UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS QUIXADÁ

BACHARELADO EM SISTEMAS DE INFORMAÇÃO

SILVINO DEOLINO NETO

MINERAÇÃO DE DADOS DE OCORRÊNCIAS CRIMINAIS PARA IDENTIFICAÇÃO DE ZONAS DE ALTA CRIMINALIDADE EM FORTALEZA E REGIÃO METROPOLITANA

QUIXADÁ

2017

SILVINO DEOLINO NETO

MINERAÇÃO DE DADOS DE OCORRÊNCIAS CRIMINAIS PARA IDENTIFICAÇÃO DE ZONAS DE ALTA CRIMINALIDADE EM FORTALEZA E REGIÃO METROPOLITANA

Monografia apresentada no curso de Sistemas de Informação da Universidade Federal do Ceará, como requisito parcial à obtenção do título de bacharel em Sistemas de Informação. Área de concentração: Computação.

Orientadora: Prof. Ma. Lívia Almada Cruz Rafael

QUIXADÁ

2017

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

D465m Deolino Neto, Silvino.
Mineração de dados de ocorrências criminais para identificação de zonas de alta criminalidade em Fortaleza e região metropolitana / Silvino Deolino Neto. – 2017.
57 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Campus de Quixadá, Curso de Sistemas de Informação, Quixadá, 2017.
Orientação: Profa. Ma. Lívia Almada Cruz Rafael.

1. Crime-Análise. 2. Mineração de dados. 3. Banco de dados. I. Título.

CDD 005

SILVINO DEOLINO NETO

MINERAÇÃO DE DADOS DE OCORRÊNCIAS CRIMINAIS PARA IDENTIFICAÇÃO DE ZONAS DE ALTA CRIMINALIDADE EM FORTALEZA E REGIÃO METROPOLITANA

Monografia apresentada no curso de Sistemas de Informação da Universidade Federal do Ceará, como requisito parcial à obtenção do título de bacharel em Sistemas de Informação. Área de concentração: Computação.

Aprovado em: __/__/____.

BANCA EXAMINADORA

Prof. Ma. Lívia Almada Cruz Rafael (Orientadora)
Universidade Federal do Ceará (UFC)

Prof. Dra. Ticiana Linhares Coelho da Silva
Universidade Federal do Ceará (UFC)

Prof. Me. Régis Pires Magalhães
Universidade Federal do Ceará (UFC)

A Deus que nunca me deixou faltar nada apesar de todas as dificuldades encontradas pelo caminho. À minha família, amigos e namorada, por suas dedicatórias e ajuda prestada nos momentos felizes e tristes, principalmente a minha mãe que me apoiou sempre.

AGRADECIMENTOS

A Deus por todas as oportunidades, vitórias e fé que me proporcionou fazendo com que eu não desistisse no meio do caminho.

À minha mãe, família e amigos que sempre depositaram fé, esperanças e, acima de tudo, forças em mim para seguir sempre adiante.

À minha namorada, Fernanda Tayla, que sempre me apoiou nas minhas decisões e sempre torceu por mim, além de dar conselhos e cobrar para que eu desse o melhor de mim nos estudos.

À minha orientadora, Lívia Almada, por todo o auxílio, paciência e ajuda prestada durante o desenvolvimento deste trabalho pois, sem ela o trabalho não teria chegado a qualidade e importância que foi alcançado.

Agradeço a todos os professores por me proporcionarem o melhor conhecimento possível além do caráter e afetividade da educação no processo de formação profissional. Por toda dedicação a mim e meus amigos de graduação, não somente por terem nos ensinado, mas por terem nos feito aprender e tornar pessoas melhores.

E à Universidade Federal do Ceará (UFC), pela qualidade de ensino proporcionada e por toda e qualquer ajuda prestada pelos seus profissionais, além da ótima estrutura e material fornecido para que eu desempenhasse o melhor papel possível como aluno da instituição.

“Talvez não tenha conseguido fazer o melhor, mas lutei para que o melhor fosse feito. Não sou o que deveria ser, mas Graças a Deus, não sou o que era antes.”

(Marthin Luther King)

RESUMO

Nos últimos anos observou-se uma alta no número de ocorrências criminais por todo o Brasil. Essas ocorrências são de variados tipos e os órgãos de segurança vêm enfrentado dificuldades no combate dos mesmos. Muitas dessas dificuldades estão diretamente relacionadas à falta de investimentos e a de materiais que auxiliam no combate direto e indireto dos crimes. A perspectiva desse trabalho é analisar ocorrências criminais, mais precisamente ocorrências da cidade de Fortaleza disponibilizados pela Secretaria de Segurança Pública e Defesa Social, e disponibilizar informações como as regiões que possuem maior densidade de ocorrências, os tipos de crimes e os bairros que fazem parte das regiões identificadas para ajudar esses órgãos no combate destes crimes. Para isso, foi utilizado técnicas como Descoberta de Conhecimento em Bases de Dados, Extração de Entidades e Mineração de Dados. Com isso, foi possível reunir informações importantes sobre as ocorrências criminais de Fortaleza e chegar a conclusões como, por exemplo, a de que cada mês obteve uma região diferente com mais densidade de crimes, mas essas regiões sempre foram próximas umas das outras.

Palavras-chave: Crime-Análise. Mineração de dados. Banco de dados.

ABSTRACT

In recent years we experienced a high number of criminal occurrences throughout Brazil. These occurrences are of various types and the security organs have been faced with difficulties in combating them. Many of these difficulties are directly related to the lack of investment and materials that help in the direct and indirect fight against crime. The perspective of this work is to analyze criminal occurrences, more precisely occurrences from the city of Fortaleza made available by the Secretariat of Public Security and Social Defense, and to provide information such as the regions that have the highest density of occurrences, types of crimes and neighborhoods that are part of the regions identified to assist these bodies in combating these crimes. For this, we used techniques such as Knowledge Discovery in Databases, Extraction of Entities and Data Mining. Thus, it was possible to gather important information about the criminal occurrences from Fortaleza and to conclude, for example, that each month obtained a different region with more density of crimes, but these regions were always close to each other.

Keywords: Crime-Analysis. Data Mining. Database.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O número de crimes realizados no mundo inteiro é alarmante e o Brasil é um dos países que contribuem com grande quantidade de ocorrências todos os anos. Só ano de 2014, cerca de 47 mil vidas foram perdidas devido a homicídios registrados no país inteiro (CERQUEIRA, 2014). Esse número é ainda mais alarmante se comparado com outros países como os Estados Unidos que tem uma população maior, mas mesmo assim possui um número de mortes menor por homicídios, que, no ano de 2014, girava em torno de 8 mil pessoas (QUEALY; KATZ, 2016). Segundo Amaral (2015), só no estado do Ceará a taxa de homicídios no ano de 2013 era de 50,8 por 100 mil habitantes o que superou em cinco vezes o índice que a Organização das Nações Unidas (ONU) considera aceitável, que é a quantia de 10 mortes por 100 mil habitantes.

A Tecnologia da Informação e Comunicação (TIC) possui alguns recursos que podem auxiliar a compreender padrões e tendências dos crimes ocorridos a partir de bases de dados de ocorrências. Dentre estes, se destaca o processo de Descoberta de Conhecimento em Bases de Dados (KDD, do inglês *Knowledge Discovery in Databases*) que, de acordo com Fayyad, Piatetsky-Shapiro e Smyth (1996), é todo o processo de preparação, seleção, limpeza, incorporação de conhecimento prévio apropriado e a interpretação apropriada dos dados.

A Mineração de Dados é a etapa do processo KDD que possui a finalidade de buscar, em grandes bases de dados, adquirir conhecimentos que antes não eram utilizados por empresas, indústrias e instituições de pesquisa. A Mineração de Dados também é responsável pela escolha dos métodos a serem utilizados para encontrar padrões nos dados e, logo depois, nas formas de representação e visualização (SILVA, 2004). Com a Mineração de Dados o presente trabalho pode, portanto, adquirir conhecimentos a partir de bases de dados com informações sobre crimes e, então, responder algumas perguntas como quais são as regiões com maior quantidade de ocorrências e quais são os crimes mais comuns nessas regiões.

Na Mineração de Dados existem vários algoritmos e técnicas, sendo esses de muitas áreas. A área mais comum é a de Inteligência Artificial, seguida por Banco de Dados e Estatística (SILVA, 2004), mas vale ressaltar que elas se relacionam entre si. Para a realização deste trabalho serão utilizadas algumas dessas técnicas da Mineração de Dados como, por exemplo, a Extração de Entidades que nos auxiliará a encontrar informações em dados textuais e Clusterização que nos permite identificar informações de ocorrências criminais de acordo com suas similaridades e semelhanças. A semelhança dos dados pode ser medida de várias formas distintas. Isso

dependerá do escopo do problema.

Neste contexto, este trabalho visa auxiliar essa falta de recursos de análise de combate ao crime, principalmente na tomada de decisão, uma vez que respostas como a localização do crime e os tipos mais comuns estarão disponíveis. Não apenas organizações interessadas em políticas de segurança pública vão ser beneficiadas, mas também a população em geral, pois como mostra Lopes et al. (2013), muitas das vítimas de insegurança sofrem, após os ocorridos, com uma série de problemas de saúde mental e emocional. Considerando que a redução de criminalidade gera um melhor convívio para a população, pode-se dizer que este trabalho indiretamente também contribui para a área da saúde e bem estar social.

A seguir são apresentados o objetivo geral e os específicos ao quais se pretende alcançar no trabalho.

1.1 Objetivos

1.1.1 Objetivo Geral

Analisar dados de ocorrências criminais do estado do Ceará na região metropolitana de Fortaleza.

1.1.2 Objetivos específicos

a) Identificar, a partir dos dados textuais de ocorrências criminais, o endereço da ocorrência, a geolocalização e o tipo de crime;
b) Identificar bairros e regiões com maior densidade de ocorrências;
c) Identificar ocorrências mais frequentes em cada bairro e região.

Espera-se, ao final da realização deste trabalho que, a partir de dados de ocorrências criminais disponibilizados pela Secretaria de Segurança Pública, ter um mapeamento das áreas de maior risco em segurança, bem como tipos de crimes mais frequentes, na região metropolitana de Fortaleza.

2 FUNDAMENTAÇÃO TEÓRICA

A seguir são descritos os conceitos que fundamentam o trabalho. A Seção 2.1 explana sobre Descoberta de Conhecimento em Base de Dados, também conhecido como KDD (termo proveniente do inglês, *Knowledge Discovery in Databases*) e como o mesmo é usado no trabalho. Em seguida, na Seção 2.2, a Extração de Entidades é abordada e explicada, de acordo com a definição de alguns autores. Por fim, na Seção 2.3 o processo de Clusterização é explicado, assim como sua importância como um dos passos para a realização do estudo proposto.

2.1 Descoberta de Conhecimento em Bases de Dados

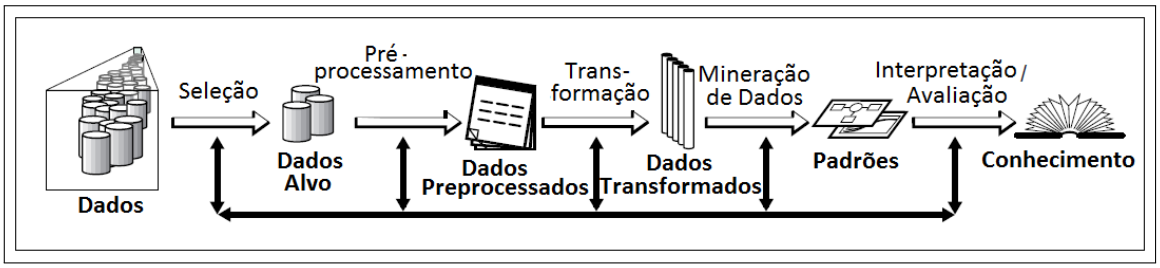
De acordo com Fayyad, Piatetsky-Shapiro e Smyth (1996), o processo KDD envolve todos os processos de preparação, seleção, limpeza, incorporação de conhecimento prévio apropriado e a interpretação apropriada dos dados. Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996), o KDD é um processo iterativo e interativo que envolve vários passos, os quais são descritos a seguir:

1. Aprender o domínio da aplicação: engloba conhecimentos prévios considerados relevantes e objetivos do projeto.
2. Criação de um grupo de dados: engloba a seleção dos dados nos quais a descoberta será realizada.
3. Limpeza e pré-processamento dos dados: engloba operações como a retirada de dados ruidosos, decisão do que fazer com campos de dados em falta.
4. Redução e projeção dos dados: engloba a busca por recursos úteis para representar os dados e usar técnicas de redução de dimensionalidade.
5. Escolha da técnica de Mineração de Dados: engloba a decisão do objetivo do modelo a partir do algoritmo de mineração.
6. Escolher o algoritmo de Mineração de Dados: engloba as técnicas de seleção a serem utilizadas para buscar padrões nos dados, bem como quais os modelos e parâmetros podem ser os mais adequados.
7. Exploração dos dados ou Mineração dos Dados: engloba a busca por padrões e similaridade nos dados que possam ser de interesse. Existem técnicas como a regressão, agrupamento e classificação que auxiliam nessa etapa do processo.
8. Interpretação: engloba toda a interpretação dos padrões descobertos, bem como a sua

visualização, além da remoção de padrões redundantes traduzindo-os em termos compreensíveis pelos usuários.

9. Uso do conhecimento adquirido: engloba a incorporação do novo conhecimento em sistemas, ou simplesmente documentando para que ele possa ser acessível por partes interessadas.

Figura 1 – Visão geral das etapas que constituem o processo KDD



Fonte – Fayyad, Piatetsky-Shapiro e Smyth (1996)

A Mineração de Dados é uma atividade do KDD que realiza descobertas de novas informações e conhecimentos em grandes bases de dados, por meio de algumas técnicas estatísticas e de inteligência artificial. Na Mineração de Dados, algoritmos recebem uma grande quantidade de dados desorganizados e, na maioria das vezes, sem qualquer informação previamente útil e retornam novos conhecimentos sobre estes dados. Seguindo os passos definidos por Fayyad, Piatetsky-Shapiro e Smyth (1996), a Mineração de Dados compõe, principalmente, no KDD o passo número 7. O tipo de conhecimento dependerá do algoritmo escolhido e dos dados existentes na base de dados (PIMENTEL; OMAR, 2006).

Para Silva (2004), o motivo de se utilizar mais de uma técnica para a descoberta de novos conhecimentos em bases de dados é a complexidade, volume e características dos dados que impõem fortes limitações em uma única abordagem, tais como o domínio de origem dos dados e o tipo de dado (número, texto, símbolos, etc). A aplicabilidade da Mineração de Dados é muito grande, podendo ser utilizada em vários contextos diferentes como saúde, *marketing*, educação e segurança (FRANÇA; AMARAL, 2013). Neste trabalho, a técnica de mineração de dados utilizada será a clusterização, explicada posteriormente na Seção 2.3.

O KDD irá guiar grande parte das atividades realizadas neste trabalho, como a seleção dos dados extraídos, a limpeza, a estruturação, a aplicação do algoritmo de Mineração de Dados em si para a incorporação do conhecimento prévio e avaliação do conhecimento encontrado e sua apresentação.

2.2 Extração de Entidades

Hoje em dia há uma grande quantidade de dados gerados diariamente por meio do tráfego da Internet, principalmente pela ascensão do uso de *smartphones* e outros inúmeros tipos de computadores existentes, além de sistemas inteligentes que a cada dia se tornam mais presentes em nosso meio de convívio, estando estes na maioria conectados à Internet. Obter conhecimento sobre essas informações pode ser um diferencial para organizações, fazendo com que esta saia na frente do mercado em relação a seus concorrentes. Entretanto extrair dados nem sempre é uma atividade rápida e fácil de ser realizada. A partir de problemas como esse, a Extração de Entidades ganha destaque fornecendo funcionalidades que possibilitam a identificação de padrões, extração, classificação e simplificação das informações trocadas na internet (RIBEIRO; MEDEIROS, 2016).

A Extração de Entidades é uma tarefa que classifica entidades como pessoas, organizações e localização contidas em textos, além de campos mais avançados como genes. Para Amaral (2013), a classificação das entidades é realizada principalmente pela aprendizagem supervisionada, ou seja, existe um conjunto de dados que são disponibilizados para que o algoritmo, através do treinamento, se ajuste no campo de pesquisa que o problema a ser solucionado se encontra, mas também existem técnicas de aprendizagem não-supervisionada na literatura, além de outras técnicas baseadas em gramáticas. Quando o treinamento é finalizado, o algoritmo pode então receber os dados para realizar a extração de entidades (AMARAL, 2013).

Manning (2012) dá uma explicação mais ampla sobre extração de entidades onde, de acordo com ele, extração de entidades é realizada em três categorias: *Hand-written regular expressions*, que utiliza técnicas de expressões regulares para extrair de textos não estruturados as entidades e é uma abordagem semelhante à utilizada no trabalho; *Using classifiers*, que utiliza algoritmos de aprendizagem supervisionada como o *Naïve Bayes* para sua execução e *Sequence models*, que é uma derivação do método de classificação, mas trata cada palavra do texto, a ser extraído, separadamente.

