

Projeto Final IA368DD_2023S1: Deep Learning aplicado a Sistemas de Buscas Student: Marcus Vinícius Borela de Castro

IndIR

Indexing Improving Information Retrieval in the Juris Dataset
(Jurisprudence Statements and Thesaurus of the
Federal Court of Accounts of Brazil - TCU)

Leonardo Pacheco & Marcus Borela

https://github.com/marcusborela/ind-ir



IndIR - Datasets and other prepared files

README.md

IndIR - Datasets and other prepared files

JURIS TCU

Collection of statements from the Selected Jurisprudence of the Federal Court of Auditors.

Files:

- doc.csv each document contains a statement from the Selected Jurisprudence.
- query1.csv 50 queries generated from the access log of the Integrated Research of the Federal Court of Auditors (most executed queries).
- query2.csv transformation of questions in query3 to search expression, removing part of the words used.
- query3.csv 50 queries, each generated by LLM from a statement among the most accessed in the access log of the Integrated Research of the Federal Court of Auditors.





O enunciado procura retratar o entendimento contido na deliberação da qual foi extraído, não constituindo, todavia, um resumo oficial da decisão proferida pelo Tribunal. Tampouco objetiva representar o posicionamento prevalecente no TCU sobre a matéria.

ACÓRDÃO: DATA DA SESSÃO: RELATOR:

Acórdão 602/2008-Plenário 11/03/2008 BENJAMIN ZYMLER

ÁREA: TEMA: SUBTEMA:

Desestatização Concessão pública Estudo de viabilidade

OUTROS INDEXADORES:

Requisito, Regulação, Leilão , Preço máximo

TIPO DO PROCESSO:

ACOMPANHAMENTO

ENUNCIADO:

A definição do valor máximo de leilão, em processo de regulação, assenta toda a viabilidade do empreendimento, de forma que seu estabelecimento deve representar o correto balanceamento entre a modicidade tarifária e a atratividade negocial.

JURIS TCU INDEX

Indexing database for the statements of the Selected Jurisprudence using terms from the External Control Vocabulary of the Federal Court of Auditors (VCE).

Files:

- . doc.csy each document contains the definition of a term from the VCE.
- query.csv each query is a statement from the Selected Jurisprudence.
- qrel.csv each record corresponds to an indexing of a statement from the Selected Jurisprudence by a term from the VCE. This indexing
 was performed by operators of the Jurisprudence system and can be observed in the Integrated Research of the Federal Court of Auditors
 Selected Jurisprudence database.

5. INDEXAÇÃO DO ENUNCIADO JURISPRUDENCIAL

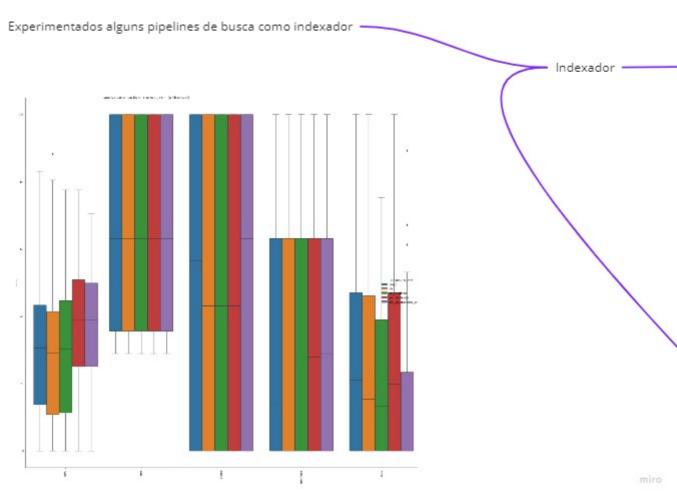
5.1 Etapas do Processo de Indexação

A indexação é uma operação que consiste em identificar os principais conceitos que caracterizam o conteúdo de um texto para obtenção de uma representação da informação relevante por meio de linguagem controlada e padronizada (termos descritores). Exemplo:

Contrato Administrativo. Aditivo. Serviço novo. Preço. Referência. Orçamento estimativo. BDI. Equilibrio econômico-financeiro. Desconto.

