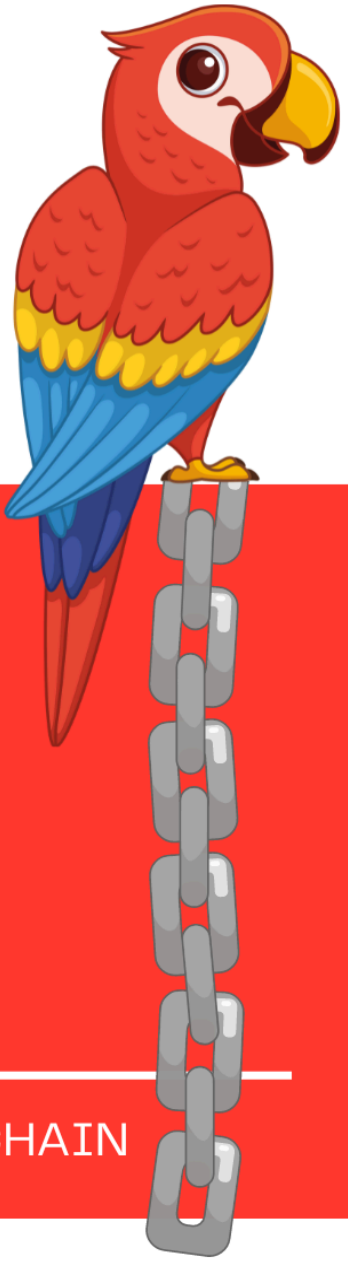
UNIVERSIDADE FEDERAL DE GOIÁS / INSTITUTO DE INFORMÁTICA



Machine Learning Operations e Large Language Models

INTEGRAÇÃO E INOVAÇÃO COM A STACK LANGCHAIN

UFG
UNIVERSIDADE FEDERAL DE GOIÁS

Gabriel da Mata Marques

UNIVERSIDADE FEDERAL DE GOIÁS (UFG)
INSTITUTO DE INFORMÁTICA (INF)

GABRIEL DA MATA MARQUES

MACHINE LEARNING OPERATIONS E LARGE LANGUAGE MODELS

Integração e inovação com a stack LangChain

Goiânia
2024

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TRABALHO DE CONCLUSÃO DE CURSO DE GRADUAÇÃO NO REPOSITÓRIO INSTITUCIONAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio do Repositório Institucional (RI/UFG), regulamentado pela Resolução CEPEC no 1240/2014, sem ressarcimento dos direitos autorais, de acordo com a Lei no 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo dos Trabalhos de Conclusão dos Cursos de Graduação disponibilizado no RI/UFG é de responsabilidade exclusiva dos autores. Ao encaminhar(em) o produto final, o(s) autor(a)(es)(as) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do Trabalho de Conclusão de Curso de Graduação (TCCG)

Nome(s) completo(s) do(a)(s) autor(a)(es)(as): **GABRIEL DA MATA MARQUES**

Título do trabalho:

MACHINE LEARNING OPERATIONS E LARGE LANGUAGE MODELS

Integração e inovação com a stack LangChain

2. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador) Concorda com a liberação total do documento [**X**] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante: a) consulta ao(à)(s) autor(a)(es)(as) e ao(à) orientador(a); b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo do TCCG. O documento não será disponibilizado durante o período de embargo.

Casos de embargo:

- Solicitação de registro de patente;

- Submissão de artigo em revista científica;

- Publicação como capítulo de livro.

Obs.: Este termo deve ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **Gabriel Da Mata Marques**, **Discente**, em 16/02/2024, às 00:34, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Fernando Marques Federson**, **Professor do Magistério Superior**, em 12/09/2024, às 11:04, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4383345** e o código CRC **D287A78C**.

Referência: Processo nº 23070.008376/2024-10 SEI nº 4383345

GABRIEL DA MATA MARQUES

MACHINE LEARNING OPERATIONS E LARGE LANGUAGE MODELS

Integração e inovação com a stack LangChain

Relatório final de Trabalho de Conclusão de Curso, apresentado à Universidade Federal de Goiás, como parte das exigências para a obtenção do título de Bacharel em Inteligência Artificial.
Orientador: Prof. Dr. Fernando Marques Federson

Goiânia
2024

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

MARQUES, GABRIEL DA MATA
MACHINE LEARNING OPERATIONS E LARGE LANGUAGE MODELS [manuscrito] : Integração e inovação com a stack LangChain / GABRIEL DA MATA MARQUES. - 2024.
58 f.

Orientador: Prof. Dr. FERNANDO MARQUES FEDERSON.
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Goiás, Instituto de Informática (INF), Inteligência Artificial, Goiânia, 2024.

1. inteligência artificial. 2. machine learning operations. 3. large language models. I. FEDERSON, FERNANDO MARQUES, orient. II. Título.

CDU 004

GABRIEL DA MATA MARQUES

MACHINE LEARNING OPERATIONS E LARGE LANGUAGE MODELS

Integração e inovação com a stack LangChain

Relatório final de Trabalho de Conclusão de Curso, apresentado à Universidade Federal de Goiás, como parte das exigências para a obtenção do título de Bacharel em Inteligência Artificial.

Data da Aprovação: 08 de fevereiro de 2024.

Prof. Dr. Fernando Marques Federson
Orientador (INF-UFG)

Prof. Dr. Aldo André Díaz Salazar
Coordenador de TCC do BIA (INF-UFG)

Prof. Dr. Vinícius Sebba Patto
Coordenador do BIA (INF-UFG)

gov.br
Documento assinado digitalmente
LEONARDO AFONSO AMORIM
Data: 09/02/2024 09:41:30-0300
Verifique em https://validar.iti.gov.br

Dr. Leonardo Afonso Amorim
(CEIA-UFG)

GABRIEL DA MATA MARQUES

MACHINE LEARNING OPERATIONS E LARGE LANGUAGE MODELS

Integração e inovação com a stack LangChain

RESUMO

Este Relatório de Conclusão de Curso tem como objetivo reunir os resultados da minha jornada para me tornar um especialista em **MLOps (Large Language Models)**. Uma ilustração e sua narrativa descrevem os períodos de trabalho. Os Apêndices contêm os Termos de Aceite de Entrega e os resultados obtidos durante cada período de trabalho.

Palavras-chave: inteligência artificial, machine learning operations, large language models.

ABSTRACT

This Course Completion Report aims to bring together the results of my journey to become an expert in **MLOps (Large Language Models)**. An illustration and its narrative describe the work periods. The Appendices contain the Delivery Acceptance Terms and the results obtained during each work period.

Keywords: artificial intelligence, machine learning operations, large language models.

Goiânia

2024

Minha Jornada

Gabriel da Mata Marques

Especialista em: MLOps (Large Language Models)

MINHA JORNADA

Nome: Gabriel da Mata Marques

Especialidade: MLOps (Large Language Models)

Objetivo deste documento

Durante o processo da disciplina Residência em IA[1], foram gerados diversos resultados na construção da minha especialização. A cada semana, um conjunto de resultados foi formalizado por um Termo de Aceite de Entrega e avaliado por uma banca, considerando o planejado e o realizado para o período. Este documento tem como objetivo descrever esses resultados obtidos, fazendo referência aos Termos de Aceite de Entrega e seus documentos associados.

Minha Jornada

Minha jornada de especialização teve um início estratégico e prático na **Semana 1**, com a tarefa de categorizar o time encarregado do tópico MLOps (*Machine Learning Operations*). Esta etapa inicial envolveu não apenas a organização dos estudos com base nas diretrizes estabelecidas pela *Conference on Computational Science and Computational Intelligence* (CSCI), mas também uma pesquisa minuciosa nos registros do evento por trabalhos relacionados ao tema MLOps. A execução desta tarefa foi essencial para estabelecer um alicerce sólido e bem informado para todo o meu aprendizado e desenvolvimento subsequente na área de especialização escolhida. Os resultados obtidos, incluindo a categorização conforme os critérios da CSCI e o levantamento dos estudos sobre MLOps publicados nos anais da CSCI de 2018 a 2022, forneceram uma base de dados valiosa e um ponto de partida claro para o meu percurso de especialização.

Na **Semana 2**, embarquei na fase inicial de um processo detalhado de revisão bibliográfica, focando nas referências que fundamentam os trabalhos do nosso grupo sobre

[1] Dez semanas, entre setembro de 2023 e janeiro de 2024.

a temática de MLOps. Este período foi dedicado a expandir a busca para além dos registros da CSCI, incluindo uma variedade de artigos e publicações de outras bases científicas. A construção de um repositório com os trabalhos encontrados e a realização de uma revisão preliminar, acompanhada de breves resumos para cada artigo, foram tarefas fundamentais que realizei. Essas atividades foram cruciais para uma compreensão mais ampla e aprofundada do campo de MLOps. Adicionalmente, a estruturação do processo de revisão bibliográfica com a ferramenta Parsifal e a criação da Table Zero para unificar, resumir e analisar os trabalhos encontrados (**APÊNDICE 1**), representaram etapas importantes na modelagem do caminho que minha jornada de aprendizado seguiria a partir de então.

Durante as **Semanas 3 e 4**, minha jornada de especialização em MLOps avançou significativamente. Na **Semana 3**, conclui o processo de revisão bibliográfica, um passo fundamental para identificar trabalhos de referência essenciais na área de MLOps. A finalização deste processo não foi apenas uma tarefa de compilação; envolveu uma análise criteriosa dos artigos, resultando na criação de um repositório com os artigos encontrados e suas respectivas análises, disponível na Table Zero (**APÊNDICE 2**). Além disso, construímos um cronograma geral de atividades a serem desenvolvidas até o final da Residência, estabelecendo um plano de referência para os próximos passos do projeto. Simultaneamente, iniciamos a elaboração de um vocabulário específico sobre MLOps, consolidando termos e definições que formariam a linguagem técnica de nosso trabalho.

Na **Semana 4**, aprofundamos ainda mais no embasamento teórico, finalizando o vocabulário com a inclusão de conceitos de DevOps, uma etapa vital para entender a origem de MLOps. Essa expansão do vocabulário reforçou minha compreensão dos conceitos-chave e terminologias, fundamentais para o entendimento dos conceitos futuramente encontrados. Também dedicamos a buscar repositórios de código(**APÊNDICE 2**) que seriam úteis para nosso projeto, selecionando recursos que melhor se alinhem com nossos objetivos. Por fim, participei ativamente no planejamento das aplicações de cada integrante do grupo, partindo de uma discussão entre todos membros sobre objetivos pessoais e seus respectivos requisitos.

Durante a **Semana 5**, minha jornada de especialização se concentrou na finalização da construção do embasamento teórico dos trabalhos que seriam desenvolvidos. Esta etapa envolveu uma revisão e busca por artigos referentes à implantação de *Large Language*

Models (LLMs), que apesar da relevância do tema, não foi possível encontrar uma base suficiente na literatura. Além disso, elaborei um plano de estudo detalhado para os itens apresentados em langchain_stack.png (**APÊNDICE 3**). O Framework LangChain foi escolhido como objeto de estudo para a sequência da Residência em IA. Este plano de estudo foi essencial para estabelecer paralelos entre os conceitos novos encontrados no LangChain e os já presentes no vocabulário que havia desenvolvido.

Na **Semana 6**, porém, enfrentei alguns desafios pessoais que impactaram minha participação em dois *Gates*. Apesar dessas complicações, mantive meu foco na continuidade do aprendizado, dedicando-me a entender conceitos fundamentais anteriores ao Framework LangChain e que são amplamente utilizados neste. Fiz um estudo aprofundado sobre LORA, resumindo os principais tópicos neste documento, e sobre Retrievers, compilando as informações relevantes em outro documento. Estes estudos, presentes em no **APÊNDICE 4**, não apenas enriqueceram minha compreensão desses conceitos essenciais, mas também me permitiram fazer conexões mais profundas com o Framework LangChain.