Em seu funcionamento, a Extração de Entidades utiliza algumas técnicas como Aprendizagem de Máquina, já citada, Processamento de Linguagem Natural, que é uma subcampo de Inteligência Artificial e o Reconhecimento de Entidades Nomeadas (NER, do inglês *Named Entity Recognition*) (RIBEIRO; MEDEIROS, 2016). NER é um subcampo de pesquisa de Extração de Informação e essa, assim como a Extração de Entidades, tem como

objetivos identificar entidades, mais precisamente entidades nomeadas, e organizá-las em categorias pré-definidas (EVANDRO et al., 2015). É dito também por Evandro et al. (2015) que NER é uma técnica extensamente usada no Processamento de Linguagem Natural. Ela consiste em identificar atributos de entidades chave que estão contidos dentro de elementos textuais.

Neste trabalho, algumas informações que estão contidas dentro de elementos textuais dos dados extraídos, como por exemplo localização, precisam ser identificados. Para isso, é necessária a aplicação de alguma técnica de Extração de Entidades. Além da localização, outras entidades contidas nos dados criminais, tais como os nomes das vítimas, nomes de suspeitos e informações sobre o veículo usado na realização do crime, serão extraídos.

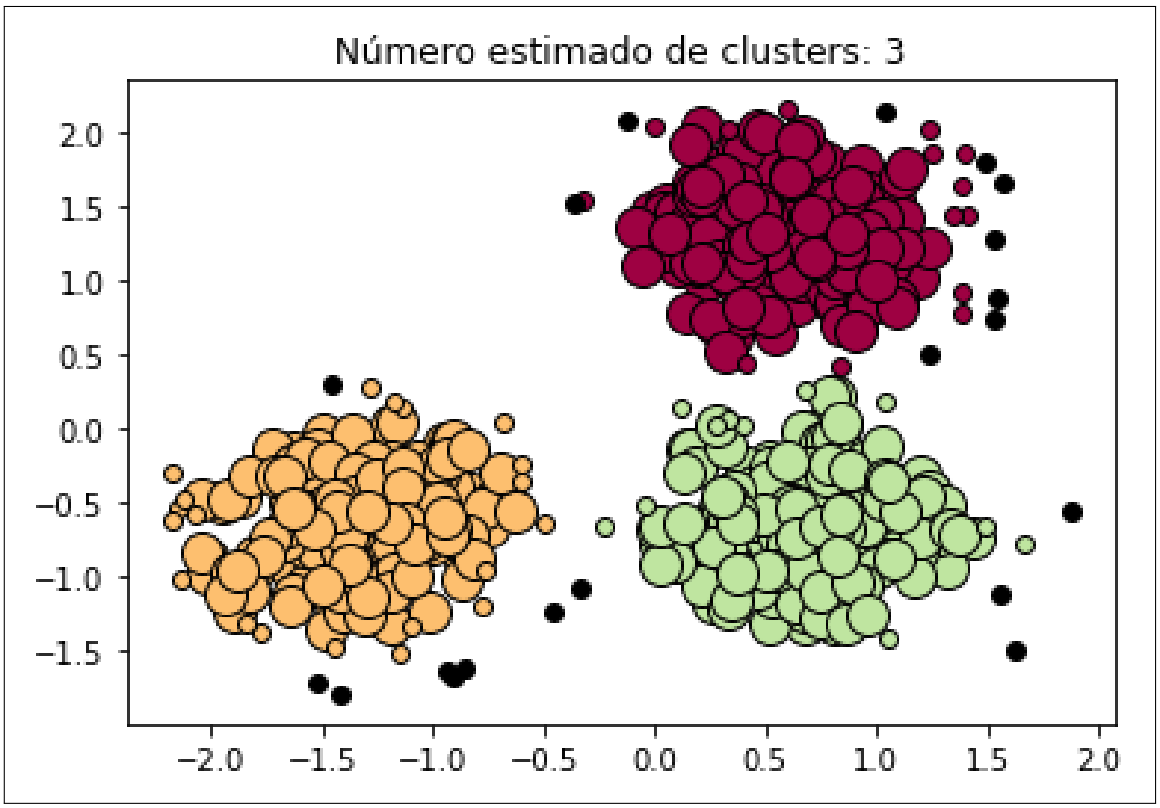
2.3 Clusterização

Clusterização é o processo de, a partir de uma base de dados, agrupar o conteúdo presente na base de tal forma que as informações mais semelhantes fiquem próximas uma das outras enquanto que as menos semelhantes fiquem mais distantes entre si (OCHI; DIAS; SOARES, 2004).

Birant e Kut (2007) definem clusterização como um dos melhores métodos para descobrimento de conhecimento em grandes bases de dados. A análise de *cluster*, de acordo com Birant e Kut (2007), é a maior ferramenta em muitas áreas de engenharia e da ciência incluindo discretização de atributos contínuos, redução de dados, reconhecimento de padrões e processamento de imagem. No campo do processo KDD, a análise de *cluster* é conhecida como um processo de aprendizado não supervisionado, uma vez que não há um conhecimento prévio sobre os dados fornecidos para a análise (BIRANT; KUT, 2007).

Aprendizado não supervisionado é uma categoria de algoritmo que não possui uma amostra de treinamento. O número de objetos de dados a serem treinados também pode não ser conhecido, porém isso não é uma regra. Por conta dessas características, ocorre o fato da clusterização ser e fazer parte de uma aprendizagem não-supervisionada. Dentre as técnicas que se enquadram nessa abordagem, se sobressaem a Associação e a Clusterização. Esta última será a técnica utilizada neste trabalho (SILVA, 2004). A Figura 2 apresenta um exemplo de agrupamento por *clusters*. Como se pode ver pela Figura 2, nem todos os dados irão necessariamente pertencer a um *cluster* específico, sendo estes conhecidos por dados ruídosos ou *outliers*.

Figura 2 – Exemplo de dados agrupados em 3 *clusters*



Fonte – Elaborada pelo Autor.

2.3.1 Medidas de Similaridade

Medidas de Similaridades se tornam um conceito fundamental para o processo de clusterização uma vez que, o processo de clusterização procura agrupar objetos semelhantes. Existem outras maneiras de medir a semelhança entre dois objetos como, por exemplo, as medidas de distância. Na medida de distância, os objetos que são mais semelhantes entre si ficam mais próximos um dos outro em relação a distância, enquanto que os menos semelhantes tendem a ficar mais distantes (NALDI, 2011).

Algumas das medidas de distância mais comuns são a Distância Euclidiana e a Distância Manhattan. Dados dois pontos p e q, a Distância Euclidiana, ilustrada na equação 2.1, é definida como sendo a soma da raiz quadrada da diferença entre q e r em suas dimensões, respectivamente. Através da Distância Euclidiana é encontrada a menor distância para se chegar de um ponto a outro. Já a Distância Manhattan, ilustrada na equação 2.2, é definida como sendo a soma das diferenças entre q e r em cada dimensão. A Distância Manhattan produz um segmento

de reta na horizontal quanto na vertical, similar a uma rota de carro (LOPES et al., 2008).

$$f(q,r) = \sqrt{\sum_{i=1}^{k}(q_i - r_i)^2} \tag{2.1}$$

$$f(q,r) = (\sum_{i=0}^{k}|q_i - r_i|^p)^{\frac{1}{p}} \tag{2.2}$$

Uma outra medida de distância é a *Haversine*. Essa medida é uma equação usada principalmente na navegação, onde é fornecida a distância entre dois pontos através de latitudes e longitudes. De acordo com Ponciano et al. (2016), a medida *Haversine* é equivalente à lei esférica dos cossenos, além de ser menos sensível a erros de arredondamento. Por conta disso, essa foi a medida escolhida neste estudo, uma vez que os pontos serão disponibilizados em dados com latitudes e longitudes. Sejam $P = (lat_1, lng_1)$ e $Q = (lat_2, lng_2)$, *Haversine* é definida como: $HaversineDist(P,Q) = R.c$, onde R é o raio da terra e c é definida na Equação 2.3.

$$\begin{aligned} c &= 2.atan2(\sqrt{a}, \sqrt{1-a}) \\ dlat &= lat_2 - lat_1 \\ dlng &= lng_2 - lng_1 \\ a &= (\sin dlat/2)^2 + \cos(lat_1).\cos(lat_2).(\sin(dlng/2))^2 \end{aligned} \tag{2.3}$$

O resultado da medida *Haversine* estará na mesma medida do valor R, ou seja, como no presente trabalho a unidade de medida utilizada foi quilômetros, o valor da distância retornado entre os dois pontos P e Q também é em quilômetros.

2.3.2 O Algoritmo DBSCAN

O algoritmo DBSCAN (Density Based Spatial Clustering of Applications with Noise), proposto por Ester et al. (1996), é um algoritmo de clusterização que agrupa pontos próximos uns dos outros baseado na densidade de cada um dos pontos.

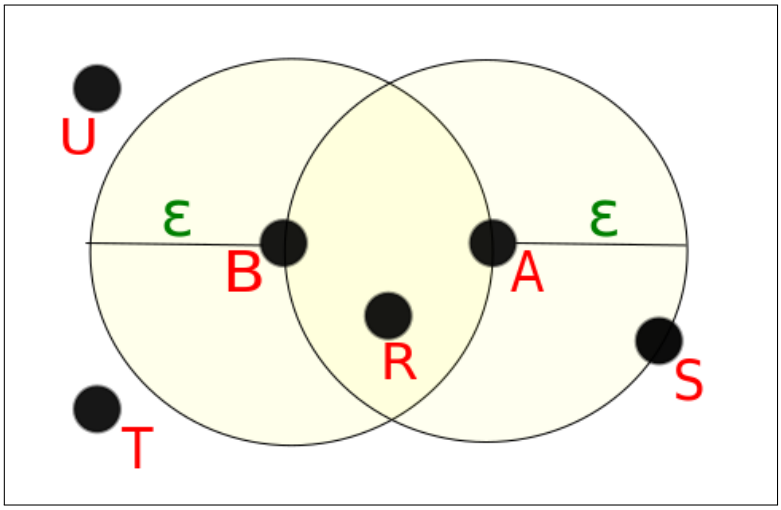
No algoritmo DBSCAN, a densidade associada com um ponto é obtida contando o número de pontos em uma região de raio, anteriormente especificado, ao redor do ponto. Pontos com uma densidade acima de um limite especificado são construídos como *clusters*, enquanto que pontos com densidade abaixo são rotulados como ruídos (BIRANT; KUT, 2007). Para

encontrar um novo *cluster*, o DBSCAN recebe como parâmetro um ponto qualquer e retorna toda a densidade de pontos acessíveis pelo ponto inicial em relação a ε-vizinhança (Eps) e ao número mínimo de pontos (MinPts). Se o ponto for central então um novo *cluster* é criado. Se o ponto não é central, ou seja, for uma borda, nenhum outro ponto é acessível por densidade a ele e o DBSCAN irá visitar o próximo ponto existente nos dados (ESTER et al., 1996).

De acordo com Gava et al. (2013), em um conjunto de dados o DBSCAN separa os objetos em três tipos:

1. Pontos centrais: Caso a ε-vizinhança de um objeto A contenha ao menos o número mínimo MinPts de outros objetos, então esse objeto A é dito como ponto central. Tomando a Figura 3 como base, A será um ponto central, por exemplo, caso o MinPts seja igual a 4, enquanto que os demais pontos não seriam pontos centrais.
2. Pontos de borda: Caso a ε-vizinhança de um objeto A contenha menos que MinPts mas contenha ao menos um ponto central, então o objeto A é dito ponto de borda. Na Figura 3, por exemplo, os pontos B, R e S são pontos de borda.
3. *Outliers* ou Ruídos: Caso um objeto A não pertença a nenhum grupo de objetos ele é dito como ruído, ou seja, um objeto qualquer que não é nem ponto central nem ponto de borda, é ruído. Na Figura 3, por exemplo, os pontos U e T são ruídos.

Figura 3 – ε-vizinhança de B e ε-vizinhança de A



Fonte – Elaborada pelo Autor.

O algoritmo DBSCAN é também muito eficiente na identificação de *clusters* arbitrários e de variados tamanhos sem qualquer informação preliminar sobre os possíveis *clusters*, ou seja, o algoritmo encontra novos padrões nas informações sem qualquer

interferência de conhecimentos anteriormente estabelecidos.

A técnica de clusterização escolhida, para descoberta de regiões com maior densidade de criminalidade, foi o DBSCAN.

3 TRABALHOS RELACIONADOS

Na literatura já existem alguns estudos sobre análise de dados de crimes, bem como técnicas que possam ajudar no combate de crimes por terceiros. A seguir, alguns estudos são apresentados como forma de esclarecimento e a relação destes com o trabalho aqui proposto.

3.1 *Crime pattern detection using data mining*

Nath (2006) estudando dados sobre crimes dos Estados Unidos, encontrou alguns padrões sobre estes crimes em uma determinada região geográfica por meio de algoritmos de clusterização. Nath (2006) também buscou encontrar similaridades entre os crimes existentes em relação a idade e o objeto utilizado na realização do ocorrido, e agrupar os crimes de acordo com estas similaridades.

Assim como em Nath (2006), este estudo pretende buscar padrões em crimes, por meio de algoritmos e técnicas de clusterização em conjunto com os processos do KDD. Logo, Nath (2006) contribui com alguns exemplos e ferramentas utilizadas para a realização deste novo estudo. Entretanto, Nath (2006) apenas clusterizou os dados de acordo com suas similaridades sobre um atributo específico, o que não é o suficiente, pois muitas outras informações úteis podem ser descobertas através de uma análise mais detalhada sobre os resultados obtidos. Este estudo, além de buscar os padrões nos dados, buscará disponibilizar informações mais relevantes e úteis do ponto de vista do usuário que as utilizam como, por exemplo, os crimes mais comuns por uma determinada região e os horários em que estes crimes ocorreram. Outro ponto de diferença deste trabalho com o de Nath (2006) é a fonte de dados utilizada, que neste estudo serão extraídas do Portal da Secretaria de Segurança Pública e Defesa Social do estado do Ceará, ou seja, que irá mostrar um pouco da realidade vivida em uma área do Brasil.

3.2 *Crime analytics: Analysis of crimes through newspaper articles*

Jayaweera et al. (2015) propõem um sistema de análise de criminalidade inteligente, com o intuito de facilitar o combate ao crime na cidade de Tucson nos Estados Unidos. De acordo com os autores, o combate ao crime é um processo cada vez mais difícil de ser realizado devido à grande quantidade de ocorrências que acontecem diariamente, ou seja, pela grande quantidade de dados gerados e também pela vasta diversidade geográfica destes.

Este trabalho utiliza algumas ferramentas e técnicas empregadas em Jayaweera et

al. (2015), tais como a extração de dados de locais onde os dados não estão bem organizados e a Extração de Entidades de textos, que no trabalho deles são textos retirados de notícias sobre crimes. Outro fator em comum são os passos necessários para atingir os resultados, como a aplicação de técnicas de limpeza de dados e a utilização de um algoritmo que busca encontrar algum padrão nos dados. O algoritmo utilizado por Jayaweera et al. (2015) é denominado Máquina de Vetores de Suporte (SVM, do inglês *support vector machine*).

Porém Jayaweera et al. (2015) faz a extração de todas as notícias existentes em portais conhecidos em sua região. Neste trabalho, essa etapa será um pouco diferente, pois os dados serão extraídos do Portal da Secretaria da Segurança Pública e Defesa Social, mais especificamente de dados disponibilizados para *download* em formato PDF contendo as informações criminais do estado do Ceará. Portanto, os dados são extraídos diretamente destes arquivos. O uso de arquivos dispensa uma etapa de Jayaweera et al. (2015): o uso de algoritmos para buscar dados duplicados, já que são retirados de um local específico para crimes. Como já foi mencionado anteriormente, esta etapa também não é realizada neste trabalho por não ser necessária.

3.3 *Detecting and investigating crime by means of data mining: a general crime matching framework*

Em Keyvanpour, Javideh e Ebrahimi (2011) foi desenvolvido um estudo sobre crimes com o foco em descobrir a relação dos crimes entre si e os criminosos que cometeram essas atividades. Apesar de ser usada uma base de dados contendo dados sobres estes casos, o problema foi centrado nos campos de textos abertos utilizados pelos agentes da lei para descrever cada crime.

O principal objetivo de Keyvanpour, Javideh e Ebrahimi (2011) era desenvolver uma abordagem inteligente de investigação de crimes que, através das ferramentas de Mineração de Dados, pudesse facilitar a complexidade da realização de análise. Sendo esta realizada manualmente por várias pessoas diferentes, ocasionava perdas de informações relevantes em alguns momentos, e isso se dava principalmente pelo fato de não haver uma revisão sobre os dados por outras pessoas. Apesar deste trabalho possuir um objetivo diferente do estudo de Keyvanpour, Javideh e Ebrahimi (2011) há coincidência em seus propósitos gerais: existir para ajudar agentes da lei em combater a criminalidade. O algoritmo utilizado por eles para encontrar a relação entre os crimes é o *Self-Organizing Map Neural Network* (SOM) que consideram o melhor para sua abordagem devido à busca de comparação na parte textual dos dados disponíveis.