Quando houver a celebração de aditivos contratuais para a inclusão de novos serviços, tanto nos regimes baseados em preço global quanto nos regimes de empreitada por preço unitário e tarefa, o preço desses serviços deve ser calculadoconsiderando as referências de custo e taxa de BDI especificadas no orçamento- base da licitação, subtraindo desse preço de referência a diferença percentual entre o valor do orçamento-base e o valor global obtido na licitação, com vistas agarantir o equilíbrio econômico-financeiro do contrato e a manutenção do percentual de desconto oferecido pelo contratado (art. 37, inciso XXI, da Constituição Federal e arts, 14 e 15 do Decreto 7.983/2013).

A indexação compreende duas etapas distintas: a análise conceitual do assunto e a tradução dos conceitos em descritores.



('SÚMULA TCU 1: Não se compreendem como vencimento, para efeito de concessão da pensão especial com fundamento na Lei n^2 3.738, de 94/84/60, as vantagens previstas no art. 184 da Lei n^2 1.711, de 28/18/52.'.

"O termo 6 "Pessoal".\nPessoal ten nota de escopo: "Designação genérica de todos os servidores ou funcionários civis pertencentes ao quadro de pessoal de un órgão ou entidade.".\nPessoal ten nota de escopo: "Tema agrupador para área de atuação do Controle Externo.".\nPessoal é uma generalização de: "Servidor público", "Funcionário público", "Pessoal civil", "Pessoal militar", "Pessoal temporário" e "Colaborador".\nPessoal tem termo relacionado: "Sisae", "Desposa con pessoal" e "HCAAF".\nPessoal tem tradução em espanhol: "Recursos humanos".\nPessoal tem tradução em inglês: "Personal" e "Human respurces".".

Tentando treinar unicamp-dl/mt5-3B-mmarco-en-pt

Thiago Soares Laitz 8h17 @Marcus Borela vocês estão treinando o t5 pra classificar se o documento é ou não relevante, é isso mesmo?



Marcus Borela 8h37

Isssso @Thiago Soares Laitz

O reranker que está no hugging face: unicamp-dl/mt5-3B-mmarco-en-pt



3 respostas Última resposta hoje à(s) 8h57

Thiago Soares Laitz 12h26 Oi @Marcus Borela, vou te mandar o código aqui, mas eu tive que adaptar bastante coisa para ficar mais facil para vocês (eu estava treinando o t5 para uma tarefa diferente) (editado)





Hoje a noite vou testar esse código no colab pra ver se não quebrei nada

Thiago Soares Laitz 12h27 Estou mandando agora caso queira ir dando uma olhada

GPU A100; Batch size=2

OutOfMemoryError: CUDA out of memory.

Tried to allocate 16.00 MiB (GPU 0; 39.56 GiB total capacity; 36.46 GiB already allocated; 16.56 MiB free; 38.23 GiB reserved in total by PyTorch)





