

FIAP

Pós-Graduação IA para Devs

Diego de Moraes Pereira Sonnenthal

Guilherme Munhoz Lima

Leonardo Fernandes de Oliveira

Fine Tuning de Modelo LLM com informações de produtos da Amazon

Relatório entregue para cumprimento da
atividade de Tech Challenge exigida na
Fase 3.

São Paulo

2024

Sumário

I.	INTRODUÇÃO.....	3
II.	MATERIAIS E MÉTODOS	4
	Dataset	4
	Seleção de Modelo para Fine Tuning	4
	Infraestrutura	5
	Versões	5
	Plus RAG	5
III.	Desenvolvimento Versão Final	5
	Inicial	6
	Versão Final com Fontes.....	7
	Plus RAG	8
IV.	RESULTADOS	8
V.	CONCLUSÃO.....	13
VI.	REFERÊNCIAS.....	14
VII.	ANEXOS	14

I. INTRODUÇÃO

O presente relatório tem como objetivo descrever o desenvolvimento do projeto proposto no Tech Challenge da fase 3, que consiste em aplicar técnicas de fine-tuning em um modelo de foundation, utilizando o dataset "The AmazonTitles-1.3MM". O desafio proposto exige a implementação de um modelo capaz de gerar respostas baseadas em perguntas formuladas pelos usuários, utilizando como contexto o arquivo "trn.json" contido no dataset.

A atividade é uma oportunidade para integrar os conhecimentos adquiridos ao longo das disciplinas desta fase, envolvendo desde a seleção e preparação de dados, até o ajuste fino de modelos avançados de aprendizado de máquina, como Llama, BERT ou MISTRAL. O fine-tuning tem o intuito de aprimorar a capacidade do modelo de compreender e gerar descrições de produtos da Amazon com base em interações textuais reais de usuários.

Este documento abrange todo o processo de desenvolvimento, incluindo a preparação do dataset, a escolha do modelo, a execução do fine-tuning e a análise dos resultados obtidos, demonstrando a evolução do modelo após o treinamento e sua capacidade de gerar respostas relevantes.

II. MATERIAIS E MÉTODOS

Dataset

O Dataset base utilizado foi o definido no próprio Tech Challenge, sendo este o [“The AmazonTitles-1.3MM”](#).

Sendo ainda mais específico, foi utilizado o arquivo “trn.json”, em que continha as informações de títulos e descrições dos produtos.

Seleção de Modelo para Fine Tuning

O método de escolha de modelo foi realizado a partir de experimentação e testes com diferentes modelos e infraestrutura, por cada integrante do time.

Combinações experimentadas:

- Mistral 7 b | Llama 7b , sem unsloth pelo google colab máquina base;
- Gemma 2b pelo google colab máquina A100;
- Opt 1.3b com Collab pago gpu simples;
- Opt 125m com collab pago;

Em todos esses casos mencionados acima, teve-se problemas devido a necessidade de uma infraestrutura ainda mais robusta ou resultados não satisfatórios.

Por fim, optamos utilizarmos de um provedor de LLM e que executa em “cloud”. Sendo este o Google através do Google AI Studio, que fornece serviços de IA Generativa com a Gemini API. As vantagens de utilização desses serviços optando pelo Gemini foi:

- Modelo ainda mais poderoso que os testados anteriormente;
- Utilização via serviço, ou seja, sem preocupação de infraestrutura;
- O Google Cloud oferece uma Tier gratuita de Project, de forma que não possua custos para executar LLM na Gemini API. O próprio serviço de Fine Tuning é livre de cobranças;

Nesse sentido, o modelo final utilizado foi “gemini-1.5-flash-001-tuning” através dos serviços e lib do Google AI Studio / Cloud para generative ai. Foi necessário realizar tratamentos e limpeza no dataset para sua utilização.

Infraestrutura

Conforme já mencionado os testes realizados no tópico anterior, a execução final do notebook Python desenvolvido para o fine tuning e teste do modelo tuned, foi feita através da máquina local e um enviroment python criado exclusivamente para o projeto.

Versões

Ainda sobre a o método de desenvolvimento com o último modelo, foi realizado duas versões. Uma inicial que serviu como prova de conceito sobre o model e continha diretamente o descritivo do produto como resposta nos treinamentos; e uma outra que, com o resultado relativamente satisfatório, foi incrementado a source do dataset na resposta em seu novo treinamento.

