
Ciência da Computação
Universidade Estadual de Mato Grosso do Sul

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

GUILHERME HENRIQUE VIEIRA PEREIRA

Dr. Rubens Barbosa Filho (Orientador)

DOURADOS – MS

2017

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

Guilherme Henrique Vieira Pereira

Trabalho de Conclusão de Curso, como requisito parcial para a obtenção do grau de Bacharel no curso de Ciência da Computação, na Área de Ciências Exatas e da Terra, da Universidade Estadual de Mato Grosso do Sul, defendido por Guilherme Henrique Vieira Pereira e aprovado pela Banca Examinadora.

Dourados, 27 de outubro de 2017.

Prof. Dr. Rubens Barbosa Filho (Orientador)

Curso de Ciência da Computação
Universidade Estadual de Mato Grosso do Sul

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

Guilherme Henrique Vieira Pereira

Outubro de 2017

Banca Examinadora:

Prof. Dr. Rubens Barbosa Filho (Orientador)
Área de Computação – UEMS

Prof. Msc. André Chastel Lima
Área de Computação – UEMS

Prof. Dr. Osvaldo Vargas Jacques
Área de Computação – UEMS

Dedico este trabalho a minha mãe Andréa Carla da Silva Vieira, batalhadora e minha heroína, que me deu apoio durante todos esses anos, sempre me incentivando nas horas mais difíceis.

E ao meu pai Luis Carlos Domingues Pereira que me deu bons conselhos, para meu crescimento pessoal e profissional, sempre muito importantes.

“In God we trust, all the others must bring data.”

W. Edwards Deming

“Don’t believe in anything you read on the net. Except this. Well, including this. I suppose

Douglas Adams

AGRADECIMENTOS

Primeiramente a Deus que permitiu o desenvolvimento deste trabalho e que tudo isso acontecesse, ao longo da minha vida, e não somente nestes anos na academia, mas que em todos os momentos é o meu maior mestre.

A Universidade Estadual de Mato Grosso do Sul, pela oportunidade de fazer o curso.

Ao meu professor Orientador, Prof. Dr. Rubens Barbosa Filho, pela orientação, apoio, confiança, pelo empenho dedicado a este trabalho, pela paciência, por suas correções, incentivos e pela oportunidade de elaboração deste trabalho.

Aos meus amigos pelo apoio em todas as etapas.

E a todos que direta ou indiretamente fizeram parte da minha formação, o meu muito obrigado.

RESUMO

Há uma grande necessidade de mais segurança no país de modo geral, e os leitores dessas notícias buscam identificar as regiões e os números dessas ocorrências para terem um cuidado maior em determinadas regiões e de compreender melhor a cidade que residem, com a intenção de se prevenirem, de certa forma, dessa estatística que só cresce. O propósito deste trabalho tem como objetivo a mineração de páginas de notícias da web, especificamente notícias de cunho policial, e com prioridade em notícias sobre delitos e infrações de trânsito que ocorrem diariamente na cidade de Dourados – MS, e estruturar, no formato .CSV essas informações para que sejam facilmente acessadas e acompanhadas pelos leitores. A necessidade de realização deste surge do fato de haver muitas informações em páginas de notícias da web. O acompanhamento dessas ocorrências é uma tarefa difícil, o que leva à falta de informações padronizadas e de difícil acompanhamento pela população. Por exemplo, pode-se citar a difícil identificação de regiões críticas e com alto índice de acidentes de trânsito na cidade de Dourados – MS. A Mineração de Dados foi a tecnologia utilizada porque é uma tecnologia que permite tomar decisões melhores e em menor tempo, aumento na chance de acertar padrões e é ela que permite obter o conteúdo para ser analisado. Este tipo de tecnologia permite estruturar e identificar essas regiões, possibilitando apontar o local da ocorrência com informações de latitude e longitude, bairro e nome da rua. Os resultados apresentam informações simples e atômicas, estruturadas em quando, onde, o que e o link com a fonte da notícia. Os resultados obtidos com os testes realizados permitem, de antemão, concluir que o uso desta tecnologia contribui para uma organização e apresentação mais adequada dos dados pesquisados; Contribuindo, desta forma, para um maior esclarecimento de quem busca este tipo de informação. Com essas informações, é possível também realizar cálculos estatísticos para identificar regiões com alto índice de riscos a população, locais com muitos acidentes que talvez estejam com pouca sinalização, determinar quais as ocorrências que possuem mais frequência e também conscientizar a população.

Palavras-Chave: Mineração de Dados, Notícias Web, Informação Policial, Scrapy

ABSTRACT

There is a great need for more security in our country, and the news readers seek to identify these occurrence regions and numbers to be more careful in certain regions and to better comprehend the city, intended to preventing, in a way, from this growing statistic. Work purpose is news web pages data mining, specifically police news, with priority to crimes and traffic infractions news in Dourados City, Mato Grosso do Sul State, structuring the information in .CSV format to be easily accessed and accompanied by the news readers. The need to perform this work arises from big amount of web news page informations. Accompany this occurrences is a hard task, which leads to lack of standardized information and difficult accompaniment by population. By example, is possible quote the hard identification of critical regions and with traffic accidents high rate in Dourados City, Mato Grosso do Sul State. Data mining technology was utilized because allows better time-consuming decisions in less time, an increase of right standards choice chance and allows to obtain analyse content. This kind of technology allows to structure and identify these regions, making possible pointing the occurrence location with latitude and longitude, neighbourhood and street name informations. Results show simple and atomic informations, structured by where, when, what and the news font link. Tests results allows, beforehand, to conclude that this technology use contributes to a researched data more adequate organization and presentation; Contributing, this way, to a bigger clarification for those who search this kind of information. With these information, is possible to statistics calculate to identify high risk index regions to population, many accidents places with maybe poor signalization, to determine which the most frequent occurrences and also raise awareness among population.

Keywords: Data Mining, Web News, Police Information, Web Scraping.

SUMÁRIO

1. INTRODUÇÃO.....	18
1.1 – Objetivos.....	19
1.1.1 – Objetivo Geral.....	19
1.1.2 – Objetivos Específicos.....	19
2. REFERENCIAL TEÓRICO.....	21
2.1 – Métodos de Mineração de Dados.....	24
2.1.1 – Mineração de Itens Frequentes.....	24
2.1.2 – Árvores de Decisão.....	25
Fonte: Própria.....	25
2.1.3 – Raciocínio Baseado em Casos.....	26
Fonte: Grupo de Sistemas Inteligentes.....	26
2.1.4 – Redes Neurais Artificiais.....	27
3. METODOLOGIA.....	28
3.1 – Mineração de Dados.....	28
3.2 – Python.....	29
3.3 – Scrapy.....	30
3.4 – Scrapinghub – Servidor de Web Scraping na Nuvem.....	31
3.5 – Regular Expression.....	32
3.6 – XPath.....	33
3.7 – Processo de Implementação.....	35
4. ANÁLISE E INTERPRETAÇÃO DOS RESULTADOS.....	39
5. CONCLUSÃO.....	47
REFERÊNCIAS.....	49

1. INTRODUÇÃO

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). Entretanto, esse volume de informações pode deixar de esclarecer as notícias para quem estiver lendo, caso essas não estejam bem classificadas, isto é, caso não haja um percentual satisfatório de informações relevantes que possam gerar informações adicionais.

Tratando especificamente sobre informações voltadas para o espaço policial, essa abrangência, ou a falta de foco, da apresentação dessas informações, torna difícil o acompanhamento e a identificação de determinadas informações que permitam retirar do texto dados que possam ser utilizados em compilações estatísticas. Com isso, os leitores ficam sem embasamento para poderem inferir sobre situações relevantes, como por exemplo, dados do bairro em que residem.

Diariamente são publicadas, aproximadamente, cerca de 3 mil notícias em mais de 80 páginas de notícias nacionais (MORENO, 2017). Esse excesso de informação, gerada praticamente a todo instante, representa uma mudança enorme no cotidiano da sociedade, e que segundo Braga (2010, p. 1) representa um grande volume notícias e pouco tempo.

Uma pergunta que se faz necessária é: Um programa que permita acessar e visualizar informações sobre delitos e infrações de trânsito, seria viável? Essa resposta será positiva ao constatar que as pessoas utilizam esse sistema com o objetivo principal de se manterem atualizadas sobre delitos e infrações de trânsito, conhecendo melhor as regiões da cidade que apresentam essas ocorrências.

Um *software* que permita o acesso a estatísticas e informações de cunho policial é imprescindível para o gerenciamento e acompanhamento da segurança pública de maneira geral, tanto para quem atua na área quanto para um leitor comum conhecerem locais mais críticos de sua cidade. Esses dados serão então analisados e apresentados de maneira direta, visando a resolução de problemas específicos de determinadas regiões, seja através de uma intervenção policial na região ou mesmo civis tomando mais cuidado, sendo mais atenciosos nessas regiões.

Softwares com essas características podem gerar dados estatísticos com aplicações no contexto policial, e apresentar análises das informações. Através desse processo foi possível que as informações fossem gerenciadas externamente, de maneira bem estruturada e com informações diretas. Por meio de testes, foi possível identificar facilmente que a maioria das informações estavam realmente relacionadas as notícias de cunho policial, sobre acidentes de trânsito, assaltos, furtos, entre outras e os locais de cada ocorrência, quando citada na notícia.

Este trabalho apresenta as informações divididas em capítulos. No capítulo 1 (um) tratamos da introdução, apresentando os objetivos deste trabalho; O capítulo 2 (dois) aborda o referencial teórico, onde é descrita a função da Mineração de Dados e um pouco de suas características, com ênfase no processo utilizado para este trabalho; No capítulo 3 (três) é descrita a metodologia utilizada no trabalho, além da apresentação dos materiais e tecnologias utilizadas no desenvolvimento do *software*; O capítulo 4 (quatro) é reservado para a análise e apresentação dos resultados, alcançados por meio das técnicas utilizadas; No capítulo 5 (cinco) é apresentada a conclusão dos estudos e; E o capítulo 6 (seis) é voltado às referências bibliográficas.

