UNIVERSIDADE ESTADUAL DE MATO GROSSO DO SUL

ALUNO: GUILHERME HENRIQUE VIEIRA PEREIRA

MINERAÇÃO DE DADOS DE PÁGINAS DE NOTÍCIAS SOBRE DELITOS E INFRAÇÕES DE TRÂNSITO USANDO PYTHON E WEB SCRAPING

ORIENTADOR: PROF. DR. RUBENS BARBOSA FILHO

DOURADOS – MS
2017

PROBLEMÁTICA

- Páginas de notícias têm como objetivo apresentar informações de cunho abrangentes, onde a preocupação com a sociedade é a de transmitir de forma imparcial e clara assuntos diversificados, passando desde notícias sobre esportes até notícias sobre cultura (BENASSI, 2007).

OBJETIVO

- Desenvolver um *software* minerador de dados com aplicações em páginas *web*, com foco na região de Dourados – MS e que busque informações de cunho policial.

OBJETIVOS ESPECÍFICOS

- Métodos de Mineração de Dados
- Analisar Técnicas de *parsing* na aquisição de informações
- Testar vários sites de notícias
- Comprovar sua eficiência

HIPÓTESE

- Diariamente são publicadas, aproximadamente, cerca de 3 mil notícias em mais de 80 páginas de notícias nacionais (MORENO, 2017).
- Um programa que permita acessar e visualizar informações sobre delitos e infrações de trânsito, seria viável?

JUSTIFICATIVA

- Um software que permita o acesso a informações estatísticas de cunho policial é imprescindível para o gerenciamento e acompanhamento da segurança pública
- Para quem atua na área
- Para os leitores

REFERENCIAL TEÓRICO

- Sartori utiliza a mineração de dados em dados da Polícia Militar de Balneário Camboriú (SARTORI, 2012).
- informações de boletins de ocorrências sobre homicídios dolosos (SILVA e ROVER, 2011).
- Mineração Textual está diretamente relacionada a obtenção de informações do tipo texto.
- A busca por informações que possam prever e determinar regiões críticas, tem como propósito auxiliar a questão de segurança pública (MELO, 2010).