12

Consider de erritair restrute servit fortearte

Department de mateir defendants e active vigatification

| 1 | | | query2.csv (LLM> | termo | DS) |
|---------|--|------------|--|--------|---|
| Preview | w (1)2 forc 31 how (11 ha.) - 1/3 61 | Providen 3 | Code - Blone 10 Lane (20 Local - 3.89 E2 | Freder | Cook Diene 5 (4:14) - 4.85 M |
| L | 080_20807_RC | 1 3 | LIEV_12/QUEV_1101 | 1 | given to great that |
| > | Linkovica e propo | 2 50 | ignoscerable exemperada de usa de bers públicas matalitácia | 1.3 | pouveel é a accelitate de locitacio edequeca pera a conomicio resoverada de aro de lena guidicos? |
| 1 | acceptas a pager | 1. 20" | scriterio valida falacidado | 1 | 200, A citação e considerada editida apia a fateriam la do responsávei se a defena ja foi apresentada? |
| | Apadition a varianto | 4 83 | Appropriate and related participated disputate minimum | 4 | 1993) processos de generales concentral é promiticia destración con a exigilarie de partitaleie. Discisio atolant |
| 4 | Aprilente e eta de regisera de preças | 4 0 | Approventa de poeticiposta petetabelo l'Aprila etalas | | povos edigência de garantia de participacio é permitide (ontenesta com a edigência ca patrimisio Lóquido miximo) |
| > | Crapt repress a superfeturemento | 5 20 | companies see cuberture contratual enits | | Mitgate six as consultatal perces respectively per existent depress an otherize a softental? |
| 7 | Aprile te ty Zar, a magnetit til statute | 7 9 | Apara de desas come e latis de conseque las perigles | 7 | 1650 people A edequade para encreation de una de Acons concretain en necessarion? |
| - | completion a suprember | e 100 | generale des deuts essentials de entreta plitters de absorbles. | | 20/ Junia romas dever se color adas se casaro das e en com casas de certo em procursos de efectaciones de generos al |
| * | Viscous in an increase she may district | 2 29 | Companyable demographism de companyable com o objeto de bissi. | 4 | MQF recorded a right on the foods improve que 12 minor au fron perthada a major facto, com a abbita de 10 Major. |
| 15 | Anthony Dispose de provincios | 10 50 | Continued and a state of the state of the same and the state of the st | 180 | puri permittire a controllar de l'attes considerates en era, co, 65 pa a co, ca pel c'alcalinat para electrolles constant |

bein de Heille; Se decreebes conserve 238/Cart e responsibilidade do gratar autoritos que da a gratar entenantor omitia o decen de prestar contact rechiscles à prossure actemporabilitànte do garror recursor oriente de avenuence 111 dans Plan matellitade de Districada del Itada para contratar constituir distribuir acconstituir à extensionada de protection 13 Migrafeto de paracria pablica-a cuada madalidade There is a remaind where the

there our activities as elle bound connectifilities arter o collect do current a at attitioning practices to currents social.

SMA contribute de conversación de capatitlativas materia en acros as correlarios nos características constitucion à limitar

- 1.4 STECHES IN CONTROL WITH CONTROL WAS ARRESTED FOR THE PARTY OF THE STEEL OF STREET Didenses der opperture communici-1.4 accresponsabilidade garrão cos repursos rundo variolasi de saáce The Contracted a same for the proceedings of charged as appropriate form the Helbardes' Underestanded CARTONICS Manageration de monagia per appeta la recognidades.
- (Courses a marrelations) Magazantina de procesa startgatólete, en Distingation. ctore permite de prepos conventes no mercado é obtigadoria en limitacles?
- 1966 on hear new well-har a metabolish de abolic d'atait ta represionidade acuatic à aca-15 livenego ifilialede e compresenta
- (3) I constitute a resource to come analysis of a security and account attachment and Principalments in thillar is authorize. Proposed to refer metallics is superior, metallic additional
- involvelue a short late and delegate properties for the control of the superficients 218 Chaus des es concuces pere cre o restabalacimento total ou personal de que futativo da una el terrormente que umbo
- 119.5 Adeletatore 2 come fector a pretendada por servicio de 16 fectos para ambalica de cartacion de statat William's a localitate Married Repaired at leading
- para qui à consent el prin étarellemete des notes a proctora de palée segre con con cara communion nels petits en bice
- Condition to believe Mark recommendable competitions as IVII.

