UNIVERSIDADE FEDERAL FLUMINENSE

MARCELLO TELLES DOS SANTOS

YASMIM DE LIMA MARTINS

WEB SCRAPING E TEXT MINING: UMA SOLUÇÃO PARA CAPTURA DOS COMPONENTES UTILIZADOS PELOS PRINCIPAIS MEDICAMENTOS CATALOGADOS

Niterói

2020MARCELLO TELLES DOS SANTOS

YASMIM DE LIMA MARTINS

WEB SCRAPING E TEXT MINING: UMA SOLUÇÃO PARA CAPTURA DOS COMPONENTES UTILIZADOS PELOS PRINCIPAIS MEDICAMENTOS CATALOGADOS

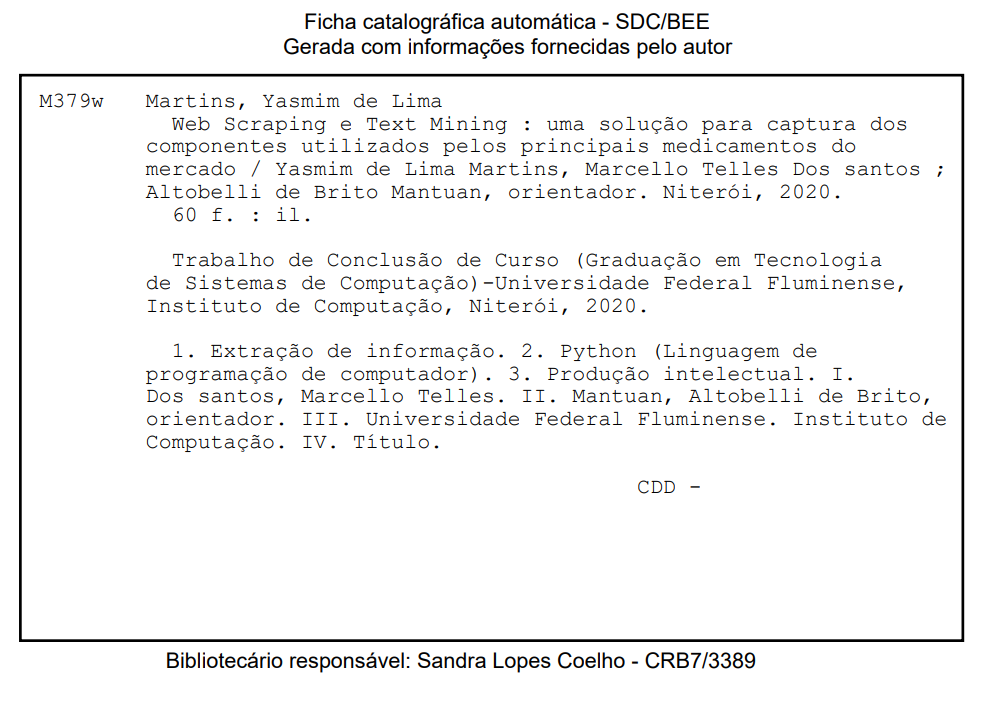
Trabalho de Conclusão de Curso submetido ao Curso de Tecnologia em Sistemas de Computação da Universidade Federal Fluminense como requisito parcial para obtenção do título de Tecnólogo em Sistemas de Computação.

Orientador(a):

ALTOBELLI DE BRITO MANTUAN

NITERÓI

2020



MARCELLO TELLES DOS SANTOS

YASMIM DE LIMA MARTINS

WEB SCRAPING E TEXT MINING: UMA SOLUÇÃO PARA CAPTURA DOS COMPONENTES UTILIZADOS PELOS PRINCIPAIS MEDICAMENTOS CATALOGADOS

Trabalho de Conclusão de Curso submetido ao Curso de Tecnologia em Sistemas de Computação da Universidade Federal Fluminense como requisito parcial para obtenção do título de Tecnólogo em Sistemas de Computação.

Niterói, \_\_\_ de \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ de 2020.

Banca Examinadora:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Prof. Altobelli de Brito Mantuan, MSc. – Orientador

UFF – Universidade Federal Fluminense

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Prof. ou Profa. <NOME>, <Título>. – Orientador ou Avaliador

<Sigla da Universidade> - <Nome da Universidade>

A dedicatória é opcional, no qual o autor presta homenagem ou dedica seu trabalho a alguém.

Dedico este trabalho a(o) minha(meu) esposa(o) e aos meus estimados filhos.

AGRADECIMENTOS

MARCELLO

A minha família, principalmente minha mãe, Simone, pelo constante apoio, preocupação e por sempre ter sido uma fonte de inspiração ao longo da minha vida.

A minha namorada, Renata, pelo estímulo para que eu fizesse um curso de graduação, e pela inspiração e apoio ao longo da jornada.

Ao meu orientador Altobelli de Brito Mantuan pelos direcionamentos e apoio concedido na confecção deste trabalho.

Aos colegas de curso pela troca ao longo dos últimos 3 anos, que enriqueceram muito a experiência de aprendizado.

Epígrafe é a inscrição colocada no início de um trabalho, de um capítulo ou partes principais. É opcional para a monografia.

Delete as caixas amarelas

“A Escola é uma arena onde grupos sociais lutam por legitimidade e poder”.

Dinair Leal da Hora

RESUMO

A obtenção de dados referentes a excipientes utilizados em medicamentos para uma pesquisa farmacêutica é um tipo de atividade que demanda tempo e esforço elevados, considerando o método de extração manual. A partir de desenvolvimento de *bots*, é possível coletar esses dados de forma automatizada através da técnica de *web scraping* e realizar a interpretação dos dados através da técnica de *text mining*. Neste estudo, expomos a implementação de um robô que simula as ações humanas para obter os dados relevantes à pesquisa do medicamento de interesse. Os requisitos do projeto foram levantados junto a um pesquisador farmacêutico especialista que indicou o procedimento e as fontes de dados utilizadas por eles para a extração das informações. O sistema desenvolvido foi testado e os resultados obtidos foram validados pelo pesquisador considerando o objetivo de definir os pontos de interesse a serem capturados.

Palavras-chaves: web scraping, text mining, excipiente, extração automática de dados e pesquisa farmacêutica.

*O resumo e abstract* devem ser digitados sem parágrafo inicial e o espaçamento entre linhas é simples. Lembre-se que já existem estilos que formatam o texto adequadamente. Remova as caixas amarelas antes de imprimir.