1.1 – Objetivos

1.1.1 – Objetivo Geral

O objetivo deste trabalho é desenvolver um *software* minerador de dados com aplicações em páginas web, com foco na região de Dourados – MS e que busque informações com cunho policial.

1.1.2 – Objetivos Específicos

- Estudar métodos de mineração de dados.
- Analisar técnicas de *parsing*¹ utilizadas para a aquisição de informações em sites de notícias.

¹ A análise sintática, ou *parsing*, é o núcleo principal do *front-end* de um compilador. O parser é o responsável pelo controle de fluxo do compilador através do código fonte em compilação. É ele quem ativa o léxico para que este lhe retorne *tokens* que façam sentido dentro da gramática especificada pela linguagem e também ativa a análise semântica e a geração de código. (MACARENO JR., [201-?])

- Desenvolver um sistema minerador que obtenha informações de categoria policial de páginas web.
- Realizar testes em vários sites de notícias, a fim de comprovar a eficiência e veracidade do sistema.

2. REFERENCIAL TEÓRICO

A Mineração de Dados é um conjunto de técnicas e processos que são voltados para a busca de padrões, relações e exploração de grandes quantidades de dados em banco de dados ou em páginas *web* (MICROSOFT, 2016). Ocasionalmente, com o crescimento de informações geradas diariamente na *web*, a produção dessa grande quantidade de informação causa um problema de organização que pode ser solucionado utilizando a técnica de mineração de dados, para então estruturar e organizar essas informações.

Um dos principais objetivos, de quem utiliza a técnica de mineração de dados, é a de analisar e explorar os dados em busca de tendências, relações entre as informações e com isso prever comportamentos futuros, onde seja possível até mesmo realizar projetos de negócios, com fins lucrativos, para organizações.

O trabalho desenvolvido por Pereira utiliza a Mineração de Dados com o intuito de encontrar padrões, onde possam prever o IDH (Índice de Desenvolvimento Humano) do Brasil, utilizando a técnica de regressão linear (PEREIRA, 2014). Neste trabalho, é utilizada a Mineração de Dados com o intuito de disseminar o conhecimento, beneficiando a sociedade de maneira geral, não objetivando a busca por padrões, mas sim a análise dos dados para gerar informações estatísticas que beneficiem a sociedade.

Existem várias maneiras de obtenção de dados por meio da mineração de dados, e entre elas, a Mineração Textual que está diretamente relacionada a obtenção de informações do tipo texto, como por exemplo, as páginas de notícias em geral e *feed RSS*².

O trabalho, intitulado Text Mining no Aperfeiçoamento de Consultas e Definição de Contextos de uma Central de Notícias Baseada em RSS, aborda essa técnica e utiliza a Análise Semântica e técnicas do processamento de Linguagem Natural (PASSARIN, 2005).

Esse método, entretanto, não realiza uma varredura em diversos sites, pois o *feed RSS* é sua principal e geralmente a única fonte de obtenção das informações, ou seja, apenas esta página será frequentemente assistida com a finalidade de obter as

²A sigla RSS tem mais de um significado. Alguns a definem como RDF Site Summary, outros a denominam Really Simple Syndication. Há ainda os que a entendem como Rich Site Summary. RSS é um padrão desenvolvido em linguagem XML que permite aos responsáveis por sites e blogs divulgarem notícias ou novidades destes. Para isso, o link e o resumo daquela notícia (ou a notícia na íntegra) é armazenado em um arquivo de extensão .xml, .rss ou .rdf, além de outros formatos. Este arquivo é conhecido como feed ou feed RSS. (ALECRIM, 2005)

informações.

Este trabalho utiliza a Mineração de Dados de uma maneira mais abrangente, pois é possível vasculhar inúmeras páginas de notícias, que possuam diferentes formatos de desenvolvimento, além de também ser possível analisar as páginas de *feed* de notícias.

Aumentando a quantidade de informações a serem estruturadas, analisadas e quantificadas. Aliado a isso, é utilizado um autômato de expressões regulares que tem enfoque mais direto, sem ter que criar regras em que leva muito tempo para ensinar computadores a entender e analisar a semântica dos textos.

Ao levarmos em consideração que o Brasil encontra-se na rota de contrabando e tráfico de drogas internacional, identifica-se que outros crimes estão diretamente relacionados a estes dois, como furtos, assassinatos, acidentes de trânsito, devido as fugas e perseguições policiais, e com isso as taxas de criminalidades tendem a aumentar.

Por esse motivo, a busca por informações que possam prever e determinar regiões dessas ocorrências, tem como propósito de auxiliar na questão de segurança pública (MELO, 2010). Melo busca mapear regiões para identificar níveis criminais de determinadas áreas geográficas, utilizando o aprendizado de máquina para auxiliar no processo, onde trará resultados sem detalhes e sem especificidade.

Neste trabalho, entretanto, buscamos determinar não apenas dados estatísticos para ter informações quantitativas relacionadas as notícias de cunho policial, mas também de conscientizar as pessoas das regiões que concentram grande número dessas ocorrências e auxiliar na gestão da segurança pública.

O Scrapy é um framework do Python utilizado para varrer a web em busca das informações que possam alimentar a *software* minerador de dados, apresentado com mais detalhes na sessão seguinte de mesmo nome (3.3 – Scrapy).

Utilizado para vasculhar a web, em busca de padrões e informações que sejam muito específicas, como por exemplo, informativo de eventos acadêmicos e até para buscar padrões na área econômica. Carilo e Silva utilizam o Scrapy com o intuito de identificar eventos dentro de uma instituição pública, beneficiando todos os acadêmicos de maneira geral.

Este trabalho procura abranger a população de maneira geral, não apenas informativo, mas que também sirva de orientação para tomadas de decisões em seu cotidiano (CARILO e SILVA, 2015).

A metodologia utilizada neste trabalho busca analisar as ocorrências de modo que

as informações sejam processadas para gerar dados que possam englobar o maior número de delitos e infrações de trânsito possíveis, para que sua credibilidade seja um fato e não apenas uma especulação e com isso transformar os dados em informações consistentes e úteis para a população.

No trabalho, O Processo de Descoberta do Conhecimento Como Suporte À Análise Criminal: Minerando Dados Da Segurança Pública De Santa Catarina, desenvolvido por Edson e Aires, os autores utilizam apenas informações de boletins de ocorrências sobre homicídios dolosos (quando o ator assume o risco de causar a morte a vítima), considerando a percepção da sociedade sobre os índices de homicídios por conta da agressividade e o choque que causam esse delito (SILVA e ROVER, 2011).

O trabalho desenvolvido por Sartori tem como objetivo, utilizando a mineração de dados em dados da Polícia Militar de Balneário Camboriú, auxiliar na tomada de decisão e de aplicação de recursos e efetivo na atividade de segurança pública. De maneira que as informações devem ser acessadas através de uma base de dados da própria Polícia Militar de Balneário Camboriú, ou seja, são informações sigilosas e que precisam da aprovação dos órgãos de segurança pública para que sejam acessadas (SARTORI, 2012).

Neste trabalho as informações são obtidas das páginas de notícias da cidade de Dourados – MS que podem ser acessadas sem restrições, o que torna o acesso menos burocrático aos dados e ainda assim possuem um alto nível de confiabilidade.

2.1 – Métodos de Mineração de Dados

Há dois métodos de Mineração de Dados tradicionais, supervisionado (preditivo) e o não-supervisionado (descritivo). Apesar dessa divisão, entretanto, existem métodos preditivos que podem ser descritivos e o mesmo vale para o contrário.

A principal diferença entre esses métodos é que o não-supervisionado não precisa de um atributo alvo para realizar o processo de mineração. Um exemplo de utilização do método não-supervisionado é a tarefa de agrupamento e associação. Para o método supervisionado, que trata de um conjunto de dados já definidos para serem analisados, tem como exemplo de sua utilização as classificações de padrões num determinado conjunto de dados (CAMILO e SILVA, 2009).

A seguir serão apresentados, com uma breve descrição, alguns dos principais métodos utilizados na mineração de dados.

2.1.1 – Mineração de Itens Frequentes

Naturalmente dividida em duas etapas, onde a primeira gera um conjunto de itens frequentes. Onde são determinados valores mínimos de frequências para os itens. Após esse processo, são geradas regras de associação através da mineração do conjunto de itens. A partir dessas regras é possível determinar um percentual de dados que se enquadram a cada uma. Um exemplo de utilização desse método, e um dos mais famosos é o algoritmo *Apriori* (AGRAWAL, 1993).

Figura 1 – Algoritmo Apriori

O Algoritmo Apriori

```

 $C_k$ : itemset candidato de tamanho k
 $L_k$ : itemset frequente de tamanho k
Pseudo-código:
1:  $L_1 = \{\text{itens frequentes de tamanho } 1\}$ 
2: for ( $k = 1$ ;  $L_k \neq \emptyset$ ;  $k++$ ) do begin
3:    $C_{k+1} = \text{apriori-gen}(L_k)$  //Gerar itemsets candidatos
4:   for each transaction  $t \in T$  do
5:      $C_t = \text{subset}(C_k, t)$  //Identificar todos os candidatos que
                                     pertencem a t
6:   for each itemset candidato do
7:      $\sigma(c) = \sigma(c) + 1$  // Incrementar o contador de suporte
8:    $L_{k+1} = \text{candidatos em } C_{k+1} \text{ com minsup}$  // Extrair os
                                     k+1-itemsets frequentes
9: return  $\cup_k L_k$ 