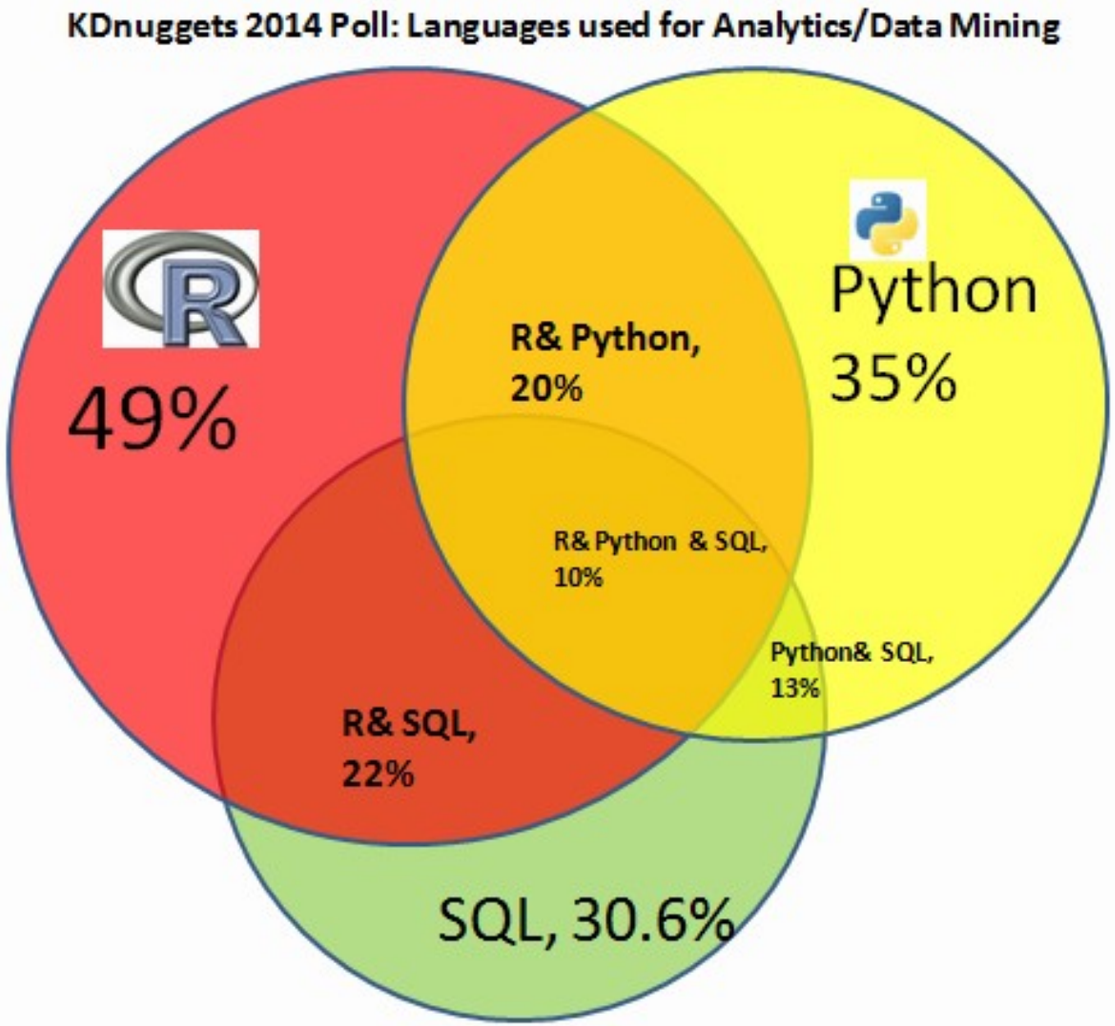
METODOLOGIA

- Estudos de técnicas e tecnologias
- Python
- Scrapy
- Scrapinghub*
- RegEx*
- Xpath*
- Googlemaps API

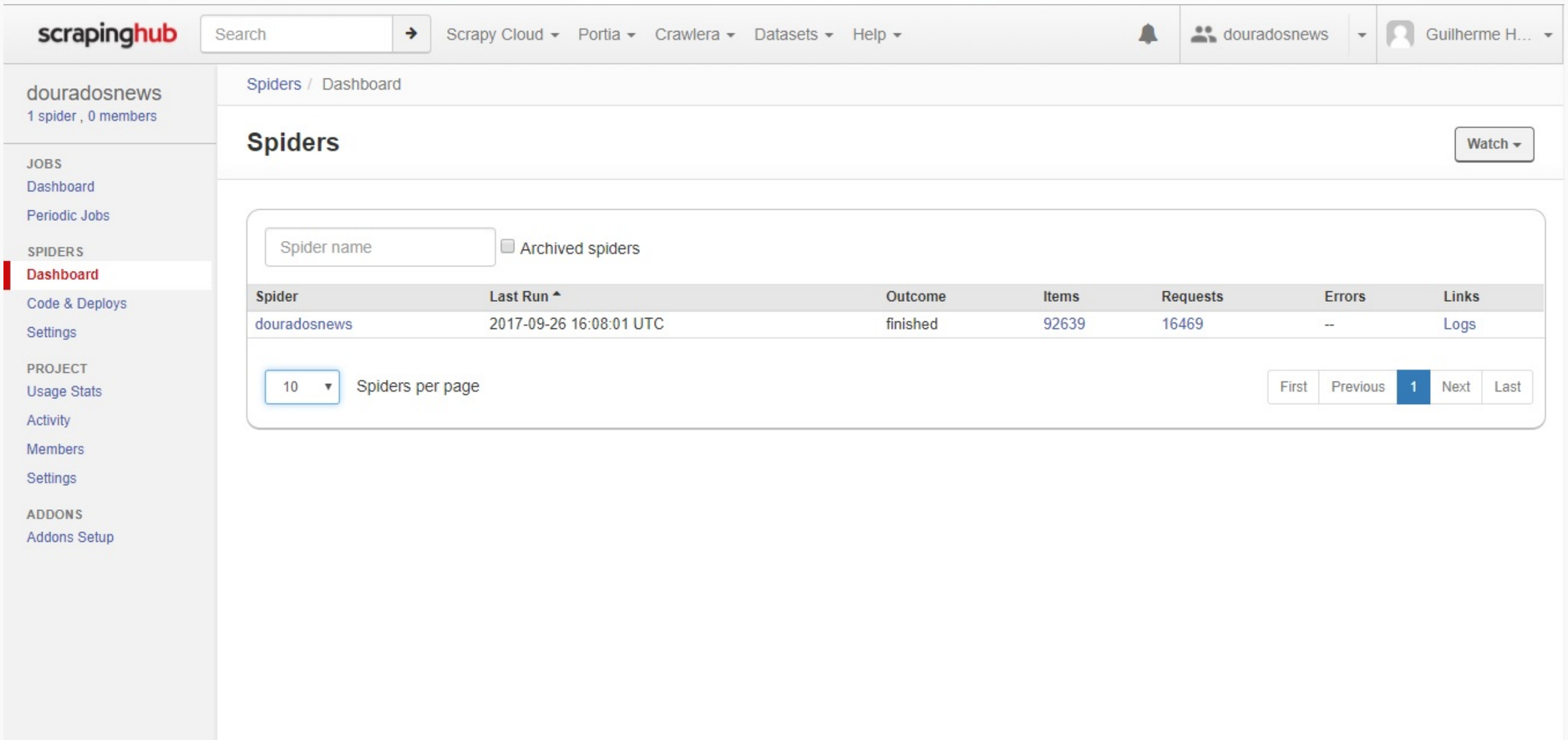
METODOLOGIA

- A Mineração de Dados é a técnica utilizada na busca, coleta e análise de dados em grande escala, com o objetivo de encontrar padrões, relação entre os dados obtidos.
- Análise preditiva, prescritiva, diagnóstica e descritiva, são os tipos de análise utilizados na mineração de dados, tendo este último tipo como o mais relacionado com o objetivo deste trabalho e portanto sendo o modelo de análise utilizado.

