INSTITUTO UNIVERSITÁRIO DE LISBOA

Processamento e Modelação de Big Data

Projeto final

29 de março de 2025

1 Introdução

Este projeto visa consolidar os conhecimentos práticos no manuseamento de dados em larga escala. Em termos de ferramentas de implementação, o projeto deve recorrer essencialmente a funcionalidades disponibilizadas pela plataforma Apache Spark e à linguagem de programação Python.

A realização do projeto será feita por grupos de trabalho constituídos por três ou quatro alunos.

2 Problema

Neste trabalho pretende-se avaliar a implementação de uma solução computacional para o estudo e análise de dados em larga escala. Um dos aspectos a considerar é a análise da complexidade computacional da solução. A análise da complexidade de um algoritmo passa por determinar as operações que são executadas pelo algoritmo e quantas vezes cada uma delas é executada. Uma vez que existe uma relação directa entre a complexidade e o tempo de execução, neste trabalho vamos avaliar o tempo de execução das diferentes operações e algoritmos.

2.1 Domínio de dados e algoritmia

Os dados a utilizar neste trabalho devem ser escolhidos a partir da lista de fontes especificada na Tabela 1. Os autores do trabalho devem selecionar uma fonte de dados e submeter a decisão na plataforma moodle até ao dia **11 de abril de 2025**. Recorda-se que, os grupos que não informarem qual o conjunto de dados (*dataset*) que irão trabalhar, passam automaticamente para a modalidade de avaliação por exame. Após esta data não serão aceites escolhas nem alterações ao *dataset*.

Os algoritmos a utilizar para a construção do modelo de análise e processamento de dados fazem parte da plataforma Apache Spark.

2.2 Formulação do problema

Em função da escolha da fonte de dados a usar, deverá formular um problema de aprendizagem não supervisionada. O problema proposto deverá motivar toda a análise subsequente, isto é, o trabalho pretende avaliar a complexidade temporal da resolução do problema proposto.

Tabela 1: Lista de fontes de informação para seleção de dados.

Dataset	URL
1	2013 American Community Survey
2	Indian Cities Weather 2010-2024: Dive In!
3	Death in the United States
4	Smart meters in London
5	eCommerce behavior data from multi category store
6	Flight status prediction

2.3 Etapas da resolução do problema

Na resolução do problema deverá ter em conta as seguintes fases:

1. ingestão dos dados (ETL);
2. exploração e a avaliação dos dados;
3. selecção das variáveis a usar (feature engineering);
4. treino do algoritmo;
5. análise e visualização dos resultados.

2.4 Análise da complexidade

Como foi referido, a analise da complexidade será feita medindo o tempo de execução do código. Pretende-se por isso que seja feito um estudo comparativo da execução das diferentes fases da resolução do problema, variando a dimensão dos dados. Considere as fases 1, 2 e 4.

3 Implementação

A implementação da solução deve ser modular, ou seja, deve ser composta por mais do que um notebook ou módulo Python. Compete aos autores do trabalho estruturar de forma criteriosa o código implementado. Por outro lado, chama-se a atenção para os seguintes aspetos, também já referidos ao longo das aulas:

- A escolha do domínio de dados e consequentemente seleção de dados, bem como a formulação do problema em estudo, são da maior importância para o sucesso do projeto como um todo. Estas fases não devem ser menosprezadas, em termos relativos.
- Por questões de produtividade, devem ser considerados dois conjuntos de dados aquando do desenvolvimento da solução. Assim, para além dos dados originais na sua íntegra, deve ser utilizado um conjunto de dados de menor dimensão (sub-conjunto dos anteriores), para o caso de tarefas intensivas e frequentes, inerentes ao próprio processo de desenvolvimento da solução.

- Cada notebook (ou módulo) deverá ser autónomo em termos de fontes de dados. Sugere-se que estruturem o código por forma a ler e gravar os dados entre cada uma das etapas do projeto. Isto é particularmente importante para a parte da visualização: a geração de um gráfico ou tabela não deverá implicar a realização da simulação/processamento no mesmo instante. Preferencialmente deverá importar os dados já processados a partir de ficheiros.

4 Material a entregar

O trabalho terá duas entregas: uma parcial e outra final.

