Coleta e Busca de Entidades Estruturadas em um Domínio

Fillipe de Menezes - fmcs3
Luís Delgado - lhds

Domínio

Sites escolhidos

- https://www.usajobs.gov/
- https://www.ziprecruiter.com/
- https://www.indeed.co.uk/
- http://www.jobs.ac.uk/
- https://www.reed.co.uk/
- https://www.totaljobs.com/

Objetivos

- Remover o HTML das páginas
- Criar arquivo invertido
- Processamento de Consulta

Performance

- Windows 10 Home Single Language
- Processador: Intel(R) Core(™) i7-5500U CPU @2.40GHz 2.40 GHz

Remover o HTML das páginas

Funciona em menos de 1min

Desvantagem: Elimina alguns indicadores de atributos

Modificação

- Para a apresentação:
  - Diante da falta de dados que iria haver para comparar, foi feito um arquivo invertido para cada site.
  - Utilizando o paradigma de threads, ambos foram rodados simultaneamente.

Criar Arquivo Invertido Medindo tamanho

	Tokenização e inferência ID	Tamanho Arquivo Invertido
Site 1	16.443.694 bytes	590.000 bytes (24,1%)
Site 2	0 bytes	0 bytes (100%)
Site 3	17.674.612 bytes	556.000 bytes (12,4%)
Site 4	10.552.118 bytes	671.000 bytes (41,4%)
Site 5	0 bytes	0 bytes (100%)
Site 6	32.859.281 bytes	601.000 bytes (15,2%)
Total	77.529.705 bytes	2.048.000 bytes (48,9%)
Tempo	1 minuto	06 horas e 53 minutos

Processamento de Consulta

A cara do programa

No console

Digite sua consulta: *Programmer*

Programmer foi encontrado nos seguintes documentos:

11 - 12 - 33 - 44 - 65, onde o primeiro dígito de cada número indica o site e o restante de dígitos indica o respectivo documento.

Assim, para 11, significa que o termo foi achado no site 1, respectivamente no seu documento 1.

Ou, caso não seja encontrado em nenhum documento, será informado:

Infelizmente não foi encontrado nenhum documento para a sua consulta.

Processamento de Consulta

Console

Digite sua consulta: *Programmer*

Programmer foi encontrado nos seguintes documentos:

62 - 63 - 64 - 65 - 66 - 67 - 68 - 69 - 610 - 611 - 612 - 613 - 614 - 615 - 616 - 617 - 618 - 619 - 620 - 621 - 622 - 623 - 624 - 625 - 626 - 627 - 628 - 629 - 630 - 631 - 632 - 633 - 634 - 635 - 636 - 637 - 638 - 639 - 640 - 641 - 642 - 643 - 644 - 645 - 646 - 647 - 648 - 649 - 650 - 651 - 652 - 653 - 654 - 655 - 656 - 657 - 658 - 659 - 660 - 661 - 662 - 663 - 664 - 665 - 666 - 667 - 668 - 669 - 670 - 671 - 672 - 673 - 674 - 675 - 676 - 677 - 678 - 679 - 680 - 681 - 682 - 683 - 684 - 685 - 686 - 687 - 688 - 689 - 690 - 691 - 692 - 693 - 694 - 695 - 696 - 697 - 698 - 699 - 6100 - 6101 - 6102 - 6103 - 6104 - 6105 - 6106 - 6107 - 6108 - 6109 - 6110 - 6111 - 6112 - 6113 - 6114 - 6115 - 6116 - 6117 - 6118 - 6119 - 6120 - 6121 - 6122 - 6123 - 6124 - 6125 - 6126 - 6127 - 6128 - 6129 - 6130 - 6131 - 6132 - 6133 - 6134 - 6135 - 6136 - 6137 - 6138 - 6139 - 6140 - 6141 - 6142 - 6143 - 6144 - 6145 - 6146 - 6147 - 6148 - 6149 - 6150 - 6151 - 6152 -

Dúvidas?

Acesso ao código em:

- https://github.com/luisdelgado/Coleta-e-Busca-de-Entidades-Estruturadas-em-um-Dominio