## 2. Fundamentação Teórica

### 2.1 Conceito de Revisão Bibliográfica

Revisão bibliográfica é um estudo e avaliação da literatura relacionado a um assunto específico (AVEYARD, 2014). Ou seja, é um estudo que apresenta o conhecimento atual existente para um tópico em particular. No termo em inglês existem várias formas de se referir à revisão bibliográfica: *systematic review*, *meta-analysis*, *rapid review*, *literature review*, *narrative review*, *research synthesis* e *structured review* (O'MALLEY, 2005).

Segundo Hart (2018) existem pelo menos 12 propósitos para uma revisão bibliográfica:

1. Distinguir o que já foi feito do que precisa ser feito
2. Descobrir variáveis relevantes para o tópico
3. Sintetizar e obter uma nova perspectiva
4. Identificar relações entre ideias e prática
5. Estabelecer o contexto para o tópico ou problema de pesquisa
6. Racionalizar a significância teórica ou prática do problema
7. Melhorando e obtendo vocabulários comuns ao assunto
8. Entender as origens e a estrutura do assunto
9. Relacionar ideias e teoria com problemas e questões
10. Identificar as principais metodologias e ferramentas de coleta de dados que têm sido usados
11. Colocar a pesquisa num contexto histórico para mostrar familiaridade com o estado da arte da pesquisa

12. Ter um corpo de conhecimento com o qual você pode relacionar com as descobertas da sua pesquisa

Existem dois principais tipos de revisão bibliográfica: a intervencionista e a escolástica (HART, 2018). A intervencionista tem o propósito de usar todos as evidências disponíveis e confiáveis para tomar uma decisão. A escolástica tem por objetivo examinar os argumentos, procurar por contradições, desafiar proposições existentes e fazer inferências.

Segundo Aveyard (2014), uma revisão bibliográfica de qualidade, mesmo sendo apenas uma seção de um artigo maior, ela por si só é um trabalho de pesquisa completo e geralmente contém as seguintes estruturas:

- Problema de pesquisa bibliográfica ou contextualização
- Métodos de pesquisa usados
- Apresentação dos resultados
- Discussão dos resultados

Lingard (2018) afirma que a seção de revisão bibliográfica deve mostrar o que ainda não é conhecido dentro daquele tópico (o que Lingard chama de lacuna, ou *gap*) permitindo ao leitor entender porque tal pesquisa é necessária. Essa estratégia faz com que a revisão bibliográfica se torne mais um argumento convincente do que apenas uma lista de fatos.

Lingard (2018) identificou os quatro tipos mais comuns de *gap*:

- Um simples déficit no conhecimento — "*ninguém nunca analisou a relação entre habilidade no xadrez e desempenho em matemática*"
- A shortcoming in the scholarship, often due to philosophical or methodological tendencies and oversights — "*os pesquisadores interpretaram x numa perspectiva cognitiva, mas ignoraram a perspectiva humanista*" ou "*até hoje nós analisamos a frequência de erros médicos cometidos por residentes, mas não exploramos as experiências subjetivas deles em tais erros*"

- Uma controvérsia — "*pesquisadores discordam na definição de profissionalismo dentro da medicina ...*"

- Uma suposição comum, mas não comprovada — "*é muito comum na literatura a afirmação de que o cérebro possui 100 bilhões de neurônios, mas existem artigos publicados na literatura que mediu e confirmou esse número?*"

## 2.2 Web Scraping

Web Scraping é a prática de coletar dados da Web que não seja acessando uma API. Isso é normalmente feito criando um programa que faz requisição para um servidor web, recebe os dados (na maioria das vezes na forma de documentos HTML) e extrai as informações necessárias. (MITCHELL, 2018).

### 2.2.1 Funcionamento Básico

Os dados na Internet são transferidos num formato bastante formal e estruturado, fácil de ser lido por computadores. Mas nas páginas web estes dados são exibidos numa forma desestruturada, em forma de textos que são fáceis de ler pelo ser humano, porém que são difíceis de serem processados por algoritmos. A prática de Web Scraping permite que estes dados desestruturados sejam extraídos em dados estruturados (BOEING et al, 2016), daí a importância das técnicas de Web Scraping.

Mitchell (2018) explica de forma didática e simples o funcionamento básico da Internet, através de um exemplo:

Alice possui um servidor web. Bob usa um computador desktop que está tentando se conectar ao servidor de Alice. Quando uma máquina quer falar com outra máquina, algo como a seguinte troca ocorre:

1. O computador de Bob envia fluxo de dados de 1 e 0 bits, representados por tensões elétricas alta e baixas no fio de transmissão. Esses bits formam algumas informações, contendo um cabeçalho (header) e um corpo (body). O cabeçalho contém um destino imediato do endereço MAC do seu roteador local, com um destino final do

endereço IP de Alice. O corpo contém sua solicitação para a aplicação do servidor web de Alice.