Plus RAG

Como parte a mais, foi realizado testes de método de RAG para verificar o resultado final em comparação com o resultado das versões oficiais com fine tuning. Foi executado através do Collab.

Para tanto, foi realizado o framework lang chain, bem como sua integração com Chroma DB, banco de dados vetorial e um modelo genérico de embeddings.

III. Desenvolvimento Versão Final

Neste projeto, foram realizadas duas versões distintas de implementação para o fine-tuning do modelo, utilizando o dataset "The AmazonTitles-1.3MM". Cada uma das versões foi desenvolvida com o objetivo de validar a viabilidade do modelo e, posteriormente, melhorar a qualidade das respostas geradas, conforme descrito a seguir.

É valido pontuar que esse tópico pode ser melhor explorado diretamente nos próprios notebooks contido no repositório referência em anexo, já que ele como um todo

está documentado e explicado passo a passo do que foi desenvolvido. Aqui se encontra um resumo dos objetivos e decisões tomadas.

Inicial

Na versão inicial do projeto final, o objetivo principal era comprovar que o modelo poderia ser ajustado para gerar respostas baseadas na descrição dos produtos, presente na coluna content do dataset. Para isso:

- O dataset foi carregado e pré-processado, removendo-se dados irrelevantes e garantindo que os prompts fossem formados com as colunas de título (title) e descrição (content).
- O modelo selecionado para esta versão foi treinado com as descrições de produtos, de forma que, ao receber uma pergunta sobre o título de um produto, ele fosse capaz de retornar uma resposta diretamente relacionada ao conteúdo daquela descrição.
- O fine-tuning foi realizado utilizando uma configuração padrão e sugerido pela documentação da Google, porém com um número limitado de épocas e parâmetros básicos, para avaliar o desempenho do modelo.
 - Modelo base = 'models/gemini-1.5-flash-001-tuning'
 - Épocas = 2
 - Devido a limitação pela quantidade de registros
 - Tamanho Batch = 4
 - Learning Rate = 0,001
- Os resultados foram satisfatórios para a prova de conceito, uma vez que o modelo conseguiu gerar respostas coerentes com as descrições de produtos. Um exemplo de resposta gerada pode ser visto no seguinte cenário:
 - Exemplo: Pergunta: " What is the Memorex Sport CD Boombox product?"
 - Resposta gerada: "The Memorex Sport CD Boombox is a portable boombox that is perfect for the active lifestyle. It features a built-in AM/FM radio, CD player, and auxiliary input for MP3 players. The boombox also has a built-in clock and alarm, and a 3.5mm

headphone jack. The Memorex Sport CD Boombox is a great way to enjoy your favorite music while on the go.”

Esse primeiro treinamento mostrou que o modelo tinha a capacidade de gerar descrições satisfatórias dos produtos a partir dos títulos, confirmando a viabilidade do ajuste fino.

É válido destacar que o processo de fine tuning é demorado, sendo que levou por volta de 43 horas para seu treinamento.

Versão Final com Fontes

Após os resultados positivos da prova de conceito, uma segunda versão do modelo foi desenvolvida, incorporando uma nova funcionalidade: além de gerar a descrição do produto, o modelo também deveria incluir a source (fonte) das informações fornecidas, de modo a melhorar a transparência e confiabilidade das respostas.

- O processo de pré-processamento foi ajustado para garantir que as fontes fossem extraídas corretamente do arquivo trn.json, presente no dataset.
- Novamente, o modelo foi treinado, agora com o ajuste adicional de incluir a source como parte da resposta gerada. Isso foi alcançado ao modificar os prompts de treinamento para adicionar a origem das descrições (ex.: o arquivo trn.json).
- O fine-tuning foi realizado utilizando uma configuração padrão e sugerido pela documentação da Google, porém com um número limitado de épocas e parâmetros básicos, para avaliar o desempenho do modelo.
 - Modelo base = ‘models/gemini-1.5-flash-001-tuning’
 - Épocas = 2
 - Devido a limitação pela quantidade de registros
 - Tamanho Batch = 4
 - Learning Rate = 0,001
- A qualidade das respostas foi relativamente mantida e a adição das fontes enriqueceu o modelo devido a menção da fonte.
 - Exemplo: Pergunta: "What is product "5 Ft Turbo Tube Slide Yellow" on Amazon?"

