Adicionar aprendizado de máquina (ML)

**visão global**

Neste laboratório, você usará a API do Google Cloud para incorporar vários serviços de aprendizado de máquina da API do Natural Language, incluindo análise de sentimento e análise de entidade, para produzir resultados significativos a partir de dados não estruturados.

[O Cloud Machine Learning Engine](https://cloud.google.com/ml-engine/docs/) traz o poder e a flexibilidade do TensorFlow para a nuvem, permitindo que você execute treinamento em grande escala em um cluster gerenciado e, em seguida, servidor escalável do seu modelo treinado para previsão.

**Objetivos**

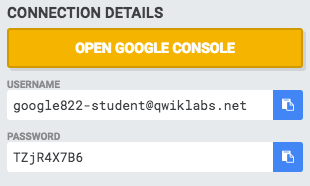
Neste laboratório, você executará as seguintes tarefas:

* Adicionar aprendizado de máquina (ML) a um aplicativo Spark

Para cada laboratório, você obtém um novo projeto de GCP e um conjunto de recursos por um tempo fixo, sem nenhum custo.

1. Certifique-se de ter entrado no Qwiklabs usando uma **janela anônima** .
2. Observe o tempo de acesso do laboratório (por exemplo, img / time.pnge certifique-se de que você pode terminar nesse bloco de tempo.

Não há recurso de pausa. Você pode reiniciar, se necessário, mas você deve começar no começo.

1. Quando estiver pronto, clique em img / start_lab.png.
2. Anote suas credenciais de laboratório. Você os usará para fazer login no Cloud Platform Console. 
3. Clique em **Abrir o Google Console** .
4. Clique em **Usar outra conta** e copie / cole credenciais para **este** laboratório nos prompts.

Se você usar outras credenciais, receberá erros ou **incorrerá em cobranças** .

1. Aceite os termos e pule a página de recursos de recuperação.

Não clique em **Terminar, a** menos que tenha terminado o laboratório ou queira reiniciá-lo. Isso limpa o seu trabalho e remove o projeto.

**Tarefa 1: preparação**

Um cluster Dataproc foi preparado para você. Se você fizer login no GCP antes de a barra de progresso informar que o "Laboratório está em execução", talvez seja necessário esperar alguns minutos para que o cluster faça a transição de "Fornecimento" para "Em execução" antes que o cluster conclua a configuração.

Você estará executando a maioria das etapas de laboratório do nó mestre do cluster em uma janela do terminal SSH.

1. No Console, no **menu Navegação** ( 7a91d354499ac9f1.png), clique em **Dataproc** > **Clusters** .
2. Localize o cluster denominado cluster de **dataproc** . Em qual região e zona ele está localizado? A região e a zona foram selecionadas automaticamente para você pela Qwiklabs.
3. Observe o intervalo de armazenamento temporário do Cloud Storage definido para este cluster. Esse intervalo tem o mesmo nome que o ID do projeto, que é uma maneira conveniente de tornar o nome globalmente exclusivo.
4. Clique no nome **dataproc-cluster** para ir para a página Detalhes do cluster.
5. A página de detalhes do cluster é aberta na guia **Visão geral** . Clique na guia denominada **instâncias de VM** .
6. Na linha da VM denominada **dataproc-cluster-m,** você verá que ela tem a função de mestre e que há um link SSH ao lado dela. Clique em **SSH** para abrir uma janela de terminal para o nó mestre.

**Tarefa 2: Preparação**

Neste laboratório, você estará executando aplicativos **PySpark** . Você percorrerá uma série de programas que desenvolverá gradualmente um aplicativo Spark Dataproc com elementos de Aprendizado de Máquina.

No Laboratório 4, Tarefa 2, você aprendeu como preparar um aplicativo do PySpark no Cloud Storage e enviá-lo da página Tarefas do Dataproc.

Este laboratório usa um script para automatizar a personalização e a preparação dos aplicativos de exemplo do Spark + Machine Learning. Você precisa definir duas variáveis ​​de ambiente antes de executar o script; BUCKET e APIKEY.

Prepare a chave da API

A chave da API nos programas não é válida. Você deve criar um novo e substituí-lo no laboratório.

1. No Console, no **menu Navegação** ( 7a91d354499ac9f1.png), clique em **APIs e Serviços** > **Credenciais** .
2. Clique em **Create Credentials** e selecione **API Key** .
3. Copie a chave da API. No terminal, crie uma variável de ambiente para recuperar facilmente a chave.

APIKEY=<your-api-key>

Verificar o intervalo a ser usado para o armazenamento temporário do Dataproc

O Dataproc pode usar um intervalo do Cloud Storage para organizar seus arquivos durante a inicialização. Você pode usar esse intervalo para preparar programas de aplicativos ou dados para uso pelo Dataproc. O bucket também pode hospedar scripts e saída de inicialização do Dataproc. O nome do intervalo deve ser globalmente exclusivo. A Qwiklabs já criou um intervalo para você que tem o mesmo nome da ID do projeto, que já é globalmente exclusiva.