```

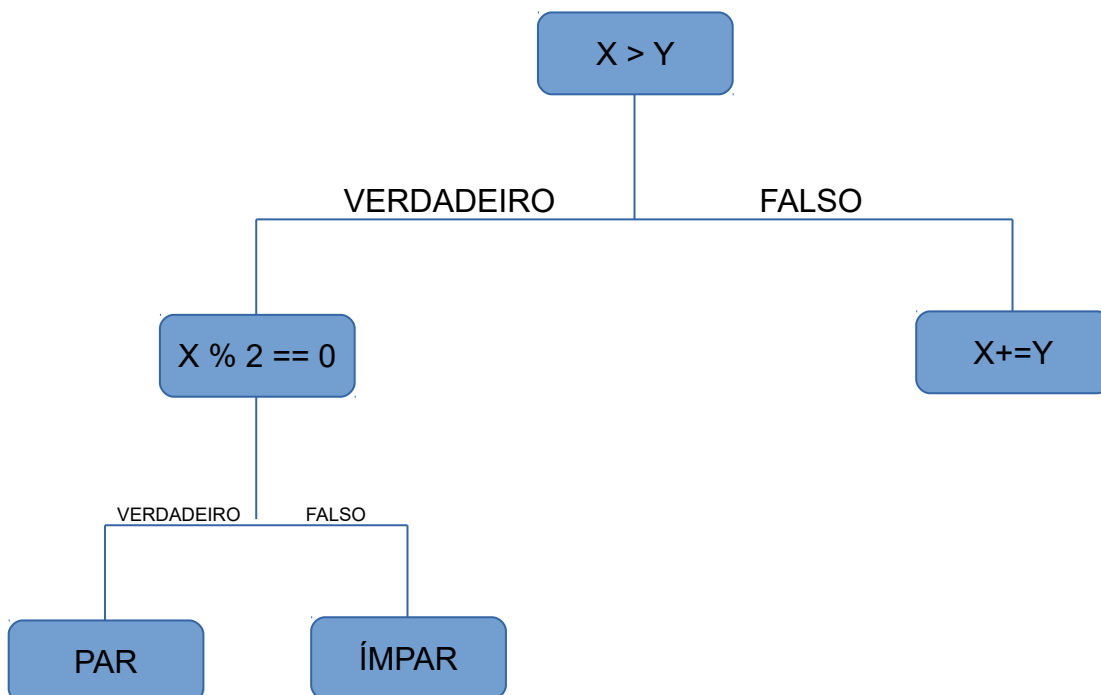
Fonte: Nunes e Guimarães, 2013.

2.1.2 – Árvores de Decisão

Este método funciona como um fluxograma em forma de árvore. Cada nó não folha representa um teste, como por exemplo a comparação entre duas variáveis $X > Y$. As ligações deste nó representam possíveis valores do teste no que o executou, e as folhas indicam a categoria que o registro pertence. Após a montagem da árvore basta seguir o fluxo na árvore, partindo do nó raiz até a folha, para a classificação de um novo registro.

A estrutura formada pelas árvores de decisões são consideravelmente fáceis de serem convertidas em Regras de Classificação. Apesar de ser uma técnica extremamente poderosa, é necessária uma análise detalhada dos dados que serão usados para garantir bons resultados. Quinlan apresenta diversas técnicas para reduzir a complexidade das árvores de decisão geradas (QUINLAN, 1986).

Figura 2 – Árvore de Decisão Simples, Editada pelo Autor.

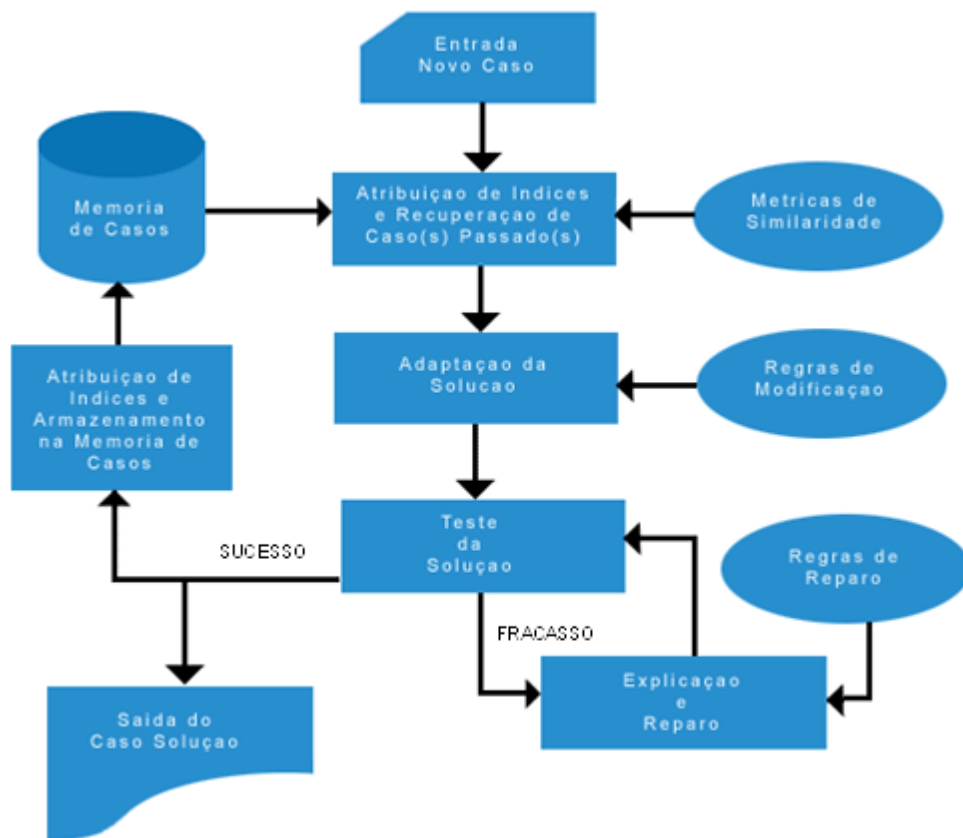


Fonte: Própria

2.1.3 – Raciocínio Baseado em Casos

Baseado no método do vizinho mais próximo, o raciocínio baseado em casos busca esse vizinho e combina os valores do mesmo para atribuir valores de classificação ou previsão. Um exemplo de utilização desse método é a de classificação e segmentação de dados. Onde os dados devem ser escolhidos, postos em funções de distância, determinar o número máximo de vizinhos e, por fim, determinar a função de combinação.

Figura 3 – Arquitetura de um Sistema a Base de Casos



Fonte: Grupo de Sistemas Inteligentes.

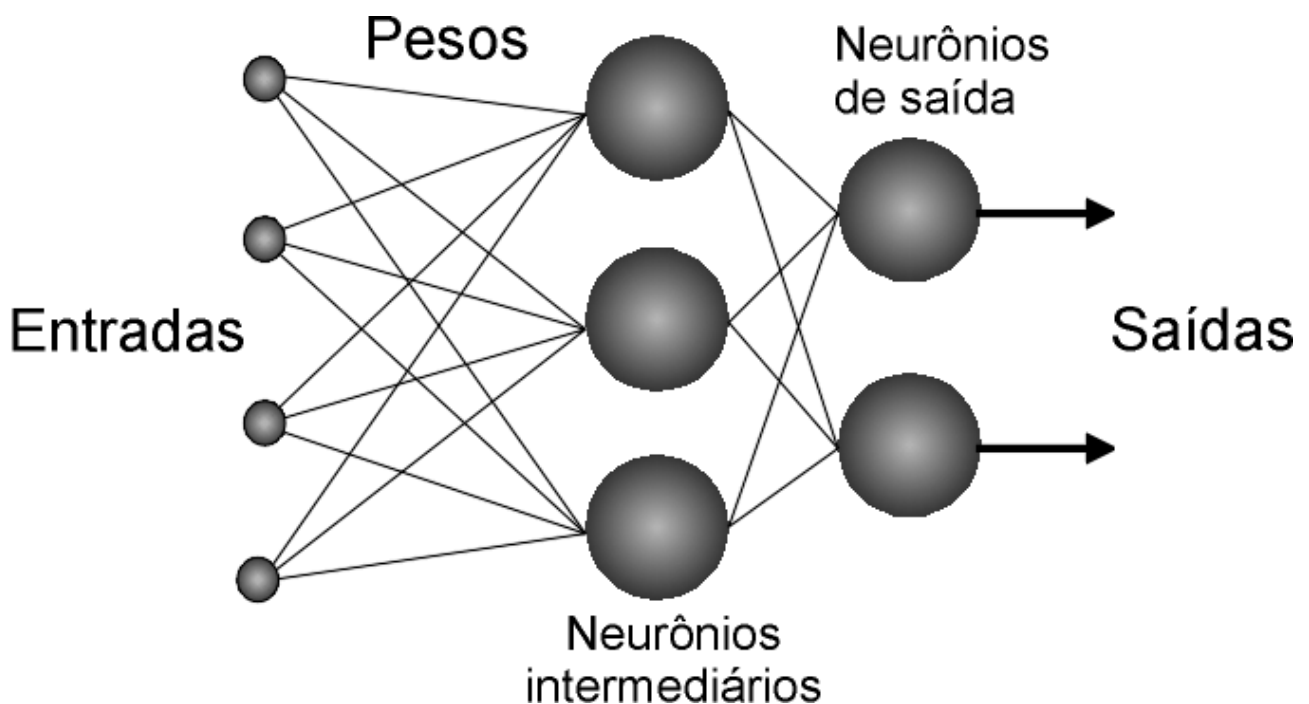
2.1.4 – Redes Neurais Artificiais

Redes neurais fazem parte de uma classe especial de sistemas que tem como objetivo modelar sistemas baseados no funcionamento do cérebro humano e que são desenvolvidos com neurônios artificiais interconectados de maneira similar aos neurônios do cérebro humano (GOEBEL e GRUENWALD, 1999).

“Como no cérebro humano, a intensidade de interconexões dos neurônios pode alterar (ou ser alterada por algoritmo de aprendizagem) em resposta a um estímulo ou uma saída obtida que permite a rede aprender” (GOEBEL e GRUENWALD, 1999, p. 23).

A técnica de redes neurais é apropriada às seguintes tarefas: classificação, estimativa e segmentação (CÔRTEZ, PORCARO e LIFSCHITZ, 2002).

Figura 4 – Redes Neural Artificial de 2 camadas, 4 entradas e 2 saídas



Fonte: Cérebro e Mente, 2002.

3. METODOLOGIA

Foi necessário para este trabalho realizar estudos aprofundados sobre técnicas e tecnologias utilizadas para o processo de implementação de um *software* minerador. E através dos estudos feitos foi possível determinar as ferramentas mais viáveis para o desenvolvimento do *software* minerador que serão, posteriormente, apresentadas nesta mesma sessão com mais detalhes de utilização.