Tabela 1 – Trabalhos Relacionados

Trabalho	Dados Geospaciais	Extração de Entidades	Clusterização	Algoritmo Utilizado
Nath (2006)	x		x	K-means
Jayaweera et al. (2015)	x	x		SVM
Keyvanpour, Javideh e Ebrahimi (2011)		x	x	SOM
Furtado et al. (2008)	x			
Presente Trabalho	x	x	x	DBSCAN

O objetivo deste trabalho é um pouco diferente da abordagem de Keyvanpour, Javideh e Ebrahimi (2011), pois aqui a parte textual irá ser destrinchada em rótulos e disponibilizadas desta forma nos dados criminais, ou seja, quando chegar na etapa de uso de algum algoritmo, todo o texto existente antes já estará disponível de forma particionada e isso também acaba refletindo no tipo de algoritmo escolhido para buscar os padrões nos dados. Uma outra diferença é que este trabalho leva em consideração a localização onde os crimes ocorreram enquanto que no trabalho de Keyvanpour, Javideh e Ebrahimi (2011) isso não foi considerado.

3.4 WikiCrimes - Um Sistema Colaborativo para Mapeamento Criminal

Furtado et al. (2008) apresenta em seu trabalho o WikiCrimes, um sistema colaborativo de mapeamento criminal. O sistema funciona através de dados da localização e informações dos crimes como, nome (furto, roubo, etc) e data. Com essas informações é disponibilizado no portal do WikiCrimes[1] um mapa geoprocessado que possibilita a pesquisa das ocorrências criminais registradas no sistema.

Assim como no trabalho de Furtado et al. (2008), o estudo aqui proposto usa informações de geolocalização dos crimes para análise e apresentação das ocorrências identificadas, porém para tal é feito um processamento maior sobre os dados, uma vez que será realizado, primeiramente, a extração de entidades e, posteriormente, a clusterização dos dados em busca de semelhanças entre as ocorrências. Além disso, os trabalhos possuem objetivos comuns como: capturar dados de crimes e apresentar estes de uma forma melhor para órgãos e pessoas interessadas em obter essas informações.

Na Tabela 1 apresenta-se um resumo das técnicas e algoritmos usados por cada trabalho apresentado nesta Seção.

[1] http://wikicrimes.org

4 PROCEDIMENTOS METODOLÓGICOS

A seguir são apresentados os procedimentos metodológicos utilizados neste trabalho.

Na Seção 4.1 é apresentada a fonte dos dados, seu formato e como ela foi manipulada no decorrer do trabalho. Na Seção 4.2 é explicado como os dados foram tratados para o melhor uso deles neste trabalho. Na Seção seguinte, 4.3, explica-se o processo de extração de dados de entidades localizadas no texto que são importantes para este trabalho alcançar os objetivos propostos. Na Seção 4.4 é explicado o porque o uso do georreferenciamento e a importância dele para o processo de clusterização. Na Seção posterior, 4.5, é apresentado como os dados foram agrupados para a realização da clusterização e análise. Na Seção 4.6 o processo de clusterização é apresentado e justificado o seu uso. Por fim, na Seção 4.7, são descritas quais foram as análises feitas no trabalho. A Figura 4 ilustra visualmente cada uma das etapas.

Figura 4 – Passos do procedimentos metodológicos

Fonte – Elaborada pelo Autor.

4.1 Obtenção dos dados

A primeira etapa do trabalho foi a obtenção dos dados das ocorrências criminais da região metropolitana da cidade de Fortaleza. As informações foram recuperadas do Portal da Secretaria de Segurança Pública e Defesa Social[1] do estado do Ceará no formato PDF e se referem ao período de janeiro à maio de 2017. Os arquivos no formato PDF foram convertidos para o formato de arquivo *Comma-Separated Values* (CSV), utilizando-se uma biblioteca chamada Tabula[2]. Tabula é uma biblioteca em Java para extrair tabelas de arquivos PDF e salvá-las em outros formatos, como CSV, JSON ou TSV. A partir dos dados extraídos, é possível obter informações sobre a fonte do crime, a natureza da ocorrência o histórico da ocorrência, bem como outras informações.

[1] http://www.sspds.ce.gov.br/
[2] http://tabula.technology/

4.2 Preparação e limpeza dos dados

Após a obtenção dos dados, foi necessário realizar a preparação e a limpeza dos dados contidos nos arquivos CSV. Este processo consistiu em tirar todos os acentos, vírgulas, *underlines*, e até mesmo padronização entre maiúsculas e minúsculas, além de outras pontuações que podem reduzir a eficiência dos algoritmos no decorrer do trabalho. Um exemplo de dado removido ocorre quando as informações criminais não contém a informação do local do crime o que impossibilita o georreferenciamento dele.

Essa etapa foi realizada logo após os dados estarem disponibilizados no formato CSV e se estende no decorrer do trabalho, dependendo das necessidades que surjam no tratamento dos dados e na forma de análise e apresentação dos mesmos.

4.3 Extração de Entidades

O objetivo desta etapa é extrair da coluna HISTÓRICO DE OCORRÊNCIA do arquivo de ocorrências, atributos como local, arma apreendida, veículo, entre outros. Dessa forma, foi implementado um algoritmo para a extração das entidades nos dados extraídos na etapa 4.1. Esta etapa é necessária porque essas informações estão em uma formato de texto não-estruturado.

Para realizar essa atividade, o formato do texto contido na coluna HISTÓRICO DE OCORRÊNCIA foi primeiramente analisado. Com esta análise identificou-se que cada entidade contida nas ocorrências vinha precedida do seu rótulo. Assim, foram identificados os rótulos mais comuns presentes nas ocorrências. Os rótulos são algumas informações pertinentes a dados como local e nome de bairros que são encontrados nos dados. Neste trabalho alguns dos rótulos encontrados foram local, suspeito, hora e armas apreendidas. O algoritmo para extração de entidades proposto recebe como parâmetro a coluna na qual se deseja extrair as entidades e uma lista de rótulos e retorna, em um novo arquivo, as informações relacionadas a cada rótulo como novas colunas.

Na extração, o algoritmo percorre cada rótulo da lista de rótulos, e verifica se aquele rótulo está presente no texto da coluna desejada. Caso o rótulo esteja presente, o algoritmo obtém a substring referente à primeira posição depois do rótulo até encontrar o primeiro ponto encontrado, quando deve se iniciar a descrição de uma nova entidade. Uma coluna é criada no arquivo referente ao rótulo, se o rótulo não tiver sido encontrado na ocorrência, o valor atribuído

à coluna é "nulo". O código da implementação deste algoritmo está disponível no GitHub[3].

A Tabela 2 demonstra um exemplo de entidades que podem ser extraídas. Em negrito estão alguns dos rótulos contidos na coluna HISTÓRICO DE OCORRÊNCIA. Nesse exemplo, ao executar o algoritmo extrator, os outros rótulos, que não foram identificados na ocorrência terão valor "nulo"na coluna equivalente, indicando que essa informação é inexistente, pois os mesmos não foram registrados nos dados daquela ocorrência.

Tabela 2 – Exemplo de dados textuais a extrair

NATUREZA DA OCORRÊNCIA	HISTÓRICO DA OCORRÊNCIA
VEICULO LOCALIZADO 06H41	**LOCAL**: RUA MONSENHOR FURTADO, BELA VISTA.,**VEÍCULO**: VOYAGE SUPER, 1984, CINZA. **PLACA**: HWG-5368-CE.
MORTE A BALA 04H24	**LOCAL**: RUA TIAGO PEREIRA, BOM JARDIM.,**VÍTIMA**: RICHARD M S. **SUSPEITO**: NAO IDENTIFICADO.

4.4 Georreferenciamento da localização dos crimes

As informações obtidas sobre localização na etapa 4.3 são o endereço da ocorrência criminal em formato de texto. Apesar desse formato ser suportado pelo algoritmo DBSCAN no processo de criação de *clusters*, o foco deste trabalho é fazer uso da localização geográfica do crime para extrair informações sobre localidades mais perigosas, ou seja com maior densidade de crimes. Portanto, nesta etapa se fez necessário o uso de ferramentas que, através de dados textuais sobre localização, nos forneça dados de geolocalização através de coordenadas, valores esses também suportados pelo algoritmo de clusterização DBSCAN.

A ferramenta Google Maps API[4] foi utilizada nesse processo. Para isso, o endereço no formato textual é enviado para o serviço do Google Maps e o serviço retorna a localização em coordenadas geográficas (latitude e longitude) do local. Após isso, cada crime foi atualizado com informações de geolocalização para posterior uso do algoritmo de clusterização.

4.5 Particionamento dos dados

Ao final da execução das etapas anteriores, os arquivos se encontram no formato CSV sendo cada um deles referente a um dia do mês. Essa etapa consiste em reunir estes arquivos extraídos pelo algoritmo de extração de entidades com os dados de geolocalização em um único lugar. Neste trabalho, os arquivos foram particionados por mês. Para a realização dessa etapa foi criado um algoritmo que faz a leitura de vários arquivos criminais e os agrupam em um único arquivo. Desta forma é possível, por exemplo, particionar dados por mês, semestre ou ano. Com

[3] https://goo.gl/FVd4tL

[4] https://developers.google.com/maps/?hl=pt-br

os dados particionados o algoritmo de clusterização pode enfim ser utilizado sobre os dados, retornando, por fim, novos padrões e, consequentemente, novas possibilidades de análise sobre os dados criminais.

4.6 Clusterização das ocorrências criminais

O objetivo desta etapa foi encontrar localidades com alta densidade de crimes. Dessa forma, a técnica de clusterização de dados foi aplicada sobre os dados das ocorrências criminais. Ao oferecer informações de tais localidades, os resultados da clusterização podem ser utilizados para tomada de decisão em políticas de segurança pública.

O algoritmo utilizado para a clusterização dos dados foi DBSCAN. Como o objetivo desta etapa é encontrar localidades de maior densidade, a medida de distância escolhida para esta tarefa é baseada na distância entre as localizações das ocorrências, assim como já explicado na Seção 2.3. A distância utilizada foi a distância *Haversine*. Dessa forma, a realização dessa etapa só pôde ser feita após as ocorrências serem georreferenciadas.

4.7 Análise dos resultados

Por fim, a última etapa do trabalho é a análise dos resultados obtidos em cada etapa. Esta análise deve permitir responder a questões como:

- Quais as regiões com maior quantidade de ocorrências criminais?
- Quais os bairros dentro de cada região com maior densidade de ocorrências criminais?
- Quais os crimes mais comuns por região?
- Quais os crimes mais comuns em um bairro específico?
- Como as ocorrências criminais variaram mês a mês?

5 RESULTADOS

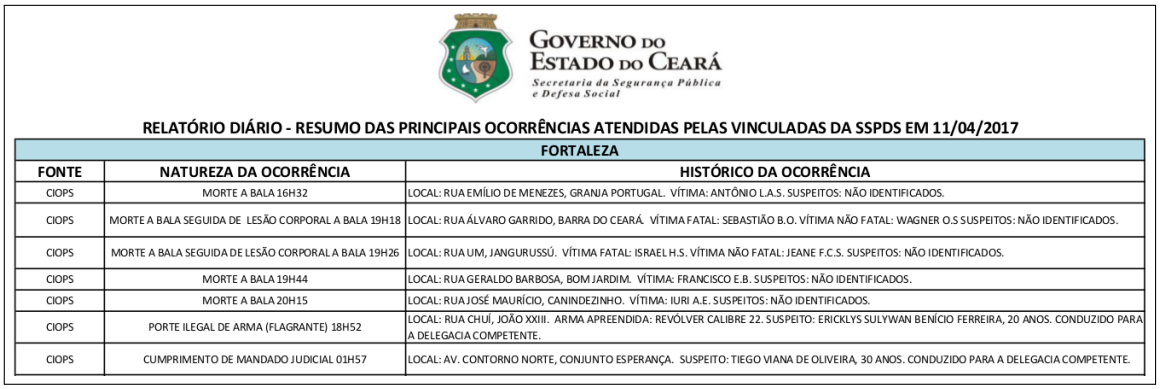
Nesta etapa são apresentados os resultados obtidos na execução do trabalho. Para melhor organização e entendimento os resultados foram divididos em sub-tópicos.

5.1 Obtenção dos dados

Primeiramente, foi feito *download* dos dados do portal da Secretaria de Segurança Pública e Defesa Social do Ceará. Os arquivos se encontram no formato PDF, mas para melhor manipulação foram convertidos para arquivos no formato CSV.

Cada arquivo contém dados sobre ocorrências de crimes de um único dia na região metropolitana de Fortaleza. Foi feito o *download* de 144 arquivos, sendo estes referentes aos meses de janeiro à maio de 2017. Porém, alguns arquivos estavam com links corrompidos e, portanto, não foi possível utilizá-los. A Figura 5 ilustra um exemplo de um desses arquivos extraídos.

Figura 5 – Dados de ocorrências criminais do dia 16 de Abril de 2017



GOVERNO DO ESTADO DO CEARÁ
Secretaria da Segurança Pública e Defesa Social

RELATÓRIO DIÁRIO - RESUMO DAS PRINCIPAIS OCORRÊNCIAS ATENDIDAS PELAS VINCULADAS DA SSPDS EM 11/04/2017

FORTALEZA		
FONTE	**NATUREZA DA OCORRÊNCIA**	**HISTÓRICO DA OCORRÊNCIA**
CIOPS	MORTE A BALA 16H32	LOCAL: RUA EMÍLIO DE MENEZES, GRANJA PORTUGAL. VÍTIMA: ANTÔNIO L.A.S. SUSPEITOS: NÃO IDENTIFICADOS.
CIOPS	MORTE A BALA SEGUIDA DE LESÃO CORPORAL A BALA 19H18	LOCAL: RUA ÁLVARO GARRIDO, BARRA DO CEARÁ. VÍTIMA FATAL: SEBASTIÃO B.O. VÍTIMA NÃO FATAL: WAGNER O.S SUSPEITOS: NÃO IDENTIFICADOS.
CIOPS	MORTE A BALA SEGUIDA DE LESÃO CORPORAL A BALA 19H26	LOCAL: RUA UM, JANGURUSSÚ. VÍTIMA FATAL: ISRAEL H.S. VÍTIMA NÃO FATAL: JEANE F.C.S. SUSPEITOS: NÃO IDENTIFICADOS.
CIOPS	MORTE A BALA 19H44	LOCAL: RUA GERALDO BARBOSA, BOM JARDIM. VÍTIMA: FRANCISCO E.B. SUSPEITOS: NÃO IDENTIFICADOS.
CIOPS	MORTE A BALA 20H15	LOCAL: RUA JOSÉ MAURÍCIO, CANINDEZINHO. VÍTIMA: IURI A.E. SUSPEITOS: NÃO IDENTIFICADOS.
CIOPS	PORTE ILEGAL DE ARMA (FLAGRANTE) 18H52	LOCAL: RUA CHUÍ, JOÃO XXIII. ARMA APREENDIDA: REVÓLVER CALIBRE 22. SUSPEITO: ERICKLYS SULYWAN BENÍCIO FERREIRA, 20 ANOS. CONDUZIDO PARA A DELEGACIA COMPETENTE.
CIOPS	CUMPRIMENTO DE MANDADO JUDICIAL 01H57	LOCAL: AV. CONTORNO NORTE, CONJUNTO ESPERANÇA. SUSPEITO: TIEGO VIANA DE OLIVEIRA, 30 ANOS. CONDUZIDO PARA A DELEGACIA COMPETENTE.

Fonte – Secretaria de Segurança Pública e Defesa Social.

Após o *download*, foi feita a conversão dos dados em PDF para arquivos CSV's. Essa etapa foi realizada com o auxílio de uma biblioteca disponível na linguagem Java[1] chamada *Tabula*, já citada na Seção 4.1. Ao total, foram recuperadas a quantidade de 2.747 ocorrências criminais, uma média de 549,4 ocorrências por mês. A Figura 6 ilustra um exemplo de arquivo CSV extraído a partir do PDF disponibilizado. Cada arquivo possui duas colunas: a coluna NATUREZA DA OCORRÊNCIA, que descreve o tipo de ocorrência e a hora em que aconteceu, e a coluna HISTÓRICO DA OCORRÊNCIA, que é uma descrição geral da ocorrência, contendo

[1] https://www.java.com

entidades como LOCAL, VÍTIMAS, ARMA, etc.