- **Entrega parcial**: submissão de um arquivo no formato **zip** (extensão zip e não outra) com os seguintes elementos de avaliação:
  1. notebooks e/ou módulos Python relativo às etapas 1 a 4. Não considere a análise de complexidade;
  2. breve relatório apenas com a descrição do problema.
- **Entrega final**: submissão de um arquivo no formato **zip** (extensão zip e não outra) com os seguintes elementos de avaliação:
  1. relatório final;
  2. notebooks e/ou módulos Python.
- O relatório final deve ser sucinto e em formato **pdf**, com o máximo de oito páginas, excluindo a capa e o índice. O documento deve:
  - Conter uma descrição do problema em estudo e respetivos dados utilizados.
  - Abordar os aspetos mais relevantes sobre as decisões tomadas.
  - Incluir informação sobre as experiências e testes realizados. Deverão incluir a comparação dos tempos de execução, recursos e capacidades de processamento utilizadas.
  - Incluir uma análise sobre os resultados obtidos, não só em termos de desempenho da solução mas, sobretudo, na perspectiva do problema formulado. Ou seja, com base no problema enunciado, indicar quais são as conclusões a retirar após a análise que é feita aos resultados obtidos.
  - Incluir informações adicionais que os autores do trabalho considerem relevantes.
- Os notebooks e/ou módulos Python constituem a solução computacional. Assume-se que os mesmos são auto-explicativos, contendo comentários com nível de detalhe apropriado.

Refira-se ainda que, de acordo com as regras de avaliação da unidade curricular, este projeto tem uma ponderação de 30% na nota final da unidade curricular. Embora o foco do trabalho seja na implementação da solução, será igualmente considerada a qualidade da solução apresentada.

5 Prazos a respeitar

O trabalho deverá obedecer aos seguintes prazos:

1. Formação dos grupos e escolha do conjunto de dados: até **11 de abril de 2025**;
2. Entrega parcial: até **23 de abril de 2025** às 9h00;
3. Entrega final: até **5 de maio de 2025** às 9h00.

O trabalho deve ser submetido de acordo com as seguintes regras:

- A submissão consiste num arquivo em formato **zip** (extensão zip e não outra) com os seguintes elementos de avaliação:
  1. Relatório.
  2. Notebooks e/ou módulos Python.
- O prazo de submissão é **9h00 de 5 de maio de 2025**, com o respetivo arquivo zip a ser submetido na plataforma de ensino Moodle. O *link* a utilizar será indicado em momento oportuno.
- **Importante**: A submissão do trabalho no Moodle não pode conter ficheiros de dados.

6 Discussão do trabalho

O trabalho será discutido presencialmente, em local e hora a indicar após submissão do mesmo e de acordo com a disponibilidade dos membros do grupo e dos docentes. A avaliação do trabalho levará em conta as seguintes componentes: relatório, código apresentado e discussão. Relembra-se ainda que o resultado da avaliação do trabalho é individual.

Política em caso de fraude

Os alunos podem partilhar e/ou trocar ideias entre si, sobre os trabalhos e/ou resolução dos mesmos. No entanto, o trabalho entregue deve corresponder ao esforço individual de cada grupo. São consideradas fraudes as seguintes situações:

- trabalho parcialmente copiado;
- facilitar a copia através da partilha de ficheiros.

Em caso de detecção de algum tipo de fraude, os trabalhos em questão não são avaliados, sendo enviados à comissão pedagógica, que decide a sanção a aplicar aos alunos envolvidos. Serão utilizadas ferramentas para detecção automática de cópias.

Recorda-se ainda que o Anexo I do Código de Conduta Académica, publicado a 25 de Janeiro de 2016 em Diário da República, 2ª Série, nº 16, indica no seu ponto 2 que: *"Quando um trabalho ou outro elemento de avaliação apresentar um nível de coincidência elevado com outros trabalhos (percentagem de coincidência com outras fontes reportada no relatório que o referido software produz), cabe ao docente da UC, orientador ou a qualquer elemento do júri, após a análise qualitativa desse relatório, e em caso de se confirmar a suspeita de plágio, desencadear o respetivo procedimento disciplinar, de acordo com o Regulamento Disciplinar de Discentes do ISCTE-Instituto Universitário de Lisboa, aprovado pela deliberação n.o 2246/2010, de 6 de dezembro"*. O ponto 2.1 desse mesmo anexo indica ainda que: *"No âmbito do Regulamento Disciplinar de Discentes do ISCTE-IUL, são definidas as sanções disciplinares aplicáveis e os seus efeitos, podendo estas variar entre a advertência e a interdição da frequência de atividades escolares no ISCTE-IUL até cinco anos"*.