Os recursos tecnológicos utilizados para o desenvolvimento do *software* minerador são: *softwares Open Source*³ ou a versão gratuita de ferramentas disponibilizadas na web.

Linguagem de Programação, Tecnologias e Ferramentas utilizadas:

- Python 2.7.12
- Scrapy 1.4
- Scrapinghub
- RegEx (Regular Expression)
- Xpath (XML Path Language)

3.1 – Mineração de Dados

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.

Na análise descritiva o principal objetivo é visualizar os dados e extrair as

³ Trata-se de *software* de código aberto, isto é, em que o código fonte está acessível para inspeção e é passível de manipulação; adicionalmente, e na maior parte dos casos, surge também com licenciamento livre de encargos. É a frequente associação destes dois aspectos, que não são típicos do *software* comercial, que confere a qualificação comum de 'livre' ao *software* de código aberto. (CORDEIRO, 2010)

informações existentes em uma base de dados ou, no caso deste trabalho, para extrair as informações de cunho policial de páginas web.

3.2 – Python

A linguagem de programação Python, foi escolhida para ser utilizada na codificação do *software* por ser de alto nível e se tratar de uma linguagem mundialmente conhecida pelo fácil manuseio, pelo suporte que a comunidade Python oferece e por possuir *frameworks*⁴ específicos para a implementação de *softwares* com a finalidade de realizar mineração de dados.

Python é uma linguagem de programação que data de 1991, desenvolvida por Guido van Rossum. Seus principais objetivos ao desenvolvê-la eram: produtividade e legibilidade. De maneira direta, Python foi desenvolvido para produzir código bom e fácil para dar suporte, de maneira rápida e prática (ANUSKIEWICKZ, 2016). Dos objetivos da linguagem, os seguintes são alguns que a compõem:

- Baixo uso de caracteres especiais, o que torna a linguagem muito parecida com pseudocódigo executável;
- O uso de indentação para marcar blocos;
- Quase nenhum uso de palavras-chave voltadas para a compilação;
- Coletor de lixo para gerenciar automaticamente o uso da memória.

Utilizada em todas as etapas do trabalho, ou seja, na requisição de páginas da web, no processo de análise das informações e na armazenagem das informações, a linguagem possui uma comunidade ativa e que está sempre dando suporte a utilização da linguagem e a todas as suas bibliotecas, de tal maneira que auxilia em todo o processo de implementação e desenvolvimento do *software*.

A versão 2.7.12, instalada e empregada neste trabalho, a linguagem Python tem suporte à Orientação a Objetos, que facilita na geração dos objetos que conterão as informações das páginas web mineradas, tais como data de publicação, *link* da página, latitude e longitude do local apresentado na notícia e, por se tratar de informação de

⁴ Um *framework* é uma arquitetura desenvolvida com o objetivo de atingir a máxima reutilização, representada como um conjunto de classes abstratas e concretas, com grande potencial de especialização.

cunho policial, o tipo de infração ou delito contido na notícia.

Python, entretanto, não realiza as tarefas de mineração de dados, apresentadas acima, sozinha e por isso é necessária a utilização de um *framework* para que seja assim possível alcançar o objetivo deste trabalho.

3.3 – Scrapy

Para a realização efetiva do processo de Mineração de Dados, o *framework* Scrapy foi escolhido como principal ferramenta de varredura, ou seja, de busca e coleta das informações com base num padrão de informação enviado para o analisador descritivo do *framework*. Esse processo é conhecido como *Web Scraping*.

Escrito em Python, Scrapy é uma aplicação *framework*, mais famoso e popularmente conhecido, para a realização de extração de dados estruturados, ou seja, *Web Scraping* (KOROBOV, 2017). Desenvolvido e mantido principalmente pela plataforma web Scrapinghub, o Scrapy é um *framework* completo e portanto necessária sua instalação utilização no desenvolvimento do *software* minerador deste trabalho.

Utilizado na versão 1.4, o Scrapy tem uma vasta configuração para o processo de Mineração de Dados, onde é possível, por exemplo, definir a quantidade de páginas para serem analisadas, quantas requisições em paralelo serão executadas, os padrões de expressão regular e caminho dos arquivos HTML das páginas rastreadas, entre muitas outras muitas configurações que podem ser feitas para que o *framework* possa trabalhar com a melhor configuração para uma tarefa específica.

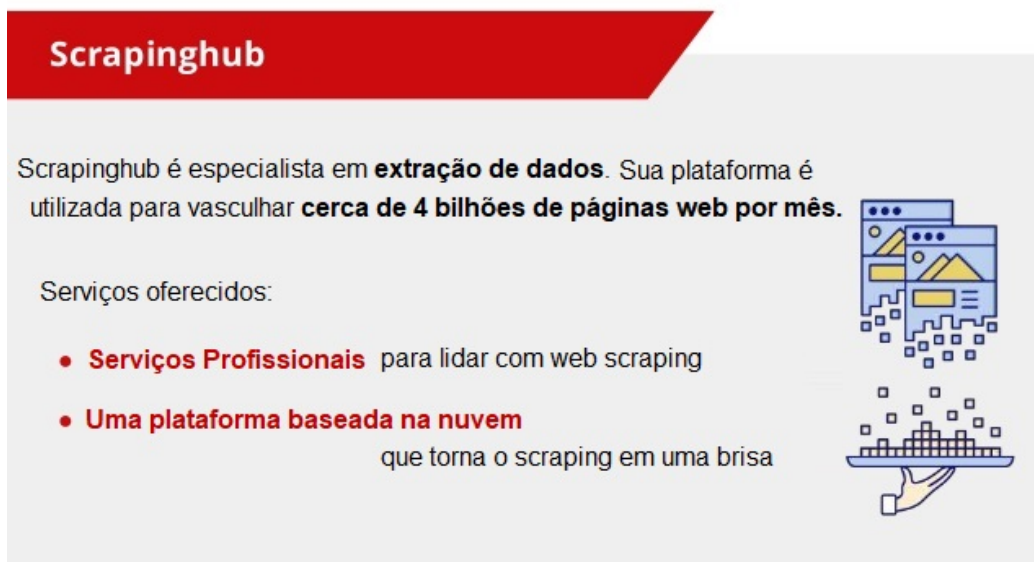
O Scrapy se responsabiliza em rastrear *links* de um ou mais sites, e extrair as informações das páginas utilizando o Python. Sua estrutura dividida em 3 partes: Definir quais *links* seguir, qual a estrutura da informação que deve ser requisitada e, por fim, retornar as informações obtidas (DORNELES, 2014).

As ferramentas utilizadas para implementar os padrões de buscas do *software* minerador, escolhidos para este trabalho são RegEx e XPath, ambos inseridos no código fonte do programa em Python e utilizados pelo *framework* Scrapy. Configurados para localizar *links* de páginas web, endereços físicos (RegEx) encontrados em cada página com as informações mineradas e o conteúdo específico contido num arquivo HTML

escrito em uma Linguagem de Marcação (Xpath).

3.4 – Scrapinghub – Servidor de Web Scraping na Nuvem

Figura 5 – Sobre Scrapinghub.



Scrapinghub

Scrapinghub é especialista em **extração de dados**. Sua plataforma é utilizada para vasculhar **cerca de 4 bilhões de páginas web por mês**.

Serviços oferecidos:

- **Serviços Profissionais** para lidar com web scraping
- **Uma plataforma baseada na nuvem** que torna o scraping em uma brisa

The image contains an illustration of a hand holding a tray with a keyboard, and another illustration of a hand holding a tray with a computer monitor displaying charts and data.

Fonte: Shane Evans, 2016.

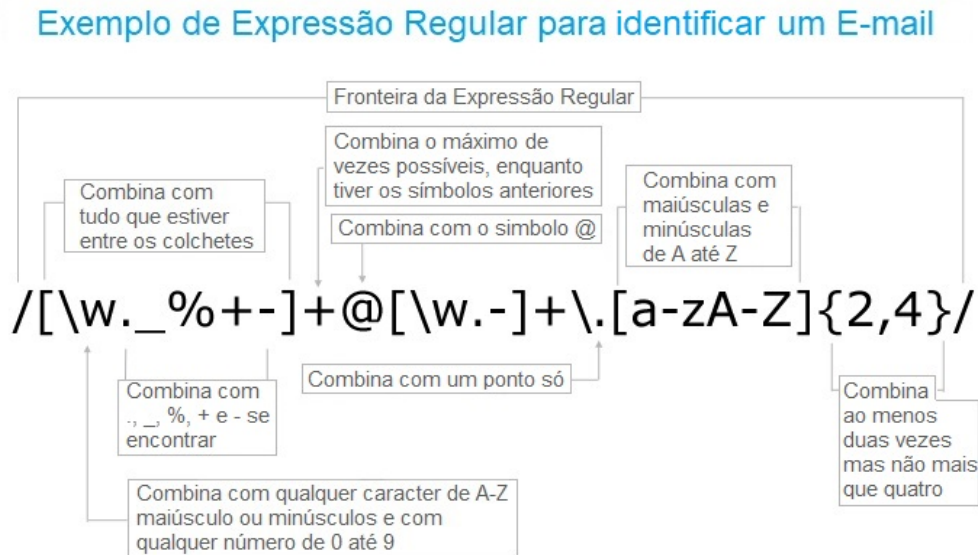
O Scrapy é *framework* mais popular para *Web Scraping*, utilizado em cerca de 4 bilhões de páginas web mensalmente, conforme a Figura 5, e o Scrapinghub é uma plataforma baseada a nuvem, onde é possível inserir os *softwares* de mineração desenvolvidos para serem executados em máquinas com configurações mais avançado do que as configurações de um computador pessoal e numa velocidade e quantidade de rastreadores maiores (SCRAPINGHUB, [201-?]).

A versão gratuita da plataforma Scrapinghub permite apenas um rastreador, que para os testes realizados neste trabalho foi o suficiente para potencializar e agilizar no processo de aquisição dos resultados.