- Resposta gerada: "The 5 Ft Turbo Tube Slide Yellow is a great way to cool off on a hot summer day. The slide is made of durable PVC and is easy to assemble. It is also easy to clean and store. \n***\n**Source:** 'trn.json' file in dataset [The AmazonTitles-1.3MM](<https://drive.google.com/file/d/12zH4mL2RX8iSvH0VCNnd3QxO4DzuHWnK/view>)"
- Ao longo desta avaliação, foi testado resposta com hiper parâmetro de temperatura mais 0, buscando diminuir a criatividade.

Essa segunda versão acaba estando mais em linha com a solicitação, em visto que as respostas geradas traziam, além da descrição do produto, a referência da fonte original dos dados, o que facilita sua validação.

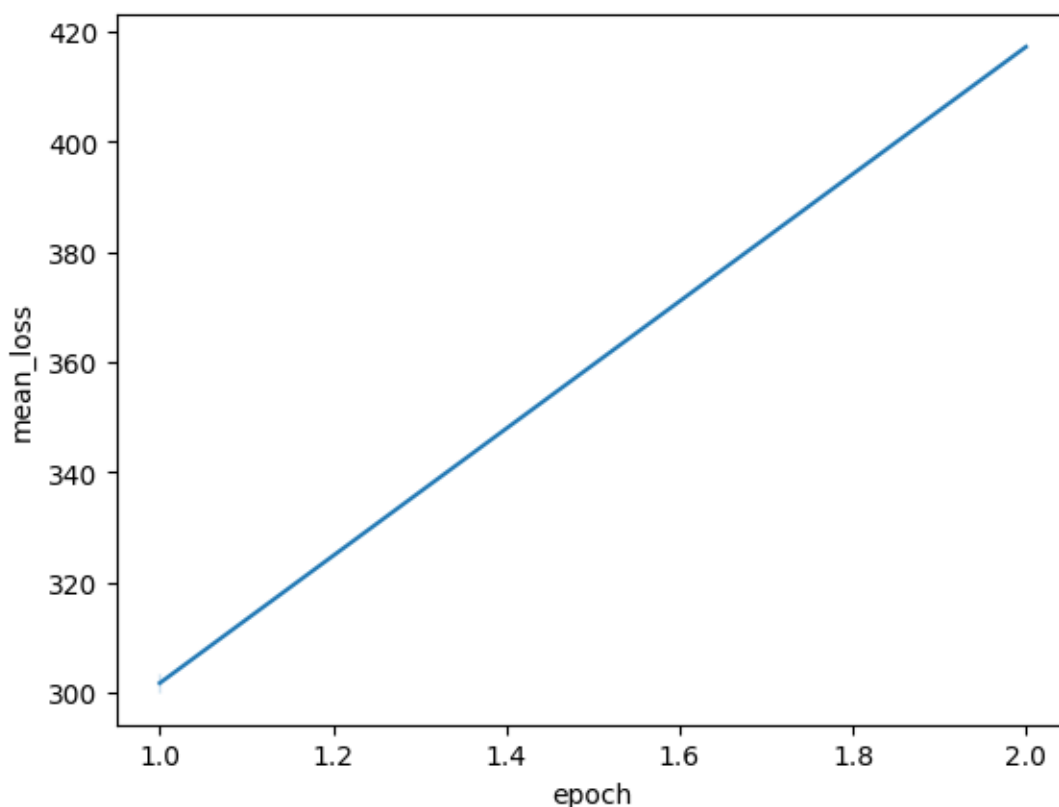
É válido destacar que o processo de fine tuning é demorado, sendo que levou por volta de 43 horas para seu treinamento.

Plus RAG

No desenvolvimento dessa entrega a mais, não foi utilizado modelo de fundacional de Inteligência Artificial Generativa, tendo em vista que a finalidade era apenas a avaliação e comparação da capacidade de recuperação de informação.

IV. RESULTADOS

Olhando para os resultados do primeiro fine tuning, com a versão sem sources, obtivemos a seguinte métrica para a Média de Perca (mean_loss) VS Épocas (epoch):



Plot de Mean Loss X Epoch para a versão sem sources na resposta

Percebe-se que a curva de perda aumenta, no entanto deve ser levado em consideração que há um número baixo de épocas, sendo apenas duas, devido a limitação do modelo em relação a quantidade de registros de treinamento solicitada no desafio.

