Trabalho Prático - Fase I

Mineração de Dados

Daniel Xavier
PG50310

Diogo Rebelo
PG50327

Lídia Sousa
PG50551

I. Introdução

Neste trabalho de **Mineração de Dados**, com o intuito de proceder à implementação de um estudo que envolva a recolha, processamento, análise e mineração de dados, optou-se por analisar o tráfego e congestionamento, utilizando dados do *Google Maps* provenientes das respetivas APIs. A ideia consiste em recolher dados das principais vias e estradas da cidade de Braga e processar estes dados identificando padrões de tráfego e congestionamento, de forma a possivelmente prever atrasos em rotas. Isto permitiria, se utilizado numa aplicação em tempo real, que utilizadores evitem rotas congestionadas e que economizem tempo nas suas viagens. O objetivo seria fornecer informações precisas e úteis, ajudando-os a evitar congestionamentos e tornando as viagens mais eficientes.

II. Motivação e objetivos

A. Motivação

A motivação para realizar um trabalho de análise de tráfego e congestionamentos, utilizando dados do *Google Maps*, especialmente focado na cidade de Braga, surge do facto de que os congestionamentos e atrasos no tráfego são um problema comum, como se comprova de seguida.

Cidade	Horas de atraso por motorista
Braga	29
Lisboa	64
Porto	40
Évora	28
Guimarães	24

TABLE I
Tempo médio gasto em deslocamentos na cidade [1]

Em Portugal, especificamente na cidade de Braga, o tráfego tem sido um grande problema nos últimos anos. De acordo com dados do estudo I de mobilidade urbana de Braga de 2022, o tempo médio gasto em deslocamentos na cidade é de 29 minutos, sendo que 63% dos deslocamentos são feitos de carro. Além disso, a cidade tem enfrentado um aumento do número de carros em circulação, o que tem contribuído para o aumento do congestionamento e da poluição.

Outro estudo realizado pela *Inrix* [1], uma empresa de dados de tráfego, apontou que a cidade de Braga teve o maior congestionamento em Portugal em 2019, com um tempo médio de congestionamento de 12 horas por motorista. Esses dados mostram a importância de ter um sistema de análise de tráfego e congestionamentos em tempo real, para ajudar os usuários a evitar rotas congestionadas e economizar tempo nas suas viagens.

Cidade	Mudança na horas de atraso por motorista (%)
Braga	+37
Lisboa	-53
Porto	-40
Évora	+20
Guimarães	-11

TABLE II
Mudança no tráfego em Braga desde 2019 (%) [1]

B. Objetivos

O objetivo principal deste trabalho consiste em fornecer informações acerca do tráfego nas principais vias e estradas. Ora, tal envolve a recolha de dados, possivelmente em tempo real, processamento e análise desses dados, e a entrega de informações úteis aos utilizadores, como a localização de congestionamentos e rotas alternativas. Além disso, deve ser possível prever possíveis atrasos e sugerir rotas alternativas.

Outro objetivo importante é relativo à apresentação de dados de tráfego de forma clara e compreensível, utilizando gráficos e mapas interativos, além de fornecer informações relevantes sobre o tráfego, como o tempo de espera estimado e a distância até o próximo congestionamento.

C. Requisitos

- O sistema deve ser capaz de recolher dados de tráfego, utilizando APIs do *Google Maps* ou outras fontes públicas de dados. Os dados recolhidos devem ser armazenados numa base de dados e processados para identificar padrões de tráfego e congestionamentos;
- O sistema deve ser capaz de analisar dados históricos de tráfego e prever possíveis atrasos nas rotas, utilizando técnicas de *machine learning* e análise de dados;
- O sistema deve ser capaz de garantir que as informações estão sempre atualizadas e precisas. O sistema também deve ser escalável, permitindo que ele seja expandido para atender a um grande número de usuários;
- Caso seja desenvolvida uma interface para a apresentação dos resultados do estudo, esta deve ser intuitiva, permitindo que os utilizadores visualizem as informações de forma clara e compreensível.

III. Fontes de dados

Para recolher dados de tráfego, como referido anteriormente vai ser usada a API do *Google Maps*, que fornece uma

variedade de informações úteis sobre o tráfego, incluindo dados de velocidade e congestionamentos em tempo real.

As vantagens desta API é o fornecimento de dados precisos (e com a possibilidade de ser em tempo real), a facilidade de integração em sistemas *web* e aplicações. No entanto, para grandes volumes de dados pode ser desvantajoso, devido a ter um custo de uso e limitações na quantidade de pedidos permitidos.