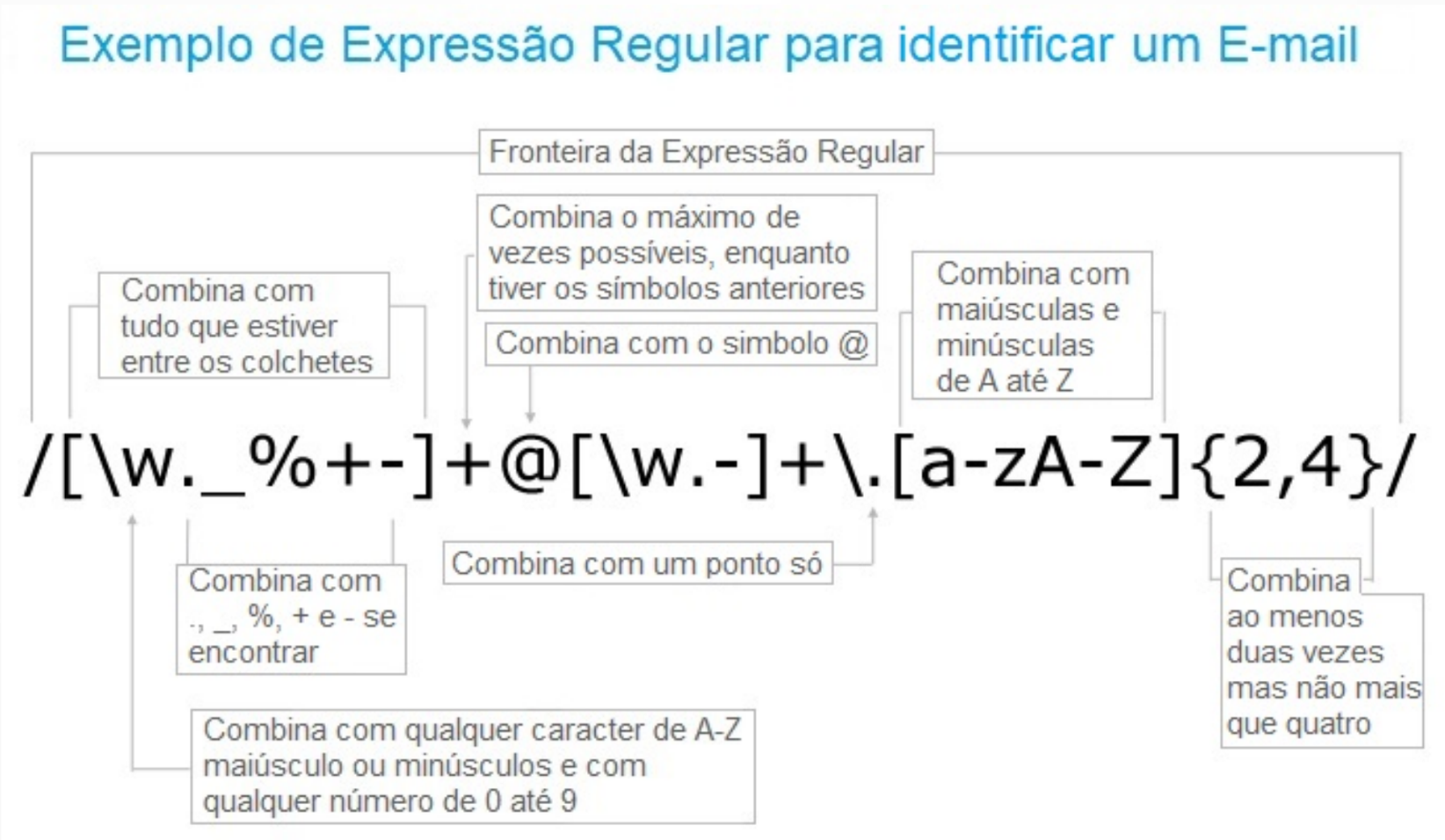
METODOLOGIA



METODOLOGIA

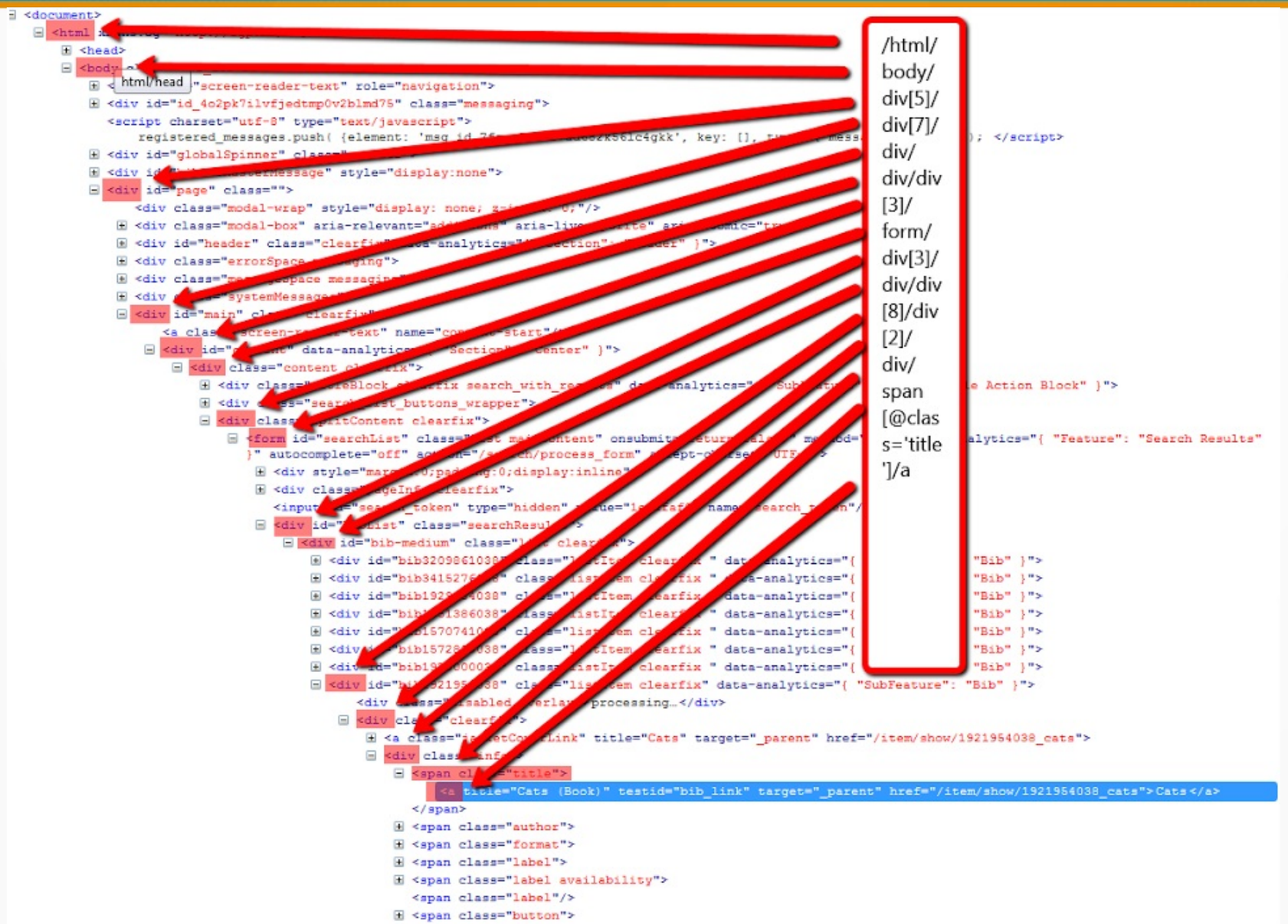


METODOLOGIA



Fonte: Computer Hope, 2017.

METODOLOGIA



XPath
Fonte:
Test Automation
For Manual Testers,
2014.

DESENVOLVIMENTO

```
### PARTE 1 ###
class QuotesSpider(scrapy.Spider): # NOME DA CLASSE PRINCIPAL
    name = "quotes" # NOME DA SPIDER
    start_urls = [
        # URLs INICIAIS - PODENDO HAVER MAIS DE UMA
        'http://quotes.toscrape.com/tag/humor/',
    ]
```

```
### PARTE 2 ###
    def parse(self, response): #FUNÇÃO parse, ONDE SÃO EXTRAÍDOS FEITAS AS CONFIGURAÇÕES DE
      EXTRAÇÃO DOS DADOS
        for quote in response.css('div.quote'): # LAÇO for QUE BUSCA, EM CADA ITERAÇÃO, TODAS
          TAGs <div class="quote"></div>
            yield {
                # O TEXT É EXTRAÍDO DA TAG <span>, INTERNA A ANTERIOR, NO FORMATO <span class
                  ="text"></span>
                'text': quote.css('span.text::text').extract_first(),
                # O AUTOR DO TEXTO ACIMA É ENCONTRADO NA SEGUINTE TAG: <span>by <small class
                  ="author" itemprop="author">Author Name</small></span>
                'author': quote.xpath('span/small/text()').extract_first(),
            }
```