Avançando para a **Semana 7**, minha jornada de especialização tomou um rumo ainda mais detalhado e técnico. Com o embasamento teórico já bem estabelecido e uma compreensão sólida dos conceitos fundamentais de MLOps e LLMs, dediquei-me à elaboração de um documento abrangente que explicasse cada sistema interagente e interdependente do LangChain (**APÊNDICE 5**). A criação deste documento envolveu uma análise cuidadosa de como cada elemento do LangChain contribui para o funcionamento do conjunto, destacando a sinergia entre os diferentes sistemas e como eles se complementam para criar um framework robusto e eficiente.

Na **Semana 8**, minha jornada de especialização em MLOps e LLMs alcançou um novo patamar com um foco intensificado nos estudos e na elaboração de documentação para LangServe e LangSmith (**APÊNDICE 6**). O LangServe, fundamental para a implantação dos LLMs, foi analisado sob a perspectiva dos conceitos de MLOps e DevOps que já havia explorado. Dediquei-me a entender e documentar como o LangServe integra esses conceitos, facilitando a implantação eficiente dos modelos de linguagem. A documentação que elaborei não apenas detalha as funcionalidades técnicas do LangServe, mas também

explica como ele se encaixa no ecossistema mais amplo de MLOps e DevOps, fornecendo *insights* valiosos para a implementação prática desses modelos em projetos reais.

Por outro lado, o LangSmith, essencial para os conceitos de monitoramento e implantação contínua dos LLMs, foi explorado com um olhar crítico sobre sua aplicabilidade e eficácia. A documentação produzida para o LangSmith aborda como esta ferramenta pode ser utilizada para monitorar e manter a saúde dos LLMs em operação, assegurando sua eficiência e eficácia ao longo do tempo. Além disso, compreendi como o LangSmith se alinha com os conceitos de MLOps, proporcionando uma abordagem sistemática e contínua para a gestão de modelos linguísticos em ambientes de produção.

Dessa forma, ao refletir sobre as oito semanas desta jornada de especialização em MLOps para LLMs, é evidente que cada etapa foi essencial para a construção de um conhecimento profundo e aplicável. Desde a definição inicial da área de especialização até o aprofundamento em ferramentas específicas como LangServe e LangSmith, cada semana representou um avanço significativo na minha compreensão e habilidade prática.

O estudo de conceitos anteriores ao LangChain, como LORA e Retrievers, adicionou camadas de compreensão, preparando o terreno para uma exploração mais profunda. A elaboração de documentação detalhada para cada sistema interagente da Stack LangChain e a imersão em LangServe e LangSmith foram etapas primordiais que demonstraram o potencial de coesão e eficácia envolvendo todo processo de desenvolvimento e implantação de aplicações de LLMs.

Em conclusão, esta experiência de especialização em MLOps e LLMs foi uma jornada de aprendizado contínuo, desafios superados e desenvolvimento pessoal e profissional. As habilidades e conhecimentos adquiridos ao longo destas oito semanas são inestimáveis e me equiparam não apenas para enfrentar os desafios atuais no campo de MLOps, mas também para inovar e contribuir significativamente para futuros desenvolvimentos nesta área dinâmica e em constante evolução.

APÊNDICE 1

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("gate") de aprovação: 19 de out. de 2023

Participantes da Entrega [matriculados em Residência em IA]:

Állan Christoffer Pereira Silva
Gabriel da Mata Marques
Heinz Felipe Cavalcante Rahmig

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

A tarefa realizada envolveu a categorização do time encarregado do tópico MLOps.

Para a realização desta tarefa, as seguintes exigências eram necessárias:

- Organizar os estudos com base nas diretrizes estabelecidas pela Conference on Computational Science and Computational Intelligence (CSCI).
- Procurar nos registros do evento por pesquisas relacionadas ao tema MLOps.

Os resultados obtidos com essa atividade estão disponíveis nos links a seguir:

- Categorização conforme os critérios da CSCI: Link para o documento
- Levantamento dos estudos relacionados ao MLOps publicados nos anais da CSCI de 2018 a 2022: Link para a planilha

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

Para a próxima data de entrega, em 26 de outubro de 2023, as atividades programadas incluem:

- Prospecção de artigos e publicações em diferentes bases de dados científicas.
- Desenvolvimento de um acervo digital contendo os trabalhos identificados.
- Análise e síntese dos estudos mais significativos.

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

Repositório do grupo de trabalho:https://github.com/AllanSilva156/mlops-residencia-ia

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

LUANA GUEDES BARROS MARTINS: Go!

Entrega Gate 19/10/2023

Introdução

Este documento faz parte da entrega referente ao gate do dia 19 de outubro de 2023. Nele está detalhada a classificação do grupo de pesquisa responsável pela temática de MLOps.

Responsáveis pela Entrega:

Állan Christoffer Pereira Silva
Gabriel da Mata Marques
Heinz Felipe Cavalcante Rahmig

Classificação segundo Conference on Computational Science and Computational Intelligence (CSCI):

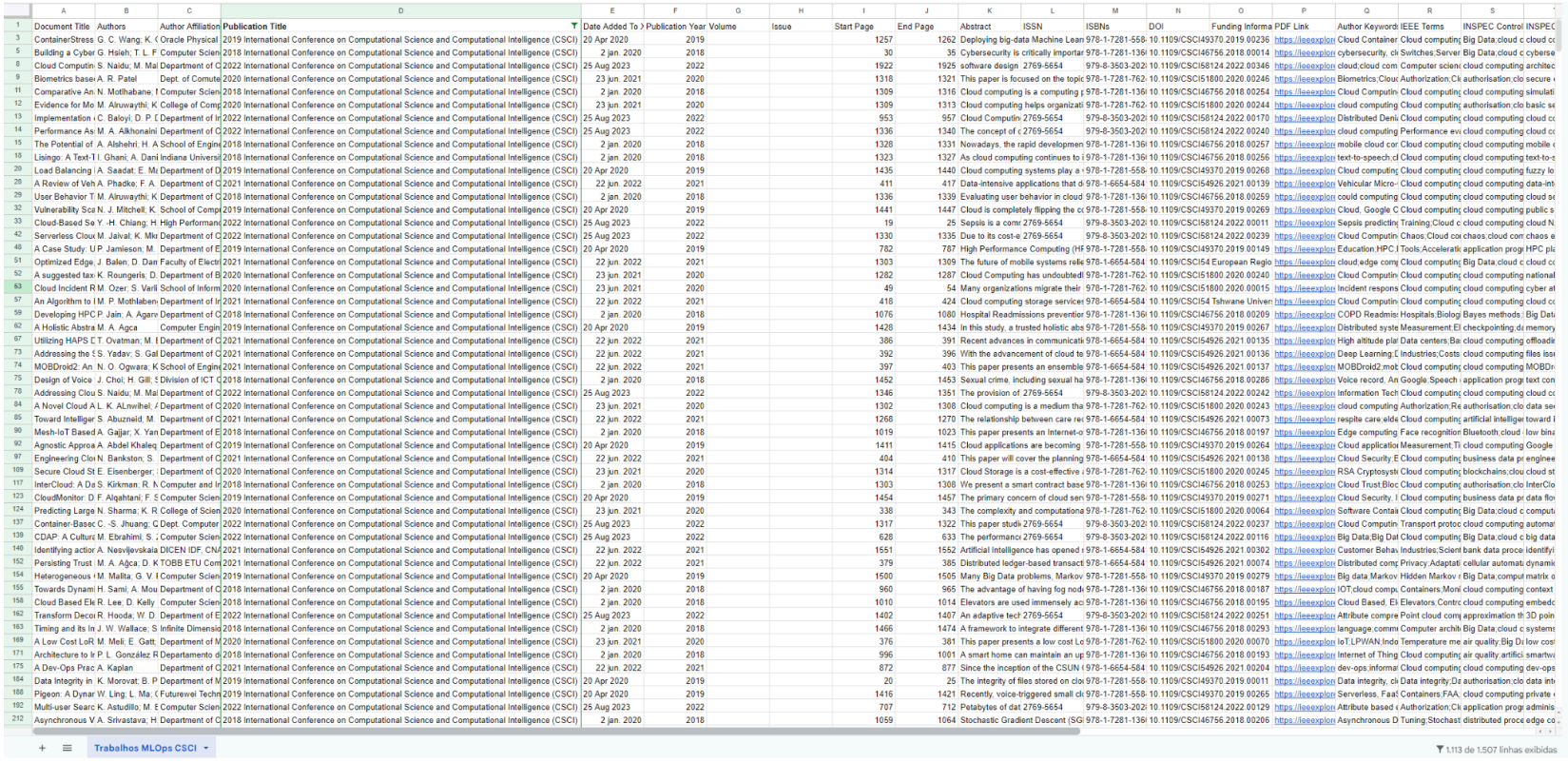
Research Track on Big Data and Data Science (CSCI-RTBD)

SECURITY & PRIVACY IN THE ERA OF DATA SCIENCE & BIG DATA:
- Privacy Preserving Big Data Collection

INFRASTRUCTURES FOR BIG DATA & DATA SCIENCE:
- Cloud Based Infrastructures (applications, storage & computing resources)
- HPC, including Parallel & Distributed Processing
- Programming Models and Environments to Support Big Data
- Software and Tools for Big Data
- Big Data Open Platforms
- Emerging Architectural Frameworks for Big Data
- Paradigms and Models for Big Data beyond Hadoop/MapReduce

BIG DATA & DATA SCIENCE MANAGEMENT AND FRAMEWORKS:
- Database and Web Applications
- Massively Parallel Processing (MPP) Databases
- Distributed Database Systems
- Distributed File Systems
- Distributed Storage Systems
- Data Preservation and Provenance
- Data Protection Methods
- Data Integrity and Privacy Standards and Policies

Levantamento dos estudos relacionados ao MLOps publicados nos anais da CSCI de 2018 a 2022:



Document Title	Authors	Author Affiliation	Publication Title	Date Added To X	Publication Year	Volume	Issue	Start Page	End Page	Abstract	ISSN	ISBNs	DOI	Funding Informa	PDF Link	Author Keywords	IEEE Terms	INSPEC Control	INSPEC
ContainerStress	G. C. Wang; K. (	Oracle Physical	2019 International Conference on Computational Science and Computational Intelligence (CSCI)	20 Apr 2020	2019			1257	1262	Deploying big-data Machine Lear	978-1-7281-558	10.1109/CSCI49370.2019.00236	https://ieeexplor	Cloud Container	Cloud computing	Big Data;cloud c	cloud cc		
Building a Cyber	G. Hsieh; T. L. F	Computer Scien	2018 International Conference on Computational Science and Computational Intelligence (CSCI)	2 jan. 2020	2018			30	35	Cybersecurity is critically importar	978-1-7281-136	10.1109/CSCI46756.2018.00014	https://ieeexplor	cybersecurity, cl	Switches;Server	Big Data;cloud c	cyberse		
Cloud Computin	S. Naidu; M. Mal	Department of C	2022 International Conference on Computational Science and Computational Intelligence (CSCI)	25 Aug 2023	2022			1922	1925	software design	2769-5654	979-8-3503-202	10.1109/CSCI58124.2022.00346	https://ieeexplor	cloud;cloud com	Computer scienc	cloud computing	architec	
Biometrics base	A. R. Patel	Dept. of Comute	2020 International Conference on Computational Science and Computational Intelligence (CSCI)	23 jun. 2021	2020			1318	1321	This paper is focused on the topic	978-1-7281-762	10.1109/CSCI51800.2020.00246	https://ieeexplor	Biometrics;Clou	Authorization;Ck	authorisation;clo	secure		
Comparative An	N. Motlhabane; I	Computer Scien	2018 International Conference on Computational Science and Computational Intelligence (CSCI)	2 jan. 2020	2018			1309	1316	Cloud computing is a computing	978-1-7281-136	10.1109/CSCI46756.2018.00254	https://ieeexplor	Cloud Computin	Cloud computing	cloud computing	simulati		
Evidence for Mo	M. Alruwaythi; K	College of Comp	2020 International Conference on Computational Science and Computational Intelligence (CSCI)	23 jun. 2021	2020			1309	1313	Cloud computing helps organizati	978-1-7281-762	10.1109/CSCI51800.2020.00244	https://ieeexplor	cloud computing	Cloud computing	authorisation;clo	basic se		
Implementation	C. Baloyi; D. P. [	Department of Ir	2022 International Conference on Computational Science and Computational Intelligence (CSCI)	25 Aug 2023	2022			953	957	Cloud Computin	2769-5654	979-8-3503-202	10.1109/CSCI58124.2022.00170	https://ieeexplor	Distributed Denia	Cloud computing	cloud computing	cloud cc		
Performance As	M. A. Alkhonaini	Department of C	2022 International Conference on Computational Science and Computational Intelligence (CSCI)	25 Aug 2023	2022			1336	1340	The concept of c	2769-5654	979-8-3503-202	10.1109/CSCI58124.2022.00240	https://ieeexplor	cloud computing	Performance eva	cloud computing	cloud cc		
The Potential of	A. Alshehri; H. A	School of Engine	2018 International Conference on Computational Science and Computational Intelligence (CSCI)	2 jan. 2020	2018			1328	1331	Nowadays, the rapid developmen	978-1-7281-136	10.1109/CSCI46756.2018.00257	https://ieeexplor	mobile cloud cor	Cloud computing	cloud computing	mobile		
Lisingo: A Text-T	I. Ghani; A. Dani	Indiana Universi	2018 International Conference on Computational Science and Computational Intelligence (CSCI)	2 jan. 2020	2018			1323	1327	As cloud computing continues to	978-1-7281-136	10.1109/CSCI46756.2018.00256	https://ieeexplor	text-to-speech;cl	Cloud computing	cloud computing	text-to-s		
Load Balancing	A. Saadat; E. Mi	Department of D	2019 International Conference on Computational Science and Computational Intelligence (CSCI)	20 Apr 2020	2019			1435	1440	Cloud computing systems play a	978-1-7281-558	10.1109/CSCI49370.2019.00268	https://ieeexplor	Cloud computing	Cloud computing	cloud computing	fuzzy lo		
A Review of Veh	A. Phadke; F. A.	Department of C	2021 International Conference on Computational Science and Computational Intelligence (CSCI)	22 jun. 2022	2021			411	417	Data-intensive applications that d	978-1-6654-584	10.1109/CSCI54926.2021.00139	https://ieeexplor	Vehicular Micro-	Cloud computing	cloud computing	data-int		
User Behavior T	M. Alruwaythi; K	Department of C	2018 International Conference on Computational Science and Computational Intelligence (CSCI)	2 jan. 2020	2018			1336	1339	Evaluating user behavior in cloud	978-1-7281-136	10.1109/CSCI46756.2018.00259	https://ieeexplor	could computing	Cloud computing	cloud computing	cloud se		
Vulnerability Sca	N. J. Mitchell; K.	School of Comp	2019 International Conference on Computational Science and Computational Intelligence (CSCI)	20 Apr 2020	2019			1441	1447	Cloud is completely flipping the c	978-1-7281-558	10.1109/CSCI49370.2019.00269	https://ieeexplor	Cloud, Google C	Cloud computing	cloud computing	public s		
Cloud-Based Se	Y. -H. Chiang; H.	High Performan	2022 International Conference on Computational Science and Computational Intelligence (CSCI)	25 Aug 2023	2022			19	25	Sepsis is a comr	2769-5654	979-8-3503-202	10.1109/CSCI58124.2022.00011	https://ieeexplor	Sepsis predictin	Training;Cloud c	cloud computing	cloud N.		
Serverless Clou	M. Jaival; K. Mik	Department of C	2022 International Conference on Computational Science and Computational Intelligence (CSCI)	25 Aug 2023	2022			1330	1335	Due to its cost-e	2769-5654	979-8-3503-202	10.1109/CSCI58124.2022.00239	https://ieeexplor	Cloud Computin	Chaos;Cloud co	chaos;cloud com	chaos e		
A Case Study: U	P. Jamieson; M.	Department of E	2019 International Conference on Computational Science and Computational Intelligence (CSCI)	20 Apr 2020	2019			782	787	High Performance Computing (HF	978-1-7281-558	10.1109/CSCI49370.2019.00149	https://ieeexplor	Education;HPC;I	Tools;Acceleratk	application progr	HPC pla		
Optimized Edge	J. Balen; D. Dan	Faculty of Electr	2021 International Conference on Computational Science and Computational Intelligence (CSCI)	22 jun. 2022	2021			1303	1309	The future of mobile systems relie	978-1-6654-584	10.1109/CSCI54	European Regio	https://ieeexplor	cloud;edge com	Cloud computing	Big Data;cloud c	cloud cc		
A suggested tax	K. Roungeris; D.	Department of B	2020 International Conference on Computational Science and Computational Intelligence (CSCI)	23 jun. 2021	2020			1282	1287	Cloud Computing has undoubtedl	978-1-7281-762	10.1109/CSCI51800.2020.00240	https://ieeexplor	Cloud Computin	Cloud computing	cloud computing	national		
Cloud Incident R	M. Ozer; S. Varli	School of Inform	2020 International Conference on Computational Science and Computational Intelligence (CSCI)	23 jun. 2021	2020			49	54	Many organizations migrate their	978-1-7281-762	10.1109/CSCI51800.2020.00015	https://ieeexplor	Incident respons	Cloud computing	cloud computing	cyber at		
An Algorithm to	M. P. Mothlabeı	Department of Ir	2021 International Conference on Computational Science and Computational Intelligence (CSCI)	22 jun. 2022	2021			418	424	Cloud computing storage services	978-1-6654-584	10.1109/CSCI54	Tshwane Univer	https://ieeexplor	Cloud Computin	Cloud computing	cloud computing	cloud cc		
Developing HPC	P. Jain; A. Agarv	Department of C	2018 International Conference on Computational Science and Computational Intelligence (CSCI)	2 jan. 2020	2018			1076	1080	Hospital Readmissions preventior	978-1-7281-136	10.1109/CSCI46756.2018.00269	https://ieeexplor	COPD Readmiss	Hospitals;Biolog	Bayes methods;	Big Data		
A Holistic Abstra	M. A. Agca	Computer Engin	2019 International Conference on Computational Science and Computational Intelligence (CSCI)	20 Apr 2020	2019			1428	1434	In this study, a trusted holistic abs	978-1-7281-558	10.1109/CSCI49370.2019.00267	https://ieeexplor	Distributed syste	Measurement;El	checkpointing;di	memory		
Utilizing HAPS E	T. Ovatman; M. I	Department of C	2021 International Conference on Computational Science and Computational Intelligence (CSCI)	22 jun. 2022	2021			386	391	Recent advances in communicati	978-1-6654-584	10.1109/CSCI54926.2021.00135	https://ieeexplor	High altitude pla	Data centers;Ba	cloud computing	offloadi		
Addressing the S	S. Yadav; S. Gal	Department of C	2021 International Conference on Computational Science and Computational Intelligence (CSCI)	22 jun. 2022	2021			392	396	With the advancement of cloud to	978-1-6654-584	10.1109/CSCI54926.2021.00136	https://ieeexplor	Deep Learning;E	Industries;Costs	cloud computing	files iss		
MOBDroid2: An	N. O. Ogwara; K	School of Engin	2021 International Conference on Computational Science and Computational Intelligence (CSCI)	22 jun. 2022	2021			397	403	This paper presents an ensemble	978-1-6654-584	10.1109/CSCI54926.2021.00137	https://ieeexplor	MOBDroid2;mot	Cloud computing	cloud computing	MOBDr		
Design of Voice	J. Choi; H. Gill; S	Division of ICT C	2018 International Conference on Computational Science and Computational Intelligence (CSCI)	2 jan. 2020	2018			1452	1453	Sexual crime, including sexual ha	978-1-7281-136	10.1109/CSCI46756.2018.00286	https://ieeexplor	Voice record, An	Google;Speech i	application progr	text con		
Addressing Clou	S. Naidu; M. Mal	Department of C	2022 International Conference on Computational Science and Computational Intelligence (CSCI)	25 Aug 2023	2022			1346	1351	The provision of	2769-5654	979-8-3503-202	10.1109/CSCI58124.2022.00242	https://ieeexplor	Information Tech	Cloud computing	cloud computing	cloud cc		
A Novel Cloud A	L. K. ALnwihel; A	Department of C	2020 International Conference on Computational Science and Computational Intelligence (CSCI)	23 jun. 2021	2020			1302	1308	Cloud computing is a medium tha	978-1-7281-762	10.1109/CSCI51800.2020.00243	https://ieeexplor	cloud computing	Authorization;Re	authorisation;clo	data se		
Toward Intelligar	S. Abuzneid; M.	Department of E	2021 International Conference on Computational Science and Computational Intelligence (CSCI)	22 jun. 2022	2021			1268	1270	The relationship between care re	978-1-6654-584	10.1109/CSCI54926.2021.00073	https://ieeexplor	respite care;elde	Cloud computing	artificial intelligei	toward I		
Mesh-IoT Based	A. Gajjar; X. Yan	Department of E	2018 International Conference on Computational Science and Computational Intelligence (CSCI)	2 jan. 2020	2018			1019	1023	This paper presents an Internet-o	978-1-7281-136	10.1109/CSCI46756.2018.00197	https://ieeexplor	Edge computing	Face recognition	Bluetooth;cloud	low bina		
Agnostic Approa	A. Abdel Khaleq	Department of C	2019 International Conference on Computational Science and Computational Intelligence (CSCI)	20 Apr 2020	2019			1411	1415	Cloud applications are becoming	978-1-7281-558	10.1109/CSCI49370.2019.00264	https://ieeexplor	Cloud applicatio	Measurement;Ti	cloud computing	Google		
Engineering Clo	N. Bankston; S.	Department of C	2021 International Conference on Computational Science and Computational Intelligence (CSCI)	22 jun. 2022	2021			404	410	This paper will cover the planning	978-1-6654-584	10.1109/CSCI54926.2021.00138	https://ieeexplor	Cloud Security;E	Cloud computing	business data pr	enginee		
Secure Cloud St	E. Eisenberger;	Department of C	2020 International Conference on Computational Science and Computational Intelligence (CSCI)	23 jun. 2021	2020			1314	1317	Cloud Storage is a cost-effective	978-1-7281-762	10.1109/CSCI51800.2020.00245	https://ieeexplor	RSA Cryptosyst	Cloud computing	blockchains;clou	cloud st		
InterCloud: A Da	S. Kirkman; R. N	Computer and Ir	2018 International Conference on Computational Science and Computational Intelligence (CSCI)	2 jan. 2020	2018			1303	1308	We present a smart contract base	978-1-7281-136	10.1109/CSCI46756.2018.00253	https://ieeexplor	Cloud Trust;Bloc	Cloud computing	authorisation;clo	InterClo		
CloudMonitor: D	F. Alqahtani; F. S	Computer Scien	2019 International Conference on Computational Science and Computational Intelligence (CSCI)	20 Apr 2020	2019			1454	1457	The primary concern of cloud ser	978-1-7281-558	10.1109/CSCI49370.2019.00271	https://ieeexplor	Cloud Security; I	Cloud computing	business data pr	data flo		
Predicting Large	N. Sharma; K. R	College of Scien	2020 International Conference on Computational Science and Computational Intelligence (CSCI)	23 jun. 2021	2020			338	343	The complexity and computationa	978-1-7281-762	10.1109/CSCI51800.2020.00064	https://ieeexplor	Software Contai	Cloud computing	Big Data;cloud c	computi		
Container-Basec	C. -S. Jhuang; C	Dept. Computer	2022 International Conference on Computational Science and Computational Intelligence (CSCI)	25 Aug 2023	2022			1317	1322	This paper studi	2769-5654	979-8-3503-202	10.1109/CSCI58124.2022.00237	https://ieeexplor	Cloud Computin	Transport protoc	cloud computing	automat		
CDAP: A Cultura	M. Ebrahimi; S.	Computer Scien	2022 International Conference on Computational Science and Computational Intelligence (CSCI)	25 Aug 2023	2022			628	633	The performanc	2769-5654	979-8-3503-202	10.1109/CSCI58124.2022.00116	https://ieeexplor	Big Data;Big Dat	Cloud computing	Big Data;cloud c	big data		
Identifying actior	A. Nesvijevskaia	DICEN IDF, CN/	2021 International Conference on Computational Science and Computational Intelligence (CSCI)	22 jun. 2022	2021			1551	1552	Artificial Intelligence has opened	978-1-6654-584	10.1109/CSCI54926.2021.00302	https://ieeexplor	Customer Beha	Industries;Scient	bank data proce	identifyi		
Persisting Trust	M. A. Ağca; D. K	TOBB ETU Com	2021 International Conference on Computational Science and Computational Intelligence (CSCI)	22 jun. 2022	2021			379	385	Distributed ledger-based transact	978-1-6654-584	10.1109/CSCI54926.2021.00074	https://ieeexplor	Distributed com	Privacy;Adaptati	cellular automati	dynamic		
Heterogeneous	M. Mallta; G. V. I	Computer Scien	2019 International Conference on Computational Science and Computational Intelligence (CSCI)	20 Apr 2020	2019			1500	1505	Many Big Data problems, Markov	978-1-7281-558	10.1109/CSCI49370.2019.00279	https://ieeexplor	Big data;Markov	Hidden Markov r	Big Data;comput	matrix q		
Towards Dynami	H. Sami; A. Mou	Department of C	2018 International Conference on Computational Science and Computational Intelligence (CSCI)	2 jan. 2020	2018			960	965	The advantage of having fog nod	978-1-7281-136	10.1109/CSCI46756.2018.00187	https://ieeexplor	IOT;cloud comp	Containers;Moni	cloud computing	context		
Cloud Based Ele	R. Lee; D. Kelly	Computer Scien	2018 International Conference on Computational Science and Computational Intelligence (CSCI)	2 jan. 2020	2018			1010	1014	Elevators are used immensely ac	978-1-7281-136	10.1109/CSCI46756.2018.00195	https://ieeexplor	Cloud Based, El	Elevators;Contrc	cloud computing	embedc		
Transform Decoi	R. Hooda; W. D.	Department of E	2022 International Conference on Computational Science and Computational Intelligence (CSCI)	25 Aug 2023	2022			1402	1407	An adaptive tech	2769-5654	979-8-3503-202	10.1109/CSCI58124.2022.00251	https://ieeexplor	Attribute compre	Point cloud com	approximation th	3D poin		
Timing and its In	J. W. Wallace; S	Infinite Dimensi	2018 International Conference on Computational Science and Computational Intelligence (CSCI)	2 jan. 2020	2018			1466	1474	A framework to integrate different	978-1-7281-136	10.1109/CSCI46756.2018.00293	https://ieeexplor	language;comm	Computer archit	Big Data;cloud c	systems		
A Low Cost LoR	M. Meli; E. Gatt;	Department of M	2020 International Conference on Computational Science and Computational Intelligence (CSCI)	23 jun. 2021	2020			376	381	This paper presents a low cost Lo	978-1-7281-762	10.1109/CSCI51800.2020.00070	https://ieeexplor	IoT;LPWAN;Indo	Temperature me	air quality;Big D	low cost		
Architecture to Ir	P. L. González R	Departamento d	2018 International Conference on Computational Science and Computational Intelligence (CSCI)	2 jan. 2020	2018			996	1001	A smart home can maintain an up	978-1-7281-136	10.1109/CSCI46756.2018.00193	https://ieeexplor	Internet of Thing	Cloud computing	air quality;artifici	smartwa		
A Dev-Ops Prac	A. Kaplan	Department of C	2021 International Conference on Computational Science and Computational Intelligence (CSCI)	22 jun. 2022	2021			872	877	Since the inception of the CSUN (	978-1-6654-584	10.1109/CSCI54926.2021.00204	https://ieeexplor	dev-ops;informa	Cloud computing	cloud computing	dev-ops		
Data Integrity in	K. Morovat; B. P	Department of M	2019 International Conference on Computational Science and Computational Intelligence (CSCI)	20 Apr 2020	2019			20	25	The integrity of files stored on clo	978-1-7281-558	10.1109/CSCI49370.2019.00011	https://ieeexplor	Data integrity; Cl	Data integrity;Da	authorisation;clo	data int		
Pigeon: A Dynar	W. Ling; L. Ma; (	Futurewei Techn	2019 International Conference on Computational Science and Computational Intelligence (CSCI)	20 Apr 2020	2019			1416	1421	Recently, voice-triggered small cl	978-1-7281-558	10.1109/CSCI49370.2019.00265	https://ieeexplor	Serverless, FaaS	Containers;FAA;	cloud computing	private		
Multi-user Searc	K. Astudillo; M. E	Computer Scien	2022 International Conference on Computational Science and Computational Intelligence (CSCI)	25 Aug 2023	2022			707	712	Petabytes of dat	2769-5654	979-8-3503-202	10.1109/CSCI58124.2022.00129	https://ieeexplor	Attribute based	Authorization;Ck	application progr	adminis		
Asynchronous V	A. Srivastava; H.	Department of C	2018 International Conference on Computational Science and Computational Intelligence (CSCI)	2 jan. 2020	2018			1059	1064	Stochastic Gradient Descent (SG	978-1-7281-136	10.1109/CSCI46756.2018.00206	https://ieeexplor	Asynchronous D	Tuning;Stochast	distributed proce	edge co		

