Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Ciência da Computação
Disciplina Introdução a Bancos de Dados (IBD)
2º Semestre de 2023

Trabalho Prático 2

Acesso, coleta, gerenciamento e análise de conjuntos de dados públicos
A ser desenvolvido em grupos de até 5 estudantes
Valor: 10 pontos
Prazo: 30/11/2023

Introdução

A Lei de Acesso à Informação (Lei nº 12.527/2011, ou LAI) é uma legislação que estabelece o direito do cidadão de obter informações públicas dos órgãos e entidades governamentais. Ela foi criada com o objetivo de garantir a transparência das informações do poder público e incentivar a participação social na fiscalização e controle das ações do governo.

Por meio da lei, qualquer pessoa pode solicitar informações sobre atos, empresas públicas, autarquias, fundações, empresas de economia mista, instituições públicas ou privadas sem fins lucrativos que recebam recursos públicos do governo. Isso inclui documentos, relatórios, estudos, pareceres, contratos, entre outros.

Os órgãos públicos têm o dever de disponibilizar as informações solicitadas por meio de canais de comunicação adequados e em prazos definidos pela lei. Caso a informação não seja fornecida ou o prazo não seja cumprido, o cidadão poderá recorrer a instâncias judiciais.

A Lei de Acesso à Informação é uma importante ferramenta de controle social e de fortalecimento da democracia, pois permite aos cidadãos o acesso a informação para fiscalizar e cobrar ações do governo. Além disso, ela incentiva a transparência e a responsabilidade dos gestores públicos.

A LAI Define como dado público toda informação produzida, coletada, organizada, armazenada, disponibilizada ou gerenciada por órgãos e entidades públicas, inclusive autarquias, fundações públicas, entidades privadas sem fins lucrativos que recebam recursos públicos, empresas públicas e sociedades de economia mista, sob qualquer suporte ou em qualquer formato.

A Lei de Acesso à informação (LAI) estabelece limites de acesso aos dados públicos nos seguintes casos:

1. Dados classificados como sigilosos em razão de interesse da segurança do Estado e da sociedade;

2. Informações protegidas por sigilo profissional ou comercial, desde que divulgá-las possa prejudicar a competitividade ou a obtenção de lucro de empresas;

3. Informações pessoais e privadas que possam comprometer a privacidade, segurança ou integridade física de indivíduos;

4. Dados que possam prejudicar a condução de investigações policiais ou processos judiciais em andamento;

5. Informações que possam prejudicar o sigilo da fonte da informação;

6. Dados que possam prejudicar a condução de pesquisas científicas, tecnológicas ou estatísticas;

7. Informações obtidas em contato com outros países ou organismos internacionais.

Além disso, a LAI permite que órgãos públicos neguem o acesso a informações públicas quando houver risco iminente à vida ou à integridade física de uma pessoa.

Em resposta às demandas da LAI, muitos órgãos governamentais e empresas públicas praticam a chamada *transparência ativa*, em que dados sobre sua atuação ou de interesse do público em geral são compilados e periodicamente publicados, ficando à disposição dos cidadãos que desejarem analisá-los. Além disso, dados que não façam parte do escopo das ações de transparência ativa, mas que estejam dentro das definições da LAI podem ser requisitados ao órgão produtor, em condições e limitações também determinadas pela Lei.

Objetivo do TP

O objetivo deste trabalho prático é promover o acesso, coleta, gerenciamento, integração e análise de conjuntos de dados públicos. Idealmente, o resultado desse trabalho poderia ser republicado, sendo que o conteúdo ainda seria formado por dados públicos, mas com as vantagens decorrentes da análise e integração que serão feitos no escopo do TP.

Etapas

1. Escolher dois ou mais conjuntos de dados em alguma fonte de dados públicos e abertos. Exemplos de fontes de dados públicos:
    a. Portal dados.gov.br
    b. Portais de dados abertos das administrações públicas estaduais e municipais. Exemplo: http://dados.pbh.gov.br
    c. Portais de dados aberto do legislativo federal, estadual e municipal. Exemplo: https://dadosabertos.camara.leg.br/