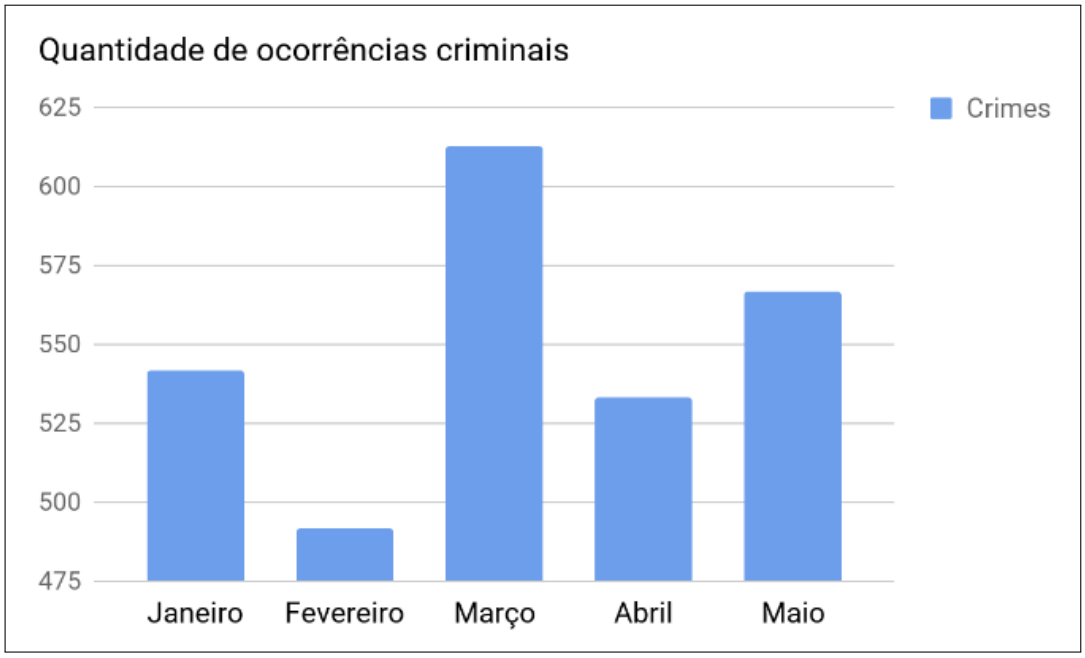
Os arquivos convertidos foram agrupados por mês, assim obteve-se 5 arquivos, referentes aos meses de janeiro à maio de 2017, e estes somam 985,5 kB de dados. A Figura 7 apresenta a distribuição da quantidade de ocorrências criminais por mês.

Figura 6 – Exemplo de um arquivo extraído do PDF para CSV

NATUREZA DA OCORRÊNCIA	HISTÓRICO DA OCORRÊNCIA
MORTE A BALA / LESÃO CORPORAL A BALA / TENTATIVA	LOCAL: RUA ARI DE SÁ CAVALCANTE, BARRA DO CEARÁ. VÍTIMAS LESIONADAS: 1a – MEN
MORTE A BALA 02H32	LOCAL: RUA HUMBERTO LOMEU, GRANJA PORTUGAL. VÍTIMA: ADRIANO F S G. SUSPEITO: I
MORTE A BALA 04H24	LOCAL: RUA TIAGO PEREIRA, BOM JARDIM. VÍTIMA: RICHARD M S. SUSPEITO: NÃO IDENTI
MORTES A BALA 06H42	LOCAL: RUA EDUARDO ARAÚJO, VILA VELHA. VÍTIMAS: 1a - JOSÉ V C 2a - ANTÔNIO R C A. I
MORTE A BALA 12H41	LOCAL: RUA BRÁS CUBAS, PLANALTO AIRTON SENA. VÍTIMA: FRANCISCO A G S. SUSPEITO
PORTE ILEGAL DE ARMA (FLAGRANTE) 05H10	LOCAL: RUA COMENDADOR FRANCISCO DE FRANCESCO DI ÂNGELO, PRAIA DO FUTURO. APREENDIDA: PISTOLA CALIBRE 40. SUSPEITO CONDUZIDO PARA A DELEGACIA COMPETE
LESÃO CORPORAL A FACA (FLAGRANTE) 00H57	LOCAL: RUIA ANA NERI, DAMAS. VÍTIMA: ANA P S S. SUSPEITO: LEANDRO DA SILVA UCHOA COMPETENTE.
ROUBO A PESSOA (FLAGRANTE) 06H26	LOCAL: AV BEIRA MAR, PRAIA DE IRACEMA. VÍTIMA: CELIO C C J. SUSPEITO: ADRIANO DA S DELEGACIA COMPETENTE.
APREENSÃO DE ENTORPECENTES 08H26	LOCAL: TRAVESSA SÃO JOÃO, VICENTE PINZON. MATERIAL APREENDIDO: MACONHA, CRA
FURTO A PESSOA (FLAGRANTE) 09H34	LOCAL: RUA ILDEFONSO ALBANO, PRAIA DE IRACEMA. VÍTIMA: FRANCISCO A O S SUSPEI CONDUZIDO PARA A DELEGACIA COMPETENTE.
LESÃO CORPORAL A OUTROS (FLAGRANTE) 10H44	LOCAL: RUA ZÉLIA CORREIA DE SOUSA, ITAPERI. VÍTIMA: MIRAMA M F. SUSPEITO: FRANCI PARA A DELEGACIA COMPETENTE.
LESÃO CORPORAL A FACA (FLAGRANTE) 15H17	LOCAL: RUA PEDRO WILSOM, ITAPERI. VÍTIMA: TADEU J S M. SUSPEITA: MARIA JOSÉ BRAG COMPETENTE.
LESÃO CORPORAL A BALA (FLAGRANTE) 20H11	LOCAL: RUA FRANCISCO MATIAS, SABIAGUABA. VÍTIMA: FRANCISCO B V S. SUSPEITO/VÍT CONDUZIDO PARA A DELEGACIA COMPETENTE.
AGRESSÃO/VIAS DE FATO (FLAGRANTE) 22H31	LOCAL: RUA DELMIRO DE FARIAS, JARDIM AMÉRICA. VÍTIMA: MARIA A M S. SUSPEITO: JOÃ PARA A DELEGACIA COMPETENTE.

Fonte – Elaborada pelo Autor.

Figura 7 – Quantidade de ocorrências criminais por mês



Fonte – Elaborada pelo Autor.

5.2 Extração de entidades

Após os dados serem transformados para arquivos no formato CSV, a extração de entidades foi realizada. Para essa tarefa foi implementado um algoritmo na linguagem Python onde, para cada arquivo CSV, o mesmo realiza uma leitura da coluna HISTÓRICO DA OCORRÊNCIA e cria um novo arquivo CSV contendo as novas colunas referentes a 12 entidades sobre crimes. As entidades são: LOCAL, BAIRRO, SUSPEITO, SUSPEITOS, VÍTIMA, VÍTIMAS, VÍTIMA FATAL, VÍTIMAS LESIONADA, ARMA APREENDIDA, MATERIAL APREENDIDO e PLACA. Quando não existe informação sobre uma determinada entidade é atribuído o texto "nulo"no campo correspondente à ela. A Figura 8 ilustra um exemplo de arquivo após a extração das entidades SUSPEITO, VEÍCULO, VÍTIMA, VÍTIMAS e ARMA APREENDIDA. A Figura 9 apresenta uma nuvem de palavras com os resultados obtidos da extração da entidade BAIRRO. A nuvem de palavras é um indício de quais bairros são os mais violentos.

Figura 8 – Exemplo de entidades após a extração

SUSPEITO:	VEÍCULO:	VÍTIMA:	VÍTIMAS:	ARMA APREENDIDA:
nulo	nulo	VALDENIR S	nulo	nulo
nulo	nulo	ANTÔNIO F	nulo	nulo
nulo	nulo	nulo	1o FRANCISCO Y	nulo
nulo	nulo	FRANCISCO C	nulo	nulo
MARIA RAYNA MONTEIRO DO NASCIMENTO, 27 ANOS	nulo	nulo	nulo	nulo
MENOR DE IDADE	nulo	nulo	nulo	nulo
nulo	ONIX 1	nulo	nulo	nulo
nulo	nulo	nulo	nulo	nulo
nulo	CG 150 TITAN ES, VERMELHA, 06/07, HXV8922	nulo	nulo	nulo
nulo	ASTRA SEDAN CD, PRATA, 2002, JWT8475	nulo	nulo	nulo
nulo	YBR 125K, PRETA, 2008, NHY3676	nulo	nulo	nulo
nulo	NXR150 BROS ESD,PRETA, 2013, ORN5362	nulo	nulo	nulo
nulo	NXR150 BROS KS, AMARELO, 2009, NQQ2836	nulo	nulo	nulo
HOMENS NÃO IDENTIFICADOS	nulo	JONATHAN B	nulo	nulo
nulo	nulo	JOÃO VU	nulo	nulo
NÃO IDENTIFICADO	nulo	SEXO MASCULINO, SEM IDENTIFICAÇÃO	nulo	nulo
1o - FRANCISCO GABRIEL DE MOURA DA SILVA, 21 ANOS	nulo	nulo	nulo	REVÓLVER CALIBRE 38
MENOR DE IDADE	nulo	nulo	nulo	REVÓLVER, CALIBRE 38
nulo	nulo	AUGUSTO A	nulo	REVÓLVER, CALIBRE 38
nulo	nulo	MENOR DE IDADE	nulo	nulo
JOÃO CARLOS LOPES LINHARES, 20 ANOS	nulo	nulo	nulo	nulo
nulo	nulo	nulo	nulo	nulo
ADLER DA SILVA RIBEIRO, 32 ANOS	nulo	MARIA L	nulo	nulo
JOÃO CHARLES GOMES MENDES, 35 ANOS	nulo	nulo	nulo	nulo
nulo	nulo	nulo	1o – RAILTON W	nulo
ROBSON GARCIA DE LIMA, 20 ANOS	nulo	nulo	nulo	nulo
nulo	nulo	MESSIAS F	nulo	nulo
MARCOS NATANAEL LAUREANO DE ALENCAR, 19 ANOS	nulo	ANA S	nulo	nulo

Fonte – Elaborada pelo Autor.

5.3 Georreferenciamento da localização dos crimes

Na quarta etapa o georreferenciamento dos locais dos crimes foi realizado. Essa informação de geolocalização foi utilizada em alguns momentos durante o trabalho, principalmente pelo algoritmo DBSCAN na busca por regiões maior densidade de criminalidade. Para que a etapa do georreferenciamento ocorresse, foi implementado um algoritmo na

Figura 9 – Nuvem de palavras com dados da entidade bairro dos meses de janeiro à maio de 2017

Fonte – Elaborada pelo Autor.

linguagem Python que lê os arquivos CSV já extraídos e que contenham os locais dos crimes em uma coluna específica e então, através de uma Interface do Google Maps API, ela retorna a latitude e a longitude do local passado à mesma. Ao final, o arquivo CSV, que acabara de ser lido, é atualizado adicionando-se novas duas colunas que correspondem a longitude e a latitude respectivamente de cada crime. A Figura 10 ilustra um exemplo das colunas LOCAL, LATITUDE e LONGITUDE de um arquivo após o georreferenciamento dos locais.

Com as informações de geolocalização também foram criados mapas de calor com a intenção de visualizar as regiões com maior quantidade de ocorrências criminais. As Figuras 11, 12, 13, 14 e 15 mostram os mapas de calor referentes aos meses de janeiro, fevereiro, março, abril e maio respectivamente. Pode-se ver com os mapas de calor que a região oeste da cidade de Fortaleza teve, na grande maioria dos meses, o maior número de ocorrências. Com essa análise é possível prever que a maioria das regiões identificadas na etapa de clusterização fará presente na região oeste. Para a criação dos mapas foi usada uma API do Google chamada Maps JavaScript API[2]. Esta API recebe como entrada dados de latitude e longitude e retorna uma imagem representando a intensidade dos dados passados.

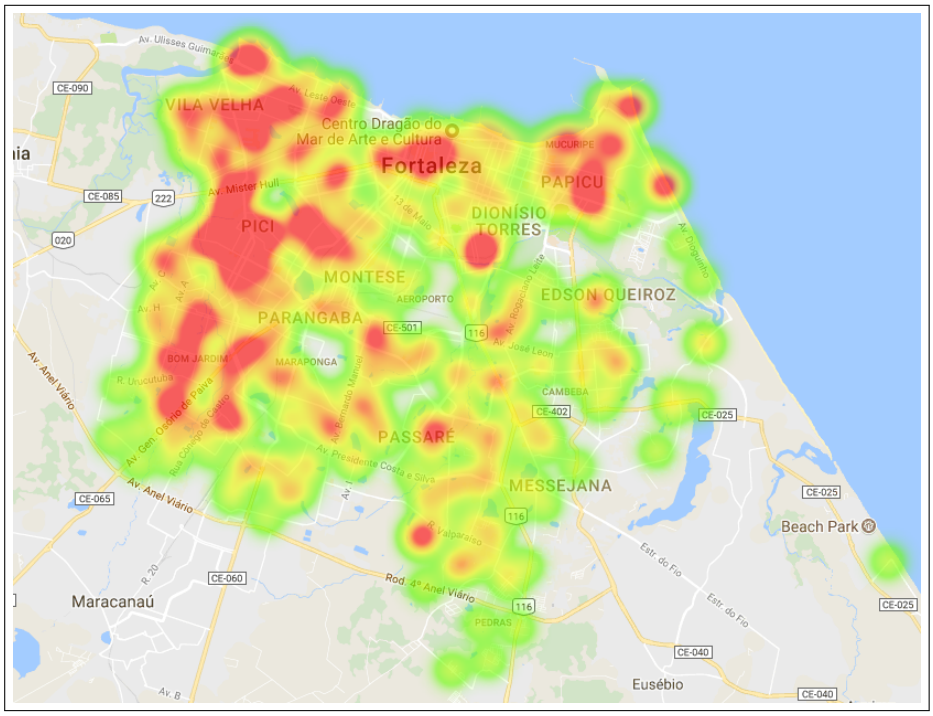
[2] https://developers.google.com/maps/documentation/javascript/heatmaplayer?hl=pt-br

Figura 10 – Exemplo de um arquivo após o georreferenciamento

LOCAL:	LONGITUDE	LATITUDE
RUA ARI DE SÁ CAVALCANTE, BARRA DO CEARÁ	-38.57497499999999	-3.7098556
RUA HUMBERTO LOMEU, GRANJA PORTUGAL	-38.6036134	-3.781712
RUA TIAGO PEREIRA, BOM JARDIM	-38.6243887	-3.8007201
RUA EDUARDO ARAÚJO, VILA VELHA	-38.6062514	-3.714283
RUA BRÁS CUBAS, PLANALTO AIRTON SENA	-38.5748912	-3.8247518
RUA COMENDADOR FRANCISCO DE FRANCESCO DI ÂNGELO, PRAIA DO FUTURO	-38.4577078	-3.7318066
RUIA ANA NERI, DAMAS	-38.5476319	-3.7487024
AV BEIRA MAR, PRAIA DE IRACEMA	-38.5114375	-3.7202598
TRAVESSA SÃO JOÃO, VICENTE PINZON	-38.479289	-3.7254643
RUA ILDEFONSO ALBANO, PRAIA DE IRACEMA	-38.5136811	-3.7346953
RUA ZÉLIA CORREIA DE SOUSA, ITAPERI	-38.5580367	-3.8155992
RUA PEDRO WILSOM, ITAPERI	-38.5477798	-3.7905289
RUA FRANCISCO MATIAS, SABIAGUABA	-38.4515946	-3.8034455
RUA DELMIRO DE FARIAS, JARDIM AMÉRICA	-38.5459073	-3.7517107
RUA RIBEIRO JÚNIOR, JÓQUEI CLUBE	-38.5760965	-3.7724507

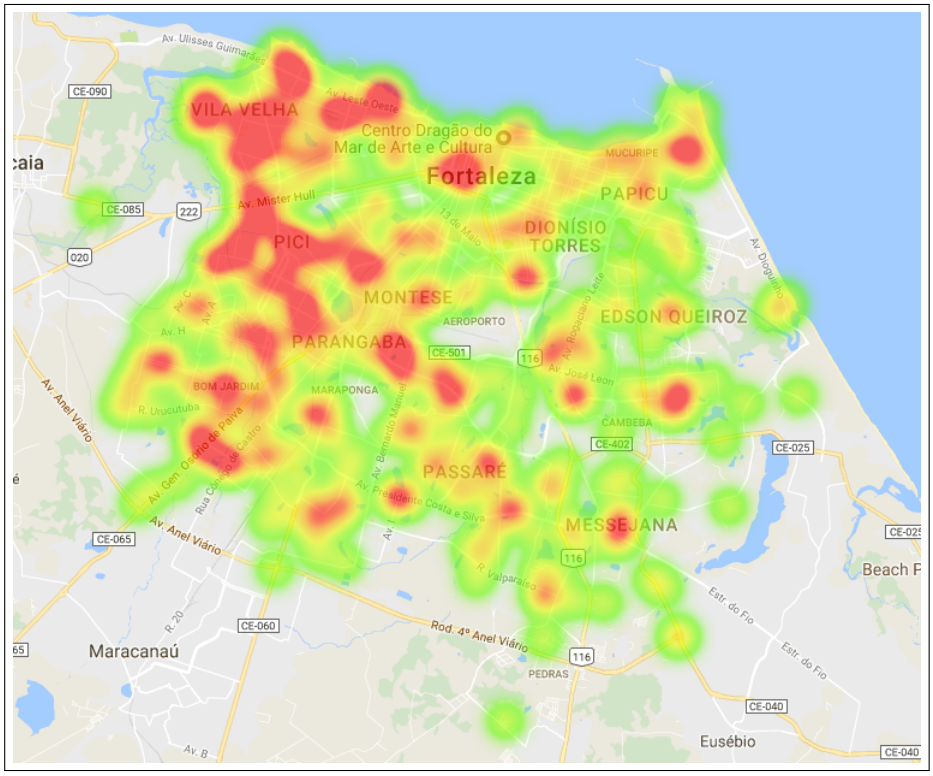
Fonte – Elaborada pelo Autor.

