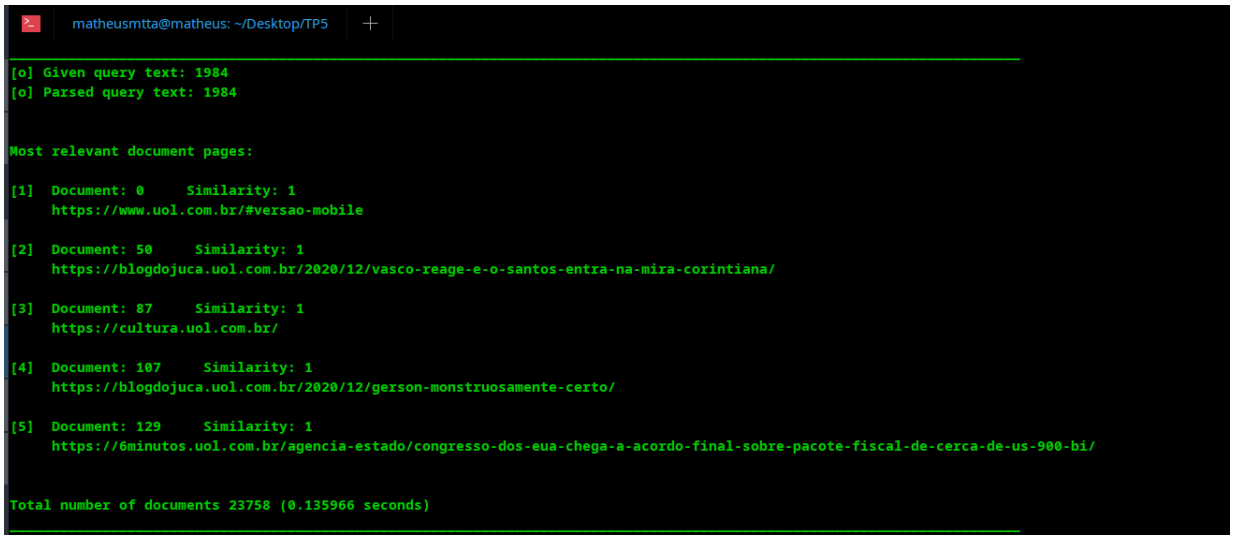

matheusmtta@matheus: ~/Desktop/TP5
[o] Given query text: Kamala Harris
[o] Parsed query text: Kamala Harris
Most relevant document pages:
[1] Document: 78543 Similarity: 1
https://agorarn.com.br/2020/11/07/
[2] Document: 90073 Similarity: 1
http://agendacapital.com.br/kamala-harris-exemplo-relevante-do-reconhecimento-feminino/
[3] Document: 644443 Similarity: 1
https://agorarn.com.br/mundo/joe-biden-escolhe-senadora-kamala-harris-como-vice-na-disputa-pela-presidencia/
[4] Document: 4522 Similarity: 1
https://aventurasnahistoria.uol.com.br/canal/historia-hoje/?page=6
[5] Document: 5174 Similarity: 1
http://atarde.uol.com.br/mundo/noticias/2150787-biden-vai-se-vacinar-na-proxima-segundafeira-contra-a-covid19
Total number of documents 273 (0.036975 seconds)


matheusmtta@matheus: ~/Desktop/TP5
[o] Given query text: coronavírus BH
[o] Parsed query text: coronavirus BH
Most relevant document pages:
[1] Document: 27 Similarity: 1
https://noticias.uol.com.br/saude/
[2] Document: 1057 Similarity: 1
http://datafolha.folha.uol.com.br/eleicoes/2020/11/1989114-70-estao-otimistas-com-cenario-da-pandemia-em-bh.shtml
[3] Document: 2289 Similarity: 1
https://carroemotos.ig.com.br/
[4] Document: 8168 Similarity: 1
https://agenciabrasil.ebc.com.br/tags/pandemia-4
[5] Document: 11263 Similarity: 1
http://broadcast.com.br/cadernos/agencia-99/?id=QXJtakNlcXFNV2RydTMrUGFybTNSUT09
Total number of documents 2361 (0.397882 seconds)