- cultivacionamento de desensas common suferencials element officials are tractor in curren-231.0 can deve um lesto se máo for posicial obter oracos referenciais non aliatean oficiale para estimatora de costos s
- 1994 normaleiro pouterado ester o portundo tidratos e a de prese es um Helinado de etan ridratos e acresid
- 12 may a improvinging or not They was to be able to prove
- Stateldoseldode de Dielborte Physical acids are accurately delited by Mealine is across 04 muchos a correspoia desarrescones como litatras a comprehencidades sus literandas en vica valcaira a praco"
- STATI consider provinces on contrate administrative, one so behavior de prese de significaries. sacificamedas a communes reconomogação de compreto adeixistrativo trimino preso de vipênci. 28
- Utiliad. Els perilles es religiade de atentode de republició electro para rentramada de armicas acionatricios para restr Maredgere, la de alestado de capacidade Physics tada de comunidade binatas que objet adequatácios conocidas d 245

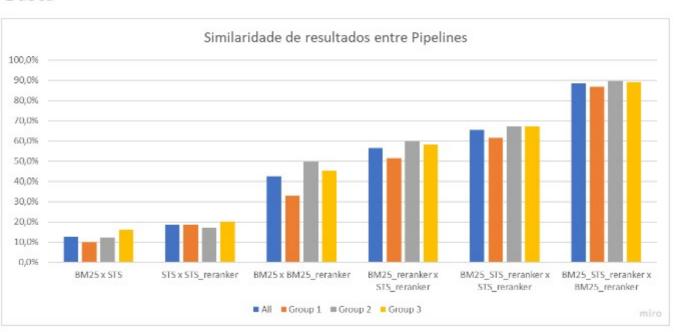
- Screenbergers cam a administration publica-Managerials at Inspetitively: Introduction providing his Warrel. costi necessária reposição de importâncias intenidemente partectoras, de top-fé, por permidores, em tintade de envolesces
- 500 977 consider a professor de como en la tra de produce as correles parte contra en este as a cabér de processor. artifectigiositique e stegateriorere Alteculación cireta risca a saura
- 18:1 abrigantela a sia diguella dan persan continuira na cidiral nos proples aura nacionale de professorana! Name of Building Street eng. The water disable and a decision on the same or called
- contentate districts. conference enclaration transport to the property of 239.5 centraturas de vilorescenso, conferenciates ou praticipas pere manaches contos de Cresimento compenhaciamento
- Michael District Common relative exclusive delevers, el constitución! allabares de contrata Warren eine de present et rentalisation.
- in il personal e an'il tratto de paragonalisto personaria e artificação do continuo acordina ao expressor?
- Campitals sandylars Alteriorensproducto excelusivo

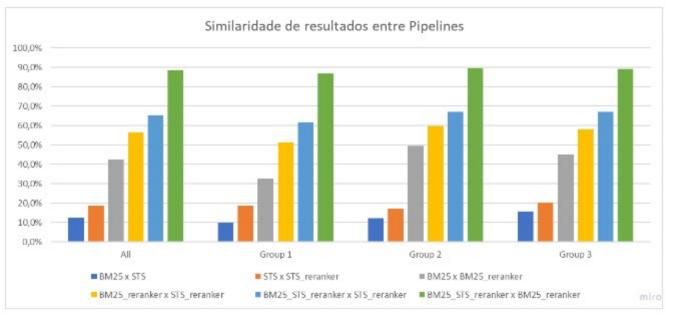
- satisficações específica a lavidaçõe cancestrante supremendo de pagamentos pares da regularidade fiscal. SIGN retendo de oguerator é permitté en usos de perde de regularidade finanti-