Figura 11 – Mapa de calor das ocorrências criminais de janeiro



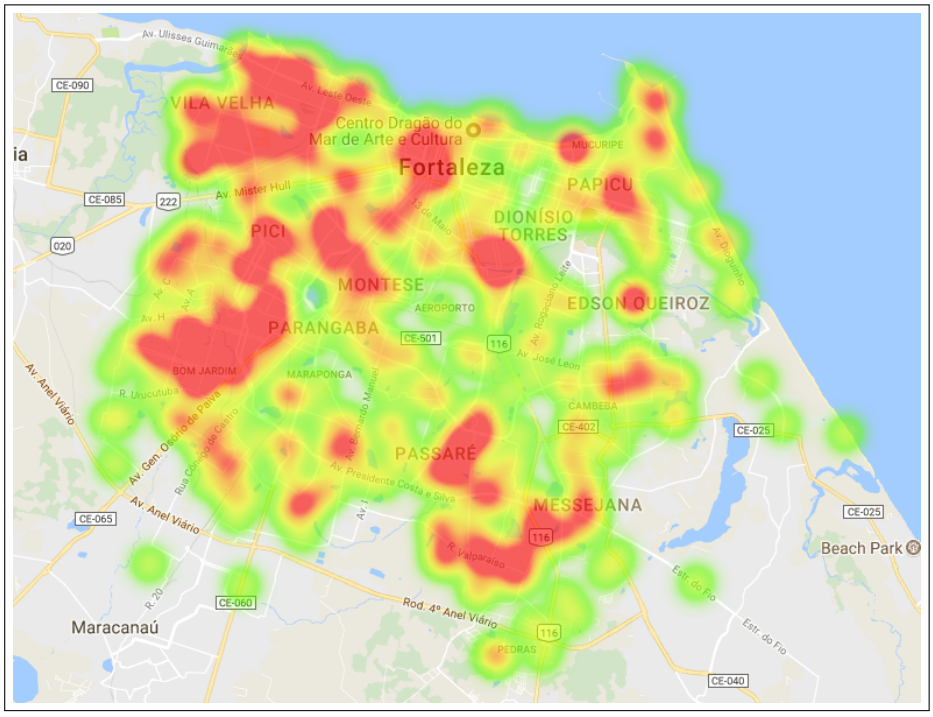
Fonte – Elaborada pelo Autor.

Figura 12 – Mapa de calor das ocorrências criminais de fevereiro



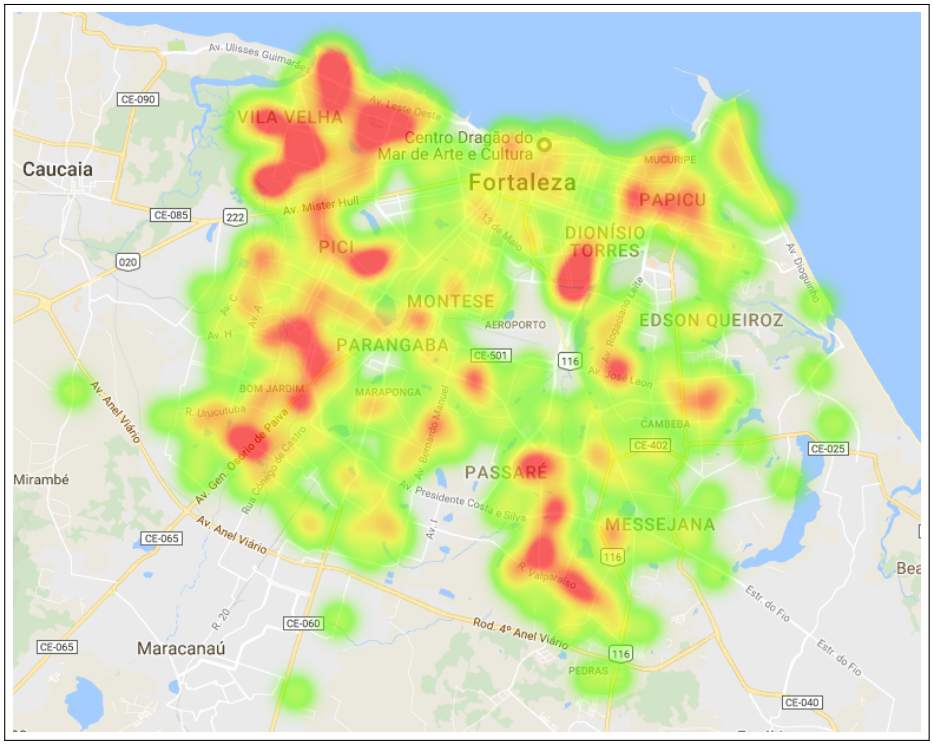
Fonte – Elaborada pelo Autor.

Figura 13 – Mapa de calor das ocorrências criminais de março



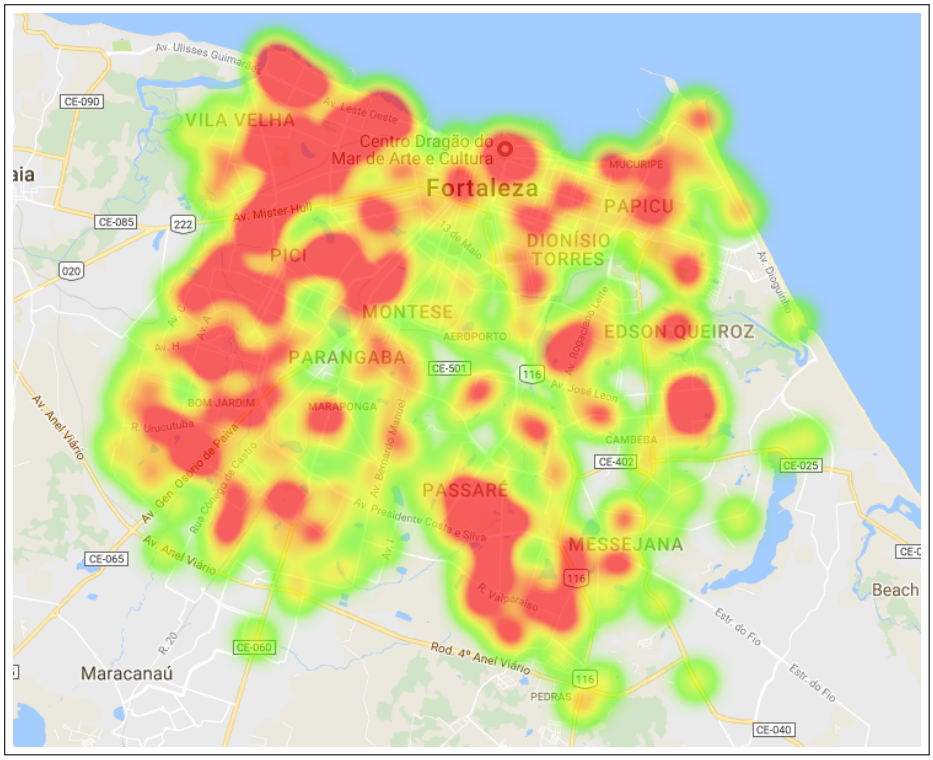
Fonte – Elaborada pelo Autor.

Figura 14 – Mapa de calor das ocorrências criminais de abril



Fonte – Elaborada pelo Autor.

Figura 15 – Mapa das de calor ocorrências criminais de maio



Fonte – Elaborada pelo Autor.

Pode-se ver que os bairros com mais ocorrências criminais na nuvem de palavras 9 correspondem com as regiões mais intensas dos mapas de calor na maioria dos meses. Alguns meses possuem regiões específicas com mais ocorrências, por conta disso os bairros dessas regiões não estão com maior intensidade na nuvem de palavras.

5.4 Clusterização dos dados criminais

Nesta etapa, os dados de ocorrências criminais foram clusterizados. A clusterização foi realizada para cada mês de janeiro à maio de 2017. Nesta etapa foi utilizado o algoritmo DBSCAN implementado e disponibilizado pela biblioteca scikit-learn[3]. O scikit-learn é uma biblioteca escrita em Python que implementa várias ferramentas para mineração e visualização de dados.

Os parâmetros escolhidos neste trabalho para o DBSCAN foram de *minPoints* = 5, métrica = *haversine* e *eps* = 0.6. Para auxiliar na métrica *haversine* o valor do *eps* foi convertido para quilômetros, ou seja, o valor de 0.6 definido representa na verdade a distância máxima de 0.6 quilômetros ou, em outra medida, 600 metros. Estes valores foram definidos após um processo de análise sobre os parâmetros do algoritmo e com os resultados obtidos nele. O valor de *minPoints* foi variado entre 3, 5 e 7 enquanto que os valores de *eps* foi variado de 0.1 quilômetros à 1.1 quilômetros. Após isto, foi decidido que os valores de *minPoints* = 5 e *eps* = 600 metros deixavam os dados mais bem distribuídos sobre regiões e bairros específicos da cidade de Fortaleza e, com isso, nosso objetivo de apresentar as regiões (*clusters*) com maior número de ocorrências será alcançado. A Tabela 3 mostra os resultados obtidos por cada mês após a clusterização.

Tabela 3 – Resultados da clusterização

Mês	**Quantidade de Clusters**	**Número de crimes**	**Outliers**
Janeiro	28	542	233
Fevereiro	31	492	209
Março	27	613	216
Abril	24	533	241
Maio	32	567	220

[3] http://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html#sklearn.cluster.DBSCAN

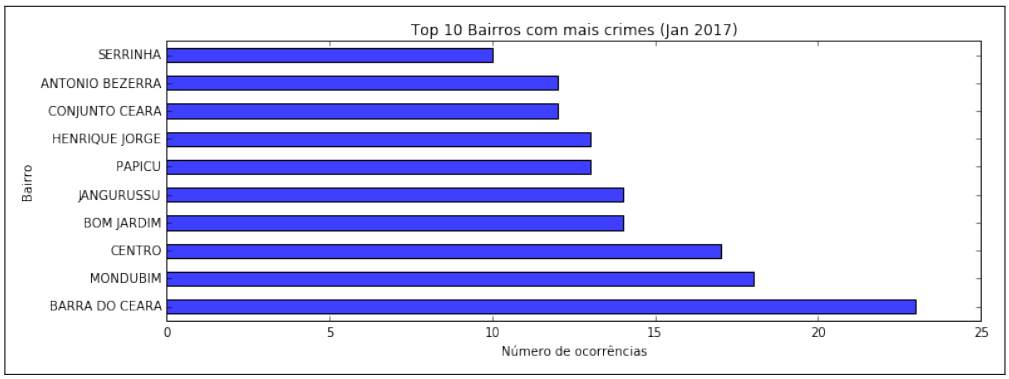
5.5 Análise dos dados

Após todo o processo de transformação e de clusterização das ocorrências criminais foi feita a análise dos resultados. A análise foi feita para cada mês e consistiu em identificar para cada mês os 10 bairros com mais ocorrências criminais; uma nuvem de palavras contendo os bairros que tiveram ocorrências criminais naquele período; a densidade de crimes em cada região identificada pela clusterização; um mapa de Fortaleza identificando as informações de cada região; os 5 principais bairros e ocorrências na região que teve a maior densidade de ocorrências criminais.

Nesta etapa os *clusters* são chamados de regiões, ou região quando se tratando de um *cluster*, uma vez que cada um deles representam as regiões identificadas na etapa de clusterização dos dados.

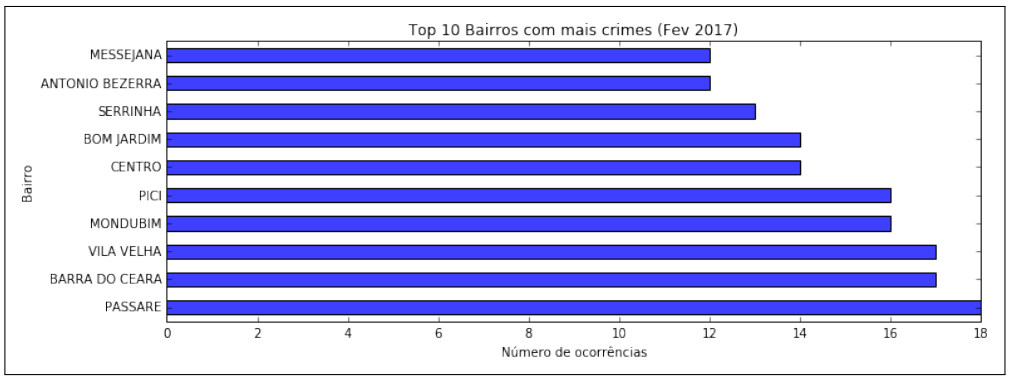
5.5.1 Os 10 bairros com mais ocorrências em cada mês

Figura 16 – 10 bairros com mais ocorrências criminais em janeiro



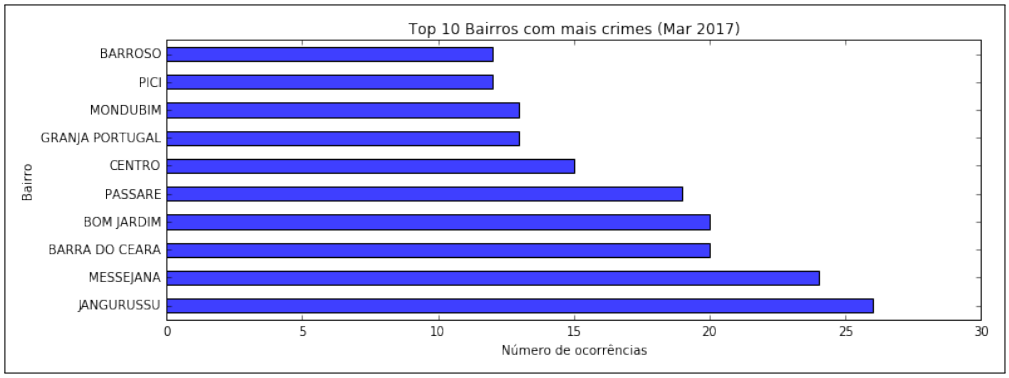
Fonte – Elaborada pelo Autor.

Figura 17 – 10 bairros com mais ocorrências criminais em fevereiro



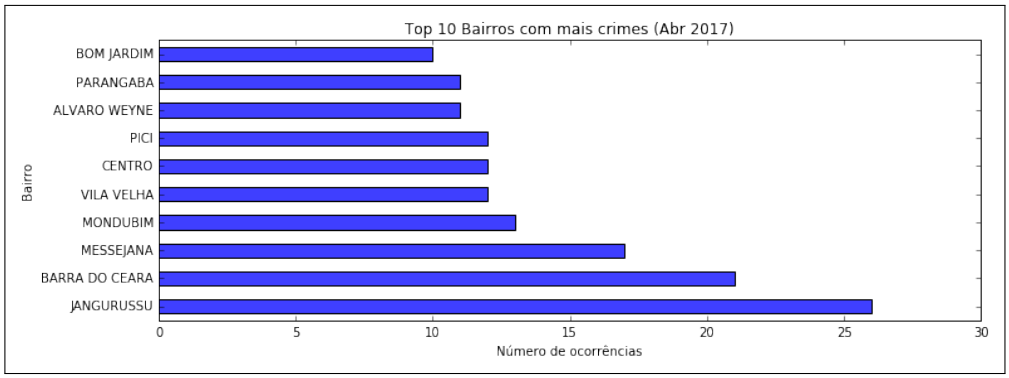
Fonte – Elaborada pelo Autor.