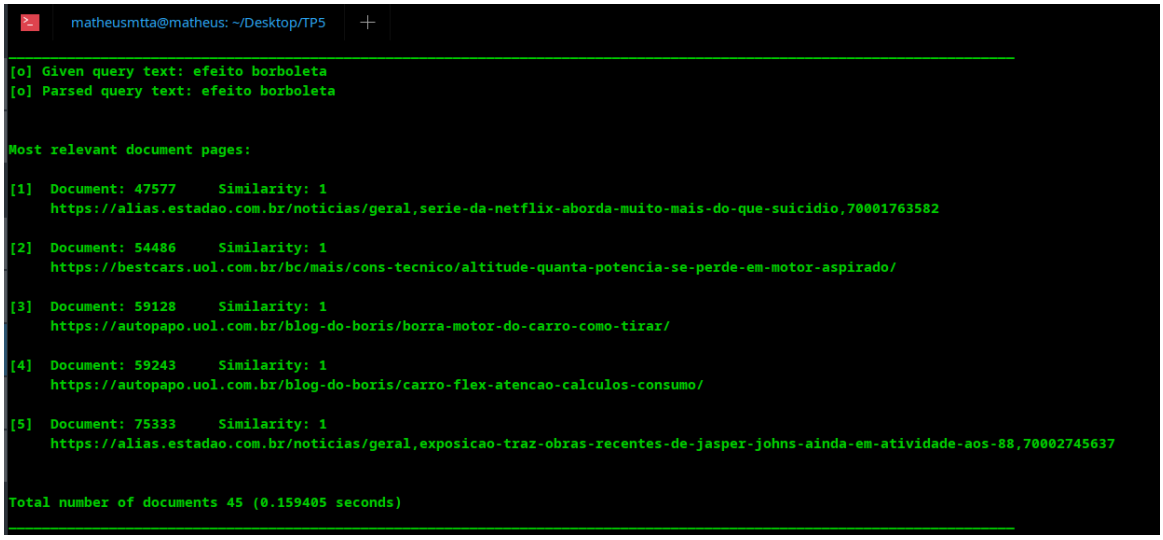

matheusmtta@matheus: ~/Desktop/TP5
[o] Given query text: Victor Hugo
[o] Parsed query text: Victor Hugo
Most relevant document pages:
[1] Document: 47 Similarity: 1
https://natelinha.uol.com.br/a-fazenda/2020/12/20/as-tretas-dos-peoes-de-a-fazenda-fora-das-cameras-da-record-156108.php
[2] Document: 75 Similarity: 1
https://caras.uol.com.br/tv/
[3] Document: 91 Similarity: 1
https://caras.uol.com.br/a-fazenda/
[4] Document: 99 Similarity: 1
https://caras.uol.com.br/atualidades/
[5] Document: 114 Similarity: 1
https://caras.uol.com.br/bebe/
Total number of documents 2699 (0.092012 seconds)