1. No console, no **menu Navegação** ( 7a91d354499ac9f1.png), clique em **Armazenamento** > **Navegador** . Verifique se o balde existe. Observe a classe de armazenamento padrão e a localização (região) desse intervalo. Você estará usando as informações desta região a seguir.

No terminal SSH **dataproc-cluster-m** , configure o BUCKET.

BUCKET=<bucket name>

Identifique um projeto

Uma variável de ambiente que você definirá é **$ DEVSHELL\_PROJECT\_ID,** que contém o ID do projeto do Google Cloud necessário para acessar recursos faturáveis.

No Cloud Shell, uma variável de ambiente é criada automaticamente, chamada DEVSHELL\_PROJECT\_ID, que contém o ID do projeto do GCP. Esta variável ainda não existe no nó mestre.

1. No console, no **menu Navegação** ( 7a91d354499ac9f1.png), clique em **Página inicial** . No painel com Informações do **projeto** , o **ID** do **projeto** é listado. Você também pode encontrar essas informações na guia Qwiklabs, em Detalhes da conexão, onde ela é rotulada como **ID do projeto do GCP** .

No terminal SSH dataproc **-cluster-m** , configure o DEVSHELL\_PROJECT\_ID.

DEVSHELL\_PROJECT\_ID=<project ID>

Verificar variáveis ​​de ambiente necessárias

1. Verifique se você tem essas variáveis ​​de ambiente definidas. **Não prossiga até que estejam definidos** .

echo $DEVSHELL\_PROJECT\_ID, $BUCKET, $APIKEY

Exporte o BUCKET e o APIKEY para que estejam disponíveis para o script de shell.

export DEVSHELL\_PROJECT\_ID

export BUCKET

export APIKEY

Copie os arquivos do aplicativo para o diretório inicial do dataproc-cluster-m

Os arquivos de amostra necessários já foram arquivados em dataproc-cluster-m. Você precisará copiá-los em seu diretório de usuários com o seguinte comando.

1. Na janela do terminal **dataproc-cluster-m** SSH.

cd

cp -r /training/training-data-analyst .

ls

**Execute o script de preparação**

No terminal, navegue até o diretório do código-fonte deste laboratório.

cd ~/training-data-analyst/courses/unstructured/

Execute o script de preparação:

./stagelabs.sh

Isso é o que o script de preparação está fazendo:

* Edita os três scripts python **01-dataprocML.py** , **02-dataprocML.py** , **03-dataprocML.py** e substitui a APIKEY, BUCKET e DEVSHELL\_PROJECT\_ID pelos valores das variáveis ​​de ambiente exportadas.
* Copia os arquivos atualizados para o seu intervalo no Cloud Storage, para que o Dataproc possa acessá-los.
* Copia arquivos de dados de amostra para o seu intervalo.

Verifique se os arquivos do aplicativo PySpark e os arquivos de dados de amostra estão no bloco.

1. No console, no **menu Navegação** ( 7a91d354499ac9f1.png), clique em **Armazenamento** > **Navegador** .

**Tarefa 3: Processamento de Linguagem Natural**

Os três programas são "instantâneos" de um processo de desenvolvimento. Cada programa baseia-se e aprimora o anterior. Examinar e executar cada programa mostra como desenvolver progressivamente um aplicativo Dataproc / Spark + Machine Learning.

Os dados da amostra são dados não estruturados. Ou seja, ou não tem estrutura, ou tem uma estrutura que não é adequada ao propósito pretendido. Neste laboratório, você usará Aprendizado de Máquina para identificar e associar os dados a valores, fornecendo estrutura e tornando os dados úteis.

Navegue até este diretório.

cd ~/training-data-analyst/courses/unstructured/

Examine **01-dataprocML.py** usando um editor como o **nano** . Não faça alterações no arquivo.

1. Este programa é apenas um programa em Python. Ele será executado no Dataproc, mas não faz uso de nenhum dos recursos de big data. O programa cria uma linha de amostra de texto na memória e, em seguida, passa-a para o serviço Natural Language Processing for Sentiment Analysis.
2. A função SentimentAnalysis()é um wrapper em torno da API REST. Esse código cria o formato estruturado da solicitação e passa a solicitação junto com a chave da API.
3. Por que a saída é impressa usando um json.dumps?
4. Você poderia fazer o pós-processamento dos dados retornados usando o Python.
5. O script stagelabs.sh que você executou na Tarefa 1 deve ter substituído o DEVSHELL\_PROJECT\_ID, o BUCKET e o APIKEY com suas informações das variáveis ​​de ambiente.

