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

- Localizar páginas relevantes

Localizar páginas relevantes

Baseline

- Decodificação UTF-8
- Verifica se tem “text/html”
- 1000 páginas por site
- Verifica robts.txt
- Otimização para remover links que nunca serão visitados
- Intervalo mínimo de acesso: 1s
- Verifica relevância da página (webanalytics)
- Verifica se o link faz parte do site domínio
- Buscar em largura

Heurística (websearch)

- Decodificação UTF-8
- Verifica se tem “text/html”
- 1000 páginas por site
- Verifica robts.txt
- Otimização para remover links que nunca serão visitados
- Intervalo mínimo de acesso: 1s
- Verifica relevância da página (webanalytics)
- Verifica se o link faz parte do site domínio
- Busca pelo melhor primeiro
    - Tenta evitar visitar um mesmo link mais de uma vez
    - Link e âncora
        - palavras-chave (location, salary) (webanalytics)
        - peso (weight)

Localizar páginas relevantes

Performance

- Windows 7 Enterprise
- Processador: AMD Phenom(™) II X4 B97 Processor 3.20 GHz
- Internet do CIn (300MB/s)

Baseline x Heurística
Harvest ratio

	Baseline	Heurística
Site 1	0,361	0,024
Site 2	0	0
Site 3	0,427	0,947
Site 4	0,660	0,973
Site 5	0	0
Site 6	0,842	0,998
Total	0,573	0,736

Dúvidas?

Acesso ao código em:

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