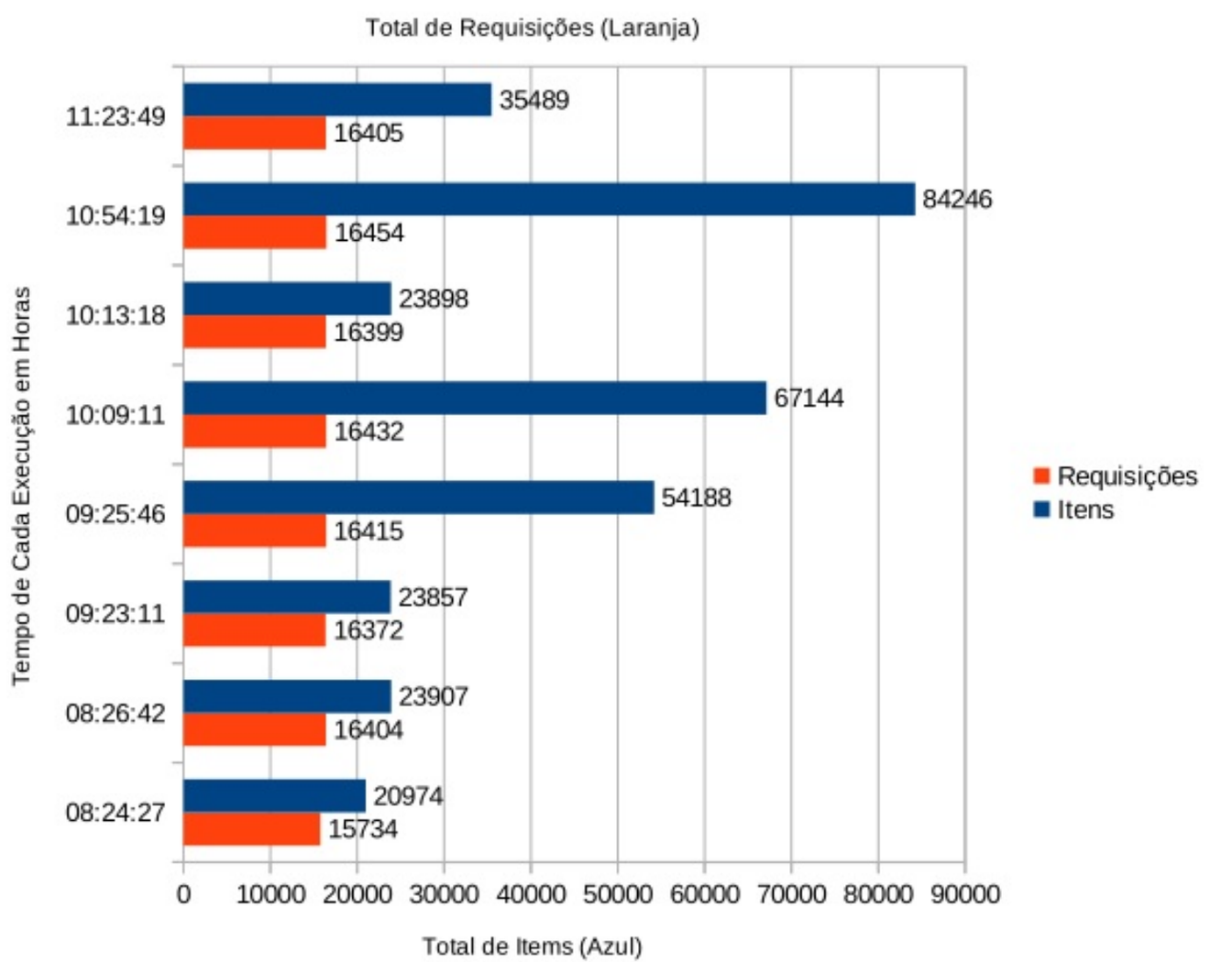
DESENVOLVIMENTO

```
### PARTE 3 ###
        next_page = response.css('li.next a::attr("href")').extract_first() # OBTÉN O link PARA
          A PRÓXIMA PÁGINA A SER REQUISITADA
        if next_page is not None: # VERIFICA SE A PÁGINA EXISTE
            yield response.follow(next_page, self.parse) # REALIZA A REQUISIÇÃO DA PÁGINA
              ENCONTRADA
```