Os resultados minerados, através do servidor, são salvos num arquivo de texto que podem ser visualizados e baixados para serem manipulados externamente. As versões pagas disponibilizam a utilização de rastreadores ilimitados, onde cada rastreador terá um custo. E tais informações podem ser encontrados na página do Scrapinghub com facilidade.

3.5 – Regular Expression

Figura 6 – Exemplo de Expressão Regular para identificar um E-mail.



Fonte: Computer Hope, 2017. Editada pelo autor.

Por Expressão Regular tem-se um conjunto de caracteres que descrevem um padrão a ser buscado.

Naturalmente em formato de texto, armazenado numa variável do tipo *string* num *software* ou *script*⁵, também conhecida como RegEx⁶, pode ser utilizada para encontrar endereços de e-mail, endereços de locais físicos, nomes de pessoas, cidades, uma frase específica, uma sequência numérica e muitas outras estruturas de textos que podem ser encontradas com sua utilização. Conforme apresentado na Figura 6.

Neste trabalho, o uso da Expressão Regular, foi escolhida para buscar padrões de *links* que são rastreados pelo Scrapy, os quais estão sempre relacionados a páginas web de cunho policial, e também para encontrar o endereço (Rua, Avenida, Travessa ou Cruzamento) informado no texto da página web, geralmente, de notícia. Além disso, é utilizado para buscar nomes de ruas.

⁵ Um script descreve uma sequência de comandos e tarefas que alguém deve executar ou, no caso de um computador, interpretar. O exemplo clássico disso são as linguagens para os terminais de comando, seja o shell/bash do Linux ou para o batch do Windows (BERNAL, 2014).

⁶ Expressão Regular. Conjunto de caracteres em formato de texto, que permite criar padrões para gerenciar textos (COMPUTER HOPE, 2017).

3.6 – XPath

XPath é uma linguagem de consulta que trata de localizar caminhos e processar nós em arquivos do tipo XML (Extensible Markup Language). Através de uma estrutura hierárquica lógica presente em arquivos que disponibilizam uma árvore de dados estruturados, onde existem nós gerando seus respectivos nós filhos, o XPath⁷ utiliza a estrutura dessa árvore para obter uma informação de forma direta.

Diferente da RegEX, apresentada anteriormente, o XPath não realiza uma busca por meio de um conjunto de caracteres, mas sim por uma estrutura de nós de uma árvore, obtida por meio de um arquivo de marcação estruturado, ou seja, XML.

Através da estrutura apresentada na Figura 7, é possível visualizar como é formada a estrutura do XPath.

⁷XPath, a XML Path Language, é uma linguagem de consulta (Query Language) para selecionar nós de um document XML. Ademais, XPath pode ser usada para computar valores (por exemplo, caracteres, números ou valores booleanos) do conteúdo de um documento XML. (WIKIPEDIA, 2015)

Figura 7 – Exemplo de uma Estrutura XPath gerada em um arquivo XML

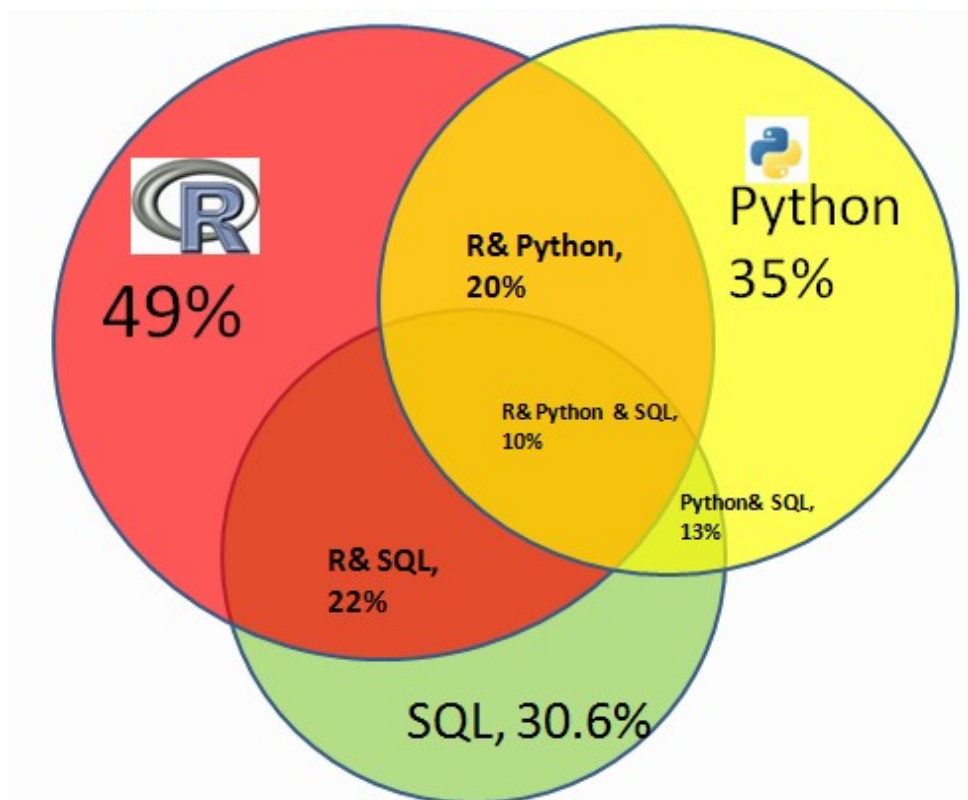


Fonte: Test Automation For Manual Testers, 2014.

3.7 – Processo de Implementação

Para desenvolver o sistema minerador de dados, a linguagem de programação Python foi a melhor opção por sua facilidade de implementação e por ser uma das principais linguagens de programação utilizadas no desenvolvimento de Mineração de Dados, conforme ilustração da Figura 8.

Figura 8 – Linguagens usadas para Análise/Mineração de Dados



Fonte: Kdnuggets, 2014.

Porém, apenas a linguagem de programação Python não foi o suficiente para realizar o método de mineração de dados da web, sendo necessário a busca por *frameworks* que pudessem auxiliar nessa função. Neste caso, o *framework* utilizado para solucionar essa questão foi o Scrapy, por ser *Open Source* e desenvolvido em Python.

Escolhido pelo fato de ser o melhor para o que propõe este trabalho, que é a realização o vasculhamento de várias páginas da web (KDNuggets, 2012). Na página oficial do *framework* (<https://scrapy.org/>) é possível encontrar a documentação e pequenos exemplos de como é o seu funcionamento e sua utilização. As Figuras 9, 10 e 11 ilustram a implementação básica, retirada da documentação oficial do Scrapy, e está

dividida em basicamente três partes que definem o Scrapy.

Na parte 1, apresentada na Figura 9, temos a definição da classe, do nome da *spider* e as urls iniciais (*start_urls*), onde são definidas as urls que formam o(s) ponto(s) de partida do vasculhamento, ou seja, a partir daquela url é possível alcançar todas as demais urls interna da página.

Figura 9 – Editada pelo Autor

```
### 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/',
    ]
```

Fonte: Scrapy at a glance, 2017.

Na parte 2 é realizado o *parse* das informações ou onde as informações são extraídas de fato, e podemos visualizar a utilização do CSS e XPath como padrões de reconhecimento para extração de informações, ou seja, com o CSS o exemplo obtém o texto de uma *tag* do tipo *span* e o XPath obtém o nome do autor desse texto conforme o caminho apresentado no comentário do código da Figura 10.

Figura 10 – Editada pelo Autor

```
### 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(),
        }
```

Fonte: Scrapy at a glance, 2017.

Na parte 3 encontramos uma variável chamada *next_page* que armazena o link para a próxima página a ser vasculhada em busca do padrão definido na parte 2. Ou seja, ele busca todos os caminhos que contenham uma tag *anchor* para um link dentro da própria página. Conforme Figura 11.

Figura 11 – Editada pelo Autor

```
### 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
```

Fonte: Scrapy at a Glance, 2017.

Este trabalho ainda utiliza os recursos da Expressão Regular para remover informações específicas das notícias, como o local da ocorrência, e com essa informação é realizada a busca pela latitude e longitude com a API do Google Maps.