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("gate") de aprovação: 26 de out. de 2023

Participantes da Entrega [matriculados em Residência em IA]:

Állan Christoffer Pereira Silva
Gabriel da Mata Marques
Heinz Felipe Cavalcante Rahmig

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

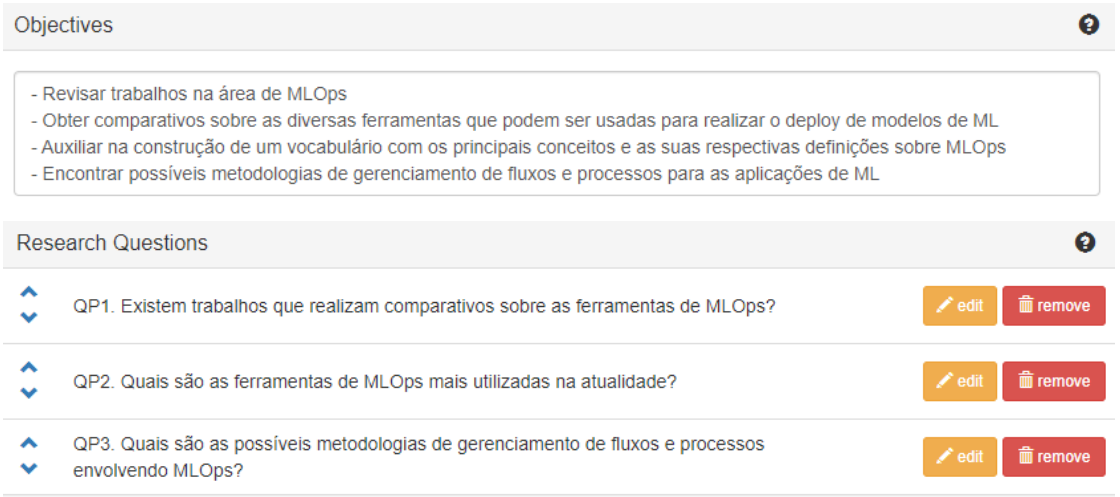
Esta entrega consistiu no início do processo de revisão bibliográfica das referências que irão fundamentar os trabalhos do grupo responsável pela temática de MLOps.

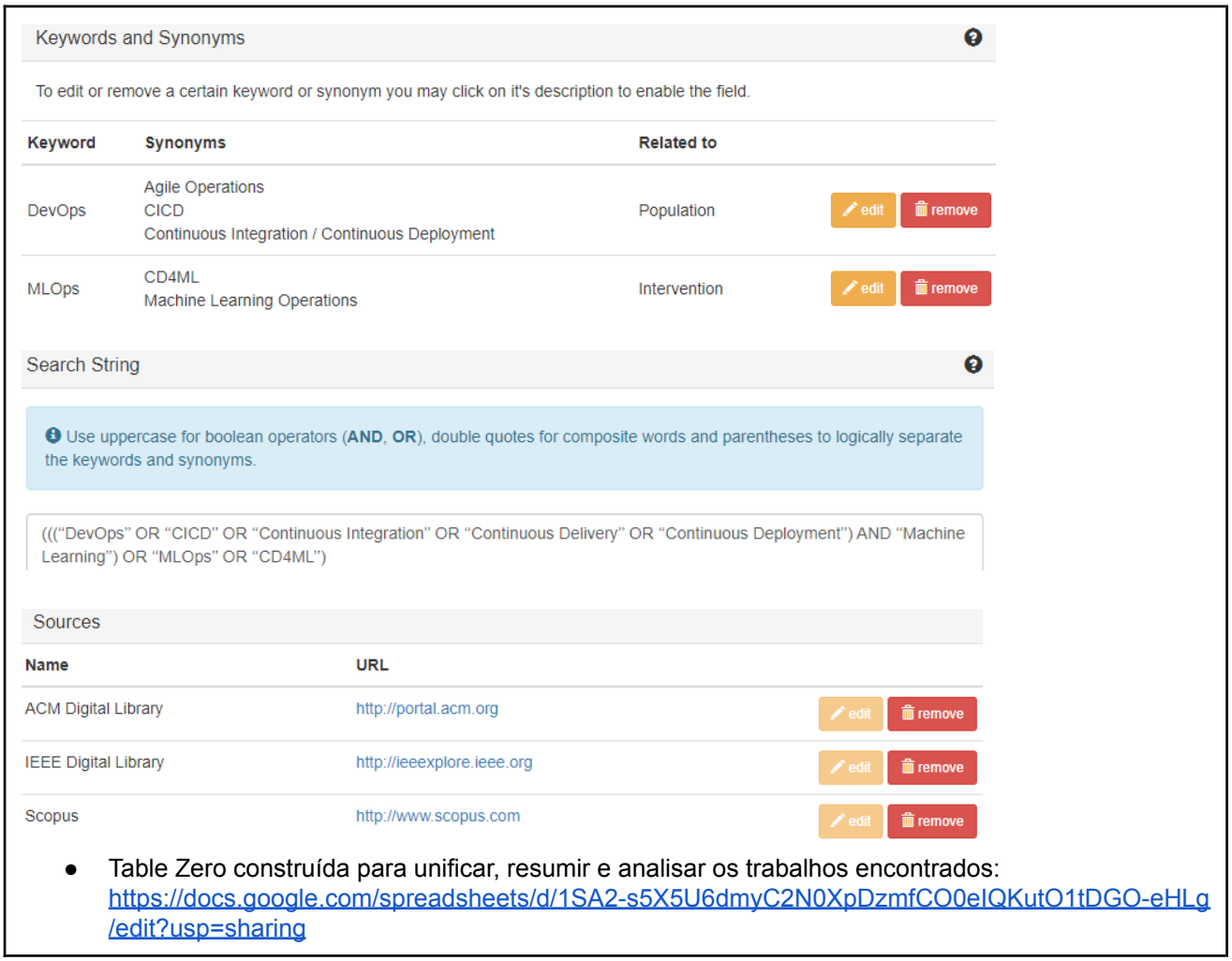
Os requisitos básicos para a entrega eram:

- Buscar artigos e publicações em outras bases científicas além da CSCI.
- Construir um repositório com os trabalhos encontrados.
- Realizar uma revisão preliminar dos trabalhos e um breve resumo sobre cada um.