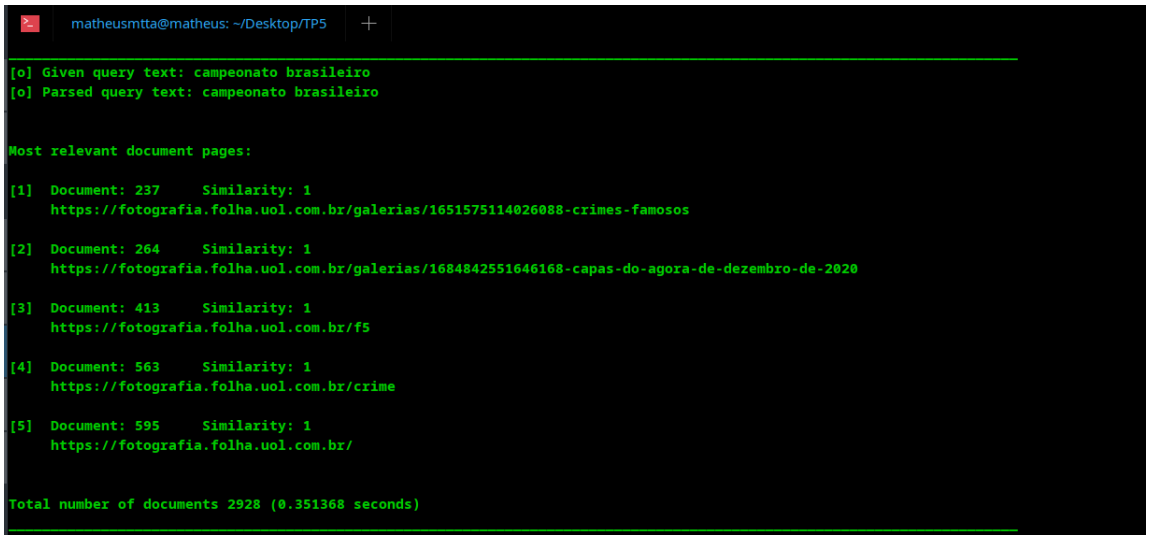

matheusmtta@matheus: ~/Desktop/TP5
[o] Given query text: golpe Myanmar
[o] Parsed query text: golpe Myanmar
Most relevant document pages:
[1] Document: 209256 Similarity: 1
https://anistia.org.br/peticao/junte-se-a-nos-por-um-brasil-para-todo-mundo/
[2] Document: 566983 Similarity: 1
https://acervo.racismoambiental.net.br/2015/11/26/nestle-admite-envolvimento-em-trabalho-escravo/
[3] Document: 893420 Similarity: 0.996102
https://agroemdia.com.br/2020/07/28/abpa-myanmar-abre-mercado-para-a-carne-suina-do-brasil/
[4] Document: 1611 Similarity: 0.94967
https://aventurasnahistoria.uol.com.br/noticias/reportagem/onde-esta-meu-irmao-sem-irma-meu-filho-sem-pai-crise-dos-refugiados.phtml
[5] Document: 934499 Similarity: 0.912635
http://agenciasn.com.br/arquivos/category/direitos-humanos/feed
Total number of documents 6 (0.097452 seconds)


matheusmtta@matheus: ~/Desktop/TP5
[o] Given query text: O Príncipe
[o] Parsed query text: O Príncipe
Most relevant document pages:
[1] Document: 3679 Similarity: 1
https://emais.estadao.com.br/noticias/gente,realeza-britanica-escolhe-fotos-de-familia-feliz-para-seus-cartoes-de-natal,70003556676
[2] Document: 66111 Similarity: 1
https://aventurasnahistoria.uol.com.br/tags/rainha-elizabeth-ii
[3] Document: 462649 Similarity: 1
http://almanaquevirtual.uol.com.br/o-pequeno-principe/
[4] Document: 462684 Similarity: 1
http://almanaquevirtual.uol.com.br/o-pequeno-principe-entrevista-com-diretor-mark-osborne-e-marcos-caruso-que-dubla-o-aviador/
[5] Document: 470884 Similarity: 1
https://acervodigital.ufpr.br/handle/1884/42328
Total number of documents 1436 (3.24715 seconds)


matheusmtta@matheus: ~/Desktop/TP5
[o] Given query text: 1984
[o] Parsed query text: 1984
Most relevant document pages:
[1] Document: 0 Similarity: 1
https://www.uol.com.br/#versao-mobile
[2] Document: 50 Similarity: 1
https://blogdojuca.uol.com.br/2020/12/vasco-reage-e-o-santos-entra-na-mira-corintiana/
[3] Document: 87 Similarity: 1
https://cultura.uol.com.br/
[4] Document: 107 Similarity: 1
https://blogdojuca.uol.com.br/2020/12/gerson-monstruosamente-certo/
[5] Document: 129 Similarity: 1
https://6minutos.uol.com.br/agencia-estado/congresso-dos-eua-chega-a-acordo-final-sobre-pacote-fiscal-de-cerca-de-us-900-bi/
Total number of documents 23758 (0.135966 seconds)