Figura 12 – Editada pelo Autor

```
from googlemaps import GoogleMaps

gmaps = GoogleMaps(api_key)
end = 'Avenida Marcelino Pires, Dourados, MS'
lat, lng = gmaps.address_to_latlng(end)
print lat, lng

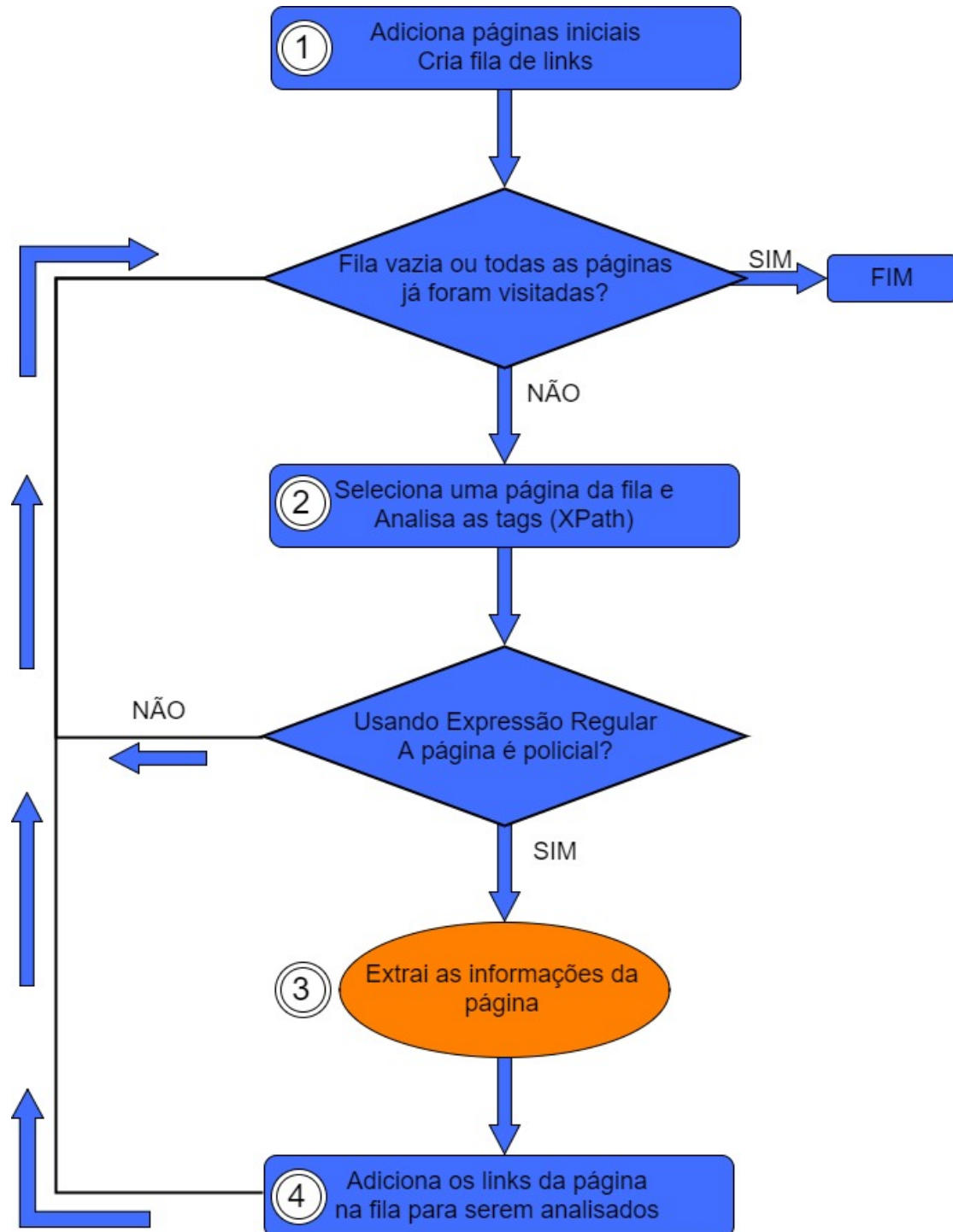
-22.2259387 -54.7951442
```

Fonte: Kleint, 2013.

A Figura 12 é um exemplo básico de utilização da API do Google Maps. Para que funcione corretamente, é necessário gerar uma chave de utilização na página do Google, que permite a busca de um número limitado de endereços (GOOGLE, 2016). Na linha 5 da Figura 11, é onde a função *address_to_latlng* busca pela latitude e longitude. Neste caso, o endereço é inserido manualmente, na linha 4, mas na mineração de dados é passada a variável que contém a localização obtida das notícias (KLEINT, 2013).

Na Figura 13 podemos acompanhar o fluxograma de desenvolvimento do minerador. Cada processo é explicado em detalhes nas subseções anteriores.

Figura 13 – Fluxograma de Desenvolvimento



Fonte: Própria, 2017.

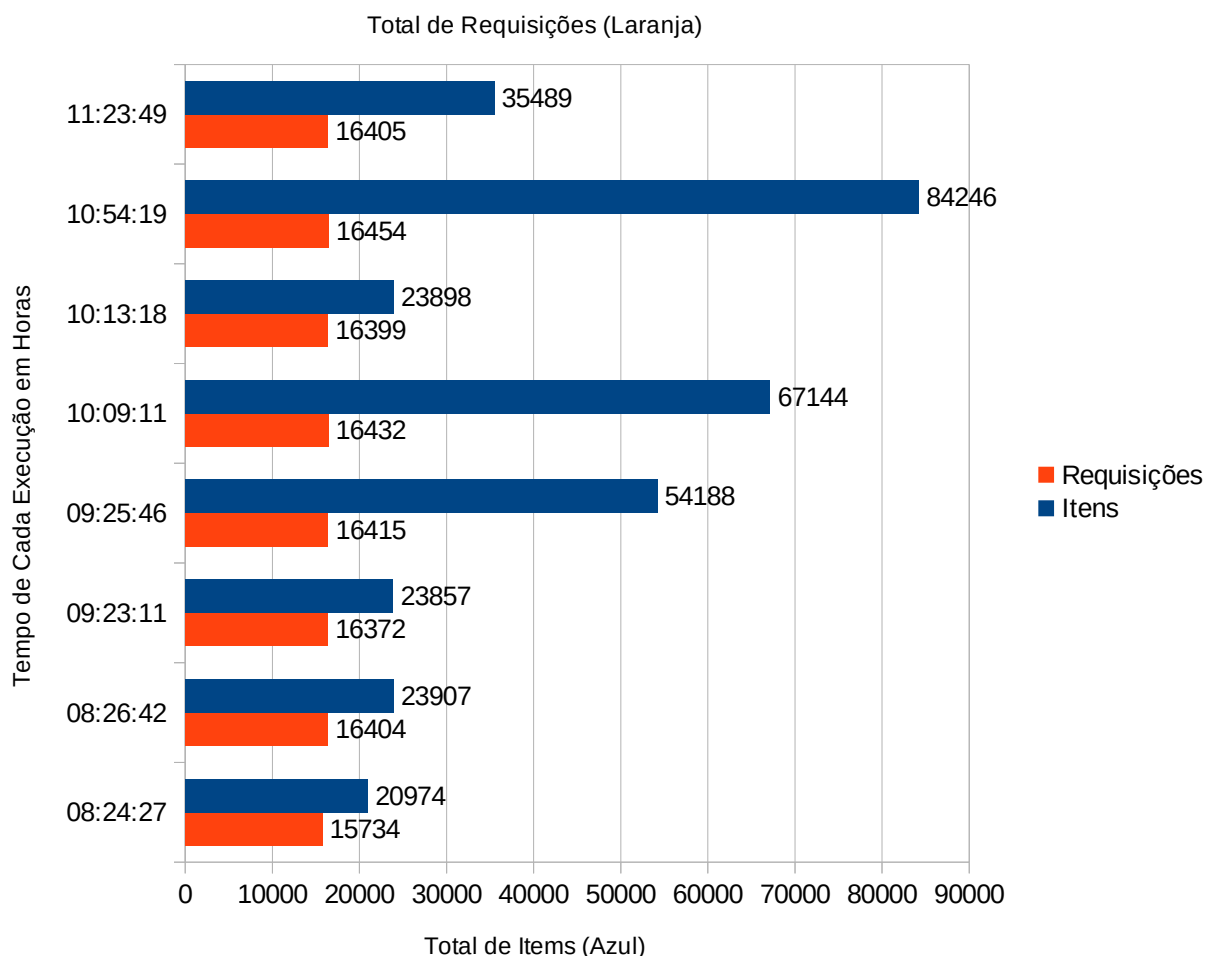
4. ANÁLISE E INTERPRETAÇÃO DOS RESULTADOS

Os procedimentos e técnicas apresentadas anteriormente, foram utilizados em páginas de notícias da cidade de Dourados. E para realizar os testes de obtenção das informações, o minerador foi executado no servidor de Scrapithub.

Para cada teste foi entregue a URL da página inicial. Após isso, todos os links da página eram analisados e em caso de um link que redirecionasse para uma notícia de cunho policial, as informações daquela página são avaliadas e salvas pelo minerador.

O tempo de cada execução varia de acordo com o tamanho da página e sua estrutura de desenvolvimento web, que pode mudar bastante de uma página para outra, além disso, é possível que uma mesma notícia esteja em vários locais e para lidar com essa situação, o Scrapy não realiza a requisição para páginas que já foram requisitadas, por meio do seu arquivo de configuração que já evita a duplicata de links acessados.

Figura 14 – Gráfico com As Requisições e os Itens Minerados no Scrapinghub da página de notícias Dourados News

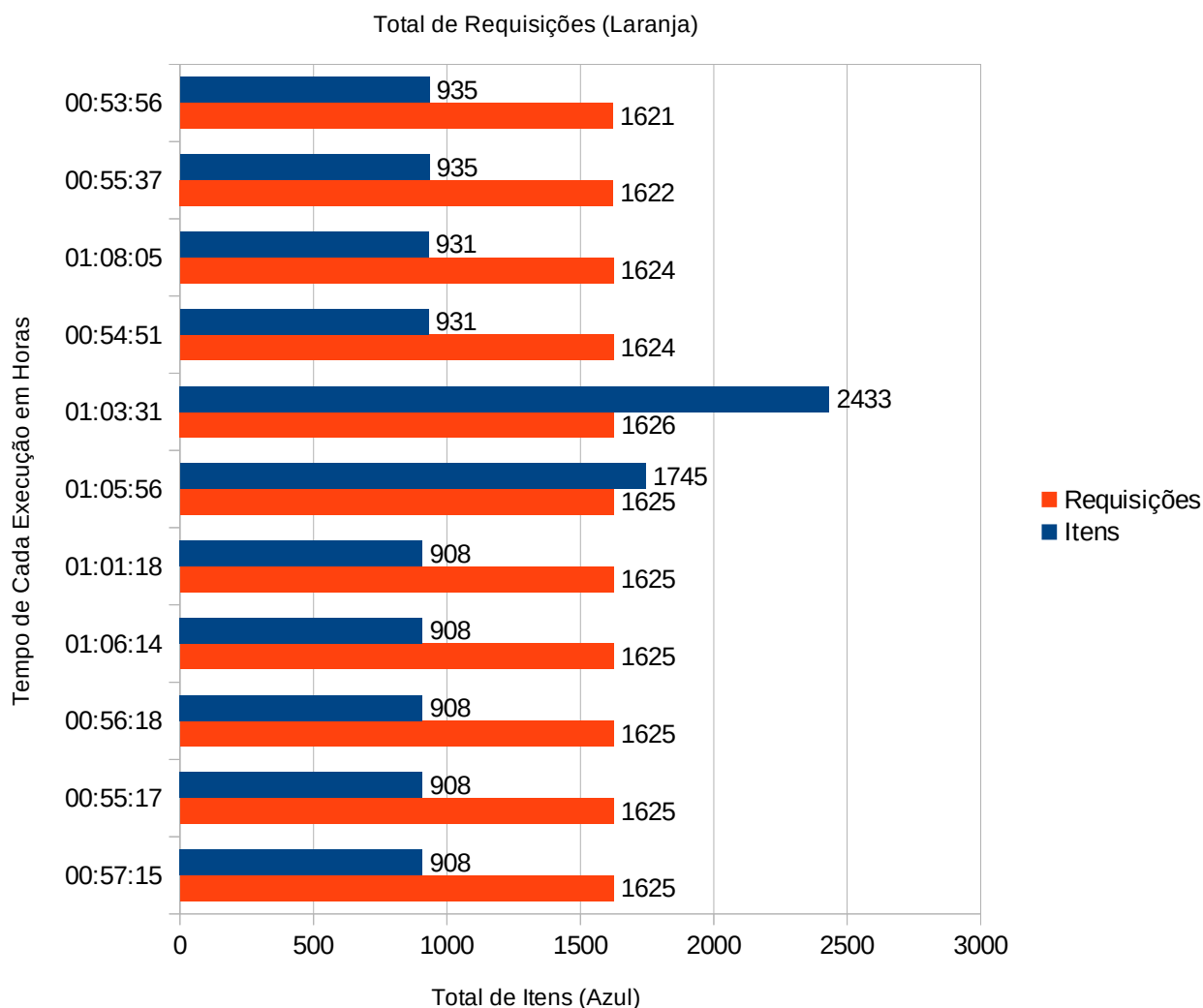


Fonte: Própria, 2017.