Os produtos gerados para esta entrega estão descritos a seguir:

- Estruturação do processo de revisão bibliográfica utilizando a ferramenta Parsifal.





Keywords and Synonyms

To edit or remove a certain keyword or synonym you may click on it's description to enable the field.

Keyword	Synonyms	Related to	
DevOps	Agile Operations CICD Continuous Integration / Continuous Deployment	Population	edit remove
MLOps	CD4ML Machine Learning Operations	Intervention	edit remove

Search String

Use uppercase for boolean operators (**AND**, **OR**), double quotes for composite words and parentheses to logically separate the keywords and synonyms.

(((“DevOps” OR “CICD” OR “Continuous Integration” OR “Continuous Delivery” OR “Continuous Deployment”) AND “Machine Learning”) OR “MLOps” OR “CD4ML”)

Sources

Name	URL	
ACM Digital Library	http://portal.acm.org	edit remove
IEEE Digital Library	http://ieeexplore.ieee.org	edit remove
Scopus	http://www.scopus.com	edit remove

- Table Zero construída para unificar, resumir e analisar os trabalhos encontrados: https://docs.google.com/spreadsheets/d/1SA2-s5X5U6dmyC2N0XpDzmfCO0eIQKutO1tDGO-eHLg/edit?usp=sharing

Planejamento: **[descrever o que pretende fazer para realizar a próxima ENTREGA]**

Para a próxima entrega do dia 09/11/2023, estão planejadas as seguintes atividades:

- Finalização do processo de revisão bibliográfica.
- Construção do cronograma de atividades a serem desenvolvidas até o final da Residência.
- Início da montagem de um vocabulário com termos e definições relacionadas à temática de MLOps.

Observação: [caso precise fazer alguma observação, de qualquer “natureza”]

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

LUANA GUEDES BARROS MARTINS: Go!

Table Zero MLOps:

ID	Nome	Ano	Base	Tópico	PDF	Resumo	Responsável	Incluído
1	Who Needs MLOps: What Data Scientists Seek to Accomplish and How Can MLOps Help?	2021	IEEE	MLOps	Link	Pesquisa feita em 63 países com 331 profissionais da área sobre a importância do MLOps para cientistas de dados. Este trabalho pode ser usado como referência para comprovar a relevância do tema no meio acadêmico.	Állan	Sim
2	Architecting MLOps in the Cloud: From Theory to Practice	2023	IEEE	MLOps	Link	Referência a um tutorial sobre como escolher entre as opções de cloud disponíveis e um caso prático de implementação. Porém, não foi encontrado o repositório referente ao tutorial.	Állan	-
3	Practitioners guide to MLOps: A framework for continuous delivery and automation of machine learning.	2021	Google	MLOps	Link	Guia prático elaborado pela equipe do Google Cloud com definições e ferramentas relacionadas à operacionalização de modelos de ML.	Állan	-
4	MLOps: Five Steps to Guide its Effective Implementation	2022	IEEE	MLOps	Link	Artigo com 5 dicas sobre como construir um ciclo de vida de soluções de ML de forma efetiva. As dicas são um pouco genéricas mas podem ser usadas como referência.	Állan	-
5	Towards MLOps: A Framework and Maturity Model	2021	IEEE	MLOps	Link	Este trabalho conta com uma revisão da literatura sobre o estado da arte em MLOps, propõe um framework validado de desenvolvimento contínuo em ML e finaliza com um modelo de maturidade para empresas com relação ao MLOps.	Állan	Sim
6	Sustainable MLOps: Trends and Challenges	2020	IEEE	MLOps	Link	Artigo fala um pouco sobre a relevância do MLOps no contexto científico e prático, traz algumas referências relacionadas a Sustentabilidade e Interpretabilidade de soluções de ML, além de relatar alguns desafios da área.	Állan	-
7	On Continuous Integration / Continuous Delivery for Automated Deployment of Machine Learning Models using MLOps	2021	IEEE	MLOps	Link	Artigo que demonstra alguns níveis diferentes de maturidade em MLOps, traz as principais diferenças entre MLOps e DevOps e relata sobre uma metodologia de desenvolvimento baseada em push e pull chamada GitOps.	Állan	Sim
8	Towards Automation for MLOps: An Exploratory Study of Bot Usage in Deep Learning Libraries	2022	IEEE	MLOps	Link	Este trabalho relata 9 tarefas em projetos de ML que podem ser automatizadas através da criação de bots e utilização em bibliotecas de Deep Learning. Apesar de parecer promissor, acredito ser bem avançado para projetos iniciais em MLOps.	Állan	-
9	An Efficient Microservices Architecture for MLOps	2023	IEEE	MLOps	Link	Artigo que trata sobre a arquitetura de microsserviços, o padrão SAGA de arquitetura, além de propor uma nova arquitetura voltada para MLOps. Os conceitos por trás da arquitetura de microsserviços podem ser bastante úteis e merecem estudos mais aprofundados.	Állan	Sim
10	MLOps for evolvable AI intensive software systems	2022	IEEE	MLOps	Link	Este trabalho traz apenas algumas relações entre DevOps e MLOps, além de uma representação visual dos processos envolvidos no MLOps. Estudo superficial e com baixo potencial de contribuição.	Állan	-
11	K2E: Building MLOps Environments for Governing Data and Models Catalogues while Tracking Versions	2022	IEEE	MLOps	Link	O artigo trata sobre alguns desafios relacionados aos dados e modelos, baseado nesses desafios é proposta uma organização chamada Knowledge To Environment (KFE), a qual aparenta ter bom potencial para estruturar sistemas com grande volume de dados e muitas versões de modelos.	Állan	Sim
12	Bridging the Gap Between Java and Python in Mobile Software Development to Enable MLOps	2022	IEEE	MLOps	Link	O artigo trata sobre a escassez de ferramentas para construção de aplicações mobile que envolvam ML, além do suporte limitado a linguagens como Python e da dificuldade em utilizar linguagens nativas como Java e Kotlin para essas aplicações. Pode ser promissor caso algum trabalho siga para a área mobile.	Állan	-
13	A Multivocal Literature Review of MLOps Tools and Features	2022	IEEE	MLOps	Link	Um dos trabalhos mais promissores e que traz uma série de termos relacionados a MLOps e suas respectivas definições. Além disso, traz uma levantamento das diferentes ferramentas e plataformas disponíveis no mercado, avaliados a partir de diferentes critérios.	Állan	Sim
14	QMP: A Cloud-native MLOps Automation Platform for Zero-Touch Service Assurance in 5G Systems	2022	IEEE	MLOps	Link	O artigo trata sobre uma arquitetura de dados para aplicações mobile e que envolvem conectividade 5G. Pode ser promissor caso algum trabalho siga para área de aplicações mobile.	Állan	-
15	Machine Learning Operations (MLOps): Overview, Definition, and Architecture	2023	IEEE	MLOps	Link	Este trabalho é do tipo survey e traz uma revisão sobre diversas referências relacionadas a MLOps, mostra como foi feito o processo de revisão (string de busca), demonstra algumas definições, papéis de stakeholders, além de uma arquitetura end-to-end de MLOps. Trabalho promissor e importante para fundamentação teórica.	Állan	Sim
16	StreamAI: Dealing with Challenges of Continual Learning Systems for Serving AI in Production	2023	IEEE	MLOps	Link	Como construir, implantar, atualizar e manter modelos dinâmicos que aprendem continuamente a partir de dados em streaming? Este artigo aborda os aspectos de industrialização dessas questões em sistemas de produção. Em ambientes em rápida mudança, as organizações enfrentam o desafio crucial de análise preditiva em moda online a partir de big data e implantação de modelos de Inteligência Artificial em escala. Aplicações incluem cibersegurança, infraestrutura de nuvem, redes sociais e mercados financeiros. Modelos de aprendizado online que aprendem continuamente e se adaptam às distribuições de dados potencialmente em evolução demonstraram eficiência para aprendizado de fluxo de big data.	Heinz	
17	StreamMLOps: Operationalizing Online Learning for Big Data Streaming & Real-Time Applications	2023	IEEE	MLOps	Link	Aprender e servir continuamente a partir de dados de streaming em evolução e servir em tempo real é um problema desafiador. Tradicionalmente, os dados são particionados e processados em lotes para treinar modelos de aprendizado de	Heinz	

Bases	CSCI, IEEE, ACM, Scopus
Período	2018 a 2023
Strings de busca	((("DevOps" OR "CICD" OR "Continuous Integration" OR "Continuous Delivery" OR "Continuous Deployment") AND "Machine Learning"

APÊNDICE 2

Termo de Aceite de Entrega

+

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("gate") de aprovação: 9 de nov. de 2023

Participantes da Entrega [matriculados em Residência em IA]:

Állan Christoffer Pereira Silva
Gabriel da Mata Marques
Heinz Felipe Cavalcante Rahmig

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Esta entrega consistiu na finalização do processo de revisão bibliográfica, o qual tinha por objetivo encontrar trabalhos de referência na área de MLOps.

Os requisitos básicos para a entrega eram:

- Finalizar o processo de revisão bibliográfica.
- Construir o cronograma de atividades a serem desenvolvidas até o final da Residência.
- Iniciar a montagem de um vocabulário com termos e definições relacionadas à temática de MLOps.

Os produtos gerados para esta entrega estão descritos a seguir:

- Repositório com os artigos encontrados e suas respectivas análises (Table Zero): https://docs.google.com/spreadsheets/d/1SA2-s5X5U6dmyC2N0XpDzmfCO0eIQKutO1tDGO-eHLg/edit?usp=sharing
- Cronograma de atividades da Residência: https://docs.google.com/spreadsheets/d/16sy4Z3gcDNV2U5fZOC_mPgE7eyiGfgc5zHQFBCrxBSc/edit?usp=sharing
- Vocabulário sobre MLOps: https://docs.google.com/document/d/1AtmA9GgHnD4mlF-bbjjbt53QLZaQAjowFyjtjQHzzLA/edit?usp=sharing

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

Para a próxima entrega do dia 16/11/2023, estão planejadas as seguintes atividades:

- Finalização do vocabulário incluindo conceitos de DevOps.
- Busca por repositórios de códigos úteis.
- Decisão sobre a aplicação e levantamento de requisitos.

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

LUANA GUEDES BARROS MARTINS: Go!