matheusmtta@matheus: ~/Desktop/TP5
[o] Given query text: irmão urso
[o] Parsed query text: irmão urso
Most relevant document pages:
[1] Document: 35065 Similarity: 1
http://atarde.uol.com.br/buscas?tag=lesao&canal=2
[2] Document: 95187 Similarity: 1
http://almanaque.folha.uol.com.br/bosi2.htm
[3] Document: 307503 Similarity: 1
https://antigo.saude.gov.br/diversus/recife-pe
[4] Document: 398684 Similarity: 1
https://ambitojuridico.com.br/cadernos/direito-penal/duracao-da-medida-de-seguranca-entre-a-intervencao-penal-e-a-saude-publica/
[5] Document: 430246 Similarity: 1
https://anamaria.uol.com.br/noticias/bem-estar-e-saude/criancas-com-medo-de-dentista-veja-10-dicas-rapidas-e-faceis-para-superar-este-problema.phtml
Total number of documents 28 (0.083038 seconds)


matheusmtta@matheus: ~/Desktop/TP5
[o] Given query text: efeito borboleta
[o] Parsed query text: efeito borboleta
Most relevant document pages:
[1] Document: 47577 Similarity: 1
https://alias.estadao.com.br/noticias/geral,serie-da-netflix-aborda-muito-mais-do-que-suicidio,70001763582
[2] Document: 54486 Similarity: 1
https://bestcars.uol.com.br/bc/mais/cons-tecnico/altitude-quanta-potencia-se-perde-em-motor-aspirado/
[3] Document: 59128 Similarity: 1
https://autopapo.uol.com.br/blog-do-boris/borra-motor-do-carro-como-tirar/
[4] Document: 59243 Similarity: 1
https://autopapo.uol.com.br/blog-do-boris/carro-flex-atencao-calculos-consumo/
[5] Document: 75333 Similarity: 1
https://alias.estadao.com.br/noticias/geral,exposicao-traz-obras-recentes-de-jasper-johns-ainda-em-atividade-aos-88,70002745637
Total number of documents 45 (0.159405 seconds)


matheusmtta@matheus: ~/Desktop/TP5
[o] Given query text: campeonato brasileiro
[o] Parsed query text: campeonato brasileiro
Most relevant document pages:
[1] Document: 237 Similarity: 1
https://fotografia.folha.uol.com.br/galerias/1651575114026088-crimes-famosos
[2] Document: 264 Similarity: 1
https://fotografia.folha.uol.com.br/galerias/1684842551646168-capas-do-agora-de-dezembro-de-2020
[3] Document: 413 Similarity: 1
https://fotografia.folha.uol.com.br/f5
[4] Document: 563 Similarity: 1
https://fotografia.folha.uol.com.br/crime
[5] Document: 595 Similarity: 1
https://fotografia.folha.uol.com.br/
Total number of documents 2928 (0.351368 seconds)


matheusmtta@matheus: ~/Desktop/TP5
[o] Given query text: Fernando Pessoa
[o] Parsed query text: Fernando Pessoa
Most relevant document pages:
[1] Document: 278 Similarity: 1
https://escolakids.uol.com.br/historia/
[2] Document: 706 Similarity: 1
https://educacao.uol.com.br/disciplinas/portugues/
[3] Document: 872 Similarity: 1
https://congressoemfoco.uol.com.br/legislativo/mais-de-cem-deputados-estao-sob-investigacao-da-justica-veja-a-lista/
[4] Document: 1616 Similarity: 1
https://brasil.estadao.com.br/naquarentena
[5] Document: 4009 Similarity: 1
https://alias.estadao.com.br/noticias/geral,na-ascensao-do-nazismo-freud-escrevia-sobre-tensao-entre-individuo-e-sociedade,70003330952
Total number of documents 2939 (0.334955 seconds)