Através do gráfico da Figura 14 acima, é possível visualizar que a quantidade de requisições (em laranja) não varia muito em função do tempo de cada execução (Eixo-x) do *software* minerador, apresentado na base de cada gráfico, mas os itens minerados (em azul) têm grande variação.

Esses dados apresentam informações do que se é esperado do *software*, uma quantidade bem próxima de requisições para cada execução, mas um número de itens minerados bem variável devido ao aumento das informações da página de notícias Dourados News e possivelmente a remoção de notícias mais antigas pelos administradores do jornal.

Figura 15 – Gráfico com as Requisições e Itens minerados no Scrapinghub da página de notícias Dourados Agora



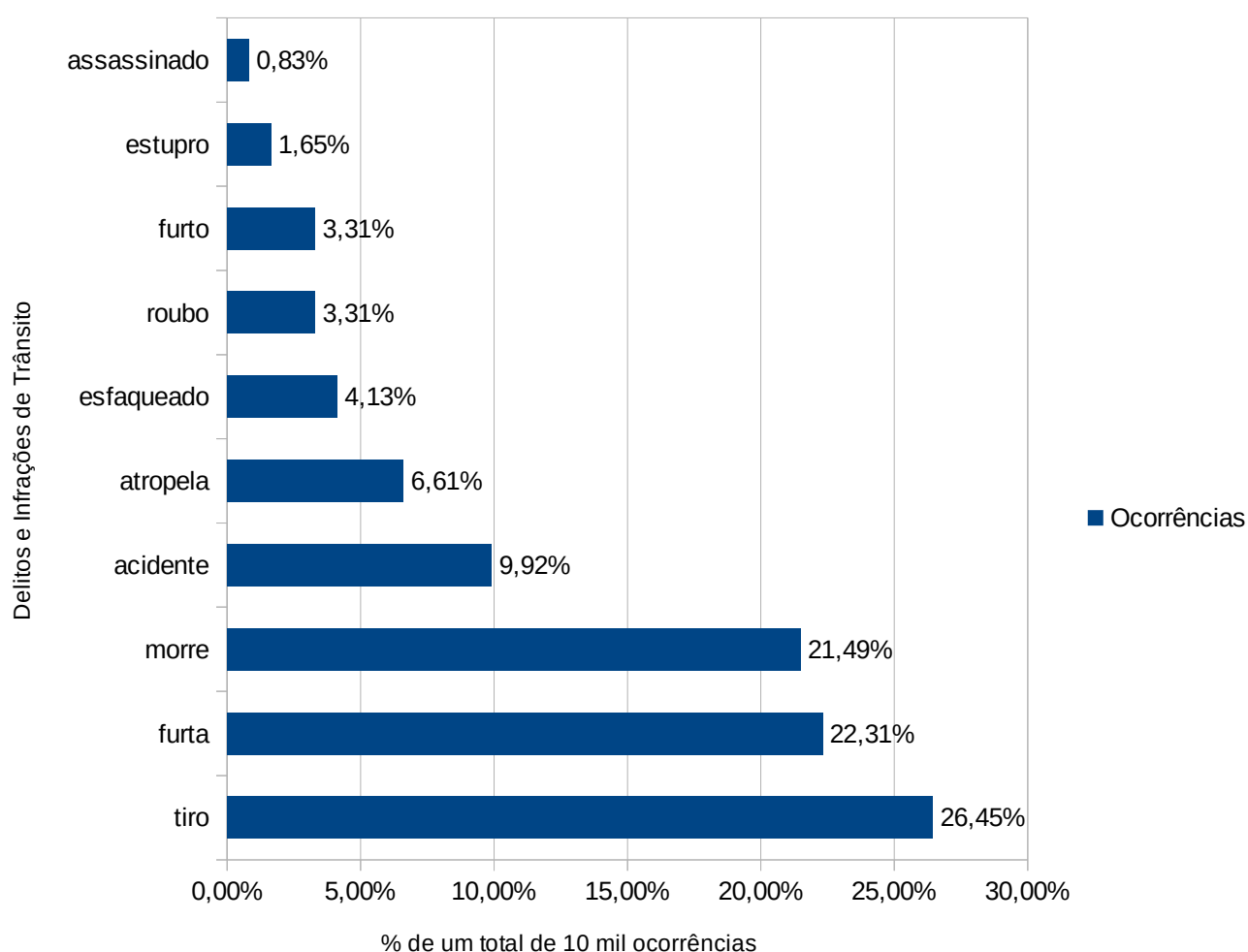
Fonte: Própria, 2017.

A diferença entre a quantidade de itens minerados da página Dourados News (Figura 14) e da página Dourados Agora (Figura 15), onde a quantidade do primeiro alcança picos de até 16 mil requisições e mais de 80 mil itens obtidos se dá devido a estrutura da página de notícias e como cada analista cuida dessas páginas, pois no caso da página Dourados News há uma quantidade maior de páginas e informação produzida diariamente, enquanto na página Dourados Agora a produção de notícias é menor. Isso apenas diz respeito a estrutura de cada página, pois ambas são informativas e possuem o objetivo de informar a população.

Com relação as informações textuais obtidas após as execuções, é possível ainda

assim que hajam páginas duplicadas entre uma execução e outra e também numa mesma execução, pois apesar de lidar com duplicatas, o Scrapy só avalia se o link informado já foi acessado, entretanto, há nessa página de notícias links com diferentes estruturas que redirecionam para o mesmo conteúdo e o Scrapy não está configurado para identificar o conteúdo das páginas, apenas os links informados.

Figura 16 – Gráfico com a porcentagem de cada ocorrência encontrada na página
Dourados News



Fonte: Própria, 2017.

As informações apresentadas no Gráfico da Figura 16 foram retiradas de um total de aproximadamente 10 mil itens minerados, uma pequena amostra do total de itens apresentados na Figura 14.

Por possuir uma maior quantidade de itens obtidos e assim uma amostra maior para trabalhar, utilizamos apenas as informações da página Dourados News para determinar as informações da Figura 16.

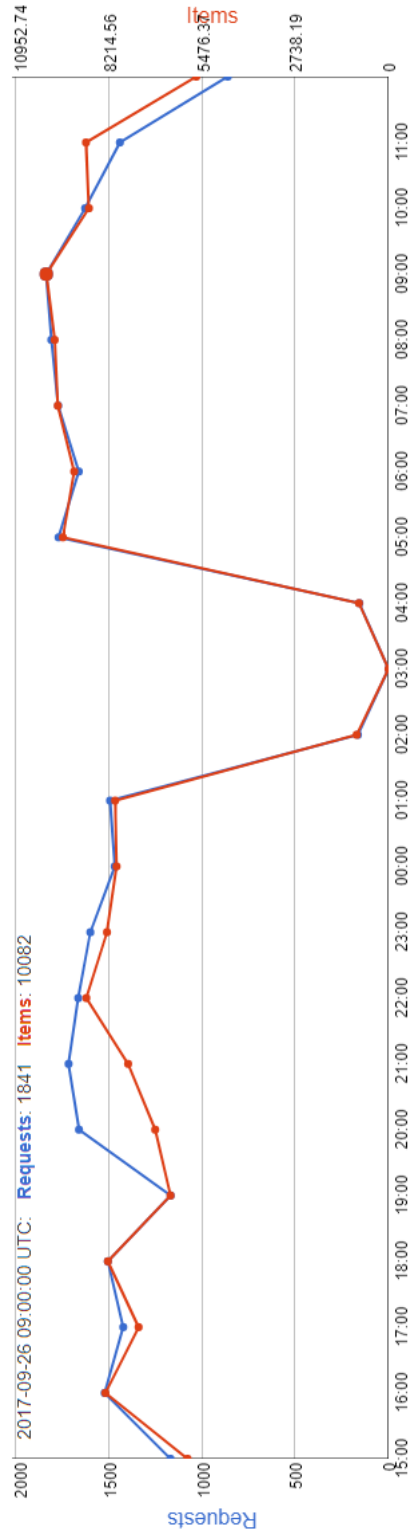
As principais ocorrências encontradas, em porcentagem, estão neste gráfico, onde é possível visualizar que mais de 25% das ocorrências correspondente ao delito que pertence ao termo abstrato *tiro*. Deve-se levar em consideração que foram utilizadas palavras abstratas para comparação entre as ocorrências e obter os delitos e também as infrações de trânsito encontradas, devido aos termos utilizados nas notícias, ou seja, *tiroteio* e *tiros* estão dentro da mesma categoria devido a composição dessas palavras.