Execute o aplicativo

1. No Console, no **menu Navegação** ( 7a91d354499ac9f1.png), clique em **Dataproc** > **Trabalhos** . O clique **ENVIAR TRABALHO** .
2. Você precisará selecionar a região onde seu cluster está localizado e o cluster, cluster de **dataproc** . O **tipo de trabalho** é o **PySpark** .
3. No campo para o **arquivo python Principal** , insira o caminho para o arquivo do aplicativo, que é algo como: **gs: // <nome do bloco> /01-dataprocML.py** , em que você substitui <nome do bloco> pelo nome do seu intervalo.
4. Clique em **Submit** . Veja a saída.

**Tarefa 4: Carregar Dados de Amostra**

1. No terminal, insira os seguintes comandos para copiar arquivos de amostra para o intervalo do Cloud Storage.

gsutil cp /training/road-not-taken.txt gs://$BUCKET/sampledata/road-not-taken.txt

1. No console, no **menu Navegação** ( 7a91d354499ac9f1.png), clique em **Armazenamento** > **Navegador** .
2. Clique no seu balde.
3. Clique em **sampledata** .
4. Alguns arquivos já foram preparados.

**Tarefa 5: Testando a Análise de Sentimento com o Spark**

Examine **02-dataprocML.py** usando um editor como o **nano** . Não faça alterações no arquivo.

1. Este programa usa o Spark RDDs. Ele lê um pequeno arquivo de amostra e o transmite para o serviço Natural Language Processing for Sentiment Analysis.
2. O pós-processamento dos dados retornados é feito no pipeline usando transformações.

Execute o aplicativo

1. No console, no **menu Navegação** ( 7a91d354499ac9f1.png), clique em **Armazenamento** > **Navegador** . clique em **Dataproc** > **Jobs** . O clique **ENVIAR TRABALHO** .
2. Você precisará selecionar a região onde seu cluster está localizado e o cluster, cluster de **dataproc** . O **tipo de trabalho** é o **PySpark** .
3. No campo para o **arquivo python Principal** , insira o caminho para o arquivo do aplicativo, que é algo como: **gs: // <nome do bloco> /02-dataprocML.py** , em que você substitui <nome do bloco> pelo nome do bloco.
4. Clique em **Submit** . Veja a saída.

**Tarefa 6: Fazendo algo útil**

Examine **03-dataprocML.py** usando um editor como o **nano** . Não faça alterações no arquivo.

1. Este programa baseia-se no anterior. Em vez de ler um poema, vai ler um livro inteiro. No entanto, poderia facilmente ler e processar uma biblioteca inteira.
2. Adiciona filtro (no pipeline) e classifica (Python).
3. Isso dá uma lista das linhas no livro com o sentimento mais forte, tanto positivo quanto negativo.
4. Agora, isso foi apenas um livro. Imagine como você poderia usar isso para classificar através de comentários de mídia social. Por exemplo, considere o feedback deixado pelos clientes em um site de compras. Você poderia usar esse tipo de análise de dados para identificar os produtos mais admirados e desprezados.

Execute o aplicativo

1. No Console, no **menu Navegação** ( 7a91d354499ac9f1.png), clique em **Dataproc** > **Trabalhos** . O clique **ENVIAR TRABALHO** .
2. Você precisará selecionar a região onde seu cluster está localizado e o cluster, cluster de **dataproc** . O **tipo de trabalho** é o **PySpark** .
3. No campo para o **arquivo python Principal** , insira o caminho para o arquivo do aplicativo, que é algo como: **gs: // <nome do intervalo> /03-dataprocML.py** , em que você substitui <nome do bloco> pelo nome do seu intervalo.

**Tarefa 7: Outros serviços ML**

O padrão geral para usar serviços ML de um aplicativo Dataproc pode ser adaptado para usar outros serviços do ML.

**Termine seu laboratório**

Quando você tiver concluído seu laboratório, clique em **Finalizar Laboratório** . O Qwiklabs remove os recursos que você usou e limpa a conta para você.

Você terá a oportunidade de avaliar a experiência do laboratório. Selecione o número de estrelas aplicável, digite um comentário e clique em **Enviar** .

O número de estrelas indica o seguinte:

* 1 estrela = muito insatisfeito
* 2 estrelas = insatisfeito
* 3 estrelas = neutra
* 4 estrelas = Satisfeito
* 5 estrelas = muito satisfeito

Você pode fechar a caixa de diálogo se não quiser fornecer feedback.

Para comentários, sugestões ou correções, use a guia **Suporte** .

Última atualização: 2018-09-14

© 2018 Google LLC Todos os direitos reservados. Google e o logotipo do Google são marcas registradas do Google LLC. Todos os outros nomes de empresas e produtos podem ser marcas registradas das respectivas empresas com as quais estão associados.