Instituto de Informática (INF) - UFG
Alameda Palmeiras, Quadra D, Câmpus Samambaia
Fone: (62) 3521-1181 / Fax: (62) 3521-1182

Vocabulário sobre MLOps

Introdução

Este documento tem como objetivo principal a documentação dos principais termos e definições relacionados à área de Machine Learning Operations (MLOps). O conteúdo contido neste trabalho serve como base teórica para estudantes e profissionais que desejam se aprofundar sobre como operacionalizar modelos de Machine Learning, isto é, realizar a implantação de modelos preditivos.

Development Operations (DevOps)

DevOps é uma combinação de filosofias, práticas e ferramentas que aumenta a capacidade de uma organização de entregar aplicações e serviços em alta velocidade. Melhorando e evoluindo produtos mais rapidamente do que organizações que utilizam processos tradicionais de desenvolvimento e gerenciamento de infraestrutura. DevOps é caracterizado pela automação e monitoramento em todas as fases do desenvolvimento de software, desde a integração, testes, liberação até a implantação e gestão de infraestrutura.

Práticas Comuns de DevOps:

1. Integração Contínua (CI): Prática que incentiva desenvolvedores a integrar código em um repositório compartilhado. Cada check-in é então verificado por uma build automatizada, permitindo que equipes detectem problemas cedo.
2. Entrega Contínua (CD): Extensão da integração contínua para garantir que o código seja seguro e que possa ser liberado a qualquer momento.
3. Monitoramento e Logging: Processos que envolvem a coleta de métricas e logs para acompanhar o desempenho das aplicações e da infraestrutura.
4. Comunicação e Colaboração: Ferramentas e práticas culturais que promovem a colaboração dentro e entre as equipes.
5. Automação de Infraestrutura: Gerenciamento e provisionamento de infraestrutura através de código e ferramentas, minimizando a intervenção manual.

Ferramentas de DevOps:

1. Docker: Ferramenta de contêinerização que permite empacotar uma aplicação com todas as suas dependências em um contêiner padronizado.
2. Jenkins: Servidor de automação open source usado para CI/CD.
3. Kubernetes: Sistema de orquestração de contêineres que gerencia aplicações construídas em contêineres.
4. Ansible/Terraform: Ferramentas que permitem aos desenvolvedores provisionar e gerenciar infraestrutura através de código.
5. Git: Sistema de controle de versão distribuído para rastrear mudanças no código fonte durante o desenvolvimento de software.
6. Nagios/Grafana: Ferramentas de monitoramento que oferecem visibilidade em tempo real sobre a saúde da infraestrutura e aplicações.

Termos Chave de DevOps:

Termo	Definição
1- Pipeline de Deploy	Sequência de passos para entregar uma nova versão de software.
2- Infrastructure as Code (IaC)	Prática de gerenciar e provisionar infraestrutura de TI através de scripts de código.
3- Micro Serviços	Arquitetura que estrutura uma aplicação como uma coleção de serviços que são executados de forma independente.
4- Orquestração de Containers	Processo de gerenciar a vida útil de contêineres, especialmente em ambientes com muitos contêineres.
5- Gerenciamento de Configuração	Processo de manter computadores, servidores e software em um estado desejado e consistente.
6 - Automação de Testes	Uso de software para controlar a execução de testes, a comparação de resultados esperados com resultados reais, e a configuração de pré-condições de testes.

7- Versionamento Semântico	Convenção para nomear e gerenciar versões de software de forma a comunicar o impacto das mudanças no código.
8- Balanceamento de Carga	Distribuição automática de tráfego de rede ou pedidos entre vários servidores.
9- Código de Infraestrutura	Código que cria e configura a infraestrutura necessária para uma aplicação.
10- Dashboard de Monitoramento	Interface visual que exibe métricas importantes da aplicação e da infraestrutura.

Machine Learning Operations (MLOps)

Machine Learning Operations (MLOps) é uma área de atuação profissional responsável por dar suporte ao modelos, ao desenvolvimento e à operacionalização do ciclo de vida de Machine Learning estruturado nos princípios e práticas de DevOps.

Um fluxo de trabalho (workflow) de MLOps muito comum é composto por:

1. Extração de Dados (Data Extraction): etapa caracterizada pela integração de dados relevantes de fontes variadas.
2. Análise de Dados (Data Analysis): etapa caracterizada pela compreensão dos dados existentes nos conjuntos de dados.
3. Limpeza de Dados, Transformação e Engenharia de Atributos (Data Cleaning, Transformation and Feature Engineering): etapa caracterizada pela divisão e preparação dos conjuntos de dados de treinamento, validação e teste.
4. Treinamento do Modelo (Model Training): etapa caracterizada pelo treinamento de modelos de Machine Learning e armazenamento dos modelos com melhor desempenho, partindo de diferentes algoritmos e configurações de parâmetros.
5. Validação do Modelo (Model Validation): etapa caracterizada pela avaliação interativa da qualidade do modelo no conjunto de dados de teste e pela constatação de que se o modelo está atendendo aos critérios de qualidade baseada nas métricas de desempenho.
6. Serviço do Modelo (Model Serving): etapa caracterizada pela implantação dos modelos nos ambientes alvo integrados a outros componentes de software.

7. Monitoramento do Modelo (Model Monitoring): etapa caracterizada pela detecção da degradação do modelo através de análises de uso, dados de entrada e desempenho.

Termo	Definição
1- Open source/Código aberto	O código do software é público e disponível para uso, modificação e distribuição.
2- Escalabilidade	A capacidade de aumentar o tamanho da carga de trabalho dentro da infraestrutura existente (hardware, software, etc.) sem impactar o desempenho.
3- Elasticidade	A capacidade de expandir ou reduzir dinamicamente os recursos da infraestrutura (computacional) conforme necessário para se adaptar às mudanças na carga de trabalho de maneira autônoma.
4- Cloud agnostic	O desempenho é consistente, independentemente da plataforma em que é implantado.
5- Extensibilidade	Defina facilmente seus próprios operadores, executores e amplie a biblioteca para que ela se ajuste ao nível de abstração adequado ao seu ambiente.
6- Gestão/Coleta de metadados	A gestão de metadados é usada para coletar dados durante todo o pipeline de ML.
7- Isolamento/Fraco acoplamento	Os componentes podem ser desenvolvidos e implantados independentemente e devem depender uns dos outros na menor medida possível.
8- CI/CD	A plataforma suporta Integração Contínua (CI) e Entrega Contínua (CD) para o pipeline completo de ML.

9- UI	Interface de Usuário ou Dashboard.
10- CLI	Interface de Linha de Comando.
11- API gateway	Em vez de chamar os serviços diretamente, os clientes podem chamar o gateway de API, que encaminha a chamada para os serviços apropriados no back-end e serve como ponto de entrada para os clientes.
12- DAGs	Grafos Acíclicos Dirigidos são usados para descrever o fluxo de trabalho ou podem ser encapsulados dentro da plataforma.
13- Data streaming (real-time)	O fluxo contínuo de dados gerados por várias fontes de dados é suportado e pode ser processado, armazenado, analisado e utilizado diretamente.
14- Data storage	Um banco de dados integrado para armazenar dados brutos, projetos e metadados.
15- Data analysis	Um componente do pipeline gera estatísticas de características tanto para dados de treinamento quanto para dados de serviço, que podem ser usados por outros componentes do pipeline.
16- Data transformation	Um componente do pipeline identifica anomalias nos dados de treinamento e de serviço e prepara os dados para tarefas de ML. O resultado deste passo são as divisões de dados.
17- Data monitoring	Os dados são monitorados para manter a qualidade e inspecionar métricas gerais.
18- API endpoint	A saída da gestão de dados pode ser acessada usando um gateway de API, que encaminha os dados, metadados ou esquema de dados.
19- Automação	O processo de gestão de dados pode ser

	executado automaticamente em produção com base em uma programação ou em resposta a um gatilho.
20- Library agnostic	Todos os principais frameworks e bibliotecas de ML são suportados.
21- Model tracking	O desempenho do modelo de ML intermediário pode ser rastreado e registrado para manter a reprodutibilidade e obter insights.
22- Model registry	Um repositório centralizado usado para adronizar a definição, armazenamento e acesso de características para treinamento e serviço, que é acessível via uma API.
23- Hyper parameter tuning	Um motor de otimização é encapsulado para o ajuste de hiperparâmetros para treinar os modelos de ML de forma eficiente.
24- Teste A/B	Testes A/B podem ser usados para rastrear diferenças entre duas versões de modelos preditivos ou modelos podem ser executados em paralelo em diferentes pontos de extremidade.
25- Detecção de anomalias	Os outliers são automaticamente identificados para revelar padrões irregulares do modelo de ML.
26- Detecção de drift	Mudanças significativas nas distribuições de dados e no desempenho da previsão são automaticamente detectadas para prevenir obsolescência e diminuição da precisão.
27- Alerta de threshold	É possível configurar alertas quando a distribuição de previsões varia significativamente dos valores esperados.
28- Monitoramento de performance	O desempenho preditivo do modelo é monitorado para potencialmente invocar

	uma nova iteração no processo de ML.

Tabela 1. Definições dos principais termos em MLOps

Infrastructure as Code (IaC)

Infrastructure as Code (IaC) é uma prática da área de DevOps que consiste na criação de documentos em linguagem de codificação descritiva de alto nível para automatizar o provisionamento da infraestrutura de TI.

IaC elimina a necessidade de configuração manual de servidores, sistemas operacionais, conexões de bancos de dados, entre outras tarefas.

As ferramentas de IaC mais conhecidas no mercado são: Ansible, Terraform, Pulumi, Azure Resource Manager (ARM) e Google Cloud Deployment Manager.

Terraform é uma ferramenta popular de Infrastructure as Code (IaC) usada para provisionar infraestrutura em várias plataformas, como AWS, Azure, Google Cloud, entre outras. Ele utiliza uma linguagem de configuração chamada HashiCorp Configuration Language (HCL) para descrever a infraestrutura desejada de forma declarativa. Abaixo está um exemplo básico de um documento de configuração do Terraform para provisionar uma instância EC2 na AWS:

```
provider "aws" {
  region = "us-west-2"
}

resource "aws_instance" "example" {
  ami           = "ami-abcdefgh"
  instance_type = "t2.micro"
}

output "ip" {
  value = aws_instance.example.public_ip
}
```

Bloco Provider:

- provider "aws" { ... }: Este bloco indica que você está usando o provedor AWS. O Terraform tem provedores para várias plataformas e serviços.
- region = "us-west-2": Especifica a região da AWS onde os recursos serão provisionados.

Bloco Resource:

- resource "aws_instance" "example" { ... }: Este bloco define um recurso, que neste caso é uma instância EC2 na AWS.
- ami = "ami-abcdefgh": Especifica a Amazon Machine Image (AMI) que será usada para lançar a instância.
- instance_type = "t2.micro": Especifica o tipo de instância que será lançado.

Bloco Output:

- output "ip" { ... }: Este bloco define uma saída que será exibida após o Terraform aplicar a configuração.
- value = aws_instance.example.public_ip: Especifica que o IP público da instância EC2 será exibido como saída.

Quando este código é aplicado usando o comando terraform apply, o Terraform cria uma instância EC2 na AWS com as especificações fornecidas. A saída do comando mostrará o IP público da instância EC2 criada.

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("gate") de aprovação: 16 de nov. de 2023

Participantes da Entrega [matriculados em Residência em IA]:

Gabriel da Mata Marques

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Esta entrega consistiu na finalização da construção do embasamento teórico dos trabalhos que serão desenvolvidos, além da realização do planejamento de atividades a serem executadas até o fim da Residência.