Pode-se levar em consideração, que para o caso de uso em específico, pode ser preferível que leve um overffiting, do que a generalização no modelo, já que o objetivo é que respondesse de fato com os dados para aquelas perguntas treinadas e não que estivéssemos treinando um comportamento propriamente dito, como sumarização por exemplo. Então aqui se faz sentido esse tipo de resultado.

Ainda em relação a primeira versão, podemos comparar algumas respostas obtidas e conteúdo original:

- **Antes do fine tuning:** 'Unfortunately, I cannot find any information about a product called "Memorex Sport CD Boombox". It's possible that:\n\n*
The product name is incorrect: There might be a slight variation in the name, or it might be a different model altogether.\n* **The product

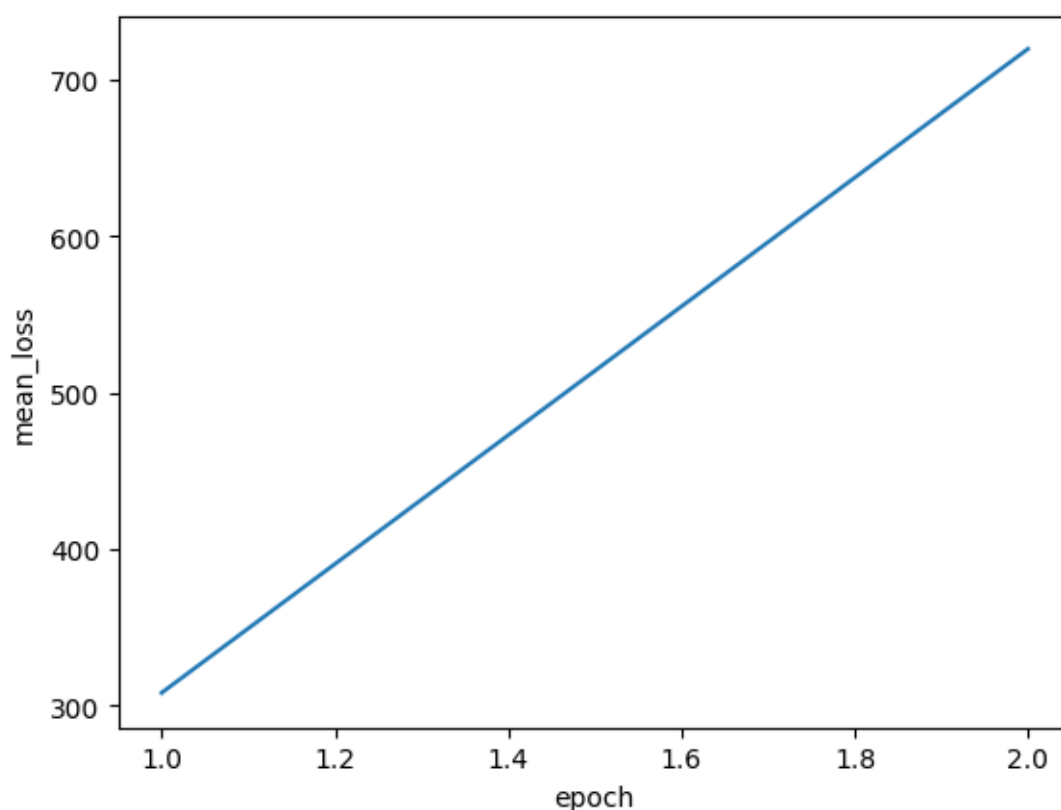
is discontinued:** Memorex, like many companies, discontinues older models, and information about them might be scarce online.\n* ****The product was region-specific:**** It's possible the product was only sold in a certain region and isn't widely known.\n\n****To find more information, I recommend:****\n\n1. ****Double-checking the product name:**** Ensure you have the correct name and try searching again.\n2. ****Searching for similar Memorex products:**** Look for other Memorex boomboxes or CD players and see if any resemble the one you are looking for.\n3. ****Checking online marketplaces:**** Sites like eBay or Amazon might have old product listings that can provide more details.\n4. ****Contacting Memorex directly:**** You can try reaching out to Memorex customer service for assistance in finding information about the product.\n\nHopefully, with a bit more investigation, you'll be able to find what you are looking for. \n'