- Webster and described described from a solventiary is relative for an enterprise file consideration from all arms.
- Mirachaie de soriente 14 Blue regular en denomina descendencia las el arteriar en a calabranda d
- precises o virta la trabalitates ou societàtic pré-adistante effeta a competitividade das libeltatata so carrama libritativile. Margin de blade e recorrera militare. Mediacally technically on any explicitly and publicate remodifications. .0
- orfelactuares a miles servantes a crace profuntosata acade parameter. 110 STER permittus el citair perforate e ses espresa citaita de esa citalitações de biseitas e presa sas base na poste de defe-
- "Utility nersy material", delse "trace à descripantificants actorizants de accompany". Michelmondina in dec. Inchina in-Man we entertial descioud langue an econocia
- Wranderija A darfardis dan menya habi Mengla Wiliafarda. Continuedoscoto extratógico. curará vedecão à sinclarão de novo documento descimedo e estaster condição de habilitação presolatante en selfilidação bias de novo documento descimedo e estaster condição de habilitação presolatante en selfilidação bias de constituição de constituição de selfilidação presolatante en selfilidaçõe en selfilidaçõe presolatante en selfilidaçõe presolatante en selfilidaçõe presolatante en selfilidaçõe presolatante en selfilidaçõe en selfilida



Busca





- pror

..

Avaliação por LLM

Human: Você é um especialista na jurisprudência do Tribunal de Contas da União com o objetivo de avaliar se um enunciado de jurisprudência responde a uma consulta.

Deve retornar um valor de score de 0 a 3, sendo: 0 - irrelevante - o enunciado não responde a consulta;

1 - relacionado - o enunciado apenas está no tópico da consulta;

2 - relevante - o enunciado responde parcialmente a consulta;

3 - altamente relevante - o enunciado responde a consulta, tratando completamente de suas nuances.

Em seguida, explique a razão para a escolha do score.

Por favor, responda no formato JSON, contendo as chaves Score e Razão; o valor de Score deve ser o valor do score atribuído; o valor de Razão deve ser a motivação da escolha do score.

Consulta: concessão remunerada de uso de bens públicos modalidade Enunciado de jurisprudência: Em regra, o prezão é a modalidade de licitação adequada para a concessão remunerada de uso

de bens públicos, com critério de julgamento pela major oferta em lances sucessivos.

...

Al:
{"Score": 2, "Razão": "O enunciado aborda a concessão remunerada de uso de bens públicos e menciona o pregão como modalidade de licitação adequada, mas não trata de todas as nuances possíveis relacionadas ao tema."}

```
for like, row in df join te25 ats remember query('BANK <- 10 and QUERY 10 -- 181').iterases():

acore, reasoning _get lim acore(row] 'QUERY TEXT'|, row| 'DOC TEXT'|)

df join he25, the remarker.low[ide, 'TIM SCORE'] -- worse

df join he25, the remarker.low[ide, 'TIM SCORETIN'] -- remember.low[ide, 'TIM SCORETIN'] -- remember.low[ide, 'TIM SCORETIN'] -- remember.low[ide]

df join he25_sts_remarker.query('RANK <- 18 and QUERY_ID -= 181').sort_values('RANK')
```