ABSTRACT

Obtaining data about excipients utilized in medicines for a pharmaceutical research is a kind of activity that demands elevated time and efforts, considering the manual extraction method. Through the development of bots, it is possible to gather these data in an automated way, using web scraping to collect and text mining to interpret the data. In this study, we expose the implementation of a robot that simulates human actions to obtain relevant data on the research of the medicine of interest. The project requirements were raised near to a specialist pharmaceutical researcher who pointed the procedure and the data sources used by them to extract the information. The project was developed and test based on these conditions, and was validated by the researcher, considering the objective of defining the points of interest to be captured.

Key words: web scraping, text mining, excipient, automatic data extraction e pharmaceutical research.

LISTA DE ILUSTRAÇÕES

[Figura 1: Dados não estruturados, semiestruturados e estruturados 22](#_Toc44518779)

[Figura 2: Estrutura básica de um documento HTML 25](#_Toc44518780)

[Figura 3: Visualização de um documento HTML em um navegador 25](#_Toc44518781)

[Figura 4: Estrutura básica de um documento HTML representado como texto 27](#_Toc44518782)

[Figura 5: Representação da estrutura básica de um documento HTML 27](#_Toc44518783)

[Figura 6: Diagrama do caso de uso Scrap Anvisa 29](#_Toc44518784)

Figura 6: Diagrama do caso de uso Text Mining Anvisa........................................ 29

[Figura 7: Diagrama de classes do projeto 34](#_Toc44518785)

[Figura 8: Fluxograma dos módulos da extração do dado 35](#_Toc44518786)

[Figura 10: Fluxograma de atividades para obter as bulas do profissional. 36](#_Toc44518787)

[Figura 11: Exemplo de lista de sugestões de nomes completos do fármaco. 36](#_Toc44518788)

[Figura 12: Fluxograma de atividades para extração dos excipientes. 38](#_Toc44518789)

[Figura 13: Exemplo de bula contemplada no CASO 1. 39](#_Toc44518790)

[Figura 14: Exemplo de bula contemplada no CASO 2. 39](#_Toc44518791)

[Figura 15: Exemplo de bula contemplada no CASO 3. 39](#_Toc44518792)

[Figura 16: Exemplo de bula contemplada no CASO 4. 40](#_Toc44518793)

[Figura 17: Exemplo de bula contemplada no CASO 5. 40](#_Toc44518794)

[Figura 18: Exemplo de bula contemplada no CASO 6. 40](#_Toc44518795)

[Figura 19: Exemplo de bula contemplada no CASO 7. 40](#_Toc44518796)

[Figura 20: Exemplo de bula contemplada no CASO 8. 41](#_Toc44518797)

[Figura 21: Exemplo de bula contemplada no CASO 9. 41](#_Toc44518798)

[Figura 22: Exemplo de bula não contemplada. 41](#_Toc44518799)

[Figura 23: Fluxograma de atividades para extração de informações do excipiente. 42](#_Toc44518800)

LISTA DE TABELAS

[Tabela 1: Descrição do caso de uso Scrap Anvisa 28](#_Toc44518626)

LISTA DE ABREVIATURAS E SIGLAS

HTML - Hypertext Markup Language

HTTP - Hypertext Transfer Protocol

CSS - Cascading Style Sheets

PDF - [Portable Document Format](https://pt.wikipedia.org/wiki/Portable_Document_Format)

JSON - JavaScript Object Notation

CSV - Comma-separated Values

XML - Extensible Markup Language

SQL – Structured Query Language

JS – JavaScript

SUMÁRIO

[RESUMO 8](#_Toc44518642)

[ABSTRACT 9](#_Toc44518643)

[LISTA DE ILUSTRAÇÕES 10](#_Toc44518644)

[LISTA DE TABELAS 11](#_Toc44518645)

[LISTA DE GRÁFICOS 12](#_Toc44518646)

[LISTA DE ABREVIATURAS E SIGLAS 13](#_Toc44518647)

[1 INTRODUÇÃO 17](#_Toc44518648)

[2 TRABALHOS RELACIONADOS 18](#_Toc44518649)

[2.1 DISCUSSÃO 19](#_Toc44518650)

[3 FUNDAMENTAÇÃO TEÓRICA 20](#_Toc44518651)

[3.1 COLETA DE DADOS 20](#_Toc44518652)

[3.2 TIPOS DE DADOS 20](#_Toc44518653)

[3.2.1 QUALITATIVOS 21](#_Toc44518654)

[3.2.2 QUANTITATIVOS 21](#_Toc44518655)

[3.2.3 PRIMÁRIOS 21](#_Toc44518656)

[3.2.4 SECUNDÁRIOS 22](#_Toc44518657)

[3.3 ORGANIZAÇÃO E ARMAZENAMENTO 22](#_Toc44518658)

[3.3.1 ESTRUTURADOS 22](#_Toc44518659)

[3.3.2 NÃO ESTRUTURADOS 23](#_Toc44518660)

[3.3.3 SEMIESTRUTURADOS 23](#_Toc44518661)

[3.4 FUNCIONAMENTO DAS PÁGINAS WEB 23](#_Toc44518662)

[3.4.1 HISTÓRIA DA INTERNET 23](#_Toc44518663)

[3.4.2 NAVEGADORES E PÁGINAS WEB 24](#_Toc44518664)

[3.5 WEB SCRAPING 26](#_Toc44518665)

[3.5.1 BUSCANDO INFORMAÇÕES EM UM DOCUMENTO HTML 26](#_Toc44518666)

[4 IMPLEMENTAÇÃO 29](#_Toc44518667)

[4.1 CASOS DE USO 29](#_Toc44518668)

[4.1.1 DIAGRAMA DO CASO DE USO SCRAP ANVISA 29](#_Toc44518669)

[4.1.2 DESCRIÇÃO DO CASO DE USO SCRAP ANVISA 30](#_Toc44518670)

[4.2 TECNOLOGIAS 32](#_Toc44518671)

[4.3 DIAGRAMA DE CLASSES 33](#_Toc44518672)

[4.3.1 DESCRIÇÃO DAS CLASSES 33](#_Toc44518673)

[4.4 FONTES 35](#_Toc44518674)

[4.4.1 ANVISA 35](#_Toc44518675)

[4.4.1.1 DOWNLOAD DAS BULAS DO PROFISSIONAL 35](#_Toc44518676)

[4.4.1.2 ANÁLISE DAS BULAS E EXTRAÇÃO DOS EXCIPIENTES 38](#_Toc44518677)

[4.4.2 HANDBOOK OF PHARMACEUTICAL EXCIPIENTS 41](#_Toc44518678)

[CONCLUSÕES E TRABALHOS FUTUROS 44](#_Toc44518679)

[REFERÊNCIAS BIBLIOGRÁFICAS 45](#_Toc44518680)

[ANEXOS 46](#_Toc44518681)

# INTRODUÇÃO

Parte inicial do texto, onde devem constar a delimitação do assunto tratado, o problema ou oportunidade, objetivos da pesquisa, motivação para fazer a pesquisa, métodos usados e resultados encontrados, quando for o caso, e outros elementos necessários para situar o tema do trabalho, bem como a estrutura do documento.