2. O roteador local de Bob recebe todos esses 1s e 0s e os interpreta como um pacote, a partir do endereço MAC do próprio Bob, destinado ao endereço IP de Alice. Seu roteador marca seu próprio endereço IP no pacote como o endereço IP de origem e o envia pela Internet.

3. O pacote de Bob percorre vários servidores intermediários, que direcionam seu pacote para o caminho correto, até o servidor de Alice.

4. O servidor de Alice recebe o pacote em seu endereço IP.

5. O servidor de Alice lê a porta de destino do pacote no cabeçalho e passa para a aplicação apropriada - a aplicação do servidor web. (A porta de destino do pacote é quase sempre a porta 80 para aplicativos da Web; a porta pode ser considerado como um número de apartamento para dados do pacote, enquanto o endereço IP é como o endereço da rua.)

6. O aplicativo do servidor da web recebe um fluxo de dados do processador do servidor. Esses dados dizem algo como o seguinte:

- Esta é uma solicitação GET.
- O seguinte arquivo é solicitado: index.html.

7. O servidor web localiza o arquivo HTML correto, agrupa-o em um novo pacote para enviar a Bob e o envia para o roteador local, para ser transportado de volta à máquina de Bob, por meio do mesmo processo.

### 2.2.2 Web Crawlers

Nos casos que é necessário coletar dados de múltiplas páginas, ou até múltiplos sites, será necessário o uso de um Web Crawler que são programas que “rasteja” (*crawl*, em inglês) ao longo da web (MITCHELL, 2018). Web Crawlers também são chamados de robôs, spiders, worms, walkers e wanderers (HEYDON et al, 1999).

Todos os motores de busca populares usam web crawlers que precisam escalar para grandes proporções da web (HEYDON et al, 1999). Porém com o avanço da Internet simples web crawlers centralizados se tornaram cada vez menos satisfatório e foi notado que é necessário paralelizar os processos de crawling usando múltiplos

servidores numa arquitetura distribuída para um web crawling eficiente (BOLDI et al, 2003).

A Google é um exemplo de sistema que usa um web crawling distribuído com múltiplas máquinas para o processo de crawling. O crawling consiste em cinco componentes funcionais rodando em diferentes processos. O *URL server process* lê as URL de um arquivo e os redireciona para múltiplos *processos crawler*. Cada processo crawler roda numa diferente máquina, é de thread única, e usa I/O assíncrono para buscar dados de mais de 300 servidores web em paralelo. Os crawlers transmite as páginas baixadas para um único *processo indexador* que extrai links do HTML e os salva num arquivo diferente. O *URL resolver process* lê o arquivo com os links, e faz um processo chamado *derelativization* das URLs e salva as URL absolutas no arquivo que é lido pelo *URL server*. Tipicamente, três a quatro máquinas crawlers são usados de forma que o sistema inteiro precise de quatro a oito máquinas (HEYDON, 1999).

A UbiCrawler é um web crawler distribuído, implementado com a linguagem de programação Java, cuja arquitetura e funcionamento foi publicado num artigo (BOLDI et al, 2003). Os autores afirmam que os principais atributos do UbiCrawlers são:

- Independente de plataforma

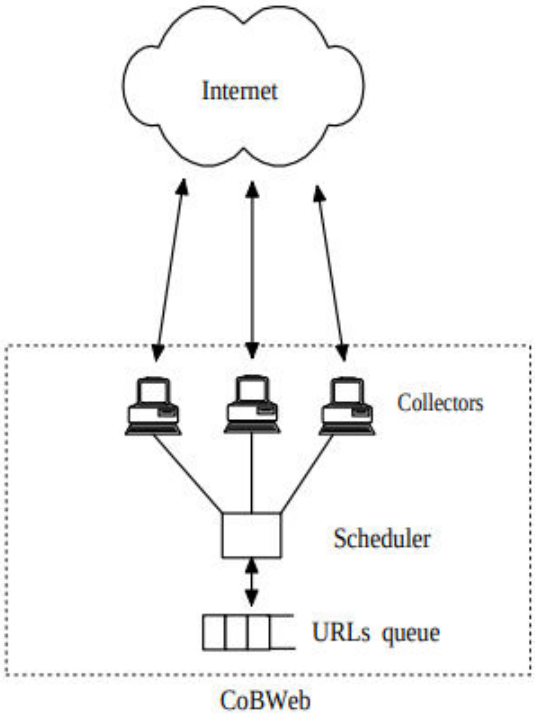
- Distribuição total de cada *task* (não tem um ponto único de falha e nem coordenação centralizada)

- Tolerância a falhas: falhas permanentes e transientes são lidados de forma eficiente

- Escalabilidade

O CoBWeb é um web crawler para coletar dados especificamente da web brasileira (que tenha *.br* no domínio). As operações do CobWeb é orientado para eficiência, robustez e parcimônia no uso de recursos compartilhados (SILVA et al, 1999).