Na possibilidade de utilizar sensores, as vantagens são o fornecimento de dados precisos e em tempo real, não havendo limitações na quantidade de pedidos permitidas. No entanto, o acesso a sensores é mais difícil também pelo facto do custo elevado de instalação e manutenção dos dispositivos.

IV. Metodologias a usar

Pela pesquisa realizada ao longo desta primeira fase de trabalho percebemos que as duas metodologias mais recomendadas para este trabalho seriam:

- Metodologia SCRUM: A metodologia SCRUM tem como objetivo a execução de projetos complexos em menor tempo e com o uso mínimo de recursos. Pelo desconhecimento da metodologia recorremos à ferramenta *ChatGPT* de forma a entender de que forma poderia ser usada num projeto de mineração de dados obtendo a seguinte resposta:

  > No contexto da mineração de dados, a metodologia SCRUM pode ser utilizada para projetos que envolvem a recolha, processamento, análise e interpretação de dados. A metodologia pode ajudar a definir as metas do projeto, identificar as fontes de dados relevantes, selecionar as ferramentas de mineração de dados apropriadas, desenvolver modelos de análise, testar e validar os resultados e entregar um produto final de alta qualidade. A metodologia SCRUM também pode ajudar o grupo a comunicar e colaborar efetivamente durante todo o processo do projeto.

- Metodologia CRISP-DM: O CRISP-DM, já conhecido e utilizado por nós no contexto de outras unidades curriculares e é uma metodologia padrão para mineração de dados e análise de dados. É composta por seis fases principais: conhecimento do negócio, entendimento dos dados, preparação dos dados, modelação, avaliação e implementação. Dado que o trabalho envolve a análise de dados, o CRISP-DM pode ser útil para orientar o processo de análise de dados, desde a obtenção dos dados até a implementação do modelo.

Por uma questão de familiarização e facilidade de entendimento da metodologia, enquanto grupo optamos, como abordagem deste projeto, utilizar a metodologia **CRISP-DM**.

V. Planeamento das tarefas

Quanto ao planeamento de tarefas, segue-se de seguida uma lista com as primeiras tarefas a cumprir:

- **Recolher dados em tempo real do *Google Maps* usando suas APIs de tráfego e direções:**
  Para recolher dados em tempo real do *Google Maps*, podemos utilizar a API de *Directions* e a API de *Traffic*. A API de *Directions* permite-nos obter informações de rotas entre dois ou mais pontos, incluindo informações sobre distância, tempo estimado de viagem e possíveis rotas alternativas. Já a API de *Traffic* fornece informações em tempo real sobre o tráfego nas estradas e vias, incluindo dados sobre congestionamentos, acidentes e obras na estrada.
  Para utilizar essas APIs, é necessário criar uma conta de desenvolvedor no *Google Cloud Platform*, que no caso já temos devido ao uso noutras unidades curriculares e gerar, de seguida, uma chave de API válida. Com essa chave, podemos fazer pedidos para as APIs selecionadas;
- **Analisar o armazenamento dos dados - Base de Dados ou Acesso Direto a API:**
  Para armazenar os dados recolhidos numa base de dados seria necessário criar uma e as respetivas tabela que armazenam as informações recolhidas. Depois utilizaríamos *Python* para fazer a conexão com a base de dados e inserir os dados recolhidos nas respetivas colunas da tabela.
  As principais vantagens de armazenar os dados numa base de dados são a **velocidade de acesso** - é possível aceder aos dados rapidamente sem a necessidade de fazer pedidos às APIs. Particularmente importante quando se trata de grandes quantidades de dados ou quando a velocidade de acesso é crucial, como em aplicações de tempo real, a **possibilidade de análise e processamento *offline*** e a **redução de custos** - dependendo da frequência de pedidos, usar diretamente as APIs pode gerar custos significativos.
  No entanto, ao utilizar diretamente os dados da API, teríamos sempre o acesso aos dados mais recentes, uma redução na complexidade do sistema - armazenar e gerir uma base de dados adiciona complexidade ao sistema e o acesso a recursos adicionais que são vantajosos (informações de tráfego em tempo real, bem como direções e rotas otimizadas);
- **Processamento e análise dos dados:**
  Para o processamento dos dados, vai-se recorrer ao *Python* e às suas bibliotecas. Vão ser aplicadas técnicas de análise de dados para identificar padrões de tráfego e congestionamentos.
  Inicialmente, proceder-se-á a uma análise exploratória de dados, usada para visualizar e entender os dados, identificando padrões e tendências. Vamos dispor a informação em gráficos, histogramas e outras visualizações que possam ser usadas para entender a distribuição dos dados e identificar padrões.
  De seguida, tentar-se-á identificar o comportamento dos dados ao longo do tempo, determinando padrões sazonais e tendências. Essa técnica é útil para prever possíveis atrasos nas rotas com base em padrões históricos.
  Por fim, proceder-se-á à aplicação de modelos para prever os possíveis atrasos nas rotas com base em dados históricos de tráfego.