A introdução pode ter a seguinte ordem:

* O que é o trabalho? (1 ou 2 parágrafos)
* Motivação para desenvolvê-lo? (1 parágrafo)
* Métodos usados, quando for o caso (linguagens, protocolos, metodologias, etc.) ( 1 ou 2 parágrafos)
* Organização do trabalho:
  + no Capítulo 2....;
  + o terceiro capítulo é dedicado ...;
  + ...; e
  + finalmente, no Capítulo “tal”, temos as conclusões e indicações para futuros trabalhos.

# TRABALHOS RELACIONADOS

Visando o entendimento e análise da importância da implementação de técnicas de coletas de dados, foram ponderados artigos cujo o conceito principal é análogo ao do presente trabalho.

O estudo [1] de autoria de Carolina Franco Zanon, apresenta a criação de um sistema, utilizando a técnica *Web Scraping*, para auxiliar na definição do perfil epidemiológico dos pacientes da Clínica de Estomatologia da FO-UFRJ. Os dados foram coletados de um sistema já existente denominado Estomato Web, um *software* incapaz de fornecer estatísticas ou exportar grande quantidade de dados de uma única vez. A técnica foi implementada fazendo a coleta dos conjuntos de dados e transferindo-os para planilhas do Excel. Foi possível agrupar os dados por categorias e fazer uma análise detalhada do perfil dos pacientes. Concluiu-se que o uso do novo sistema, chamado Estomato Web Scraper, facilitou a extração e qualidade dos dados, e, assim sendo, poderá ser utilizado para diversos estudos na área de estomatologia.

Um segundo artigo [2], que tem Dhaniel Nunes Mazini e Renato Cesar Sato como autores, e aborda uma análise da rentabilidade dos dividendos das empresas que compõem o IBrX 50. Foi implementado um *Web Scraper* a partir da linguagem Python [29]. As informações foram extraídas do site Yahoo!Finanças [4], onde os dados não estão disponíveis para uma coleta em massa. As análises feitas para a área do mercado financeiro têm como característica uma grande massa de dados, fazendo-se necessária extração em diferentes fontes. Utilizando a tecnologia proposta, foi possível a automatização da coleta em uma escala de tempo reduzida. Os elementos obtidos foram armazenados em um arquivo que favorece a manipulação, possibilitando a realização de uma análise de regressão múltipla para chegar no objetivo proposto pelo estudo.

Por fim, o trabalho [3] de autoria de Thiago da Cunha Borges e Zeus Olenchuk Ganimi discorre sobre a implementação de um *Web Scraper* para capturar dados de valores de automóveis, extraídos do site da FIPE (Fundação Instituto de Pesquisas Econômicas). A ferramenta permite a coleta automatizada dos dados semiestruturados do módulo de consulta da página da FIPE e persistência das informações estruturadas em banco de dados PostgreSQL [30]. A análise do conteúdo estruturado pode ser utilizada, por exemplo, para otimização de recursos de empresas adquirentes de veículos, e sua visualização foi possibilitada através do software de Tableau Desktop [31].

## DISCUSSÃO

A partir dos trabalhos supracitados é possível analisar que existem diversas aplicações da tecnologia de *Web Scraping* para a obtenção de dados. Constatou-se que a implementação da técnica é capaz de auxiliar a automação, assiduidade e estudo dos dados obtidos, além de dar suporte para distintas áreas de estudo.

Uma das áreas na qual foi identificada a oportunidade de automatização é a farmacêutica. Parte do procedimento de proposta de um novo fármaco envolve a consulta aos ingredientes mais utilizados pela comunidade para se compor a receita, tendo como parâmetro de busca o princípio ativo do medicamento. Não há um veículo de comunicação oficial que permita aos pesquisadores acompanharem essas tendências de utilização de ingredientes eficientemente, de forma que as informações são extraídas de sites especializados da área.

A aplicação de *Web Scraping* nesse contexto de extração das informações de sites farmacêuticos especializados através da busca por elementos-chave que compõe a página web proverá agilidade em uma etapa importante da criação de novos medicamentos, principalmente nessa época na qual a nossa sociedade demanda por soluções rápidas e eficientes que combatam os agentes infecciosos recém descobertos.

# FUNDAMENTAÇÃO TEÓRICA

Com objetivo de disponibilizar embasamento teórico para uma acertada análise e interpretação do presente estudo, neste capítulo serão abordados os principais conceitos e definições encontrados na literatura sobre tipos de dados, técnicas de coleta e de análise e armazenamento.

## COLETA DE DADOS

Ao iniciar uma pesquisa é necessário um estudo intensivo do conjunto de dados que gere informações relevantes para a análise a ser feita. A obtenção desses dados possibilita um posterior exame pormenorizado de diferentes fontes para o estudo da área de interesse. [18]

De acordo com a finalidade da pesquisa e tipo de dado interessante a mesma, a coleta poderá ser realizada através de questionários, entrevistas, observações, experimentos ou estudo e revisões de documentos já disponíveis, além de poder ser necessária a utilização de mais de um método para explorar com profundidade as informações desejadas. No mesmo sentido, cabe realçar que independentemente do tipo de coleta utilizado, a prioridade desta será a garantia da precisão, coerência e confiabilidade nas informações obtidas.

## TIPOS DE DADOS

Para compreendermos melhor o desenvolvimento do trabalho, que será discorrido no capítulo a seguir, é importante definirmos com clareza os tipos de dados que existem. Estes são divididos em dois grupos, quanto a sua característica: qualitativos e quantitativos, e em dois outros grupos, quanto a sua origem: primários e secundários.

### QUALITATIVOS

Os dados qualitativos possuem a função de rotular e atribuir uma identificação ao objeto analisado. Como exemplos de dados qualitativos, temos: nível de escolaridade, estado civil ou nome [7].