    d. Sites de órgãos e empresas públicas, com destaque para os órgãos tipicamente associados à produção de dados públicos e abertos, como o IBGE e outros. Exemplo: https://www.ibge.gov.br/estatisticas/sociais/populacao.html
    e. Dados publicados por institutos de pesquisa e ONGs, voltados para consolidar e analisar dados primários gerados por órgãos e empresas públicas. Exemplo: Atlas da Violência do IPEA, https://www.ipea.gov.br/atlasviolencia/downloads
    f. Comunicar a escolha ao professor por e-mail, indicando as fontes, sua correlação e a expectativa de realização de análise envolvendo as fontes.
2. Carregar esses dados em um gerenciador de bancos de dados relacional, como o PostgreSQL ou outro, e combinar esses dados, de modo a permitir análises integradas
    a. Naturalmente, será necessário escolher conjuntos de dados que mostrem aspectos diferentes de um mesmo problema.
    b. Por exemplo: atendimentos hospitalares pelo SUS e dados de cobertura vacinal; consumo de energia por município e PIB municipal.
3. Realizar uma análise exploratória nos dados (ver abaixo)
    a. Recuperar um esquema conceitual (ER/UML) dos dados obtidos
    b. Recuperar ou produzir um dicionário de dados a respeito dos dados obtidos
    c. Recuperar ou organizar um conjunto de metadados sobre os dados obtidos, incluindo fonte, data de obtenção, órgão produtor, data de referência (atualização), limitações registradas, cobertura (região geográfica de referência), etc.
4. Apresentar uma análise crítica das fontes de dados utilizadas
    a. Indicar dificuldades, omissões, limitações, problemas de atualização ou de qualidade dos dados percebidos durante a obtenção e análise.
    b. Exemplo: excesso de valores nulos, atributos preenchidos esparsamente, inconsistências, valores faltantes, desatualização, etc.
5. Apresentar análises referentes à combinação ou integração de dados.
    a. Como os dados das múltiplas fontes estão relacionados entre si, produzir uma análise integrada, mostrando correlação entre os dados obtidos.

Sobre a análise exploratória

A análise exploratória de dados é um processo fundamental para entender a estrutura e características dos dados de uma determinada fonte. Essa análise inclui a identificação de padrões e tendências, bem como a identificação de valores discrepantes e lacunas nos dados. Durante a análise exploratória, os dados são visualizados e estatisticamente analisados para determinar a sua distribuição, correlação e outras propriedades.

Para realizar uma análise exploratória de dados, é preciso seguir os seguintes passos:

1. Preparação dos dados: Antes de começar a análise, é preciso garantir que os dados estejam limpos, consistentes e bem estruturados. Isso inclui remover dados duplicados, corrigir erros de ortografia, padronizar datas e valores e preencher valores faltantes. Nesta etapa, caso dados sejam obtidos em formatos como CSV, XLS ou outros, pode-se realizar a importação para tabelas em um SGBD.

2. Definição dos objetivos: É preciso definir claramente o objetivo da análise exploratória, incluindo quais questões buscamos responder e quais hipóteses queremos testar.

3. Análise descritiva: A análise descritiva é a parte mais importante da análise exploratória. Essa análise envolve a visualização dos dados em gráficos e tabelas para determinar a distribuição, correlação e outras propriedades dos dados. Sugere-se usar SQL para obter as características dos atributos, como distribuição de valores, quantidade de dados faltantes, calcular estatísticas básicas (ex. valor médio, quantidade de valores distintos, distribuição de valores, etc.)

4. Identificação de valores discrepantes: A identificação de valores discrepantes (ou *outliers*) é outro aspecto crítico da análise exploratória de dados. É importante identificar valores discrepantes, que são valores que se afastam significativamente do padrão geral dos dados. Caso necessário, dados discrepantes podem ser excluídos das análises.

5. Análise de correlação: A análise de correlação é uma técnica que ajuda a entender as relações entre diferentes variáveis. Através dessa análise, é possível determinar se duas variáveis estão altamente correlacionadas, o que significa que apenas uma delas pode ser usada em análises.

6. Conclusões: A partir da análise exploratória, podemos tirar conclusões sobre os dados e gerar insights importantes que possam ser utilizados para tomada de decisão.

Produtos

1. Relatório contendo itens referentes às etapas acima.
2. Dados utilizados no trabalho, com eventuais transformações, correções ou melhoramentos, publicados em um repositório público. Incluir no texto do relatório um link para os dados trabalhados. Exemplos: GitHub, Zenodo.