DESENVOLVIMENTO

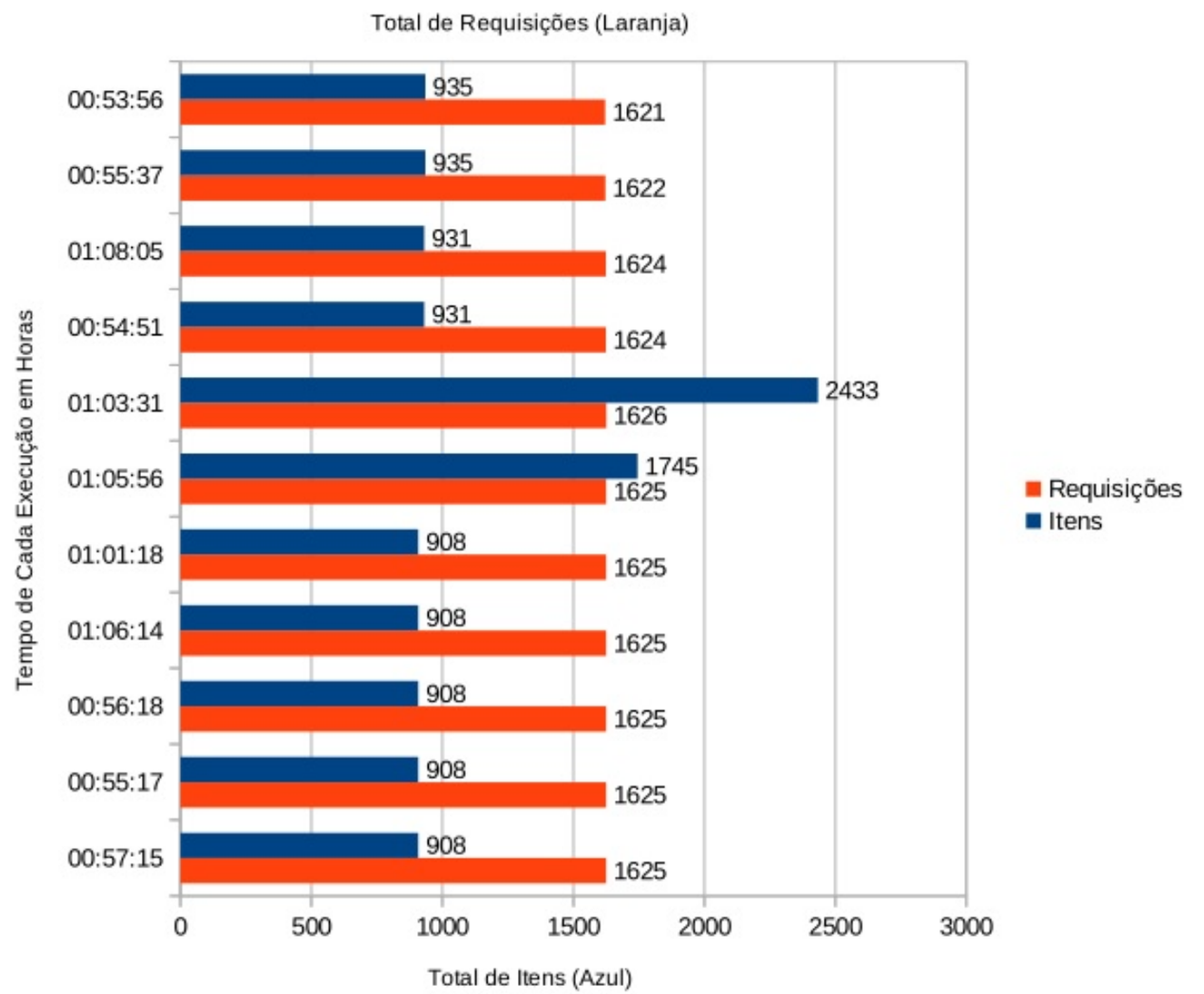
```
1  from googlemaps import GoogleMaps
2
3  gmaps = GoogleMaps(api_key)
4  end = 'Avenida Marcelino Píres, Dourados, MS'
5  lat, lng = gmaps.address_to_latlng(end)
6  print lat, lng
7
8  -22.2259387 -54.7951442
```

RESULTADOS – Dourados News



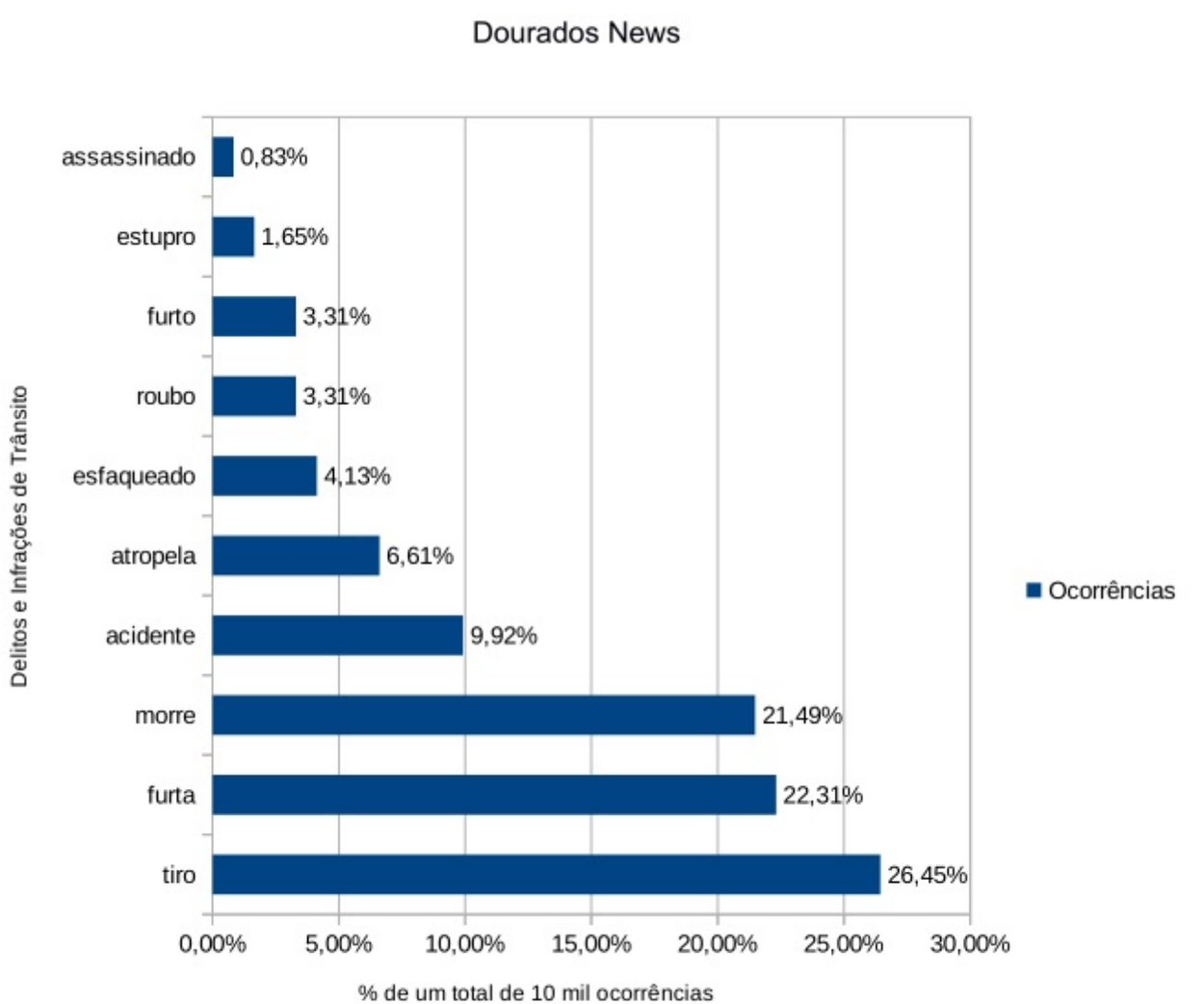
Fonte: Própria, 2017.