### QUANTITATIVOS

Os dados quantitativos são numéricos e, como o nome sugere, possuem o objetivo de quantificar as amostras, além de serem amplamente utilizados em modelos estatísticos. São considerados discretos, quando assumem valores dentro de um intervalo enumerável (por exemplo, minutos inteiros em uma hora), e contínuos, quando assumem valores reais que não podem ser enumerados [7].

### PRIMÁRIOS

Os dados denominados primários são aqueles coletados diretamente de sua fonte. A obtenção de dados primários é planejada especialmente para o estudo de interesse e, portanto, são mais adequados aos seus objetivos. Em contrapartida, há a desvantagem de o esforço de coleta ser alto. São exemplos de dados primários as respostas de entrevistas e questionários dadas diretamente pelo entrevistado [7].

### SECUNDÁRIOS

Dados secundários são obtidos a partir de coletas previamente realizadas. Por já terem sido publicados, comumente os dados secundários não possuem o formato ou a integridade mais adequada ao estudo de interesse. Desta forma, apesar do menor esforço de obtenção, em comparação aos dados primários, os dados secundários requerem maiores esforços na definição de fontes e na metodologia empregada em seu manejo [7].

## ORGANIZAÇÃO E ARMAZENAMENTO

Posterior a coleta dos dados, os mesmos têm de ser armazenados de acordo com sua organização que pode ser estruturada, não estruturada ou semiestruturada. É priorizada formas acessíveis e a escolha leva em consideração a análise e nível de controle necessário dos dados.

### ESTRUTURADOS

São dados que possuem um formato específico e bem definido. Nesse tipo de organização, cada registro possui características obrigatórias que devem ser respeitadas (figura 1) [7].



Figura : Dados não estruturados, semiestruturados e estruturados

### NÃO ESTRUTURADOS

São dados que não possuem estrutura fixa, como arquivos de texto comuns, cujo conteúdo pode conter inúmeras variações em formato, idioma, formatação, entre outros (figura 1) [7].

### SEMIESTRUTURADOS

Os dados semiestruturados apresentam um avanço com relação aos dados não estruturados, pois são dotados de regras que orientam a organização do seu conteúdo, mesmo ainda possuindo considerável liberdade com relação à composição como um todo (figura 1) [7].

## FUNCIONAMENTO DAS PÁGINAS WEB

A técnica de coleta de dados proposta no presente trabalho terá como base fundamental a navegação através da estrutura de sites para a captura dos dados desejados. Portanto, é de suma importância a compreensão dos conceitos envolvendo as páginas *web* e os navegadores que conhecemos.

### HISTÓRIA DA INTERNET

As motivações que culminaram na estruturação da rede que hoje conhecemos como internet datam-se nas primeiras décadas da Guerra Fria. No ano de 1962, os militares americanos, com o apoio do governo, concluíram o desenvolvimento da rede que ficou conhecida como ARPANET. Concebida como uma resposta ao Projeto Sputnik da URSS, lançado alguns anos antes, a rede era limitada foi utilizada principalmente para a comunicação entre universidades de ponta e institutos de pesquisa da época [5].

Até o final da década de 80, a Net foi majoritariamente um meio de comunicação entre militares e acadêmicos, porém já não mais para fins necessariamente relacionados à Defesa dos EUA [5]. Em 1979, concretizou-se a visão comercial da rede através do primeiro vendedor de serviços on-line americano. Dez anos depois, o pesquisador Tim Berners-Lee, que posteriormente ficou conhecido como o pai da internet, imaginou o que chamou de *World Wide Web* (ou apenas web). Os alicerces da proposta de Berners-Lee são os mesmos da rede de ampla utilização que conhecemos hoje. Segundo KRAEMER [5], ao final do século XX, estabeleceu-se a tríade informação, educação e entretenimento na web, porém, novas possibilidades de uso seguem sendo exploradas até os dias de hoje.

### NAVEGADORES E PÁGINAS WEB

É de suma importância a compreensão da forma básica das páginas estruturadoras dos websites que compõe a web, além da responsabilidade dos navegadores na interpretação do conteúdo das mesmas. Todo site, independente da natureza de seu conteúdo, é composto de uma ou mais páginas que são relacionadas entre si através de *links* (referências) [11]. Estejam aplicadas as devidas práticas de desenvolvimento *web*, frequentemente a navegação entre as páginas é fluida o suficiente para que os usuários percebam um único contexto em todas as páginas que compõe o mesmo site [11].

Toda as páginas *web* são arquivos compostos por dados semiestruturados, mais precisamente através de uma linguagem de marcação denominada *HyperText Markup Language* – HTML [8]. A estrutura do HTML é reconhecível através das marcações realizadas através das *tags* que, por sua vez, podem agrupar outras *tags* diversas. Os dois principais agrupamentos representam o cabeçalho e o corpo do documento (figura 2).



Figura : Estrutura básica de um documento HTML

Dadas as características das páginas que compõe a internet, entra em evidência outro elemento também essencial ao funcionamento da web: os navegadores. Os navegadores são interpretadores dos diversos elementos que podem compor as páginas, como imagens, folhas de estilo CSS [13], scripts em Javascript [14] e a estrutura HTML previamente citada. Dentre os navegadores mais conhecidos estão o Google Chrome [15], Mozilla Firefox [16] e *Internet Explorer* [17].

Na figura 3, podemos visualizar o resultado da página estruturada na figura 2, quando interpretada por um navegador. Vale observar que os comentários adicionados no documento foram ignorados pelo navegador, conforme citado anteriormente.



Figura : Visualização de um documento HTML em um navegador

Apesar de os navegadores serem a forma mais conhecida e utilizada pelos usuários para acessar páginas *web*, outros programas de computador também podem requisitar e interpretar as mesmas páginas, como é o caso de algumas bibliotecas e *frameworks* de linguagens de programação. A partir destas implementações, é possível extrair o código HTML de uma página web através de poucas linhas de programação.

Dadas as características de páginas *web* previamente citadas, além das possibilidades relacionadas ao seu consumo, seja através de navegadores ou outros meios, existem diversas técnicas empregadas na análise, interpretação, busca e recuperação de dados contidos em páginas *web*.

