Caio Petrucci dos Santos Rosa RA: 248245
Henrique dos Santos Karpischek RA: 217760
Isabella Jeske Rosa RA: 237040

Projeto Final: Proposta e Base de Dados
MC886 A - 2022S2

Questão #1

Qual é o problema (IA para Bem Social) que o seu grupo investigará? Por que é interessante? Descreva o problema.

Resposta:

O problema investigado pelo nosso grupo será a detecção de incêndios florestais de forma automática, portanto, nosso objetivo é construir um modelo de classificação capaz de identificar a existência de um incêndio através de uma base de dados que contenha imagens de florestas que podem, ou não, conter incêndios.

Esse problema é de extrema importância ecológica, pois auxilia no monitoramento e combate ao incêndio e queimadas nas florestas brasileiras. Tendo em vista que a extensão territorial das florestas brasileiras é grande demais para ser monitorada de forma humana, ter um modelo que possibilite a detecção automática dessas catástrofes é de extrema importância para a manifestação e preservação da flora e fauna nacional.

Com esse tipo de ferramenta, também seria possível combater o desmatamento ilegal que utilizam de queimadas, provocadas principalmente pela agricultura e pecuária, que avança, em geral, pelo cerrado e amazônia em busca de terras para produzir.

Por fim, a luta contra esses fenômenos é importante por diversos motivos, entre eles: a preservação da riqueza da fauna e flora nacionais, a diminuição do avanço do aquecimento global e a preservação das terras protegidas indígenas.

Questão #2

Quais são os dados que o seu grupo usará? Descreva o conjunto de dados.

Resposta:

O *dataset* utilizado será *Dataset for Forest Fire Detection* disponível na Mendeley Data. Esse conjunto se trata de um banco de imagens com 1900 figuras dividido igualmente entre imagens com incêndio e sem incêndio, ou seja, 950 imagens para cada classe. Estas são coloridas - com 3 canais de cores -, com resolução de 250x250 e, em sua maioria, apresentam apenas elementos pertinentes ao estudo, contendo apenas a região de incêndio ou não, ou seja, não possui pessoas, maquinários e etc. Além disso, devido à pesquisa dos próprios autores do *dataset*, o conjunto já está dividido em conjuntos de treinamento/validação e teste, em uma proporção de 80:20. Por fim, todas as imagens estão previamente anotadas, configurando este problema em um voltado à aprendizado supervisionado. Acesso: https://data.mendeley.com/datasets/gjmr63rz2r/1.

Datasheets for Datasets
Motivação

Questão #1

Para qual propósito a base de dados foi criada? Havia alguma tarefa específica em mente? Havia alguma lacuna a ser preenchida? Por favor, forneça uma descrição.

Resposta:

A base de dados foi criada com o objetivo de estudar e solucionar o problema de detecção de incêndios florestais, portanto, ela já foi criada com o intuito de ser utilizada em um problema de *machine learning*. Como se trata de um problema binário, supervisionado e que não temos dados sem anotação, não temos lacunas a serem preenchidas.

Questão #2

Quem criou a base de dados (por exemplo, qual equipe, grupo de pesquisa) e em nome de qual entidade (por exemplo, empresa, instituição, organização)?

Resposta:

Temos poucas informações quanto à origem dos dados. Do Mendeley Data, tiramos que os autores do *dataset* são Hassan Bilal e Khan Ali e a publicadora foi a própria Mendeley Data. Sobre a montagem do conjunto, na descrição está registrado que as imagens foram coletadas através de buscas em diversos motores de busca e posteriormente foram tratadas e anotadas.

Questão #3

Quem financiou a criação da base de dados? Se houver um número de financiamento associado (por exemplo, no Brasil poderia ser CAPES, CNPq, FAPESP), forneça o nome e o número do financiamento.

Resposta:

Não há informações sobre.

Composição

Questão #1

O que representam as instâncias que compõem a base de dados (por exemplo, documentos, fotos, pessoas, países)? Existem diversos tipos de instâncias (por exemplo, filmes, usuários e classificações; pessoas e interações entre eles; nós e bordas)? Por favor, forneça uma descrição.

Resposta:

As instâncias que compõem a base de dados são imagens de paisagens, que podem ou não conter incêndios, dos mais variados biomas. Existem imagens contendo objetos como florestas, lagos e pastos em diversos relevos (montanhas, morros, planícies e etc). Além disso, também é possível perceber que, pelas paisagens retratadas, existem instâncias representando todas as estações do ano

(primavera, verão, outono e inverno) e diferentes tipos de climas, que podem ser resultado de paisagens em diferentes partes do mundo.

Questão #2

Quantas instâncias existem no total (de cada tipo, se for o caso)?

Resposta:

A base de dados contém 1900 instâncias no total e, por se tratar de uma base balanceada, existem cerca de 950 instâncias para cada classe (há incêndio ou não há incêndio).

Questão #3

A base de dados contém todas as instâncias possíveis ou é uma amostra (não necessariamente aleatória) de instâncias de uma base maior? Se a base de dados é uma amostra, qual é a base maior? A amostra é representativa da base maior (por exemplo, cobertura geográfica)? Em caso afirmativo, descreva como essa representatividade foi validada/verificada. Se não for representativo da base de dados maior, descreva por que não (por exemplo, para cobrir uma gama mais diversificada de instâncias, porque as instâncias foram retidas ou indisponíveis).

