INSTITUTO UNIVERSITÁRIO DE LISBOA

Processamento de Big Data

Projeto final

16 de maio de 2024

1 Introdução

Este projeto visa consolidar conhecimentos práticos no desenho e implementação de uma solução computacional para dar resposta a um problema de análise 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 quatro membros. A título exceptional, um grupo de trabalho pode ser constituído por três membros.

2 Contextualização do problema

Pretende-se que seja implementada uma solução computacional para estudo e análise de dados em larga escala. Nesse sentido, deverá ser construído um modelo de análise e processamento de dados baseado em métodos e algoritmos referidos nas aulas.

A escolha do domínio de dados e respetivo conjunto de dados a utilizar, bem como a formulação do próprio problema em estudo, será da responsabilidade dos autores do trabalho.

2.1 Domínio de dados e algoritmia

Os dados associados ao problema a formular devem ser obtidos a partir de um dos seguintes portais de informação:

- `https://github.com/otto-de/recsys-dataset`
- `https://www.kaggle.com/datasets/jeanmidev/smart-meters-in-london`
- `https://www.kaggle.com/datasets/erikbiswas/higgs-uci-dataset`
- `https://www.kaggle.com/datasets/dasgroup/rba-dataset`
- `https://www.kaggle.com/datasets/skeller/2021-us-federal-award-data`
- `https://www.kaggle.com/datasets/katerpillar/meteonet`
- `https://www.kaggle.com/datasets/giobbu/belgium-obu`

- `https://www.kaggle.com/datasets/rosenthal/citi-bike-stations`
- `https://www.kaggle.com/datasets/ajohrn/bikeshare-usage-in-london-and-taipei-network`
- `https://www.kaggle.com/datasets/pigment/big-sales-data`
- `https://www.kaggle.com/datasets/PROPPG-PPG/hourly-weather-surface-brazil-southeast-region`

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

A dimensão dos dados obtidos a partir da fonte de informação não deve ser inferior a 2 GB. Os autores do trabalho devem selecionar uma fonte de dados e informar o docente sobre a decisão tomada, até ao dia **26 de maio de 2024**. Os trabalhos de grupos que não informaram qual o dataset que irão trabalhar não serão aceites. Não serão igualmente aceites alterações ao dataset após esta data.

3 Implementação

A implementação da solução deve ser modular, ou seja, deve ser composta por mais do que um notebook (ou módulo) Python. Os notebooks ou módulos deverão obedecer à estrutura que se descreve a seguir.

3.1 Importação e depuração dos dados

Neste notebook deverão fazer as operações relativas ao chamado ETL (*extract, transform and load*), que consiste num processo de integração de dados que combina, limpa e organiza dados de várias fontes num único conjunto de dados consistente para armazenamento numa plataforma de *big data*. No caso em apreço, os dados serão guardados localmente mas no formato `parquet`.

3.2 Análise exploratória dos dados

Esta etapa corresponde à análise exploratória dos dados tendo em conta o problema em questão. Os dados a serem analisados deverão ser os resultantes da etapa anterior.

3.3 Treino e afinação do modelo

Após a análise dos dados, construa um `pipeline` que treine o(s) modelo(s) para o problema em questão. No final do notebook o modelo treinado assim como o `pipeline` deverão ser guardados em disco, tal como foi feito nas aulas. Considere as operações necessárias à afinação do modelo.

3.4 Aplicação do modelo

Neste notebook pretende-se aplicar o modelo treinado a novos dados. Para isso deverá começar por ler o modelo do disco, importar os dados de validação e correr o algoritmo. O(s) resultados(s) deverão igualmente ser guardados no disco, para posterior visualização.

3.5 Visualização e conclusões

Este notebook servirá como relatório do projeto. Nele deverão apresentar o problema, explicar o pré-processamento realizado aos dados, explicar a análise exploratória e os modelos escolhidos e finalmente visualizar os resultados e apresentar as conclusões. Todos os dados relevantes que foram apresentados, como números, deverão ser obtidos automaticamente. Considere guardar esses números em disco durante a execução das etapas anteriores.

Sejam sucintos na construção do notebook e organizem-no por secções. Incluam um índice no início bem como um link para a fonte de dados escolhida.

3.6 Considerações gerais

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.
- Reforça-se que cada notebook (ou módulo) deverá ser autónomo em termos de fontes de dados. 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 e conclusão: 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.
- O vosso código deverá correr sem erros: a partir da fonte de dados em bruto deverá ser possível obter o relatório final.

4 Material a entregar

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

- A submissão consiste num arquivo em formato **zip** (extensão zip e não outra) contendo os Notebooks e/ou módulos Python, **com o resultado da execução**, isto é, **com o output das células de código após ter sido feito um** `Run All`.
- O prazo de submissão é **9h00 de 11 de junho de 2024**, com o respetivo arquivo zip a ser submetido na plataforma de ensino Moodle. O *link* a utilizar será indicado em momento oportuno.

- 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.
- **Importante**: A submissão do trabalho no Moodle não pode conter ficheiros de dados.

Refira-se ainda que, de acordo com as regras de avaliação da unidade curricular, este projeto tem uma ponderação de 40% na nota final da unidade curricular.

5 Apresentação do trabalho

O trabalho será apresentado oralmente, em local e hora a indicar após submissão do mesmo e de acordo com a disponibilidade dos membros do grupo e dos docentes. 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"*.