## WEB SCRAPING

A coleta de dados da *web* moderna, conhecida como *Web Scraping* [18], é uma prática que utiliza técnicas de programação e tecnologias. É baseada no desenvolvimento de um *software* capaz de se comunicar com navegadores e simular, de forma automatizada, a interação de busca de dados de um usuário. O programa, também conhecido como *bot* ou *web crawler*, consulta servidores *web* fazendo uma requisição HTTP e, em seguida, faz uma análise minuciosa pela estrutura do documento HTML recebido afim de retornar o fragmento onde o dado de interesse foi encontrado.

### BUSCANDO INFORMAÇÕES EM UM DOCUMENTO HTML

Um documento HTML é um arquivo de texto simples e possui uma estrutura definida, portanto é possível aplicar técnicas de busca no mesmo. Há várias formas de fazer a busca, como por expressões regulares que formam um padrão de texto, por palavras-chave encontradas no documento e utilizando a estrutura do documento.



Figura : Estrutura básica de um documento HTML representado como texto

A partir da estrutura do HTML é viável a utilização de ferramentas que auxiliem a extração de dados no documento como a biblioteca *Beautiful Soup* [19]. Essa biblioteca fornece métodos que facilitam a navegação e pesquisa do conteúdo fazendo o *parsing* dos sites.

O *parsing*, que pode ser compreendido como um tradutor, é uma análise sintática onde lê-se uma entrada e monta-se uma estrutura de árvore da sua composição. Inicialmente o *parsing* faz uma análise léxica do documento tornando sequência de caracteres em palavras-chaves que podem ser analisadas futuramente. Em seguida é feito uma análise semântica conferindo se está em conformidade com as regras de composição de um arquivo HTML. Por fim, é gerado o código que pode ser utilizado para execução de comandos levando em consideração a árvore obtida ou para responder uma saída de interesse.



Figura : Representação da estrutura básica de um documento HTML

Portanto é possível extrair o título do documento representado na Figura 4 a partir do seguinte caminho: html; head; title.

# IMPLEMENTAÇÃO

Conforme definições realizadas em reunião que contou com a presença de um farmacêutico pesquisador na área de interesse do presente trabalho (detalhada no Anexo A), foram levantados os requisitos necessários para a implementação da técnica de *web scraping*, visando a automatização da busca e coleta de dados, em fontes específicas, relacionados aos princípios ativos de determinados medicamentos e seus demais constituintes.

O assunto deste capítulo serão os detalhes técnicos da implementação aplicada, incluindo: diagrama e descrição dos casos de uso observados, informações sobre a tecnologia optada, um diagrama de classes provendo uma visão geral da arquitetura implementada e a definição das fontes de dados a serem exploradas, conforme orientação do especialista farmacêutico consultado.

## CASOS DE USO

### DIAGRAMA DO CASO DE USO SCRAP ANVISA



Figura : Diagrama do caso de uso Scrap Anvisa

### DESCRIÇÃO DO CASO DE USO SCRAP ANVISA

Tabela : Descrição do caso de uso Scrap Anvisa

|  |  |
| --- | --- |
| **ID:** | **UC01** |
| **Objetivo:** | Extrair bulas dos profissionais do bulário Anvisa. |
| **Requisitos:** | Ter uma conexão ativa à Internet. |
| **Atores:** | Usuário. |
| **Pré-condições:** | As dependências (compilador/interpretador da linguagem de programação, *framework* e bibliotecas auxiliares) devem estar instaladas no ambiente. |
| **Pós-condições:** | Os arquivos obtidos deverão estar salvos em disco. |
| **Fluxo principal:** | ***Scrap* Anvisa**   1. O ator executa a ferramenta de linha de comando, fornecendo os parâmetros de busca para a fonte Anvisa; 2. A ferramenta recupera os resultados da base de dados Anvisa, conforme o filtro passado como parâmetro; 3. A ferramenta gera um arquivo contendo o histórico da execução e os arquivos de bulas dos resultados da busca no disco; |
| **Fluxo alternativo:** | Não há. |
| **Erros/Exceções:** | O tratamento de erros (como requisições inválidas e timeouts) é feito pelo *framework* utilizado (mensagens de erro podem ser exibidas para o usuário).  Em caso de falha(s) na execução, as pós-condições não serão verdadeiras (os arquivos de resultado não serão obtidos). |
| **Mensagens:** | *Logs* de saída contendo detalhes de execução da ferramenta. |
| **Regras de negócio:** | **[RN01]** O endereço <http://www.anvisa.gov.br/datavisa/fila_bula/frmResultado.asp> deve aceitar requisições com o método HTTP POST, contendo os seguintes parâmetros e seus respectivos significados:   * “txtMedicamento”: nome do princípio ativo * “txtEmpresa“: empresa registrante * “txtNuExpediente”: número de expediente * “txtDataPublicacaoI”: filtro de data inicial de publicação * “txtDataPublicacaoF”: filtro de data final de publicação * “txtPageSize”: número de registros por página   Nenhum dos parâmetros listados acima é obrigatório. Caso o parâmetro “txtPageSize” não seja especificado, o resultado trará 10 registros por página.  **[RN02]** O resultado da requisição descrita em **RN01** deverá ser uma página HTML válida, contendo um elemento do tipo “*table*” com ID “tblResultado”, que representa a listagem de resultados da busca pelo medicamento. Além disso, a tabela citada deverá conter um elemento do tipo cabeçalho de tabela (“th”) contendo o texto “Bula do Profissional”. A célula (“td”) posicionada na mesma coluna do cabeçalho citado, deverá conter um elemento de *hyperlink* (“a”) com o seu atributo “*onclick*” fazendo uma chamada de função Javascript denominada “fVisualizarBula” e que recebe dois parâmetros no formato de texto, mas que são valores numéricos válidos.  **[RN03]** O endereço <http://www.anvisa.gov.br/datavisa/fila_bula/frmVisualizarBula.asp> deve aceitar requisições com o método HTTP POST, contendo os seguintes parâmetros e seus respectivos significados:   * “pNuTransacao”: número de transação * “pIdAnexo”: identificador numérico do anexo   **[RN04]** O resultado da requisição descrita em **RN03** deverá ser um arquivo PDF válido, que representa a bula do profissional. |

### DIAGRAMA DO CASO DE USO TEXT MINING ANVISA



Figura 7: Diagrama do caso de uso Text Mining Anvisa

## TECNOLOGIAS

A biblioteca Python Scrapy [20] é um *framework* desenvolvido para extrair dados de páginas web de forma sistematizada. São disponibilizadas diversas funcionalidades para coordenar as requisições HTTP, controlar navegação, fazer o parse em HTML e representar os dados em arquivos com formatos JSON [21], CSV e XML [22].