Através da avaliação da Figura 16, podemos pressupor que há mais ocorrências de delitos envolvendo armas de fogo do que ocorrências de acidentes de trânsito. Claro que é apenas uma estimativa, mas pode-se afirmar que com base nas informações mineradas durante este os testes deste trabalhos, houve um alto índice de delitos envolvendo armas de fogo.

No Scrapinghub é possível determinar um período entre cada execução de busca, ou seja, entre cada mineração de dados. O gráfico da Figura 17 apresenta uma aproximação entre as requisições e os itens obtidos pelo minerador.

No intervalo entre as 09:00 UTC e as 10:00 UTC é possível ver que houve um pico de mineração de ambas as páginas, que juntas somam até o momento mais de 10 mil itens minerados. No intervalo entre as 01:00 UTC e as 05:00 UTC o sistema encontra-se em um estado de espera entre as execuções.

Figura 17 – Gráfico com a linha do tempo das requisições e dos itens minerados no Scrapinghub Dourados News e Dourados Agora



Fonte: Própria (2017)

Figura 18 – Execução do Minerador de Dados da página Dourados News

```

2017-09-27 01:08:53 [scrapy.core.scrapy] DEBUG: Scraped from <200 http://www.douradosnews.com.br/dourados/
dourados-sedia-nesta-terca-rodade-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-perseguiçao-policia-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=AIZA5qAha1A0t65M-VERUo HTTP/1.1" 200 1665
2017-09-27 01:08:58 [scrapy.core.scrapy] 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=AIZA5qAha1A0t65M-VERUo HTTP/1.1" 200 1665
2017-09-27 01:08:59 [scrapy.core.scrapy] 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)

```

Fonte: Própria (2017).

A Figura 18 apresenta o minerador em execução no terminal do Linux, utilizando a API do *googlemaps* para coletar as informações. É possível visualizar no corpo da imagem, em destaque, as informações coletadas no formato **{'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-presos-apos-sequencia-de-furtos-em-dourados'}**, que contém a causa, ou seja, o delito ou a infração de trânsito, o local da ocorrência com o nome do mesmo e sua latitude e longitude e o link para a notícia.

5. CONCLUSÃO

Analisar os dados e poder estruturá-los de modo que seja possível identificar regiões com ocorrências policiais, delitos ou infrações de trânsito, é uma maneira simples e prática de auxiliar na segurança da população de maneira geral, pois com essas informações pode-se ter um conhecimento melhor das regiões da cidade de Dourados – MS.

Dos itens obtidos na mineração de dados das páginas web, foi possível identificar que há um alto índice de delitos envolvendo armas de fogo, pois com base nos itens minerados, mais de 25% das ocorrências possuem a palavra abstrata *tiro*, e com essa informação é possível observar que há uma necessidade de melhora na segurança pública e de conscientização da população do índice elevado desse tipo de delito.

Entretanto, não dando foco apenas ao delito mencionado anteriormente, existem outros que afetam a segurança pública de maneira geral, sejam delitos ou infrações de trânsito, que somados elevam os índices de informações policiais que são publicadas diariamente nas páginas de notícias da região.

A Mineração de Dados somado a todas as demais tecnologias utilizadas, proporcionaram o desenvolvimento efetivo deste trabalho, os conceitos estudados sobre a mineração mostraram que o foco deste trabalho é o de analisar e buscar padrões que possam transformar as informações obtidas em material que sustente a ideia principal de informar.

Com as expressões regulares e o xpath foi possível determinar quais informações seriam buscadas e salvas pelo *software*. Gerar os padrões que formam o vasculhamento mais objetivo foi uma técnica muito importante pois assim apenas as informações, das páginas de notícias da web, de cunho policial foram assistidas, sem conter informações além desse contexto.

Durante o processo de desenvolvimento do *software* e o acompanhamento do seu funcionamento, foi possível observar a necessidade que existe sobre o controle dessas informações para auxiliar na questão de segurança pública. O índice de violência do país é um agravante que gera insegurança na população, e com essas informações bem estruturadas é possível identificar quais medidas possam ser tomadas e determinar onde

devem aplicadas essas medidas.

Para trabalhos futuros, sugere-se que sejam feitas pesquisas sobre outras técnicas de mineração textual, além da utilização de expressão regular e xpath, um levantamento estatístico sobre quais são as ocorrências mais frequentes. Sugere-se também que sejam utilizados em trabalhos futuros, outras tecnologias e algoritmos. Um exemplo é a utilização do R apresentada na Figura 5 como uma das linguagens de programação mais utilizadas para esse tipo de trabalho.

Ainda para o meio acadêmico, pode-se ter este trabalho como base para desenvolver mineração de dados utilizando as tecnologias do mesmo que possa se encaixar em outros conceitos dentro da questão de segurança pública e de estruturação da grande quantidade de informação que só cresce a cada dia.

REFERÊNCIAS BIBLIOGRÁFICAS

AGRAWAL, R; SRIKANT, R. **Fast algorithms for mining association rules**. 20Th International Conference on Very Large Data Bases, p. 487–499, 1994.

ALECRIM, Emersom. **O que é feed RSS?** 2005, Atualizado em 2011. Disponível em: <<https://www.infowester.com/rss.php>>. Acessado em: 05 set. 2017

ANUSKIEWICZ, Neil. **History and License: History of The Software**. 2017. Disponível em <<https://docs.python.org/3/license.html>>. Acessado em: 05 set. 2017.

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.

BERNAL, Guilherme. **O que é uma linguagem script?**. 2014. Disponível em: <<https://pt.stackoverflow.com/questions/17082/o-que-%C3%A9-uma-linguagem-de-script>>. Acessado em: 05 set. 2017.

BRAGA, Dr. Ryon. **O Excesso de Informação – A Neurose do Século XXI**. 2010. Disponível em: <<http://www.mettodo.com.br/pdf/O%20Excesso%20de%20Informacao.pdf>>. Acessado em: 28 ago. 2017.

CARILO, Alberto Silveira; SILVA, Gabriela Moreira da. **EVENTOS UNIFAL: UM APLICATIVO PARA OBTER INFORMAÇÕES ACERCA DE AULAS E DEMAIS EVENTOS DA UNIFAL-MG ATRAVÉS DE COMANDOS DE VOZ**. 2015. 67 p. Universidade Federal de Alfenas – Instituto de Ciências Exatas, Alfenas, 2015.

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.

CORDEIRO, Maria Inês. **Código aberto e livre acesso: uma nova cultura na gestão de recursos?** 2007. 9 p. Biblioteca Nacional de Portugal, Lisboa, 2007.

DORNELES, Elias. **Web Scraping com Scrapy – Primeiros Passos**. Disponível em: <<https://pythonhelp.wordpress.com/2014/08/05/web-scraping-com-scrapy-primeiros-passos/>>. 2014. Acessado em: 07 set. 2017.

CÔRTEZ, Sérgio da Costa; PORCARO, Rosa Maria; LIFSCHITZ, Sérgio. **Mineração de Dados – Funcionalidades, Técnicas e Abordagens**. 2002, p. 35. Pontifícia Universidade Católica do Rio de Janeiro. Rio de Janeiro, 2002.

EVANS, Shane. **Big Data at Scrapinghub**. 2016. 34 p. Presentation from Shane Evans, co-founder of Scrapinghub, Corky Big Data & Analytics Group. Corky, 2016.

GOEBEL, M.; GRUENWALD, L. **A survey of data mining and knowledge discovery software tools**. *SIGKDD Explorations*, v. 1, p. 20-33, 1999.

GOOGLE. **Google API Client Libraries > Python**. 2016. Disponível em: <<https://developers.google.com/api-client-library/python/auth/api-keys>>. Acessado em: 20 set. 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.

KLEINT, John. **Google Maps and Local Search APIs in Python**. 2013. Disponível em: <<http://py-googlemaps.sourceforge.net/>>. Acessado em: 20 set. 2017.

KOROBOV, Mikhail. **Scrapy at a glance**. Disponível em: <<https://github.com/scrapy/scrapy/blob/1.4/docs/intro/overview.rst>>. 2017. Acessado em: 05 jul. 2017.

MACARENO JR., Aleardo. **Construção de Compiladores: Capítulo 3 – Análise Sintática**. Disponível em: <<https://www.dcce.ibilce.unesp.br/~aleardo/cursos/compila/cap03.pdf>>. Acessado em: 10 ago. 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](https://msdn.microsoft.com/pt-br/library/ms174949(v=sql.120).aspx)>. Acessado em: 05 ago. 2017.

MINERAÇÃO DE DADOS. **Mineração de Dados**. Disponível em: <http://www.din.uem.br/~intersul/intersul_arquivos/documentos/mineracao.pdf>. Acessado em 29 set. 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-che-car-os-fatos>>. Acessado em: 15 ago. 2017.

NUNES, Rosângela; GUIMARÃES, Norton. **Regras de Associação – Mineração de Dados**. 2013. Instituto de Informática – Universidade Federal de Goiás. Goiás, 2013.

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.

PIATETSKY, Gregory. **Four main languages for Analytics, Data Mining, Data Science**. Kdnuggets. 2014. Disponível em: <<http://www.kdnuggets.com/2014/08/four-main-languages-analytics-data-mining-data-science.html>>. Acessado em: 12 set. 2017.

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO. Sistema Maxwell. **Laboratório de Automação de Museus, Bibliotecas Digitais e Arquivos do Departamento de Engenharia Elétrica**. Rio de Janeiro, 2006, p. 33 – 46.

QUINLAN, J. R. **Simplifying decision trees**. Technical report, Massachusetts Institute of Technology, 1986.

RACIOCÍNIO EM IA. **Raciocínio Baseado em Casos.** Disponível em <<http://www.din.uem.br/ia/intelige/raciocinio2/RacBasCasosArquit.html>>. Acessado em: 30 out. 2017.

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.

SCRAPINGHUB. **Cloud-based web crawling platform and data as a service.** 2017. Disponível em: <<https://scrapinghub.com/>>. Acessado em: 08 set. 2017.

SCRAPY. **Scrapy at a glance.** 2016. Disponível em <<https://docs.scrapy.org/en/latest/intro/overview.html#scrapy-at-a-glance>>. Acessado em 18 set. 2017.

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.