Recuperação de Informação na Web 2020/2

Construtor de um Coletor de Propósito Geral
20 pontos [1]

Com o objetivo de estudar e aprender a arquitetura de um coletor simples para Web, você deverá fazer um coletor de propósito geral além de fazer uma análise do impacto dos parâmetros deste coletor.
Os integrantes do trabalho devem ser os mesmos do Coding Dojo.

Sobre a 'Equipe'

O projeto deve ser desenvolvido por grupos de no máximo três pessoas. Caso o grupo exceda este valor, ele perderá 2 pontos por pessoa excedente. Você deverá cadastrar o grupo antes da primeira atividade avaliativa (até dia 10/09) por meio do link: https://bit.ly/3hZc0ic . Os alunos podem fazer o coletor individualmente, porém, as discussões devem ser em grupo. Caso opte por fazer individual, deve-se marcar na planilha (em amarelo) com qual grupo irá fazer as discussões. Alunos ausentes nas aulas de discussão deverão entregar uma atividade individual.

Entrega Final

Deverá ser entregue dia 07/10/2020. Com apresentações parciais (ver cronograma0.

A entrega deverá conter os seguintes itens em um arquivo comprimido:

a) Relatório do trabalho **obrigatoriamente em Jupyter Notebook**
b) Código fonte
c) Lista de URLs coletadas. Não é necessário salvar os arquivos.

Ver detalhes a seguir

Política de atraso

Será descontado 1 ponto por dia de atraso.

Tarefas

Para fazer um coletor é necessário:

1) O escalonador deverá possuir:
   (a) Uma fila de páginas, sendo coletado através de uma busca em largura
   (b) Não permitir que a mesma url seja coletada mais de uma vez
   (c) Armazenar a última vez que um servidor foi acessado. Pois, um servidor poderá ser acessado de 30 em 30 segundos
2) Multiplas threads para coletar as páginas (os *Page Fetchers*):
   (a) Coleta a página que o escalonador organizou
   (b) Dado a página coletada, extrair seus links e inserir na fila do escalonador todas as páginas coletadas
      → A classe ColetorUtil poderá auxiliar no caso de urls relativas, pois os links devem ser sempre adicionados no seu formato completo na fila
   (c) Levar em consideração páginas html mal-formadas. Você poderá usar uma API para isso.
   (d) Levar em consideração o encoding da página. A classe ColetorUtil pode auxiliar neste processo
   (e) Caso a página não exista, a mesma é ignorada

Você deverá fazer um coletor que obedeça no mínimo os seguintes requisitos:

1) Obedecer os protocolos de exclusão de robôs:

1 Ver plano didático para detalhes sobre a pontuação

(a) critérios pertencentes no robots.txt
- Pode ser usado uma API para isto

(b) Critérios “noindex” e “nofolow” das metatags de cada html extraído
ps: noindex: não é permitido coletar. Nofolow: não é permitido seguir os links por meio desta página

(c) Obedecer o prazo de, no mínimo, 30 segundos entre requisições em um mesmo servidor (*hostname*)

→ **Caso não obedeça esses critérios o grupo perderá 5 pontos**

2) Criar um nome no “*User agent*” (finalizando como bot) e uma página pessoal com descrição do coletor, nome dos membros dos grupos, datas das coletas e propósito das coletas além de um e-mail de contato. Deverá ser explicitado que este coletor baixa apenas páginas públicas sempre levando em consideração a política de exclusão de robôs (robots.txt). Esta página deverá ficar online durante o semestre todo. O endereço da página pessoal deverá ser definida no User agent. Exemplo: “meuBot (wordpress.org/infoMeuBot)”.

→ **O grupo perderá 5 pontos** caso não crie a página e/ou não utilize o nome no “User Agent” devidamente e de acordo com o especificado acima

→ **Exemplo de páginas de informação de robôs:**
https://support.apple.com/en-us/HT204683

2) Utilizar os seguintes parâmetros no coletor:
→ Número máximo de páginas (50.000 páginas)
→ Profundidade por domínio (6 páginas)
→ Número de *threads* utilizado

3) Utilizar as sementes de acordo com os integrantes do grupo, usando tabela de sementes que está apresentado no Moodle em documento separado

4) Produzir um relatório **em jupyter** a ser definido na próxima seção

5) Armazenar a lista de URLs coletadas

Conteúdo do Relatório

O relatório deverá feito em outro arquivo Jupyter e é uma parte importante do trabalho e será levado muito em consideração. O relatório deve possuir os seguintes tópicos:

a) Principais desafios, decisões e arquitetura utilizada
b) URLs sementes utilizadas
c) Como foi feito, faça referencias à classes e métodos do código fonte:
→ Os critérios de exclusão de robôs e quantidade de tempo entre requisições à um mesmo servidor
d) O impacto na velocidade de coleta (quantidade de páginas por segundo) ao aumentar o número de threads 10 a 100, de 20 em 20
e) Link para a página descrevendo o coletor criado

Não coloque muito código fonte no relatório – crie arquivo separado com o código e, no relatório, apenas poucas linhas de código para gerar os gráficos/tabelas.

Bibliotecas utilizadas

Vocês poderão usar bibliotecas para os seguintes propósitos (segue também algumas sugestões de APIs)

→ Extração dos links de páginas HTML mal-formadas:
→ Parser do protocolo de exclusão de robôs

Critério de avaliação

Para a entrega final, o trabalho será avaliado considerando:

1) Legibilidade, comentários e organização do código
2) Funcionamento do coletor
3) Uso do protocolo de exclusão de robôs

4) Página de informação do coletor
5) Lista das páginas coletadas
6) Conteúdo do relatório com todos os itens requisitados

Caso não seja feito (ou feito incorretamente) o item (3) e/ou (4) o grupo perderá 5 pontos. O conteúdo do relatório é tão importante quanto o funcionamento do coletor.

Plágio não será tolerado e, caso identificado, o grupo (quem forneceu e quem utilizou) terá seu trabalho zerado.