Um projeto Scrapy tem a arquitetura construída em classes de capturas de dados, conhecidas como *spiders* [24]. Essas classes possuem funcionalidades básicas sobre como extrair informações estruturadas da web, e é nelas que definimos o comportamento específico sobre como capturar e converter as páginas de um site em particular.

Para a tradução dos termos farmacêuticos a serem pesquisados no livro especializado utilizamos a biblioteca Python googletrans [26] e, para suportar a colheita e estruturação dos metadados de execução do projeto, utilizamos a biblioteca Pandas [25].

## DIAGRAMA DE CLASSES

O diagrama de classes do projeto traduz a arquitetura aplicada, demonstrando a organização em módulos e, quando há, o relacionamento entre as classes utilizadas, com base no framework Scrapy e na biblioteca Pandas [25].

O projeto foi projetado para funcionar através de módulos independentes, de forma que não há comunicação direta entre eles e, portanto, não há relacionamento entre as suas respectivas classes. A sequência de execução se dá através da troca de mensagens e os insumos são recebidos e lidos através de arquivos gravados no próprio sistema de arquivos do computador hospedeiro.

No caso das classes responsáveis por capturar as informações da fonte web, utilizamos o conceito de herança, presente no paradigma de Programação Orientada à Objetos, para herdar todos os atributos e métodos básicos da classe primária e implementamos as especificidades de cada fonte em questão.

### DESCRIÇÃO DAS CLASSES

* **Spider**: componente do *framework* *Scrapy* [20] que já implementa funcionalidades básicas para a captura e conversão de dados oriundos de páginas web.
* **CrawlerRunner**: componente do *framework* *Scrapy* [20] responsável por gerenciar e manter rastreabilidade da execução de diferentes *spiders*.
* **AnvisaAutocompleteSpider:** especificação da classe base *Spider*, cuja responsabilidade é a de capturar as todas as sugestões possíveis de nomes de medicamentos fornecidas pela Anvisa e salvar o resultado em um arquivo texto.
* **AnvisaBularioSpider:** especificação da classe base *Spider*, que recebe um parâmetro de busca em texto do usuário, faz a busca no bulário Anvisa, verificando as sugestões de medicamentos correspondentes, e percorre a lista completa de resultados, baixando todos os PDFs de bula do profissional disponíveis.
* **Stats:** classe que concentra a lógica de verificação dos produtos de processamento das demais classes e estrutura metadados sobre a execução do projeto.
* **BulaParser:** classe responsável por interpretar o contéudo das bulas extraídas do bulário Anvisa, identificar as informações de interesse (formulação e excipientes) e salvar os resultados em arquivos JSON.
* **TranslationManager:** classe responsável por gerenciar as traduções realizadas a partir do conteúdo extraído das bulas, fazendo a interface com a biblioteca googletrans [26].



Figura 8: Diagrama de classes do projeto

## FONTES

Afim de desenvolver a extração do dado corretamente, é executada uma sequência de operações em três módulos (figura 9) que estarão detalhados nesta seção separados pelas fontes utilizadas.



Figura 9: Fluxograma dos módulos da extração do dado

### ANVISA

#### DOWNLOAD DAS BULAS DO PROFISSIONAL

A implementação inicia com uma sucessão de atividades, figura 10, afim de fazer o download de bulas do profissional do medicamento disponíveis no bulário eletrônico da Anvisa.



Figura 10: Fluxograma de atividades para obter as bulas do profissional.

Detectou-se que para a busca das bulas do medicamento, o site do bulário eletrônico da Anvisa retorna sugestões, figura 11, conforme o usuário insere as letras do nome do medicamento. Contudo, a correlação entre o nome fornecido e os registros da base de dados do órgão é feita apenas pelas primeiras letras dos medicamentos da base e então não são retornadas as bulas de todas as sugestões, somente das que possuem os nomes que iniciam com o trecho pesquisado.



Figura : Exemplo de lista de sugestões de nomes completos do fármaco.

Devido ao cenário citado e para atender ao objetivo de recuperar todos os resultados de bulas conforme o parâmetro fornecido, o script inicia fazendo uma requisição GET para a URL <http://www.anvisa.gov.br/datavisa/fila_bula/funcoes/ajax.asp?opcao=getsuggestion&ptipo=1>, que contém a lista com todas as sugestões de nomes de medicamentos, trata os dados recebidos e salva-os no arquivo “medicamentos.txt”. Assim que é inserido o nome do medicamento desejado, todas as sugestões que o contém são salvas para garantir que nenhum resultado será ignorado.

Em seguida para cada sugestão é feito uma requisição ao endereço <http://www.anvisa.gov.br/datavisa/fila_bula/frmResultado.asp> passando os seguintes parâmetros através do método POST:

* “txtMedicamento” com uma das sugestões de nomes do medicamento;
* “hddPageSize” contendo o número de registros por página desejado. Conforme alguns testes realizados, e levando em conta relações como o equilíbrio entre o tempo necessário para a consulta e riscos como a qualidade da internet do usuário, optou-se por manter o padrão de 10 registros por página.
* “hddPageAbsolute” contendo o valor 1 para a primeira página e sendo incrementado conforme a necessidade do script de avançar para as próximas páginas.

É retornada a página de resultados que exibe uma página HTML contendo uma tabela que armazena as bulas. Vale ressaltar que a tabela também pode estar vazia. Por fim, com o objetivo de fazer o download de todas as bulas do profissional exibidas na tabela “tblResultado”, faz-se uma busca pela coluna “Bula do Profissional” no cabeçalho das colunas da tabela. Em cada célula dessa coluna existe a tag <a>, que representa um link HTML, que possui o atributo “onclick” chamando uma função JavaScript denominada “fVisualizarBula”, que fornece dois parâmetros numéricos que são capturados e armazenados para uso posterior. Finalmente é feita uma requisição para <http://www.anvisa.gov.br/datavisa/fila_bula/frmVisualizarBula.asp> passando os parâmetros obtidos no passo anterior, e recebendo o arquivo PDF da bula do profissional como resposta.

#### ANÁLISE DAS BULAS E EXTRAÇÃO DOS EXCIPIENTES

A análise dos PDFs das bulas do profissional requer uma série de etapas, descritas no fluxograma da figura 12, com objetivo de extrair as formulações e os nomes dos excipientes do medicamento.