Resposta:

A base de dados se trata de uma amostra de imagens capturadas de paisagens e localizações geográficas, que podem ou não contém incêndios/queimadas, já que é praticamente impossível construir um conjunto de dados que contenha todas as instâncias de paisagens existentes e todas as instâncias de queimadas que ocorrem em florestas e outros ecossistemas. Porém, é possível afirmar que a amostra é representativa da base maior pois, ao analisar o conjunto de imagens coletadas, fica claro que diversos tipos de paisagens, de relevos, de estações e, portanto, localizações geográficas foram consideradas. Apesar disso, não foi possível quantificar o quanto essa base de dados é representativa em relação à todas as paisagens existentes na Terra, mas dado que o objetivo do projeto é detectar incêndios e queimadas florestais, principalmente em biomas que estão contidos dentro do Brasil, o grupo considerou a representatividade e cobertura geográfica do *dataset* satisfatória.

Questão #4

Em que dados consiste cada instância? Dados "brutos" (por exemplo, texto ou imagens não processados) ou features? Em ambos os casos, forneça uma descrição.

Resposta:

Cada instância consiste em uma imagem de 250x250 pixels, com 3 canais de cores, ou seja, com canais RGB. Isso se deve ao fato de que os dados do conjunto já foram tratados e padronizados anteriormente pelos autores do *dataset*.

Questão #5

Existe um rótulo associado a cada instância? Em caso afirmativo, forneça uma descrição.

Resposta:

Para cada instância, existe uma anotação associada que indica se a imagem da instância contém ou não contém um incêndio. Essa anotação se dá pela organização de pastas do *dataset* quando baixado, ou seja, não é algo indicado na imagem diretamente, mas sim na organização já feita do conjunto.

Questão #6

Falta alguma informação em instâncias individuais? Em caso afirmativo, forneça uma descrição, explicando por que essa informação está ausente (por exemplo, porque não estava disponível). Isso não inclui informações removidas intencionalmente, mas pode incluir, por exemplo, texto redigido.

Resposta:

Não falta nenhuma informação em instâncias individuais.

Questão #7

Os relacionamentos entre instâncias individuais da base de dados são explícitos (por exemplo, classificações de filmes dos usuários, links de redes sociais)? Em caso afirmativo, descreva como essas relações são explicitadas.

Resposta:

Não há relacionamentos explícitos entre instâncias individuais da base de dados.

Questão #8

Existem splits de dados recomendados (por exemplo, treinamento, validação, teste)? Em caso afirmativo, forneça uma descrição desses splits, explicando a lógica por trás delas.

Resposta:

Não há splits “recomendados”, porém, como está detalhado na descrição da página do *dataset* no Mendeley Data, o conjunto já está separado em conjuntos de treinamento/validação e teste, em uma proporção 80:20. Como dito na página, essa separação foi feita para a própria pesquisa dos autores mas não foi explicada a lógica por trás.

Questão #9

Existem erros, fontes de ruído ou redundâncias na base de dados? Em caso afirmativo, forneça uma descrição.

Resposta:

A base de dados não contém erros ou fontes de ruídos em sua maioria, devido ao fato de que as imagens foram tratadas e curadas pelos autores. Porém, existe um pequeno número de instâncias, por volta 20-30 de 1900, que possuem alguns ruídos, como pessoas, animais, maquinários agrícolas e fios de torres de comunicação. Mesmo assim, os ruídos encontrados se mostraram muito pequenos

e difíceis de perceber, o que provavelmente deve ser a razão pela qual os autores do *dataset* deixaram tais objetos passarem despercebidos durante o tratamento dos dados.

Questão #10

A base de dados é autocontida ou está vinculada ou depende de recursos externos (por exemplo, sites, tweets, outros conjuntos de dados)? Se estiver vinculada ou depender de recursos externos, a) existem garantias de que eles existirão e permanecerão constantes ao longo do tempo; b) existem versões de arquivo oficiais na base de dados completa (ou seja, incluindo os recursos externos que existiam no momento em que a base de dados foi criada); c) existem restrições (por exemplo, licenças, taxas) associadas a algum dos recursos externos que podem se aplicar a um consumidor da base de dados? Forneça descrições de todos os recursos externos e quaisquer restrições associadas a eles, bem como links ou outros pontos de acesso, conforme apropriado.

Resposta:

A base de dados é autocontida nesse caso e, portanto, não está vinculada a recursos externos.

Questão #11

A base de dados contém dados que podem ser considerados confidenciais (por exemplo, dados protegidos por privilégio legal ou por confidencialidade médico-paciente, dados que incluem o conteúdo de comunicações não públicas de indivíduos)? Em caso afirmativo, forneça uma descrição.

Resposta:

Não há nenhum dado que possa ser considerado confidencial, pois o conjunto contém apenas paisagens sem nenhum outro tipo de anotação além de se há ou não há incêndio, como metadados de localização e etc.

Questão #12

A base de dados contém dados que, se visualizados diretamente, podem ser ofensivos, ofensivos, ameaçadores ou causar ansiedade? Em caso afirmativo, descreva o motivo.

Resposta:

A base de dados contém apenas dados de queimadas e incêndios, que, dependendo da pessoa, podem causar alguma sensação de mal-estar. Fora isso, não contém nenhum outro tipo de dado.