Web Crawler

Carlos Henrique P. da Silva
Lucas Gabriel de Moraes Martins
Bacharelado em Ciência da Computação
Centro Universitário SENAC - Campus Santo Amaro (SENAC-SP)
Av. Engenheiro Eusébio Stevaux, 823 – Santo Amaro, São Paulo – CEP 04696-000 – SP – Brasil
lucas.gmmartins@gmail.com e carloshpds@gmail.com

Resumo

Os web crawlers, também chamados de indexadores ou bots, são utilizados para percorrer a world wide web para realizar a extração de dados e meta-dados de conteúdos tanto estáticos quanto dinâmicos, que usualmente são páginas web.

1. Introdução

O Web Crawler, considerado um agente de software, tem como função navegar na WWW (*World Wide Web*),rastreando informações nas páginas web e devolvendo para o usuário.

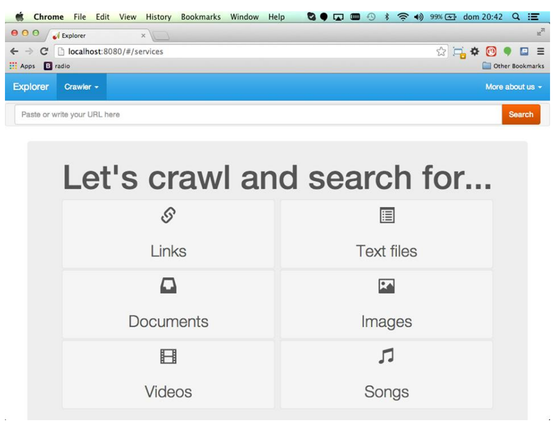
2. Objetivos

O objetivo do Web Crawler é a coleta automática e sistemática de documentos na web, que são indexados e consultados pela máquina de busca.

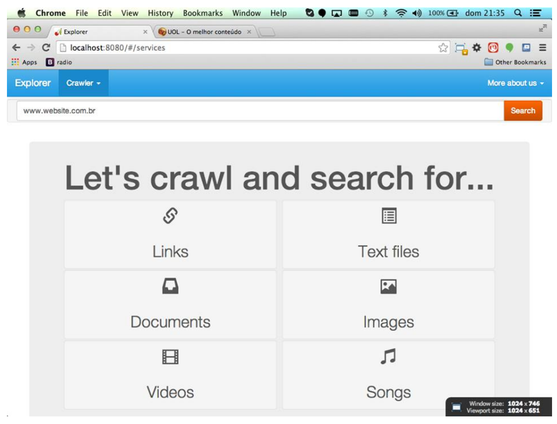
3. Metodologia

Na criação do Web Crawler for realizado na linguagem Python 2.7 e com o sistema de gerenciamente de banco de dados SQLite.
O python é uma linguagem de programação que possue estruturas de dados de alto nível.
Python se torna ideal para desenvolvimento rápido de aplicações de diversas áreas e na maioria das plataformas.
OSQLite é um banco de dados que cabe por completo em sua aplicação, sem precisar de um servido, como os demais.Seu uso é muito popular em diversas aplicações para armazenamento de dados na máquina de um usuário
Após semanas, aproximandamente, de estudados da linguagem e do banco de dados, demos inicio as discusões do que mostraríamos no nosso Web Crawler, e chegamos a seguinte conclusão:

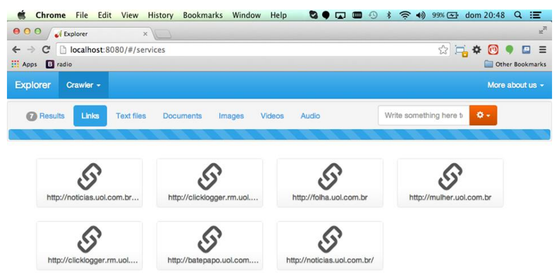
- Links;
- Vídeo;
- Imagens;
- Text file;
- Sons;
- Documentos.



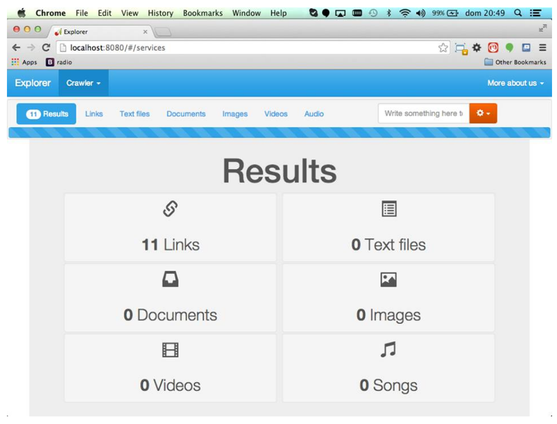
Ao digitar um link,



o crawler faz uma busca por todo o site desejado obtendo todas as informações contidas nele:



e separando nas categorias já definadas, para uma melhor visualização do usuário.



Para a realização do front-end foi utilizado a linguagem CoffeeScript/JavaScript, estilização de páginas com SASS/CSS/CSS3, markup com HTML/HTML5, com a utilização Yeoman para scaffolding, modern workflow e gerenciamento de pacotes juntamente com o Bower, além da incorporação de frameworks como: AngularJS, UnderscoreJS, GruntJS, Bootstrap (Twitter), Compass e JQuery

4. Resultados e Discussão

No processo de realização do crawler encontramos dificuldade em alguns aspectos, mas que foram resolvidas com mais estudo, uma verificação do código mais detalhada e testes. Com isso fomos resolvendo o que precisavamos e chegamos ao resultado esperado.

5. Conclusão

Chegamos no final do projeto com os objetivos compridos. COnseguimos chegar no resultado esperado que imaginavamos no começo do projeto.

Referências