RESULTADOS – Dourados Agora



Fonte: Própria, 2017.

RESULTADOS



Fonte: Própria, 2017.

RESULTADOS

```
2017-09-27 01:08:53 [scrapy.core.scraper] DEBUG: Scraped from <200 http://www.douradosnews.com.br/dourados/
dourados-sedia-nesta-terca-feira-roda-de-conversa-visibilidade-lgbt->
{'cause': u'morre', 'local_lat_lng': {u'Centro\n': {u'lat': -22.223561, u'lng': -54.8125486}}, 'link': u'ht
tp://www.douradosnews.com.br/policia/ciclista-bate-em-veiculo-parado-cai-e-morre-atropelado-por-carreta'}
2017-09-27 01:08:53 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://www.douradosnews.com.br/dourados/
dourados-mais-estacionamentos-menos-canteiros-> (referer: http://www.douradosnews.com.br/dourados/homem-per
de-controle-de-veiculo-e-bate-em-poste-na-marcelino-pires)
2017-09-27 01:08:57 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://www.douradosnews.com.br/dourados/
interno-do-semiaberto-e-baleado-a-caminho-do-presidio> (referer: http://www.douradosnews.com.br/dourados/me
nor-baleado-apos-perseguicao-policial-e-liberado-e-diz-ter-sido-contratado)
2017-09-27 01:08:58 [urllib3.connectionpool] DEBUG: https://maps.googleapis.com:443 "GET /maps/api/geocode/
json?address=rua+Dom+Jo+dourados+ms&key=AIzaSyCVMO84F5-qadD3-5qAha1AOt65M-vERUo HTTP/1.1" 200 1665
2017-09-27 01:08:58 [scrapy.core.scraper] DEBUG: Scraped from <200 http://www.douradosnews.com.br/dourados/
interno-do-semiaberto-e-baleado-a-caminho-do-presidio>
{'cause': u'furto', 'local_lat_lng': {u'rua Dom Jo': {u'lat': -22.2410875, u'lng': -54.8213549}}, 'link': u
'http://www.douradosnews.com.br/dourados/velho-conhecido-da-policia-e-preso-apos-sequencia-de-furtos-em-dou
rados'}
2017-09-27 01:08:59 [urllib3.connectionpool] DEBUG: https://maps.googleapis.com:443 "GET /maps/api/geocode/
json?address=rua+Dom+Jo+dourados+ms&key=AIzaSyCVMO84F5-qadD3-5qAha1AOt65M-vERUo HTTP/1.1" 200 1665
2017-09-27 01:08:59 [scrapy.core.scraper] DEBUG: Scraped from <200 http://www.douradosnews.com.br/dourados/
interno-do-semiaberto-e-baleado-a-caminho-do-presidio>
{'cause': u'furto', 'local_lat_lng': {u'rua Dom Jo': {u'lat': -22.2410875, u'lng': -54.8213549}}, 'link': u
'http://www.douradosnews.com.br/dourados/velho-conhecido-da-policia-e-preso-apos-sequencia-de-furtos-em-dou
rados'}
2017-09-27 01:08:59 [scrapy.extensions.logstats] INFO: Crawled 24 pages (at 24 pages/min), scraped 9 items
(at 9 items/min)
```

RESULTADOS

www.douradosnews.com.br/dourados/velho-conhecido-da-policia-e-preso-apos-sequencia-de-furtos-em-dourados/1066666/