Figura : Fluxograma de atividades para extração dos excipientes.

A partir das bulas obtidas no bulário eletrônico da Anvisa, inicia-se o processo percorrendo e abrindo cada uma afim de convertê-las de PDF para arquivo de texto. Para isso, foi utilizada a biblioteca PDFMiner [27] que, em resumo, auxilia na extração e análise de texto de documentos PDF. Foi necessário a utilização de ferramentas de configuração da biblioteca para manter o *layout* do arquivo original.

Verificou-se que a estrutura do documento das bulas segue um padrão e os excipientes desejados para esse estudo são encontrados no tópico “COMPOSIÇÃO”, portanto é feito uma busca, com o método find() em Python, pela palavra-chave afim e delimitar o início do texto que contém os dados almejados. E para determinar o fim da seção de composição e otimizar a pesquisa, foi necessário buscar também pelas palavras-chave “INFORMAÇÕES TÉCNICAS” ou “INDICAÇÕES” que correspondem aos títulos possíveis para o próximo tópico da bula.

Com o texto do tópico “COMPOSIÇÃO” em memória, foi possível encontrar a formulação do medicamento uma vez que, analisando o padrão da bula foi identificado que ela aparece entre as palavras “cada” e “contém” ou “contem”.

Apesar de existir um padrão na estrutura dos documentos, foram identificadas diferentes formas de especificar os excipientes nas bulas. Então, por último, foi necessário tratar cada caso de uma maneira e acordado contemplar os seguintes modelos de apresentação da composição dos excipientes nas bulas:

**CASO 1**: Busca-se no intervalo entre a palavra “excipientes:” e o próximo ponto final.



Figura : Exemplo de bula contemplada no CASO 1.

**CASO 2**: Busca-se no intervalo entre a palavra “excipiente:” e o próximo ponto final.



Figura : Exemplo de bula contemplada no CASO 2.

**CASO 3**: Busca-se a palavra “excipientes\*” e em seguida no intervalo entre “\*” e o próximo ponto final.



Figura : Exemplo de bula contemplada no CASO 3.

**CASO 4**: Busca-se no intervalo entre a palavra “veículos:” e o próximo ponto final.



Figura : Exemplo de bula contemplada no CASO 4.

**CASO 5**: Busca-se a palavra “excipiente\*\*” e em seguida no intervalo entre “\*\*” e o próximo ponto final.



Figura : Exemplo de bula contemplada no CASO 5.

**CASO 6**: Busca-se a palavra “excipientes” e em seguida no intervalo entre “(” e “)”.



Figura : Exemplo de bula contemplada no CASO 6.

**CASO 7**: Busca-se a palavra “excipiente” e em seguida no intervalo entre “(” e “)”.



Figura : Exemplo de bula contemplada no CASO 7.

**CASO 8**: Busca-se a palavra “veículo” e em seguida no intervalo entre “(” e “)”.



Figura 20: Exemplo de bula contemplada no CASO 8.

**CASO 9**: Busca-se no intervalo entre “quantidade suficiente de” e “como”.



Figura : Exemplo de bula contemplada no CASO 9.

Conforme os dados são extraídos, os mesmos são guardados em um arquivo JSON que relaciona as formulações aos excipientes. Os casos que não são contemplados como os da imagem 21, o sistema retornará vazio.



Figura : Exemplo de bula não contemplada.

### HANDBOOK OF PHARMACEUTICAL EXCIPIENTS

Para concluir a implementação e alcançar o objetivo proposto, foi necessário elaborar a sequência de ações representadas no fluxograma da imagem 22.



Figura : Fluxograma de atividades para extração de informações do excipiente.

Utilizou-se o PDF do *Handbook of Pharmaceutical Excipients* [28] que trata-se de um livro tido como referência mundial pela área farmacêutica em seu conteúdo sobre formulações de medicamentos, incluindo detalhes sobre forma de utilização de cada excipiente, como casos de uso, proporções, patentes e outras informações.

A princípio, já que as páginas do documento que se encontra o sumário são conhecidas, foi necessário busca-las, converte-las para um arquivo texto e trata-las para que fosse possível criar um arquivo JSON relacionando o nome do excipiente a página do início da sua seção afim de facilitar a busca posterior.

Uma vez que muitos medicamentos possuem a mesma substância como veículo nas bulas, na lista de excipientes extraídos anteriormente havia nomes replicados, portanto, precisou tratar e eliminar as repetições. Em sequência, a nova lista foi traduzida para o inglês, usando a biblioteca googletrans em Python, em razão do livro estar nesse idioma. E então foi possível pesquisar os excipientes da lista no arquivo do sumário e encontrar a página com o início da sua seção no livro.

Logo em seguida é feito a busca pela página de interesse e extraído o texto referente ao tópico “7 *Applications in Pharmaceutical Formulation or Tecnology*”, tendo como início o título citado e como marcador do fim o título do próximo tópico “8 *Decription*”. Sendo assim, busca-se pelas ocorrências dos seguintes caracteres: “%”, “w/w”, “w/v” ou “mL”, que indicam unidades de medida, e ao serem encontrados, extraímos toda a frase que os contém afim de obter as funções e quantidades utilizadas dessas substâncias no medicamento de interesse.

# TESTES

CONCLUSÕES E TRABALHOS FUTUROS

Neste capítulo devem ser colocadas as conclusões que o aluno obteve durante a elaboração do trabalho, bem como o que pretende após sua conclusão (especialização, mestrado, aplicar os conhecimentos em alguma área...).