Os requisitos básicos para a entrega eram:

- Finalizar o vocabulário incluindo conceitos de DevOps.
- Buscar por repositórios de códigos úteis.
- Decidir sobre a aplicação de cada integrante e realizar o levantamento de requisitos.

Os produtos gerados para esta entrega estão descritos a seguir:

- Estruturação do processo de revisão bibliográfica utilizando a ferramenta Parsifal.

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

Para a próxima entrega do dia 23/11/2023, estão planejadas as seguintes atividades:

- Revisão e busca por artigos referentes à implantação de Large Language Models
- Elaboração de plano de estudo dos itens de langchain_stack.png.
- Estabelecer paralelos entre conceitos novos encontrados em LangChain e os presentes no vocabulário.

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

Gate parcialmente elaborado em conjunto com Állan e Heinz

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

LUANA GUEDES BARROS MARTINS: Go!

Instituto de Informática (INF) - UFG
Alameda Palmeiras, Quadra D, Câmpus Samambaia
Fone: (62) 3521-1181 / Fax: (62) 3521-1182

APÊNDICE 3

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("gate") de aprovação: 23 de nov. de 2023

Participantes da Entrega [matriculados em Residência em IA]:

Gabriel da Mata Marques

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Esta entrega consistiu na finalização da construção do embasamento teórico dos trabalhos que serão desenvolvidos, além da realização do planejamento de atividades a serem executadas até o fim da Residência.

Os requisitos básicos para a entrega eram:

- Revisão e busca por artigos referentes à implantação de Large Language Models
- Elaboração de plano de estudo dos itens de langchain_stack.png.
- Estabelecer paralelos entre conceitos novos encontrados em LangChain e os presentes no vocabulário.

Os produtos gerados para esta entrega estão descritos a seguir:

- Plano de estudo dos itens de langchain_stack.png.

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

Para a próxima entrega do dia 30/11/2023, estão planejadas as seguintes atividades:

- Elaborar artigo básico (estilo Medium) sobre LangChain e Protocol

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

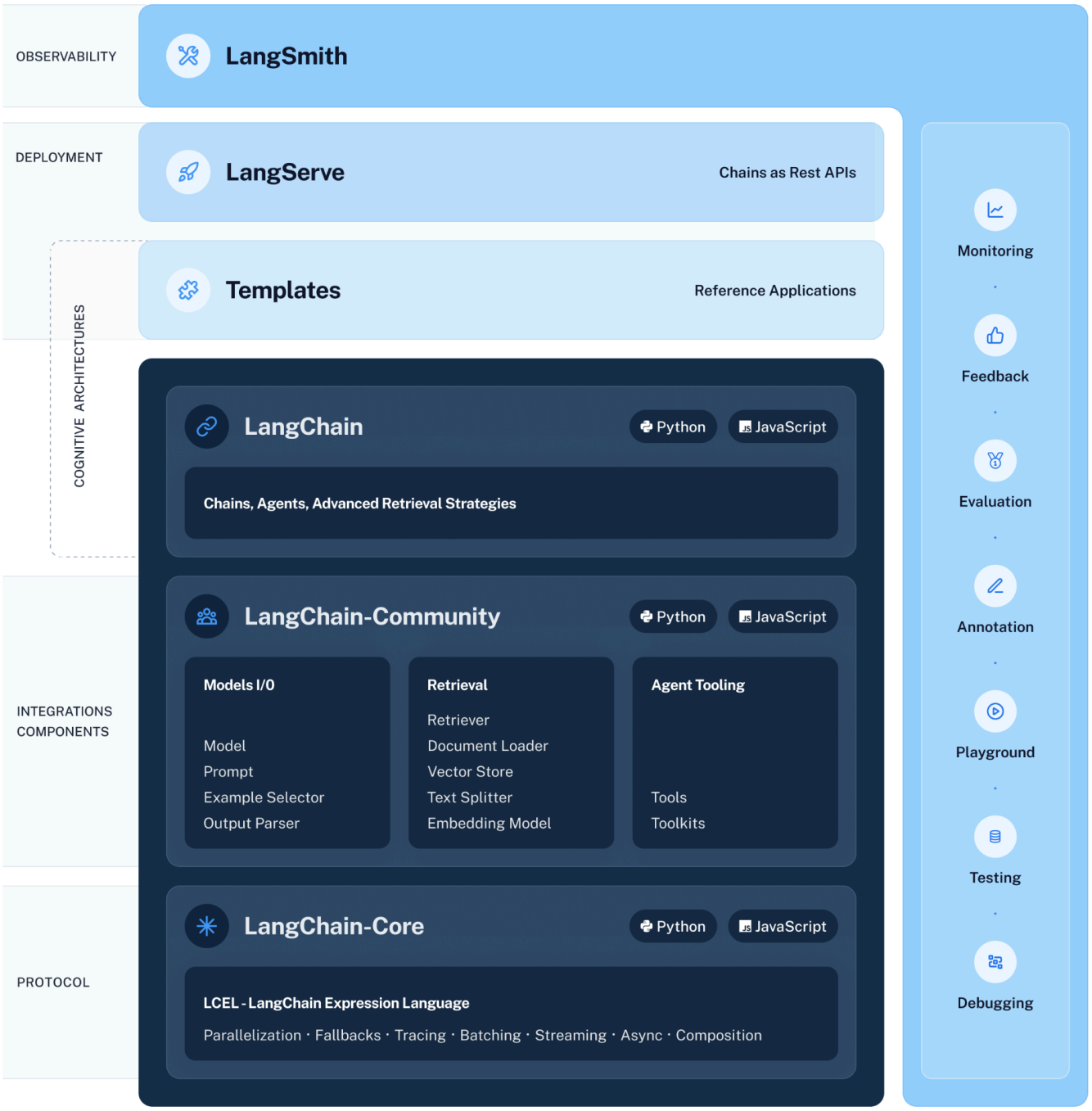
ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

LUANA GUEDES BARROS MARTINS: Go!

Instituto de Informática (INF) - UFG
Alameda Palmeiras, Quadra D, Câmpus Samambaia
Fone: (62) 3521-1181 / Fax: (62) 3521-1182

Stack LangChain



APÊNDICE 4

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("gate") de aprovação: 21 de dez. de 2023

Participantes da Entrega [matriculados em Residência em IA]:

Gabriel da Mata Marques

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Recapitulando:

- Devido a complicações pessoais não pude participar dos últimos Gates.
- Gate 23/11/2023: ***Plano de estudo dos itens de langchain_stack.png.***

O que foi planejado e desenvolvido para essa "semana":

- **Entender conceitos anteriores ao Framework do LangChain:**
  - **a.** LORA - LORA
  - **b.** Retrievers - Retrievers

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

- Explorar https://github.com/tensorchord/Awesome-LLMOps (Sugestão Állan)
- Continuar planejamento apresentado no gate 23/11/2023:
  - *a. Elaborar artigo básico (estilo Medium) sobre LangChain - Protocol*

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

Instituto de Informática (INF) - UFG
Alameda Palmeiras, Quadra D, Câmpus Samambaia
Fone: (62) 3521-1181 / Fax: (62) 3521-1182

LUANA GUEDES BARROS MARTINS: Go!

Instituto de Informática (INF) - UFG
Alameda Palmeiras, Quadra D, Câmpus Samambaia
Fone: (62) 3521-1181 / Fax: (62) 3521-1182

Low Rank Adaptation (LoRA)

Summary

LORA is a technique for adapting large pre-trained models efficiently to new tasks (fine-tuning).

It focuses on the **self-attention layers** of transformer models like BERT or GPT.

Key Features:

- Introduces **low-rank matrices** to the self-attention mechanism.
- Updates **only these matrices** during fine-tuning, not the entire set of parameters.
- Significantly **reduces the number of parameters** to be trained.
- **Freezes the original weights** of the model during fine-tuning, so that only the low-rank matrices, which are much smaller, are updated.

Benefits:

- Allows for **efficient fine-tuning** with **fewer computational resources.**
- Maintains competitive performance on NLP tasks.

Application:

- Implemented by modifying the attention calculation with additional low-rank matrices.
- Typically used when computational efficiency is a priority.

Explanation

What is **rank**? Rank = the number of **linearly independent** columns (or rows, whichever is lower) in a matrix. A low rank matrix has a lot of **"redundant" columns,** which means it can be represented using fewer parameters and be nearly as effective.

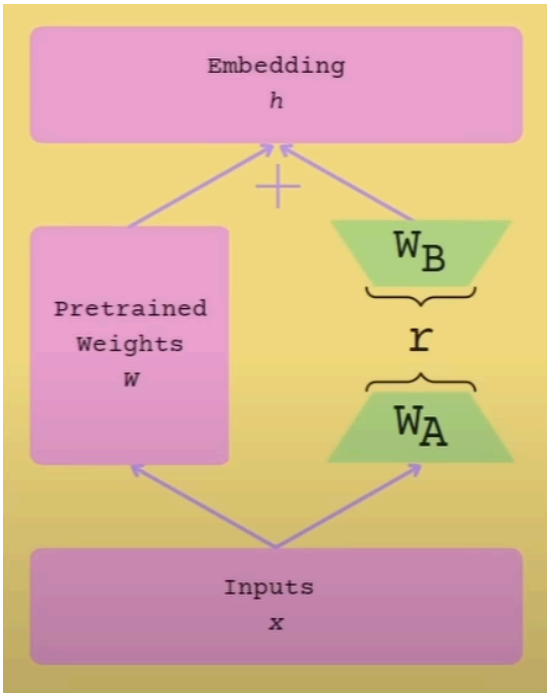
LoRA's paper hypothesized that weight matrices have a **low "intrinsic rank"** during adaptation, or rather, that the adaptation can be effectively captured with a **low-rank approximation** of the change that should be applied to the original weight matrix. That means that it can be represented using a low rank matrix (hence the name LoRA).

Matrix decomposition: representing a matrix W(d x k) as a multiplication of two smaller matrices B(d x r) * A(r x k). In this case, BA would be ΔW, or the weight update matrix. In that case, **"r" is a hyperparameter** that we must tune. [Figure 1]

In the LoRA paper, the focus is on the **attention weights**, notably **Q and V weight matrices**, so the rest of the Transformer's architecture is not touched.

Using this strategy, since the only thing that you add to the original model is the ΔW matrix, if you fine-tune a base model for, let's say, 6 different tasks using this approach, to switch from one to the other you only need to **switch the ΔW matrix**, which can even be done **at inference time**.