BUSCA

DOURADOS NEWS
A FONTE DA INFORMAÇÃO

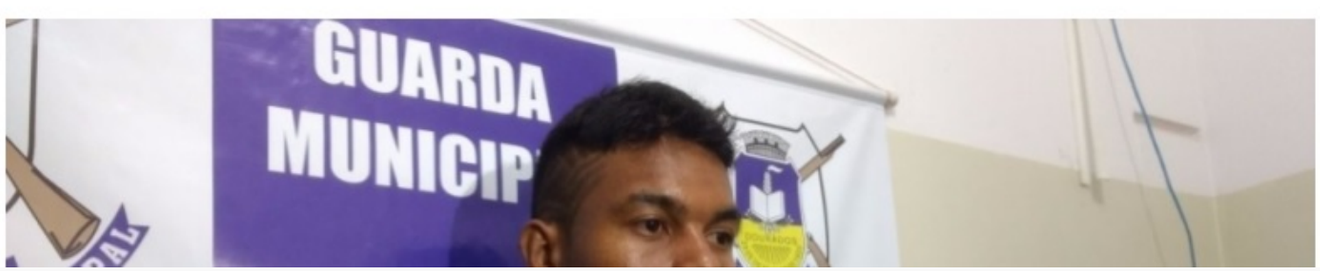
(67) 9860-3221

Página Inicial | Notícias | Colunistas | Fotos & Eventos | Vídeos | Dourados | Região | Polícia | Serviços | Cadernos | Institucional

FLAGRANTE

"Velho conhecido" da polícia é preso após sequência de furtos em Dourados

26 Setembro 2017 - 08h36



ÚLTIMAS NOTÍCIAS

08h13 BENEFÍCIO
UFGD abre nesta segunda-feira inscrições para Auxílio Transporte

CONCLUSÃO

- Analisar os dados e poder estruturá-los de modo que seja possível identificar regiões com ocorrências policiais.
- Alto índice de delitos envolvendo armas de fogo.
- 25% das ocorrências analisadas possuem a palavra “*tiro*” em sua composição.
- De maneira geral, delitos ou infrações de trânsito, somados elevam os índices policiais da região.

CONCLUSÃO

- Com as expressões regulares e o xpath foi possível determinar quais informações seriam buscadas e salvas pelo software.
- O índice de violência do país é gera insegurança na população.
- Essas informações podem auxiliar na identificação das medidas a serem tomadas.
- Outras técnicas, Linguagem de Programação. Aplicações em sistema mobile, web.

BIBLIOGRAFIA

- BENASSI, Maria Virginia Brevilheri. **O gênero "notícia": uma proposta de análise e intervenção.** In: CELLI – COLÓQUIO DE ESTUDOS LINGUÍSTICOS E LITERÁRIOS. 3, 2007, Maringá. Anais... Maringá, 2009, p. 1791-1799.
- COMPUTER HOPE. **Free computer help since 1998.** 2017. Utah – Salt Lake City. Disponível em: <https://www.computerhope.com/jargon/r/regex.htm>. Acessado em: 20 ago. 2017.
- KDNUGGETS. **Best Python modules for data mining.** 2012. Disponível em: <http://www.kdnuggets.com/2012/11/best-python-modules-for-data-mining.html>. Acessado em: 12 set. 2017.
- MELO, Marcelo Damasceno de. **Um Processo de Mineração de Dados para Predição de Níveis Criminais de Áreas Geográficas Urbanas.** 125 p. Dissertação apresentada como requisito parcial para obtenção de grau de Mestre em Ciência da Computação. Universidade Estadual do Ceará. Fortaleza, Ceará. 2010.
- MICROSOFT. **Conceitos de mineração de dados.** 2016. Disponível em: <https://msdn.microsoft.com/pt-br/library/ms174949(v=sql.120).aspx>. Acessado em: 05 ago. 2017.
- MORENO, Ana Carolina. **A importância de checar os fatos.** G1. 2017. Disponível em: <http://g1.globo.com/e-ou-nao-e/noticia/a-importancia-de-checar-os-fatos>. Acessado em: 15 ago. 2017.
- PEREIRA, Jorge Luís. **Análise Preditiva em Sistemas de Informação no Contexto do Big Data.** 2014. 72 p. Trabalho de Conclusão de Curso – Curso de Bacharelado em Sistemas de Informação. Centro Universitário Eurípedes de Marília. Marília, 2014.

BIBLIOGRAFIA

- SARTORI, Ricardo. **MINERAÇÃO DE DADOS DA POLÍCIA MILITAR DE SANTA CATARINA NO MUNICÍPIO DE BALNEÁRIO CAMBORIÚ PARA GERAÇÃO DE INFORMAÇÃO E CONHECIMENTO NA ÁREA DE SEGURANÇA PÚBLICA.** 2012. Universidade do Vale do Itajaí – Centro de Ciências Tecnológicas da Terra e do Mar. Itajaí, Santa Catarina, 2012.
- SILVA, Edson Rosa Gomes da; ROVER, Aires José. **O PROCESSO DE DESCOBERTA DO CONHECIMENTO COMO SUPORTE À ANÁLISE CRIMINAL: MINERANDO DADOS DA SEGURANÇA PÚBLICA DE SANTA CATARINA.** 2011. Programa de Pós-graduação em Engenharia e Gestão do Conhecimento – Universidade Federal de Santa Catarina Brasil. Santa Catarina, 2011.
- TEST AUTOMATION FOR MANUAL TESTERS. **About Learning Selenium Test Automation.** 2014. Vancouver – British Columbia. Disponível em: <http://testable.blogspot.com.br/2014/06/find-web-elements-with-xpath.html>. Acessado em: 10 set. 2017.

OBRIGADO