REFERÊNCIAS BIBLIOGRÁFICAS

1. ZANON, Carolina Franco. **Aplicação de um programa de extração de dados através da técnica de web scraping para determinação do perfil epidemiológico dos pacientes da clínica de estomatologia da faculdade de odontologia da ufrj cadastrados no software estomato web**, 2017. Dissertação (Mestrado em Odontologia) – Curso de Pós-graduação em Odontologia, UFRJ, Rio de Janeiro, RJ.
2. MAZINI, Dhaniel Nunes; SATO, Renato Cesar. **Extração de dados financeiros com um web scraper: um estudo sobre a rentabilidade dos dividendos**, 2019. Dissertaação (Artigo na WAIAF) - Workshop of Artificial Intelligence Applied to Finance, ITA, São José dos Campos, SP.
3. BORGES, Thiago da Cunha; GANIMI, Zeus Olenchuk. **Extração de dados com web scraping para análise da variação de preço de veículos automotores**, 2018. Dissertação (Tecnólogo em Sistemas de Computação) - Curso de Tecnologia em Sistemas de Computação, UFF, Niterói, RJ.
4. **Yahoo! Finanças** <<https://br.financas.yahoo.com/>> Acesso em 09 de Maio de 2020
5. ABREU, Karen Kraemer. **História e usos da Internet**, 2009. Biblioteca On-line de Ciências da Comunicação, p. 2-4
6. KABIR MUHAMMAD SAJJAD, Syed. **Basic Guidelines for Research: An Introductory Approach for All Disciplines**, Edition: First, Chapter: 9, Publisher: Book Zone Publication, Chittagong-4203, Bangladesh, p. 201-275.
7. ZOZUS NAHM, Meredith. **The data book collection and management of research data**-Chapman and Hall CRC Press, 2017, p. 36.
8. BERNERS-LEE, Tim. **WWW: past, present, and future**, 1996. Publicação acadêmica – Computer, vol. 29, no. 10, pp. 69-77.
9. LONGMAN WESLEY, Addison. **A History of HTML**, 1998. <https://www.w3.org/People/Raggett/book4/ch02.html> Acesso em 24 mai. 2020.
10. BERNERS-LEE, Tim. **Information Management: A Proposal**, 1989. <https://www.w3.org/History/1989/proposal.html> Acesso em 24 mai. 2020.
11. LAWSON, Bruce; SHARP, Remy. **Introducing HTML5, 2nd Edition**, 2011. Livro. New Riders Press, Berkeley, Canadá.
12. **What is the difference between webpage, website, web server and search engine?** <<https://developer.mozilla.org/en-US/docs/Learn/Common_questions/Pages_sites_servers_and_search_engines>> Acesso em 24 mai. 2020.
13. **CSS: Cascading Style Sheets** <<https://developer.mozilla.org/en-US/docs/Web/CSS> > Acesso em 24 mai. 2020.
14. **Javascript** <<https://developer.mozilla.org/en-US/docs/Web/JavaScript>> Acesso em 24 mai. 2020.
15. **Google Chrome** <https://www.google.com/chrome> Acesso em 24 mai. 2020.
16. **Mozilla Firefox** <https://www.mozilla.org/en-US/firefox/new> Acesso em 24 mai. 2020.
17. **Microsoft Internet Explorer** <https://www.microsoft.com/pt-br/download/internet-explorer.aspx> Acesso em 24 mai. 2020.
18. OLSEN, Wendy; **Coleta de Dados**: Debates e Métodos Fundamentais em Pesquisa Social. 1. ed. Porto Alegre: Penso, 2015. p. 15.
19. RICHARDSON, Leonard. Documentação Beautiful Soup. <<https://www.crummy.com/software/BeautifulSoup/bs4/doc.ptbr/>> Acesso em 08 jun. 2020.
20. **Scrapy** < <https://scrapy.org/>> Acesso em 12 jun. 2020.
21. **JSON – JavaScript Object Notation**. <http://www.json.org/> Acesso em 12 de jun. 2020.
22. **XML – Extensible Markup Language**. Acesso em 21 de jun. 2020.
23. **PDFMiner** <<https://pypi.org/project/pdfminer/>> Acesso em 29 jun. 2020.
24. **Scrapy Spiders** <<https://docs.scrapy.org/en/latest/topics/spiders.html>>. Acesso em 29 de jun. 2020.
25. **Pandas** <<https://pandas.pydata.org/>>. Acesso em 29 de jun. 2020.
26. **Googletrans** < <https://pypi.org/project/googletrans/>>. Acesso em 29 de jun. 2020.
27. **PDFMiner** <<https://pypi.org/project/pdfminer/>> Acesso em 29 jun. 2020.
28. ROWE, Raymond C; SHESKEY, Paul J; QUINN, Mirian E. **Handbook of Pharmaceutical Excipientes**, 6th edition. 2009 London, UK: Pharmaceutical Press e Washington, DC: American Pharmacists Association.
29. **Python** <<https://www.python.org/>>. Acesso em 02 jul. 2020.
30. **PostgreSQL** <<https://www.postgresql.org/>>. Acesso em 02 jul. 2020.
31. **Tableau** <<https://www.tableau.com/pt-br>>. Acesso em 02 jul. 2020.

ANEXOS

ANEXO A – CONTEÚDO DA REUNIÃO COM O ESPECIALISTA

O objetivo do projeto envolve a utilização de engenharia reversa para, a partir dos medicamentos já produzidos, possamos extrair estatísticas de utilização dos seus constituintes e produzir novos fármacos com essas formulações.

O documento abaixo apresenta o conteúdo da reunião realizada no dia 21/05/2020, que envolveu Altobelli de Brito (orientador), Lúcio Cabral (especialista farmacêutico), Marcello Telles (aluno) e Yasmim Martins (aluna).

***Manual da busca por bulas***

A primeira fonte de informação a ser consultada é a base de dados da Anvisa já que o órgão deve publicar as bulas de todos os medicamentos registrados em seu bulário.

Abaixo temos demonstradas as etapas de busca e coleta das bulas.

1. Ao acessar a página principal de consulta ao bulário (<http://www.anvisa.gov.br/datavisa/fila_bula/index.asp>), é apresentada a tela representada na **Figura 1**.



**Figura 1**: Tela da página principal do Bulário Eletrônico da Anvisa

1. Neste momento são preenchidos os campos desejados para filtrar a busca. Caso nenhum campo seja preenchido, o resultado listará todas as bulas de todos os medicamentos, contando com 10 registros por página. No exemplo na Figura 2, buscamos por “dipirona”.



**Figura 2**: Tela de resposta da busca por “dipirona”.

1. O resultado exibe informações como nome do medicamento, empresa responsável pela produção, expediente, data de publicação e os arquivos no formato PDF com versões da bula do paciente e do profissional. Nosso alvo para este projeto é a bula do profissional para todos os registros que compõe o resultado, e que pode ser baixada com um simples clique no símbolo do arquivo como mostrado na Figura 3.



**Figura 3**: Localização da Bula do Profissional na tela de repsosta da busca.

INCLUIR INFORMAÇÕES SOBRE OS BULÁRIOS DO FDA E EUDRA

ANEXO B – DOCUMENTO DE ACEITAÇÃO ASSIANDO PELO ESPECIALISTA

**\*INCLUIR IMAGEM DO DOCUMENTO\***

**Figura 1** – Documento de aceitação por parte do especialista.