| | | | | | | | () | () |
|--------|----------|------|--------|------------------------|--|--|-----------|---|
| | QUERY ID | RANK | DOC ID | ENGINE | QUERY TEXT | DOC_TEXT | LLM SCORE | LLM REASONING |
| 45851 | | | 2845 | (BM25[S15]+Recenter | Qual é a modalidade de licitação adequada para | hm regra, o pregão é a modalidade de licitação | | O enunciado responde diretamente à pergunts, |
| 41428 | | | 17360 | (BM2S[STS]+Reranker | Qual é a modafidade de ficitação adequada para | É recomendável a utilização de progão para a c | | O enunciado responde parcialmente à pergunta |
| 647904 | | | 5714 | (BM25 STS +Retarker | Qual é a modalidade de licitação adequada para | A cessão das áreas comerciais de centrais públ | | O enunciado responde diretamente à pergunta, |
| 46534 | | | 58598 | (BM2S[STS] + Renantive | Qual é a modafidade de ficitação adequada para | Ne ficitação que tem por objeto a conossão re | | O enunciado está relacionado ao tópico da perq |
| 19717 | 101 | | 14862 | (DM25[STS]+Reranker | Qual é a modalidade de licitação adequada para | O contrato administrativo de concessão remuner | | O enunciado aborda o tema de concessão remuner |
| 60024 | | | 149533 | (BM25[STS] / Recenter | Qual é a modalidade de licitação adequada para | baste amparo legal à utilização do modelo de | | O enunciado de jurisprudência aborda a utiliza |
| 74510 | 101 | | 18865 | (BM2S[STS]+Reranker | Qual é a modafidade de ficitação adequada para | É cabível a utilização do pregão para concessõ… | 2.0 | O enunciado aborda a modalidade de licitação p |
| 41179 | | | | (BM25[S15]+Recurker | Qual é a modalidade de licitação adequada para | Para a aquisição de bens comuns a Administração. | | D enunciado de jurapriidência aborda a modalid |
| 41001 | 101 | | 19399 | (BM2S[STS]+Reranker | Qual é a modafidade de licitação adequada para | Deve ser offizado a modafidade progão para aq | 1.0 | O enunciado aborda o tema de modefidades de fi |
| 22636 | | | | (BM25 STS]+Reconter | Qual é a modalidade de licitação adequada para | Nas licitações para a concessão de serviços pú | | O enunciado está relacionado ao tema de licita. E C |
| | | | | | | | | |

```
query_df = 'QUERY_ID == 181 and (RAMK == ' + ' or RAMK == '.join([str(r) for r in list(np.random.choice(780, 3, replace=False) + 381)]) + ')'
  query_df
√ 0.0b
COLLEA ID
            101 and (RANK
                               873 or BANK
                                              400 or RANK
   for 1, (fix, row) in enumerate(df_bm25.query(query_df).iterrows()):
       store, reasoning get lim store(ros['QUERY TEXT'], ros['DOC TEXT'])
      df be25.loc[idx, 'LLM SCORE'] - score
      df_bm25.loc[idx, 'LLM_REASONING'] = reasoning
  df bm25[df bm25]"LLM SCURL"].notmull()].sort values(["QULW ID", "NAWK"])
        QUERY_ID RANK DOC_ID ENGINE
                                                                            QUERY_TEXT
                                                                                                                           DOC TEXT LLM SCORE
                                                                                                                                                                               LLM_REASONING
                      400
                                       8M25 Qual é a modalidade de licitação adequada para...
                                                                                                                                                   O enunciado de jurisprudência aborda o tema de...
                             16858
                                                                                           Diante de constatação que possa levar à negati...
                                       BM25 Qual é a modalidade de licitação adequada para...
                                                                                          Não é necessário desconsiderar a personalidade...
                                                                                                                                               0.0 O enunciado de jurisprudência não aborda a con...
 76157
                                      BM25 Oual é a modalidade de licitação adequada para...
                                                                                            È causa de impedimento pera participar de lici...
                                                                                                                                               0.0 O enunciado de jurisprudência aborda o tema de...
```

Cronograma

IndIR - Indexing Improving Information Retrieval in the Juris Dataset (Jurisprudence Statements and Thesaurus of the Brazil Federal Court of Accounts - TCU)

Leonardo Pacheco e Marcus Borela

https://github.com/marcusborela/ind-ir

| | | | | | Se | 1 | 5em | ana | 6 | | |
|---|-------------------|-----------|---------|---------|---------|---|-----|-----|------|-----|-----------|
| TAREFA | ATRIBUÍDO PARA | PROGRESSO | INÍCIO | TÉRMINO | 21 q | | 24 | | 26 2 | 1 6 | 18 2 q |
| Planejamento | | | | | | П | П | П | Т | Т | 1 |
| Plano básico | LM | 100% | 18/5/23 | 2/6/23 | | | | | | | |
| 1a Apresentação | LM | 100% | 22/5/23 | 24/5/23 | | | | | | | |
| Definir base de dados | LM | 100% | 23/5/23 | 31/5/23 | | | | | | | |
| Definir métricas | LM | 100% | 25/5/23 | 18/6/23 | | | | | | | |
| Construção do Dataset de Indexação (Juris-TCU-Index) | | | | | | | | | | | |
| Carga dos dados do sistema Juris (incluindo indexações) | LM | 100% | 27/5/23 | 2/6/23 | | | | | | | |
| Explorar dados do sistema Juris | LM | 100% | 1/6/23 | 5/6/23 | | | | | | | |
| Explorar dados de indexações | LM | 100% | 5/6/23 | 7/6/23 | | | | | | | |
| Tratar dados | LM | 100% | 27/5/23 | 13/6/23 | | | | | | | |
| Gerar dataset (com grels) | M | 100% | 7/6/23 | 14/6/23 | | | | | | mi | |
| Validar formato dataset | M | 100% | 7/6/23 | 9/6/23 | | | | | | | |