Figura 18 – 10 bairros com mais ocorrências criminais em março



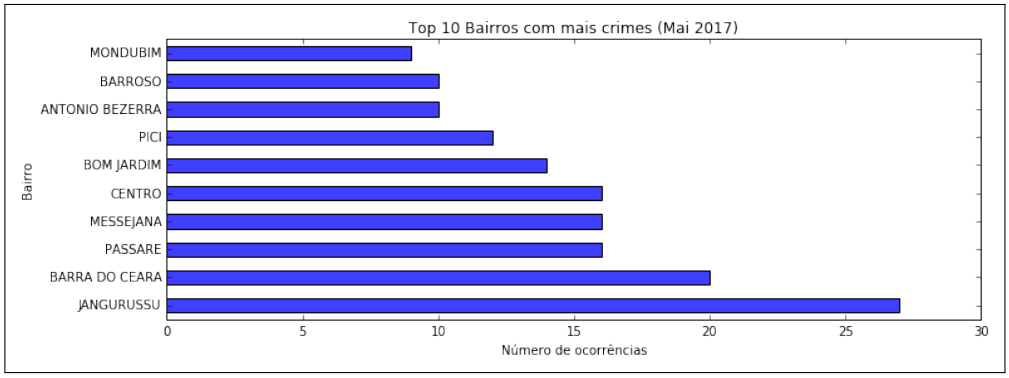
Fonte – Elaborada pelo Autor.

Figura 19 – 10 bairros com mais ocorrências criminais em abril



Fonte – Elaborada pelo Autor.

Figura 20 – 10 bairros com mais ocorrências criminais em maio



Fonte – Elaborada pelo Autor.

Como pode-se ver nas Figuras 16, 17, 18, 19, 20, alguns bairros como BARRA DO CEARA, JANGURUSSU e BOM JARDIM aparecem entre os 10 bairros com mais ocorrências durante todos os meses analisados. Também pode-se ver que o bairro que mais aparece em primeiro na análise é o JAGURUSSU com 3 aparições. Com essas informações já pode-se tomar o bairro JANGURUSSU como sendo um dos mais perigosos da cidade de Fortaleza nos primeiros 5 meses do ano de 2017, sem fazer nenhum tipo de análise com auxílio de algoritmos como a clusterização feita aqui neste trabalho. O bairro BARRA DO CEARA apareceu em todos os meses analisados entre os 3 primeiros, isso também já indica que esse bairro está constantemente com um alto número de ocorrências criminais.

5.5.2 Nuvem de palavras dos bairros com ocorrências criminais em cada mês

Figura 21 – Nuvem de palavras dos bairros com ocorrências criminais em janeiro

Fonte – Elaborada pelo Autor.

Figura 22 – Nuvem de palavras dos bairros com ocorrências criminais em fevereiro

Fonte – Elaborada pelo Autor.

Figura 23 – Nuvem de palavras dos bairros com ocorrências criminais em março

Fonte – Elaborada pelo Autor.

Figura 24 – Nuvem de palavras dos bairros com ocorrências criminais em abril

Fonte – Elaborada pelo Autor.

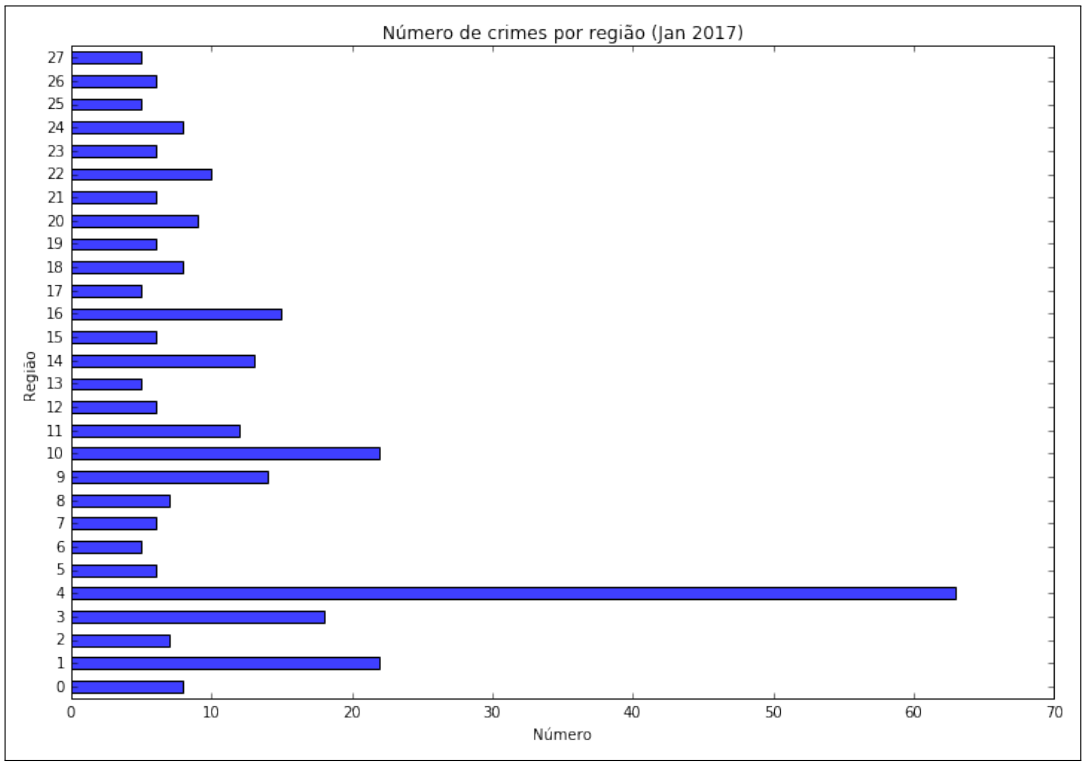
Figura 25 – Nuvem de palavras dos bairros com ocorrências criminais em maio

Fonte – Elaborada pelo Autor.

Nas Figuras 21, 22, 23, 24, 25, referentes as nuvens de palavras de janeiro, fevereiro, março, abril e maio, respectivamente, pode-se ver que os bairros com mais ocorrências ficam mais centralizados e com letras maiores, enquanto que os com menos ocorrências ficam mais distantes do centro e com letras menores. Também é possível ver que, assim como na análise sobre os 10 bairros com mais ocorrências para cada mês, o bairro JANGURUSSU aparece 3 vezes como o que obteve mais ocorrências criminais, uma vez que está centralizado na imagem e com as letras maiores que os demais bairros.

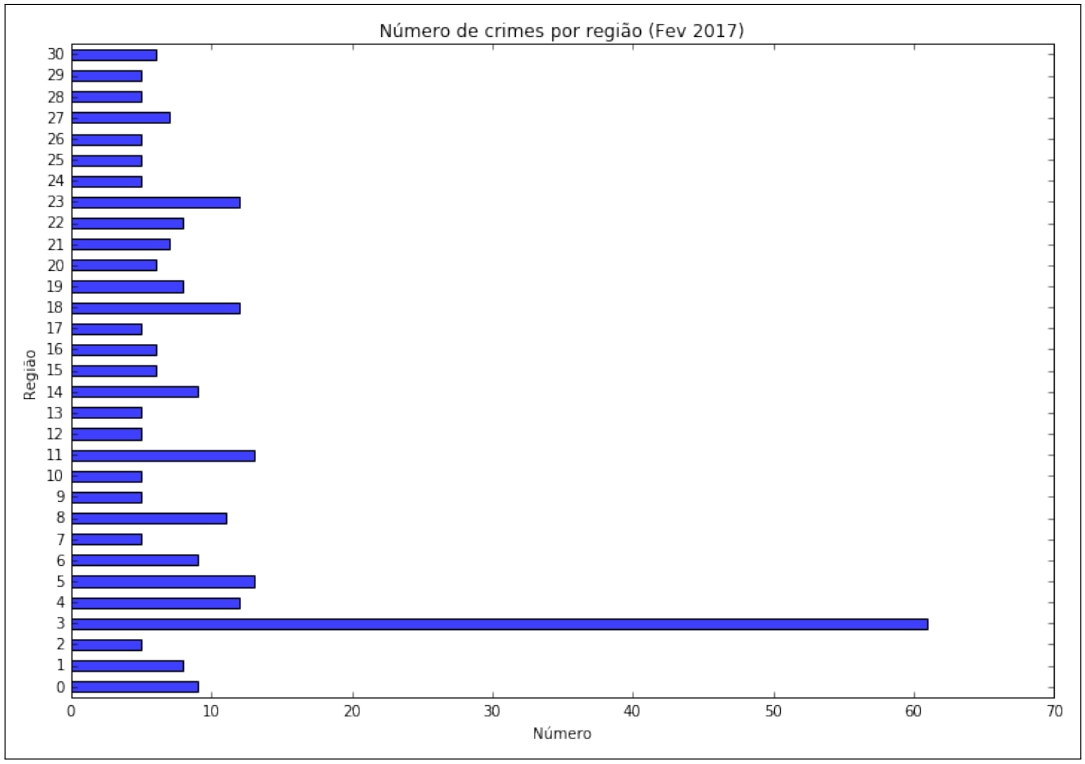
5.5.3 Regiões com maior densidade de ocorrências criminais em cada mês

Figura 26 – Quantidade de crimes por região em janeiro



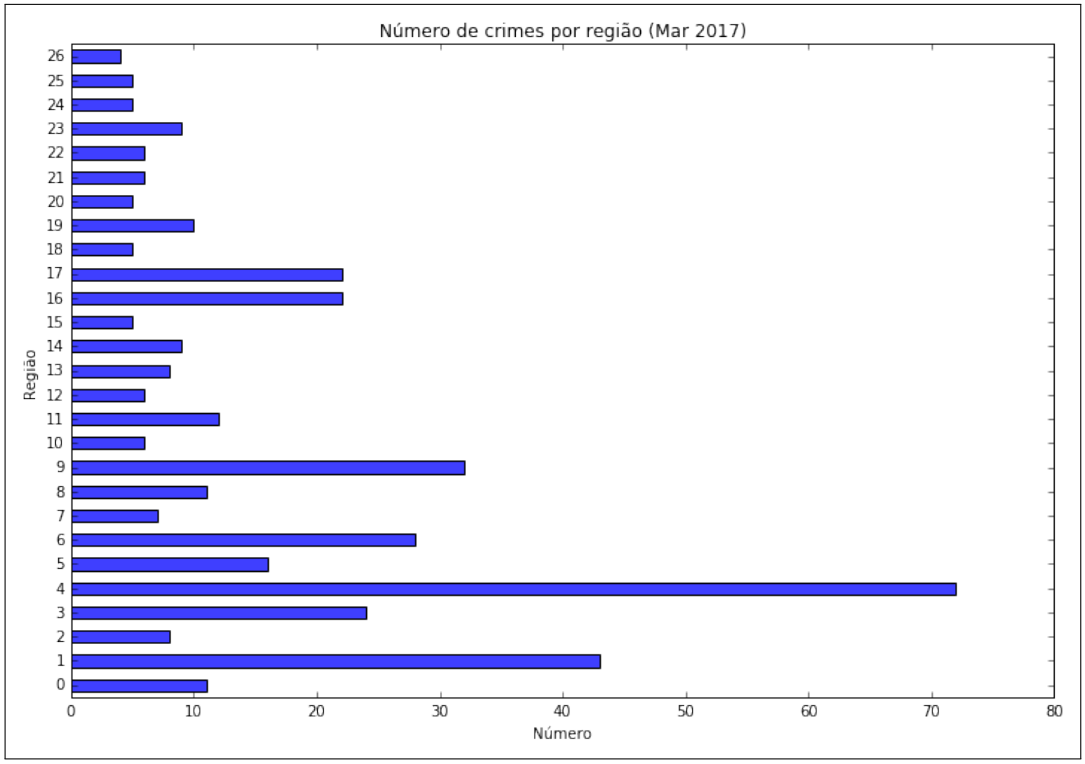
Fonte – Elaborada pelo Autor.

Figura 27 – Quantidade de crimes por região em fevereiro



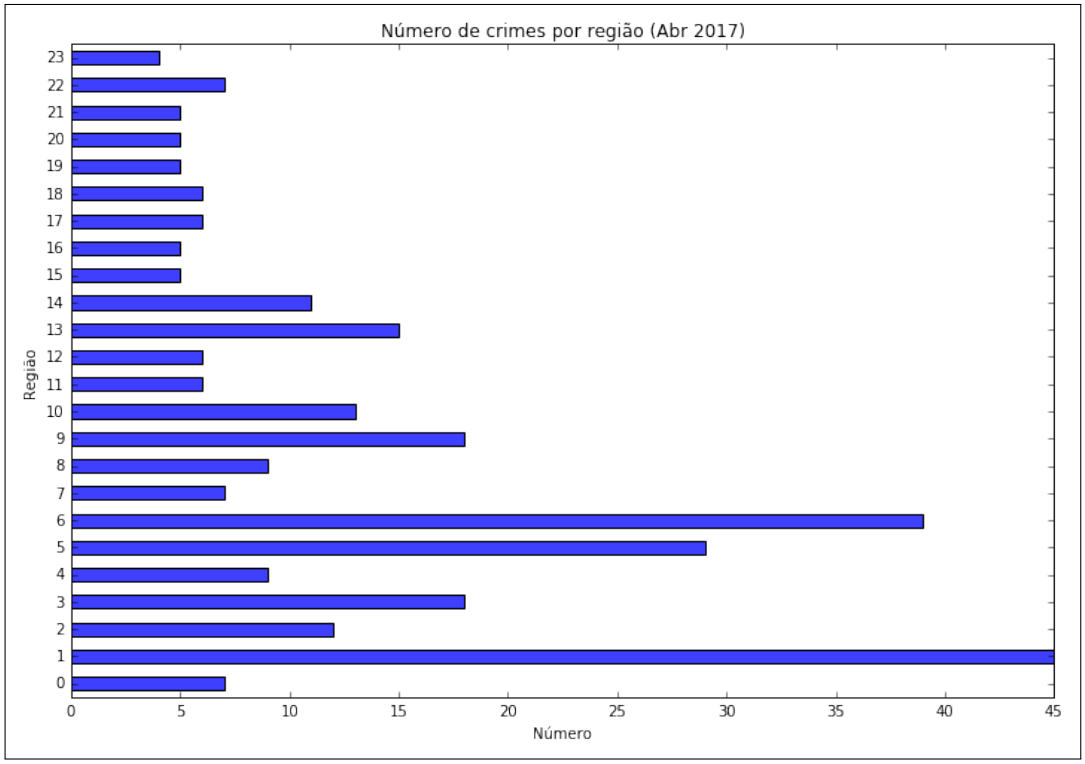
Fonte – Elaborada pelo Autor.

Figura 28 – Quantidade de crimes por região em março



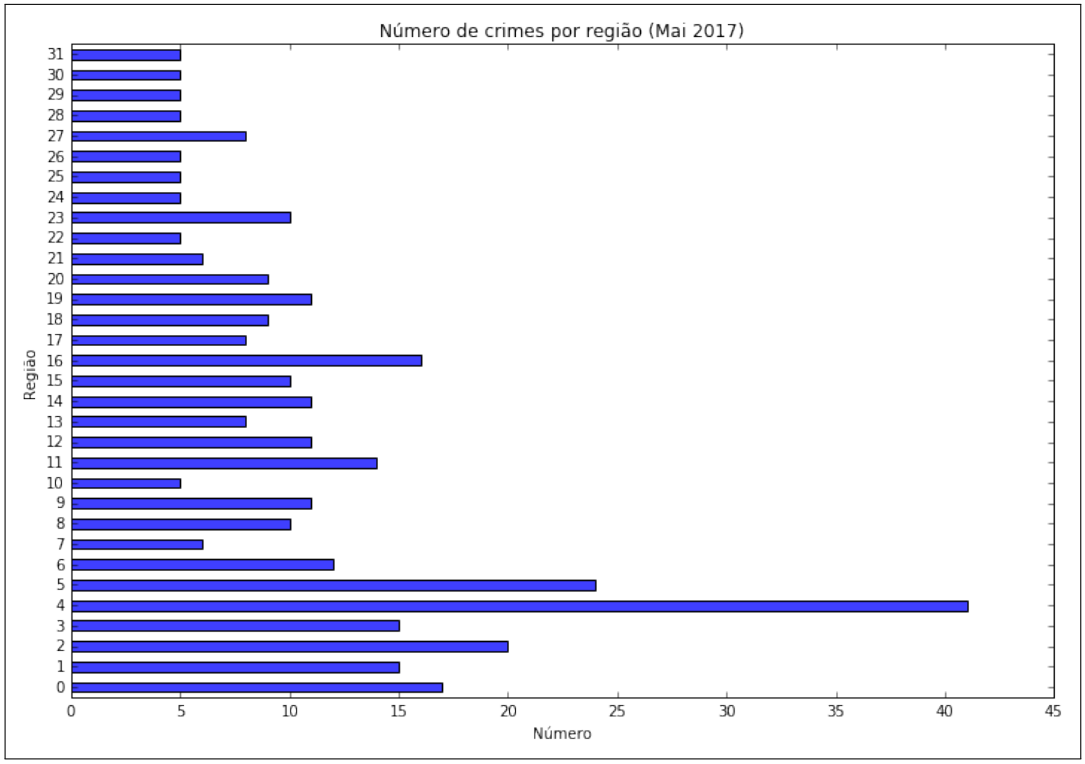
Fonte – Elaborada pelo Autor.

Figura 29 – Quantidade de crimes por região em abril



Fonte – Elaborada pelo Autor.