Figura 1 - Arquitetura do CoBWeb




Fonte - SILVA et al (1999)

### 2.2.3 Ferramentas de programação disponíveis

Uma busca no site *github.com* pelo termo “web scraping” mostra que a linguagem de programação Python é a mais utilizada para tal tarefa (busca realizada em 19 de abril de 2019 no site https://github.com). O Python possui um módulo para buscar dados da web de forma fácil e prática, chamado de *urllib.request* (docs.python.org, acessado em 19 de abril de 2019).

Figura 1. Exemplo de código para buscar dados a partir de uma URL.

```
import urllib
params = urllib.urlencode({'spam': 1, 'eggs': 2, 'bacon': 0})
f = urllib.urlopen("http://www.musi-cal.com/cgi-bin/query?%s" % params)
print f.read()
```

Fonte - docs.python.org (acessado em 19 de abril de 2019).

A biblioteca BeautifulSoup é usada para fazer o parsing de documentos HTML de forma fácil na forma de objetos Python e o seu nome é em homenagem a um

poema recitado por um personagem da estória Alice nos Países da Maravilha (MITCHELL, 2018).

Figura 3. Exemplo de código BeautifulSoup em que a partir do HTML retornado é extraído um elemento H1.

```
from urllib.request import urlopen
from bs4 import BeautifulSoup

html = urlopen('http://www.pythonscraping.com/pages/page1.html')
bs = BeautifulSoup(html.read(), 'html.parser')
print(bs.h1)
```

Fonte - Mitchell (2018).

Outra ferramenta existente é o framework Scrapy. No site oficial do Scrapy a ferramenta é definida como (acessado em 19 de abril de 2019):

Um framework de código aberto e colaborativo para extrair dados que você precisa dos sites web de forma rápida, simples e extensível.

Figura 4. Exemplo de programa Scrapy que percorre três sites do Wikipédia exibindo o texto do primeiro elemento H1 que encontrar em cada página.

```
import scrapy

class ArticleSpider(scrapy.Spider):
    name='article'

    def start_requests(self):
        urls = [
            'http://en.wikipedia.org/wiki/Python_'
            '%28programming_language%29',
            'https://en.wikipedia.org/wiki/Functional_programming',
            'https://en.wikipedia.org/wiki/Monty_Python']
        return [scrapy.Request(url=url, callback=self.parse)
            for url in urls]

    def parse(self, response):
        url = response.url
        title = response.css('h1::text').extract_first()
        print('URL is: {}'.format(url))
        print('Title is: {}'.format(title))
```

Fonte - Mitchell (2018).

O PySocks é um módulo do Python que consegue rotear o tráfego HTTP para servidores de proxy e funciona muito bem com o navegador Tor permitindo a atividade de web scraping de forma anônima.

Figura 5. Exemplo de código em PySocks que busca os dados de uma URL passando pelo proxy definido na porta 9150, que é a porta padrão onde roda o serviço Tor.

```
import socks
import socket
from urllib.request import urlopen

socks.set_default_proxy(socks.SOCKS5, "localhost", 9150)
socket.socket = socks.socksocket
print(urlopen('http://icanhazip.com').read())
```

Fonte - Mitchell (2018).

### 2.2.4 Exemplo de aplicações do Web Scraping

Boeing e Wadell (2016) usaram técnicas de Web Scraping para coletar dados do site de anúncios Craiglist para fazer uma análise sobre os aluguéis de imóveis nos Estados Unidos.

Landers et al (2016) usaram programas desenvolvidos em Python para demonstrar como o web scraping pode ser usado para coletar dados do *big data* para uso em pesquisas de psicologia.

Haddaway (2015) usou web scraping para pesquisar por literatura cinzenta, que são publicações não convencionais difíceis de encontrar em canais tradicionais de distribuição, com controle biográfico ineficaz pois não recebem numeração internacional, e não são depósito de objetos legais em muitos países sendo normalmente não incluídas em bibliografias e catálogos (BOTELHO, 2015).

Arora et al (2013) usou web scraping para coletar dados de sites de divulgação de empresas pequenas e médias de grafeno para analisar como as especializações dessas empresas impacta o mercado geral de grafeno.

Youtie et al (2012) analisou os websites de empresas de nanotecnologia para analisar quais características influenciam na transição de empresas apenas de descobertas para uma empresa de comercialização.