| Construção do Dataset de Avaliação de Busca (Juris-TCU) | | | | | |
|---|----|------|---------|---------|-------------|
| Explorar dados de log da busca | L | 100% | 1/6/23 | 14/6/23 | |
| Criar queries | L | 100% | 5/6/23 | 15/6/23 | |
| Definir prompt LLM para grel | LM | 90% | 15/6/23 | 17/6/23 | |
| Montar base de avaliação (qrel) | L | 40% | 16/6/23 | 25/6/23 | 8998 BOLD B |
| Construção do pipeline de indexação | | | | | |
| Criar estrutura busca (elastic search, etc) | М | 100% | 4/6/23 | 7/6/23 | |
| Criar estrutura de registro e avaliação (métrica) | М | 100% | 4/6/23 | 12/6/23 | |
| Realizar experimentos: indexador como pipeline de busca | LM | 100% | 12/6/23 | 21/6/23 | |
| Treinar modelo | М | 5% | 16/6/23 | 23/6/23 | |
| (ou: experimentar LLM como estágio 3 dos pipes) | М | 0% | 16/6/23 | 26/6/23 | |
| Frente pipeline de busca | | | | | |
| Criar estrutura busca (elastic search, etc) | М | 90% | 17/6/23 | 26/6/23 | |
| Criar estrutura de registro e avaliação (métrica) | М | 90% | 17/6/23 | 26/6/23 | |
| Experimentar: indexadores como expansão (cadastro) | LM | 0% | 20/6/23 | 25/6/23 | 4000 |
| Frente Avaliação e Fechamento | | | | | |
| Relatório | LM | 0% | 16/6/23 | 30/6/23 | 10000 |
| Preparar apresentação final | LM | 0% | 24/6/23 | 30/6/23 | - BA |
| Publicar datasets com greis (indexação e busca da jurisprudência) | LM | 60% | 4/6/23 | 30/6/23 | |

R.Nogueira, Z.Jiang, R.Pradeep, and J.Lin. Document ranking with a pretrained sequence- to sequence model. In Findingsof the Association for Computational Linguistics: EMNLP 2020, pages 708–718. Disponível em: https://arxiv.org/pdf/2010.06467.pdf. Acesso em 25 maio 2023 2020

Referências

BRASIL. Tribunal de Contas da União. Vocabulário de controle externo do Tribunal de Contas da União – 3.ed. rev. e ampl. – Brasília : TCU, Instituto Serzedello Corrêa, Centro de Documentação. Disponível em:

https://portal.tcu.gov.br/data/files/F8/04/8E/5E/A0B3071068A7C107F18818A8/VCE_TCU.pdf.
Acesso em 25 maio 2023, 2019.

_____. Portaria-TCU - 85, de 06 de junho de 2022. Aprova o *Manual* de Sistematização e Divulgação da Jurisprudência do Tribunal de Contas da União. Disponível em:

https://pesquisa.apps.tcu.gov.br/#/documento/atonormativo/Ac%25C3%25B3rd%25C3%25A3o%2520n%25C2%25BA%25202800%252F2022/% 2520/score%2520desc/1/%2520. Acesso em 25 maio 2023. 2022.

mire