Figura 30 – Quantidade de crimes por região em maio



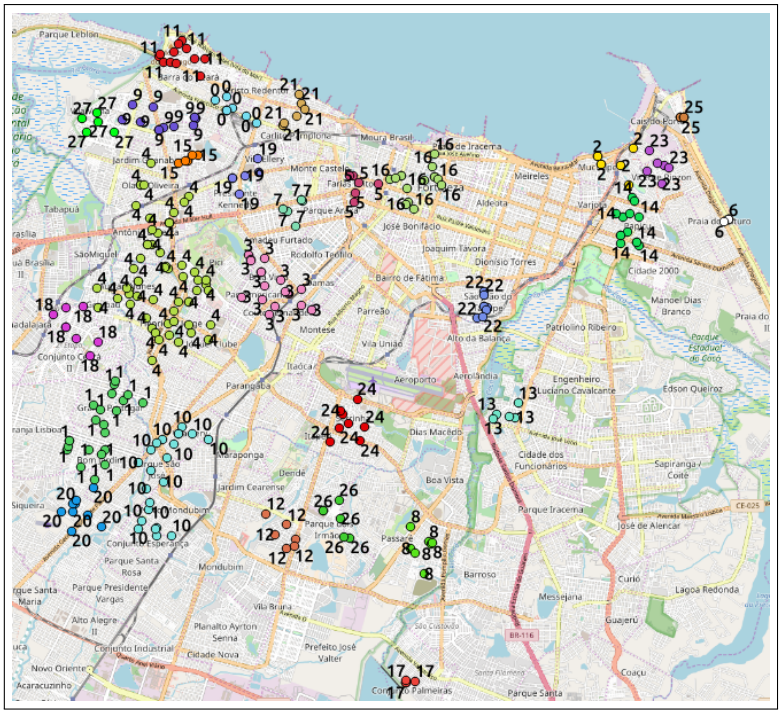
Fonte – Elaborada pelo Autor.

As Figuras 26, 27, 28, 29, 30, apresentam as regiões identificadas após a clusterização dos dados em cada mês. Como pode-se ver nas imagens, para cada mês, a região com mais ocorrências tinha, pelo menos, 6 ocorrências a mais que a região que vinha logo atrás da mesma. No geral, as regiões possuem uma quantidade similar de ocorrências. Também pode-se ver que o mês que teve a região com mais ocorrências foi março, pois neste mês a região 4 teve mais de 70 ocorrências identificadas. As regiões que chegaram mais próximo de alcançar esse número foi a 4 do mês de janeiro e a 3 do mês de fevereiro com mais de 60 ocorrências cada uma.

A seguir, na Seção 5.5.4, é apresentado mapas de Fortaleza identificando, para cada mês, todas as regiões encontradas, bem como a região que teve o maior número de ocorrências.

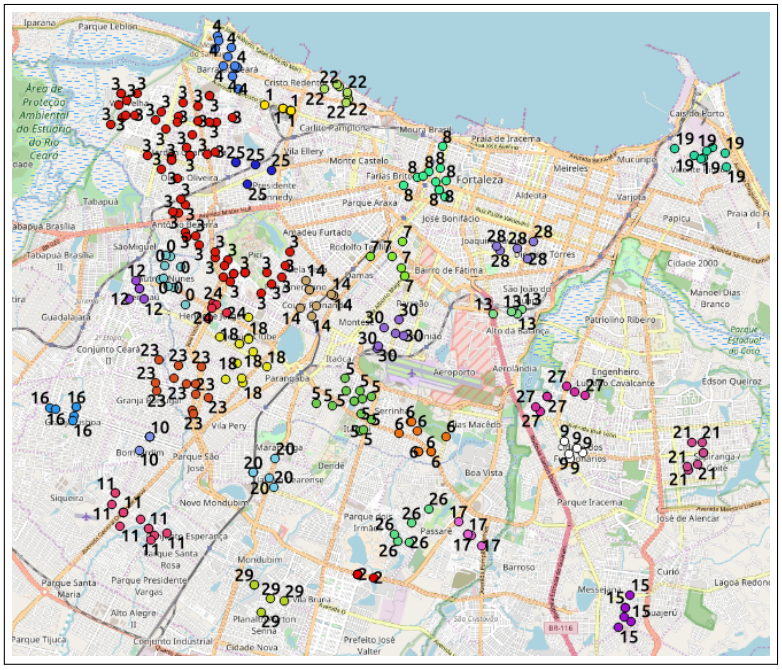
5.5.4 Mapas com as regiões identificadas em cada mês

Figura 31 – Mapa de janeiro com as regiões identificadas



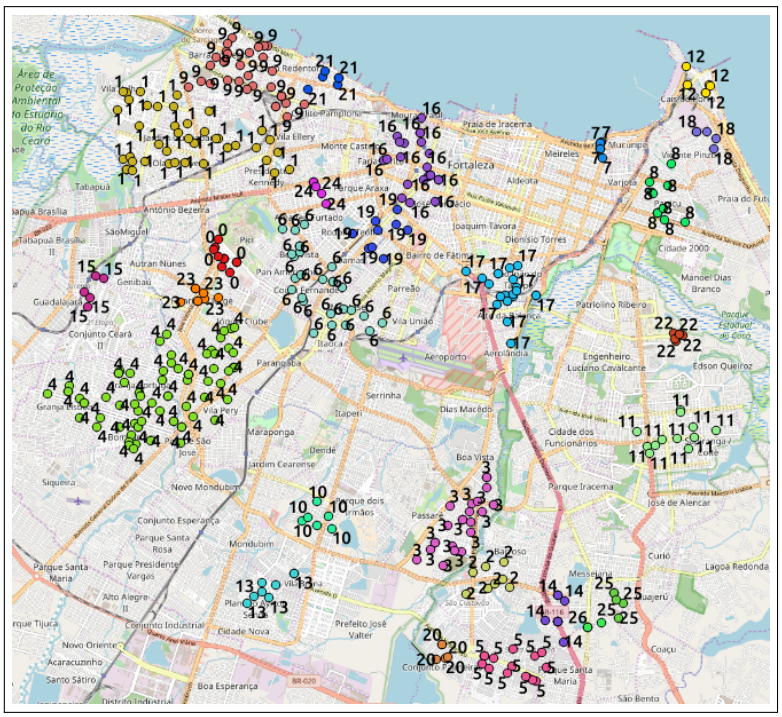
Fonte – Elaborada pelo Autor.

Figura 32 – Mapa de fevereiro com as regiões identificadas



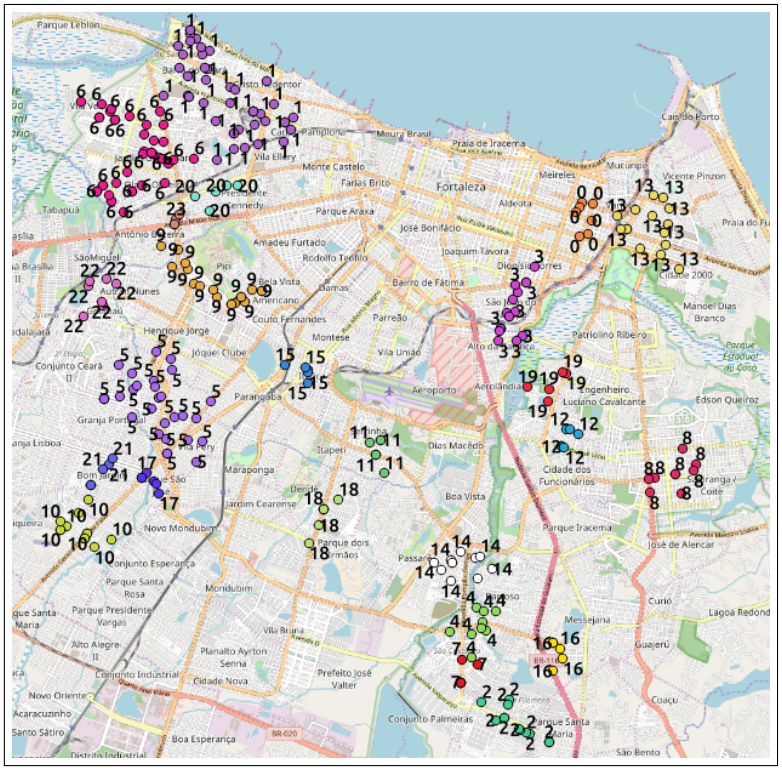
Fonte – Elaborada pelo Autor.

Figura 33 – Mapa de março com as regiões identificadas



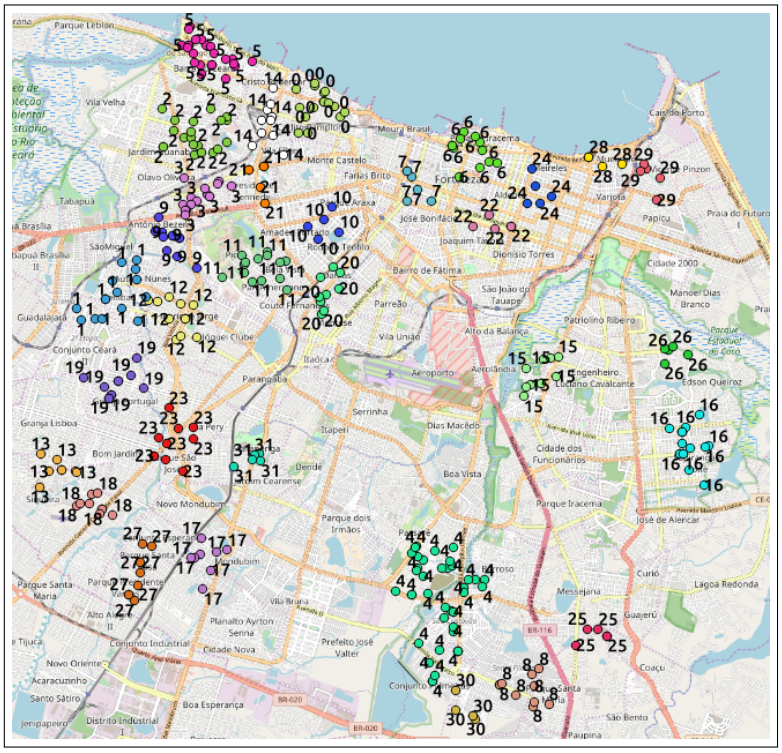
Fonte – Elaborada pelo Autor.

Figura 34 – Mapa de abril com as regiões identificadas



Fonte – Elaborada pelo Autor.

Figura 35 – Mapa de maio com as regiões identificadas

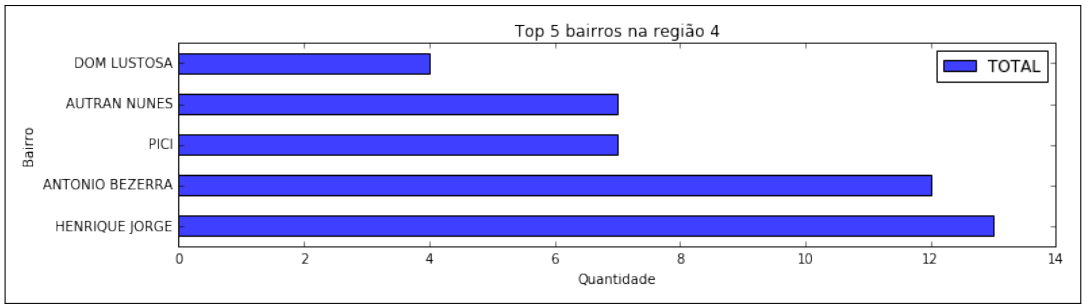


Fonte – Elaborada pelo Autor.

Com as Figuras 31, 32, 33, 34, 35, é possível agora visualizar onde exatamente fica cada região, identificada pela clusterização, no mapa de Fortaleza. Para melhor visualização, foi removido os *outliers* dos mapas, ou seja, as ocorrências que não pertencem a nenhuma região, pois dificultavam a identificação e visualização das regiões.

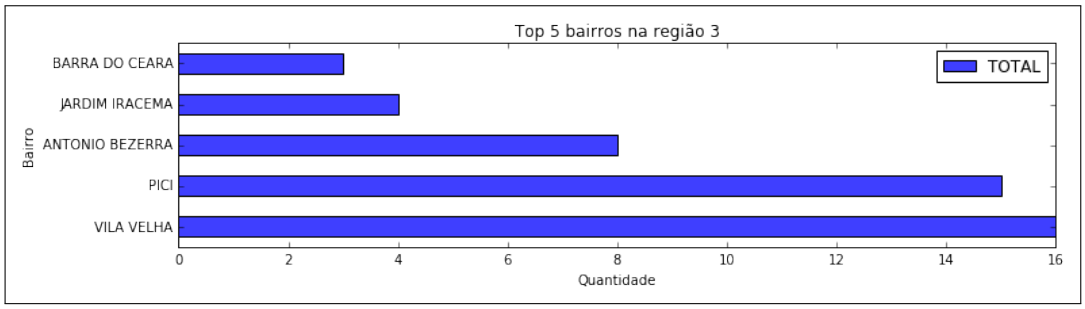
5.5.5 Top 5 bairros com mais crimes da região com mais ocorrências em cada mês

Figura 36 – 5 bairros com mais ocorrências na região 4 janeiro



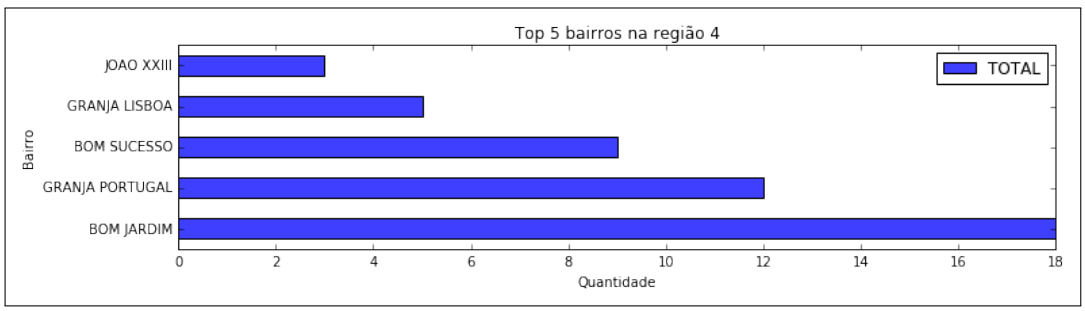
Fonte – Elaborada pelo Autor.

Figura 37 – 5 bairros com mais ocorrências na região 3 fevereiro



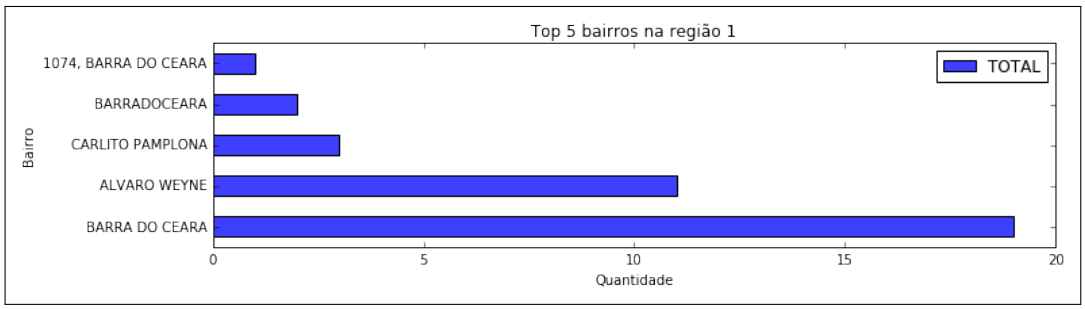
Fonte – Elaborada pelo Autor.

Figura 38 – 5 bairros com mais ocorrências na região 4 março



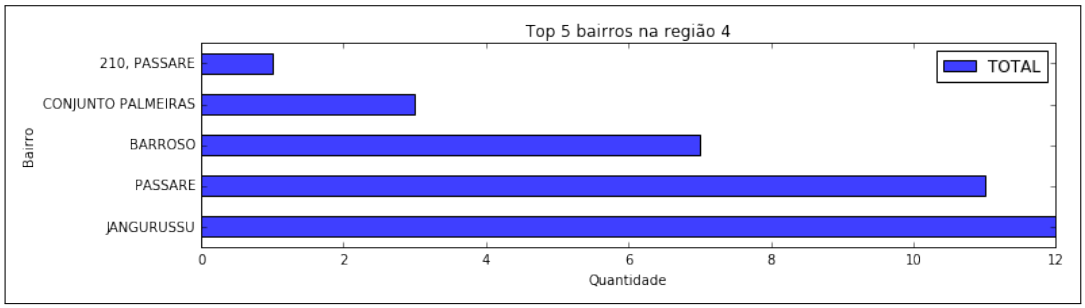
Fonte – Elaborada pelo Autor.

Figura 39 – 5 bairros com mais ocorrências na região 1 abril



Fonte – Elaborada pelo Autor.