- **Depois do fine tuning:** 'The Memorex Sport boombox brings a sleek styling look with quality stereo sound. Featuring a double cassette tape deck, radio and CD player with CDMP3 compatibility, the Memorex Sport offers many entertainment options in one unit. The Sport features a high-power output speaker that provides a powerful sound experience for up to 100 hours of play time. It also features an adjustable bass and treble control and an LCD display for radio stations and clock. The Sport has a 3.5mm audio input that lets you connect your favorite MP3 players and digital audio players to the Sport via the auxiliary line at the back. In addition, the Sport can be set to play radio, CD and tape at random or play CD/tape repeat. It also has an alarm function that includes a sleep timer and a timer for CD, tape and radio. You can also utilize its 5-band equalizers for a clear and powerful sound. The Sport is easy to use with its front panel controls and display. Other features include a headphone jack and a detachable AC/DC power cord.'
- **Depois do fine tuning com temperature 0:** 'The Memorex Sport CD Boombox is a portable boombox that is perfect for the active lifestyle. It features a built-in AM/FM radio, CD player, and auxiliary input for MP3 players. The boombox also has a built-in clock and alarm, and a 3.5mm

headphone jack. The Memorex Sport CD Boombox is a great way to enjoy your favorite music while on the go.'

- **Content Original:** 'This uniquely designed, sporty boombox features a rugged outer shell, perfect for toting to outside sporting events, the beach, or your yard. Memorex Portable Sport CD Boombox features an AM/FM radio with digital readout, a programmable CD player, and an audio input for connecting your MP3 player or other digital audio device. MP3851BLK includes a bass boost system, AC/DC operation, and has CD, CD-RV-RW playback.'

Já olhando para a última versão com sources na resposta, obteve-se um resultado de métrica com comportamento parecido para seguinte métrica para a Média de Perca (mean_loss) VS Épocas (epoch):

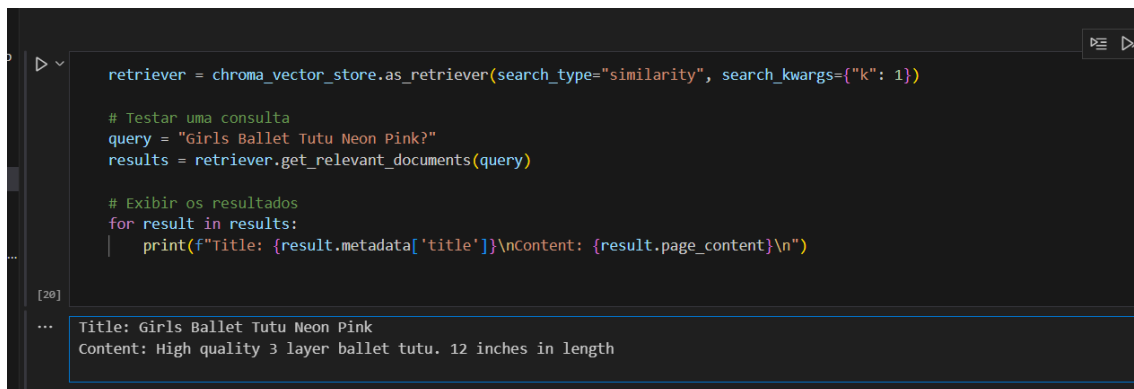


Plot de Mean Loss X Epoch para a versão com sources na resposta

Podemos comparar para esta versão também algumas respostas obtidas e conteúdo original:

- **Antes do fine tuning:** 'I do not have access to real-time information, including product details from online retailers like Amazon. To find information about the "5 Ft Turbo Tube Slide Yellow" on Amazon, you'll need to:\n\n1. **Go to Amazon's website.**\n2. **Search for "5 Ft Turbo Tube Slide Yellow".** This will likely bring up several results.\n3. **Look through the results** to find the specific product you're interested in. \n4. **Read the product description and customer reviews** to get more information about it.\n\nRemember, product availability and prices can change frequently. \n'
- **Depois do fine tuning:** "Play on all day long with the slide from 5' turbo slide with a smooth glossy slide. The slides are made of a heavy gauge 3/4 in PVC material that resists impact from slides or rocks, and the slide is resistant to fading. Slides are sold as a complete kit and include all hardware, a PVC spray bottle, and instructions. All kits include slide, galvanized metal hardware, slide poles, end caps, slide stakes, and installation instructions. Slides do not include water hose attachments. \n*** \n**Source:** 'trn.json' file in dataset [The AmazonTitles-1.3MM](<https://drive.google.com/file/d/12zH4mL2RX8iSvH0VCNnd3QxO4DzuHWnK/view>)"
- **Depois do fine tuning com temperature 0:** "'The 5 Ft Turbo Tube Slide Yellow is a great way to cool off on a hot summer day. The slide is made of durable PVC and is easy to assemble. It is also easy to clean and store. \n*** \n**Source:** 'trn.json' file in dataset [The AmazonTitles-1.3MM](<https://drive.google.com/file/d/12zH4mL2RX8iSvH0VCNnd3QxO4DzuHWnK/view>)'"
- **Content Original:** "The new 5' Turbo Tube Slide is full of turning twisting fun. This slide will take most play sets with a 5' deck to an entirely new level adding a new slide component to your child's fun and enjoyment. This thermo-formed HDPE slide comes with a lifetime warranty and will fit most 5' decks. Additional lumber purchase may be necessary depending on the configuration of your play set."

Para finalizar os resultados, é interessante mencionar a parte complementar feita em RAG, a qual obteve um bom resultado na recuperação de informações.

A screenshot of a Jupyter Notebook cell with a dark background. The code in the cell defines a retriever, sets a query to "Girls Ballet Tutu Neon Pink?", and prints the results. The output below the code shows the retrieved title and content.

```
retriever = chroma_vector_store.as_retriever(search_type="similarity", search_kwargs={"k": 1})

# Testar uma consulta
query = "Girls Ballet Tutu Neon Pink?"
results = retriever.get_relevant_documents(query)

# Exibir os resultados
for result in results:
    print(f>Title: {result.metadata['title']}\nContent: {result.page_content}\n")
```

[20]

... Title: Girls Ballet Tutu Neon Pink
Content: High quality 3 layer ballet tutu. 12 inches in length

Imagem da célula de resultado com RAG

V. CONCLUSÃO

Conclui-se que o trabalho de fine tuning realizado de fato preparou o foundation model ou base model, como é chamado pelo Google, para responder sobre o conteúdo dos produtos encontrados na dataset, tendo em vista que executando antes do fine tuning o modelo respondia não ter informações sobre o que se era perguntado, e posterior ao treinamento eles tiveram a capacidade de formar respostas que continha as características procuradas pela pergunta.

Embora a loss tenha se elevado, é importante destacar que devido as limitações do modelo devido a alta quantidade de registro solicitada no desafio, ainda foi possível atender o objetivo. E pode se inferir que para o caso de uso em específico, pode ser preferível que mesmo que leve um overffiting, o que explicaria a loss alta, do que a generalização no modelo, já que o objetivo é que respondesse de fato com os dados para aquelas perguntas treinadas e não que estivéssemos treinando um comportamento propriamente dito, como sumarização por exemplo. Então aqui se faz sentido esse tipo de resultado.

Sendo assim, o objetivo de se ter um modelo fine tuned que responda a descritivo dos produtos do dataset da Amazon foi atendido. Destacando que pode se ter resultados ainda melhores com um dataset mais equilibrado e possibilidade de executar mais épocas.

Pode-se destacar também que com o teste realizado em RAG, possa ser uma alternativa mais interessante de custo benefício devido o resultado obtido.

VI. REFERÊNCIAS

Google AI for Developers - Gemini API Docs, Fine-tuning with the Gemini API.
<https://ai.google.dev/gemini-api/docs/model-tuning> . Visto em Setembro de 2024.

Google AI for Developers - Gemini API Docs, Authentication with OAuth quickstart.
<https://ai.google.dev/gemini-api/docs/oauth> . Visto em Setembro de 2024.

Google AI for Developers - Gemini API Docs, Fine-tuning tutorial. <https://ai.google.dev/gemini-api/docs/model-tuning/tutorial?lang=python> . Visto em Setembro de 2024.

VII. ANEXOS

Código fonte - Repositório: <https://github.com/LeonardoFOliveira/ia-para-devs-tech-challenge-fase3>

Vídeo explicativo: <https://youtu.be/5tk9BoeW3zQ>