Figure 1



References

GitHub do LoRA

Paper do LoRA

[YouTube] Low-rank Adaption of Large Language Models: Explaining the Key Concepts Behind LoRA

[YouTube] Low-rank Adaption of Large Language Models Part 2: Simple Fine-tuning with LoRA

Retrievers

RAG (Retrieval-Augmented Generation)

- Combina geração de linguagem com recuperação de documentos.
- Utiliza um modelo de linguagem para gerar respostas com base em documentos relevantes recuperados.
- Paper
- Artigo da MetaAI

DPR (Dense Passage Retrieval)

- Especializado em encontrar passagens de texto relevantes em uma grande coleção de documentos.
- Usa representações densas para recuperação de informações, diferenciando-se de abordagens baseadas em palavras-chave.
- Paper
- Artigo da MetaAI

BM25 (Best Matching 25)

- Um algoritmo clássico baseado em termos para recuperação de informações.
- Funciona bem para recuperação de documentos com base em correspondências exatas de palavras-chave.
- Artigo na Wikipedia

T5 (Text-to-Text Transfer Transformer) para Retriever

- Originalmente um modelo de linguagem, adaptável para funções de recuperação.
- Converte tarefas de recuperação em um problema de texto para texto.
- Paper
- Artigo do Google Research

REALM (Retrieval-Augmented Language Model)

- Integra recuperação de conhecimento no pré-treinamento do modelo de linguagem.
- Atualiza continuamente sua base de conhecimento durante o treinamento.
- Paper
- Artigo do Google Research

ORQA (Open Retrieval Question Answering)

- Focado em questões e respostas, utilizando um retriever para buscar informações relevantes.
- Usa um modelo de linguagem para gerar respostas com base nas informações recuperadas.
- Paper

ColBERT (Contextualized Late Interaction over BERT)

- Utiliza BERT para recuperação eficiente de documentos.
- Emprega interação tardia entre consulta e documentos para melhorar relevância.
- Paper

Contriever (Contrastive Retriever)

- Baseado em aprendizado contrastivo para aprimorar recuperação de documentos.
- Foca em aprender representações ricas e discriminativas para consultas e documentos.
- Paper

APÊNDICE 5

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("gate") de aprovação: 11 de jan. de 2024

Participantes da Entrega [matriculados em Residência em IA]:

Gabriel da Mata Marques

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Recapitulando:

- Devido a complicações pessoais não pude participar dos últimos Gates.
- Gate 23/11/2023: ***Plano de estudo dos itens de langchain_stack.png.***
- **Entender conceitos anteriores ao Framework do LangChain:**
  a. LORA - LORA
  b. Retrievers - Retrievers

O que foi desenvolvido para essa "semana":

- Documento langchain - overview.

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

LUANA GUEDES BARROS MARTINS: Go!

LangChain - Framework:

LangChain é um framework desenvolvido para a criação de aplicações baseadas em modelos de linguagem. Este framework se destaca por permitir a criação de aplicações que são:

- Context-aware: Conectam um modelo de linguagem a fontes de contexto, como instruções de prompt, exemplos de poucos disparos, conteúdo para fundamentar suas respostas, etc.
- Capazes de raciocinar: Dependem de um modelo de linguagem para raciocinar sobre como responder com base no contexto fornecido, quais ações tomar, etc.

Componentes do LangChain

O LangChain é composto por várias partes interdependentes e interativas:

LangChain Libraries: Bibliotecas em Python e JavaScript que contém interfaces e integrações para uma variedade de componentes, um runtime básico para combinar esses componentes em cadeias e agentes, e implementações prontas para uso de cadeias e agentes.

LangChain Templates: Uma coleção de arquiteturas de referência facilmente replicáveis para uma grande variedade de tarefas.

LangServe: Uma biblioteca para implantar cadeias LangChain como uma API REST.

LangSmith: Uma plataforma de desenvolvimento que permite depurar, testar, avaliar e monitorar cadeias construídas em qualquer framework LLM, integrando-se perfeitamente ao LangChain.

Ciclo de Vida da Aplicação com LangChain

Estes produtos simplificam todo o ciclo de vida da aplicação:

- Desenvolvimento: Desenvolva suas aplicações em LangChain/LangChain.js, começando rapidamente com Templates para referência.
- Produção: Utilize LangSmith para inspecionar, testar e monitorar suas cadeias, permitindo melhorias constantes e implantações confiáveis.
- Deploy: Transforme qualquer cadeia em uma API com LangServe.

LangChain Libraries

Os principais benefícios dos pacotes LangChain são:

- Componentes: Ferramentas e integrações para trabalhar com modelos de linguagem. Os componentes são modulares e fáceis de usar, seja utilizando o restante do framework LangChain ou não.
- Cadeias Prontas para Uso: Montagens integradas de componentes para realizar tarefas de nível superior.

LangChain Expression Language (LCEL)

LCEL é uma maneira declarativa de compor cadeias. Foi projetado para suportar a colocação de protótipos em produção, sem alterações de código, desde a cadeia mais simples “prompt + LLM” até as cadeias mais complexas.

Módulos do LangChain

LangChain fornece interfaces padrão e extensíveis e integrações para os seguintes módulos:

- Model I/O: Interface com modelos de linguagem.
- Retrieval: Interface com dados específicos da aplicação.
- Agents: Permite que modelos escolham quais ferramentas usar, dadas diretrizes de alto nível.

Conclusão

LangChain representa uma abordagem inovadora e abrangente para o desenvolvimento de aplicações baseadas em modelos de linguagem, oferecendo uma estrutura modular e extensível que facilita desde o desenvolvimento até a implantação de soluções complexas.

Finalizado

Bônus

Awesome-LLMOps

Link citado no último gate

Apresenta uma grande variedade de repositórios a respeito não somente de LLMOps, mas também de tópicos relacionados à visão computacional e processamento de áudio e voz. Nas subseções Large Model Serving e Frameworks/Servers for Serving são apontados repositórios open source de desenvolvimento, deploy e monitoramento de LLMs. Mas como previsto na proposta deste trabalho, os itens citados que não estão diretamente implementados no LangChain possuem alternativas ou conexões externas, colaborando com um ecossistema que engloba todo pipeline a respeito da criação de serviços/produtos de/com LLMs.

APÊNDICE 6

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("gate") de aprovação: 17 de jan. de 2024

Participantes da Entrega [matriculados em Residência em IA]:

Gabriel da Mata Marques

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Recapitulando:

- Documento langchain - overview .

O que foi planejado e desenvolvido para essa semana:

- Estudo LangServe e Lang smith, aprofundados para - LangServe:Smith

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Em análise!

LUANA GUEDES BARROS MARTINS: Go!

Documentação do LangServe

LangServe é uma biblioteca do LangChain que auxilia desenvolvedores a implantar runnables e cadeias do LangChain como uma API REST. Integrada com FastAPI e utilizando pydantic para validação de dados, o LangServe oferece uma série de funcionalidades e características importantes.

Características Principais

- **Esquemas de Entrada e Saída Automáticos**: Inferidos do seu objeto LangChain e aplicados em cada chamada da API, com mensagens de erro detalhadas.
- **Documentação da API**: Página de documentação da API com JSONSchema e Swagger.
- **Endpoints Eficientes**: Endpoints `/invoke/`, `/batch/` e `/stream/` para suportar muitas requisições simultâneas em um único servidor.
- **Streaming de Logs**: Endpoint `/stream_log/` para streaming de todas (ou algumas) etapas intermediárias da sua cadeia/agente.
- **Playground**: Página `/playground/` com saída de streaming e etapas intermediárias.
- **Rastreamento Integrado**: Rastreamento opcional para LangSmith, adicionando apenas sua chave de API.
- **SDK do Cliente**: Para chamar um servidor LangServe como se fosse um Runnable local.

Limitações

- **Callbacks de Cliente**: Ainda não são suportados para eventos originados no servidor.
- **Documentação OpenAPI**: Não será gerada ao usar Pydantic V2 devido à incompatibilidade com FastAPI.

Segurança

- **Vulnerabilidade**: Versões 0.0.13 a 0.0.15 tinham uma vulnerabilidade no endpoint de playground que permitia acesso a arquivos arbitrários no servidor. Resolvida na versão 0.0.16.

Instalação

Para instalar o LangServe, tanto para cliente quanto para servidor, use:

pip install "langserve[all]"

Ou instale apenas o necessário para cliente ou servidor com `langserve[client]` ou `langserve[server]`.

Uso do LangChain CLI

O CLI do LangChain pode ser usado para iniciar rapidamente um projeto LangServe:

pip install -U langchain-cli
langchain app new ../path/to/directory

Exemplos de Uso

O LangServe oferece vários exemplos para ajudar a iniciar rapidamente sua instância, incluindo modelos mínimos de LLMs, servidores simples de retriever, implementações de agentes baseados em ferramentas OpenAI, e mais.

Playground e Widgets

O LangServe fornece um playground interativo com suporte a widgets personalizados, permitindo testar runnables com diferentes entradas.

Implantação

O LangServe pode ser implantado em várias plataformas, incluindo AWS, Azure e GCP, com instruções detalhadas para cada uma:

Deploy to AWS

You can deploy to AWS using the AWS Copilot CLI
copilot init --app [application-name] --name [service-name] --type 'Load Balanced Web Service' --dockerfile './Dockerfile' --deploy

Deploy to Azure

You can deploy to Azure using Azure Container Apps (Serverless):
az containerapp up --name [container-app-name] --source . --resource-group [resource-group-name] --environment [environment-name] --ingress external --target-port 8001 --env-vars=OPENAI_API_KEY=your_key

Deploy to GCP

You can deploy to GCP Cloud Run using the following command:
gcloud run deploy [your-service-name] --source . --port 8001 --allow-unauthenticated --region us-central1 --set-env-vars=OPENAI_API_KEY=your_key

Autenticação e Tipos de Entrada/Saída Personalizados

LangServe suporta a adição de lógica de autenticação e permite definir tipos de entrada e saída personalizados para runnables.

Conclusão

LangServe é uma ferramenta poderosa e flexível do LangChain, projetada para facilitar a implantação de runnables e cadeias como APIs REST, oferecendo uma ampla gama de funcionalidades para desenvolvedores.

Documentação do LangSmith

LangSmith é uma ferramenta desenvolvida pela equipe do LangChain, projetada para aprimorar a confiabilidade e eficiência de aplicações baseadas em modelos de linguagem de grande escala (LLMs). Este guia tático oferece uma visão geral de como utilizar o LangSmith para maximizar seus benefícios.

Principais Características e Funcionalidades

- **Rastreamento Automático**: No LangChain, o rastreamento do LangSmith é ativado por padrão, registrando todas as chamadas para LLMs, cadeias, agentes, ferramentas e retrievers. Isso é crucial para o diagnóstico de problemas e a otimização do desempenho.
- **Depuração Eficiente**: Lang Smith facilita a identificação de problemas comuns em chamadas LLM, como entradas inesperadas e resultados finais. Ele fornece uma visualização clara das entradas e saídas exatas em todas as chamadas LLM.
- **Playground Integrado**: Ao examinar uma chamada LLM, você pode usar o recurso "Open in Playground" para modificar e reexecutar a solicitação, observando as mudanças no resultado.
- **Visualização de Sequências de Eventos**: Para cadeias e agentes complexos, LangSmith oferece uma visualização da sequência de eventos, incluindo as chamadas realizadas e suas respectivas entradas e saídas.
- **Análise de Latência e Uso de Tokens**: Lang Smith rastreia a latência de cada etapa e o uso total de tokens, ajudando a identificar componentes que podem estar atrasando o processo ou consumindo recursos excessivamente.

- **Depuração Colaborativa**: Com um botão "Compartilhar", LangSmith permite que cadeias e execuções de LLM sejam facilmente compartilhadas para colaboração e resolução de problemas.
- **Coleta de Exemplos**: Lang Smith permite adicionar exemplos de entrada/saída a um conjunto de dados, facilitando a construção de benchmarks para testar versões futuras da cadeia.
- **Testes e Avaliação**: Lang Smith simplifica o upload de conjuntos de dados e oferece ferramentas para testar mudanças em prompts ou cadeias, permitindo avaliações manuais e automáticas.
- **Avaliação Humana**: Além das métricas automáticas, LangSmith facilita a revisão manual e a anotação de execuções através de filas de anotação.
- **Monitoramento**: Após a implantação em produção, LangSmith continua sendo uma ferramenta valiosa para monitorar a aplicação, registrando todas as execuções, visualizando estatísticas de latência e uso de tokens, e facilitando a solução de problemas específicos.

Conclusão

LangSmith é uma ferramenta essencial para desenvolvedores que trabalham com LLMs, oferecendo recursos avançados para depuração, teste, avaliação e monitoramento de aplicações, garantindo assim a confiabilidade e eficiência necessárias para o uso em produção.