Figura 40 – 5 bairros com mais ocorrências na região 4 maio



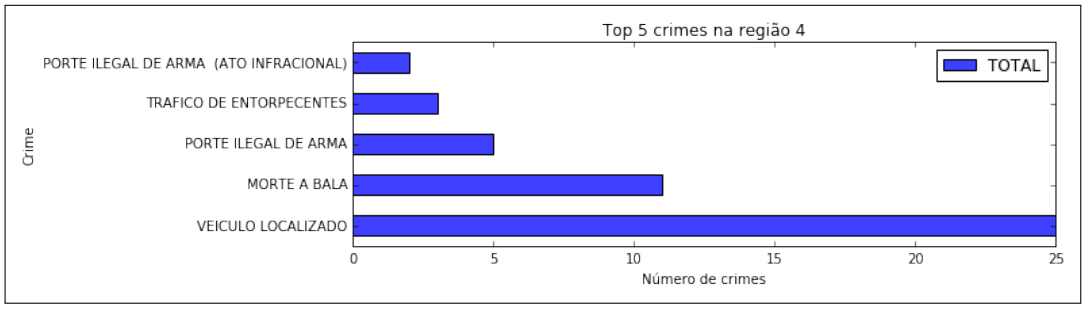
Fonte – Elaborada pelo Autor.

Como pode-se ver nas Figuras 36, 37, 38, 39, 40, os bairros com mais ocorrências em cada mês são HENRIQUE JORGE, VILA VELHA, BOM JARDIM, BARRA DO CEARA e JANGURUSSU. Dentre estes bairros, somente BARRA DO CEARA apareceu novamente entre os 5 bairros com mais ocorrências. Para ser mais preciso, isto ocorreu no mês de fevereiro, estando ele na última posição do Top 5 com 3 ocorrências. Alguns outros meses mostram-se presentes por mais de um mês no Top 5, mas nunca como primeiros colocados como, por exemplo, PICI e ANTONIO BEZERRA. Isso indica que estes bairros possuem uma quantidade de ocorrências frequente.

Como podemos ver, na análise do mês de abril, bairros como "1074, BARRA DO CEARA"e "BARRADOCEARA"são, provavelmente, os mesmos bairros, porém, por conta da falta de padrão nos dados, acabam sendo identificados como distintos pelo algoritmo na análise dos dados.

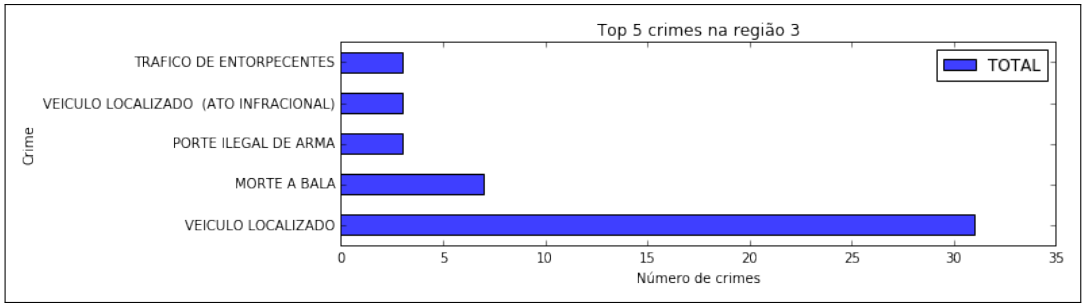
5.5.6 Top 5 crimes da região com mais ocorrências em cada mês

Figura 41 – Principais ocorrências da região 4 em janeiro



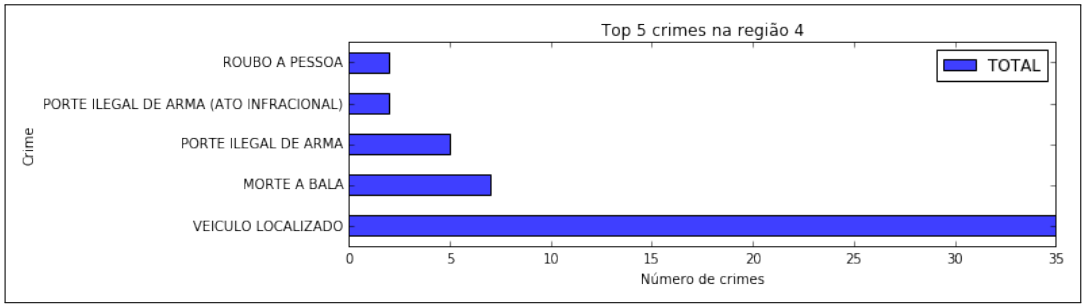
Fonte – Elaborada pelo Autor.

Figura 42 – Principais ocorrências da região 3 em fevereiro



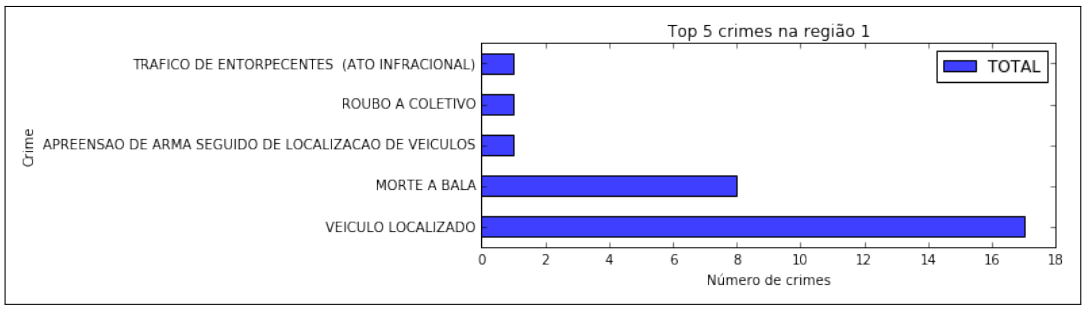
Fonte – Elaborada pelo Autor.

Figura 43 – Principais ocorrências da região 4 em março



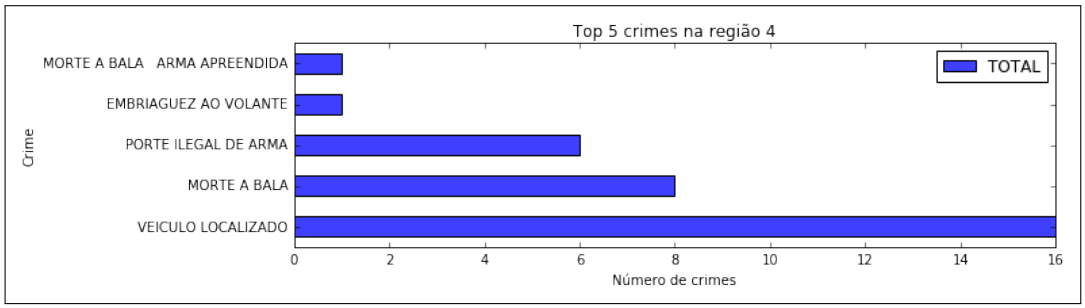
Fonte – Elaborada pelo Autor.

Figura 44 – Principais ocorrências da região 1 em abril



Fonte – Elaborada pelo Autor.

Figura 45 – Principais ocorrências da região 4 em maio



Fonte – Elaborada pelo Autor.

Como pode-se ver nas Figuras 41, 42, 43, 44, 45, as ocorrências mais comuns em todas regiões são VEICULO LOCALIZADO e MORTE A BALA. Também pode-se ver que PORTE ILEGAL DE ARMA se faz presente em quase todas as regiões analisadas. Alguns outros crimes como ROUBO A PESSOA e TRAFICO DE ENTORPECENTES também aparecem dentre as principais ocorrências, mas com poucos casos se comparado com MORTE A BALA e VEICULO LOCALIZADO.

O mês de maio foi o que MORTE A BALA mais se aproximou de VEICULO LOCALIZADO. Isso pode ter acontecido porque foi o mês que teve, na região analisada, o menor número de ocorrências para VEICULO LOCALIZADO se comparado com as regiões analisadas nos demais meses.

Com o grande número de ocorrências para VEICULO LOCALIZADO e MORTE A BALA, pode-se dizer que os esforços iniciais de combate ao crime nas regiões identificadas deve ser para estes crimes, pois ambos estavam presentes em primeiro e segundo lugares, respectivamente.

5.6 Discussão

O resultados apresentados na Seção anterior nos mostram algumas informações importantes. Dentre elas destacamos os crimes que tiveram os maiores números de ocorrências, se somados todos os meses analisados. Estes crimes foram VEÍCULO LOCALIZADO e MORTE A BALA, pois em todos os meses ambos estavam em primeiro e segundo lugares respectivamente. Essas ocorrências podem indicar, por exemplo, que a maior parte dos esforços de combate, nessas regiões, devem ser voltadas a estes tipos de crimes.

Também foi possível identificar regiões que continham bairros específicos de Fortaleza que, através de uma análise comum, seria bastante difíceis de serem identificadas.

Uma outra informação interessante que podemos tirar a partir da análise é que todas as regiões identificadas pelo algoritmo estão em pontos e bairros diferentes da cidade, apesar de que alguns fiquem próximos uns dos outros. Dentre as 5 regiões com maior ocorrência em cada mês, os 5 bairros com mais ocorrências de cada região foram, respectivamente, HENRIQUE JORGE, VILA VELHA, BOM JARDIM, BARRA DO CEARA e JANGURUSSU. Com essa informação também identificamos que a maioria das ocorrências ocorrem na região oeste do mapa de Fortaleza, assim como mostram os mapas de calor na Seção 5.3.

Por fim, caso não seja de interesse analisar as ocorrências criminais por região ou por tipo, é possível ver quais são os bairros que tiveram mais ocorrências em cada mês, assim, os esforços de combate à criminalidade podem ser centradas nesses bairros.

6 CONCLUSÕES E TRABALHOS FUTUROS

A criminalidade está espalhada pelo mundo inteiro e na cidade de Fortaleza, cidade essa que serviu de objeto de análise deste estudo, isso não é diferente. Para que seja realizado o combate a estes crimes é necessário muito investimento em segurança e em materiais de apoio correlacionados, além de mão-de-obra por parte dos agentes de segurança. Porém, através da análise de dados este alto investimento pode ser reduzido, uma vez que a análise propicia uma ação mais inteligente sobre as ocorrências criminais com base nas informações obtidas em ocorrências passadas, fazendo assim que no combate aos crimes todos os recursos como, mão-de-obra, materiais como armas, munição, coletes, dentre outros, sejam melhor alocados.

Levando em consideração esses fatos, esse trabalho apresentou uma análise sobre ocorrências criminais da cidade de Fortaleza através de dados disponibilizados pela Secretaria de Segurança Pública e Defesa Social do estado do Ceará. A análise apresenta informações sobre crimes e os bairros onde estes ocorrem. Também foi realizada a clusterização dos dados a partir das informações dos locais dos crimes e apresentadas quais as regiões que possuem mais ocorrências e quais são as mesmas. Com isso, ao fim desse estudo, foi possível concluir que a análise dos dados pode fornecer informações bastante úteis para órgãos de segurança pública e, também, para a população em geral que deseja obter informações como, por exemplo, os crimes mais comuns em determinadas regiões na cidade de Fortaleza.

Muitas outras análises podem ser feitas com estes dados disponibilizados pela Secretaria de Segurança Pública. Como trabalhos futuros, podem ser realizadas análises como predição de crimes, além de outras formas de estruturação dos dados para o uso por terceiros. Uma outra proposta seria criar um portal para disponibilizar toda a informação gerada pela análise, fazendo assim que esse conteúdo fique acessível para qualquer pessoa com acesso à Internet. Esse portal seria diferente do WikiCrimes porque trataria de informações exclusivas da cidade de Fortaleza, além de obter dados de boletins de ocorrências reais registrados pela Secretaria de Segurança Pública e Defesa Social. Também é possível fazer análises de similaridade entre as ocorrências em si, além de outras dimensões de análise.

REFERÊNCIAS

AMARAL, D. O. F. D. **O reconhecimento de entidades nomeadas por meio de conditional random fields para a língua portuguesa**. Tese (Doutorado) — Pontifícia Universidade Católica do Rio Grande do Sul, 2013.

AMARAL, J. E. d. **Uma análise do efeito dos gastos públicos estaduais em segurança pública, assistência social e educação sobre a criminalidade no Ceará para o período de 2010 a 2013**. Tese (Doutorado) — Universidade Federal do Ceará, 2015.

BIRANT, D.; KUT, A. St-dbscan: An algorithm for clustering spatial–temporal data. **Data & Knowledge Engineering**, Elsevier, v. 60, n. 1, p. 208–221, 2007.

CERQUEIRA, D. R. d. C. **Causas e consequências do crime no Brasil**. Rio de Janeiro: BNDES, 2014. v. 1.

ESTER, M.; KRIEGEL, H.-P.; SANDER, J.; XU, X. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: **Kdd**. München: Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining, 1996. v. 96, n. 34, p. 226–231.

EVANDRO, B. F.; GABRIEL, C. C.; RENATA, V.; ALINE, A. V. **Reconhecimento de Entidades Nomeadas para o Português Usando o OpenNLP**. [S.n: S.l], 2015.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The kdd process for extracting useful knowledge from volumes of data. **Communications of the ACM**, ACM, v. 39, n. 11, p. 27–34, 1996.

FRANÇA, R. S. de; AMARAL, H. J. C. do. Mineração de dados na identificação de grupos de estudantes com dificuldades de aprendizagemno ensino de programação. **RENOTE**, v. 11, n. 1, 2013.

FURTADO, V.; AYRES, L.; VASCONCELOS, J.; ALVES, R.; OLIVEIRA, M. D. Wikicrimes-um sistema colaborativo para mapeamento criminal. **Proc. 35th InfoBrasil. Brazil**, 2008.

GAVA, É. M.; FELIPPE, G.; MADEIRA, K.; PALHANO, M. B.; GARCIA, M. C. de M.; MARTINS, P. J.; SIMÕES, P. W. T. de A. O algoritmo density-based spatial clustering of applications with noise (dbscan) na clusterização dos indicadores de dados ambientais. **Anais SULCOMP**, v. 6, 2013.

JAYAWEERA, I.; SAJEEWA, C.; LIYANAGE, S.; WIJEWARDANE, T.; PERERA, I.; WIJAYASIRI, A. Crime analytics: Analysis of crimes through newspaper articles. In: IEEE. **Moratuwa Engineering Research Conference (MERCon), 2015**. Moratuwa, Sri Lanka, 2015. p. 277–282.

KEYVANPOUR, M. R.; JAVIDEH, M.; EBRAHIMI, M. R. Detecting and investigating crime by means of data mining: a general crime matching framework. **Procedia Computer Science**, Elsevier, v. 3, p. 872–880, 2011.

LOPES, A. P.; SENA, J.; TORRES, K.; LOPES, A. O transtorno de estresse pós-traumático e a violência urbana. **Caderno de Graduação-Ciências Biológicas e da Saúde-UNIT-ALAGOAS**, v. 1, n. 2, p. 21–33, 2013.

LOPES, G. F.; PEDROSA, G. V.; FERNANDES, H. C.; BARCELOS, C. A. Eliminação de ruídos e medidas de similaridade-aplicações em melanomas. **Horizonte Científico**, v. 2, n. 1, 2008.

MANNING, C. **Information Extraction and Named Entity Recognition**. California: Stanford University, California, 2012.

NALDI, M. C. **Técnicas de combinação para agrupamento centralizado e distribuído de dados**. Tese (Doutorado) — Universidade de São Paulo, 2011.

NATH, S. V. Crime pattern detection using data mining. In: IEEE. **Web Intelligence and Intelligent Agent Technology Workshops, 2006. WI-IAT 2006 Workshops. 2006 IEEE/WIC/ACM International Conference on**. Hong Kong, 2006. p. 41–44.

OCHI, L. S.; DIAS, C. R.; SOARES, S. S. F. Clusterização em mineração de dados. **Instituto de Computação-Universidade Federal Fluminense-Niterói**, 2004.

PIMENTEL, E. P.; OMAR, N. Descobrindo conhecimentos em dados de avaliação da aprendizagem com técnicas de mineração de dados. In: **Anais do Workshop de Informática na Escola**. Campo Grande: XXVI Congresso da SBC, 2006. v. 1, n. 1.

PONCIANO, J. R. et al. **ToPI-uma abordagem online para identificar locais de interesse utilizando fotografias geo-referenciadas**. [S.l.]: Universidade Federal de Uberlândia, 2016.

QUEALY, K.; KATZ, M. S. **Gun homicides in Germany are about as common as deaths from thrown or falling objects in the United States**. New York, 2016. Disponível em: <https://www.nytimes.com/2016/06/14/upshot/compare-these-gun-death-rates-the-us-is-in-a-different-world.html>. Acesso em: 07 jul. 2017.

RIBEIRO, P. B.; MEDEIROS, R. P. Extração de entidades nomeadas com maximização de entropia (opennlp). **Caderno de Estudos Tecnológicos**, v. 4, n. 1, 2016.

SILVA, M. P. d. S. Mineração de dados: Conceitos, aplicações e experimentos com weka. **Sociedade Brasileira de Computação**, v. 1, 2004.