Procedeu-se, após a idealização das tarefas a realizar, à realização do Diagrama de Gantt 1, estabelecendo prazos e tarefas bem definidas de forma a realizar o projeto com sucesso.

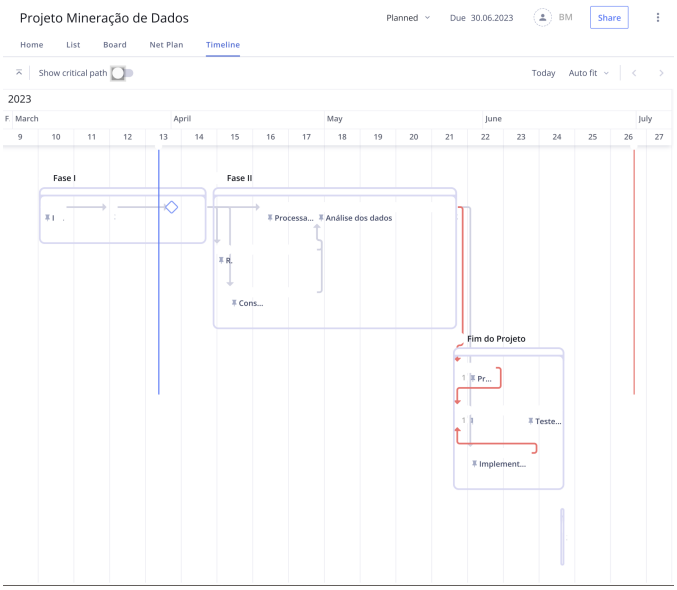

Fig. 1. Diagrama de Gantt

O diagrama de Gantt é constituído pelas seguintes tarefas e respetivos prazos:

Task	Checklist	Responsible	Status	Start - End date
Fase I				
Identificar tema do trabalho			Done	Mar 06 - Mar 08, 2023
Identificar fontes de dados			Done	Mar 09 - Mar 09, 2023
Analisar viabilidade			Done	Mar 10 - Mar 10, 2023
Documento de detalhe da fase I			Done	Mar 20 - Mar 20, 2023
Apresentação Fase I			Planned	Mar 31, 2023
Fase II				
Recolha dos Dados			Planned	Apr 09 - Apr 12, 2023
Construção da BD			Not set	Apr 12 - Apr 18, 2023
Processamento dos dados			Planned	Apr 19 - Apr 28, 2023
Análise dos dados			Planned	Apr 29 - May 25, 2023
Apresentação Fase II			Planned	May 26 - May 26, 2023
Fim do Projeto				
Produção de documento com resultado...			Planned	May 26 - Jun 02, 2023
Design do Sistema			Planned	May 26 - May 29, 2023
Implementação Sistema			Not set	May 29 - Jun 09, 2023
Teste do Sistema			Planned	Jun 09 - Jun 16, 2023
Entrega do Projeto			Planned	Jun 16 - Jun 16, 2023

Fig. 2. Tarefas a realizar e respetivos prazos

VI. Conclusão

Conclui-se então que a implementação de uma aplicação de análise de tráfego em tempo real utilizando dados do *Google Maps* seria uma solução útil e viável para ajudar a evitar congestionamentos e tornar as viagens mais eficientes na cidade de Braga.

A escolha de tecnologias apropriadas, como as APIs do *Google Maps* e as bibliotecas de análise de dados do *Python* vão ser fundamentais para a implementação bem-sucedida do sistema. Com este trabalho podemos demonstrar a importância da análise de dados e do uso de tecnologias adequadas para a resolução de problemas do mundo real.

A. *Trabalho futuro*

Como trabalho futuro, propõe-se a execução das tarefas discutidas no planeamento delineado e a aplicação os métodos e as ferramentas selecionadas, aprofundando e experimentando vários algoritmos possíveis, várias explorações de dados diferentes.

References

[1] "Global traffic scorecard — inrix global traffic rankings."