3º.Ciclo de Encontros: 4 x PLN

Capítulo 2 - Volume 2
Processamento de Linguagem Natural BPLN

O papel dos dados no pré-treinamento de Grandes Modelos de Linguagem

Miguel Carpi - IME / USP
06/05/2026

Sumário

- [ ] Por que e quando pré-treinar?
- [ ] Conjuntos de dados em Português
- [ ] Curadoria de dados
- [ ] *Corpus* Carolina

Por que e quando pré-treinar?

Fases de treinamento

- **pré-treinamento -** estrutura da língua, conhecimentos gerais do mundo
  - Modelo de Linguagem Causal - GPT
  - Modelo de Linguagem Mascarado - BERT

- **ajuste fino -** adaptar o modelo para uma tarefa específica
  - Classificar se um texto é de uma determinada autora
  - Classificar palavras (Ex: nome de lugar, nome de pessoa)

- **pré-treinamento** continuado - adaptar o modelo para um domínio específico
  - Médico
  - Legal

Exemplos interessantes

LLM = Tokenizador + Rede Neural

Tokenizador -> Quebra o texto em (sub)palavras

Tokenizador também é treinado!

Pode ser necessário re-adequar o tokenizador: Ex:

Domínio biomédico: PubMedBERT

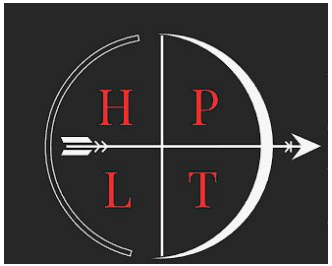
naxolona -> [`na, ##lo, ##xon, ##a`

Domínio jurídico brasileiro LegalBERT

Critérios de avaliação	Ajuste fino / pré-treinamento continuado	Pré-treinar do zero
1. Domínio e Dados		
1.1 Vocabulário	Linguagem padrão, jargão limitado. Tokenizador existente funciona bem.	Léxico vasto, único e especializado. Tokenizador padrão não fragmenta bem os termos.
1.2 Dados disponíveis	*Dataset* rotulado pequeno/médio.	*Corpus* com pelo menos 1B tokens.
1.3 Conhecimento geral	Conhecimento de mundo é útil.	Conhecimento de mundo não é necessário.
2. Tarefa e Capacidades		
2.1 Novidade da tarefa	Tarefa padrão de PLN em um novo domínio.	Tarefa exige raciocínio novo e complexo. Conhecimento médico, jurídico, ou matemático.
2.2 Teto de desempenho	Ajuste fino gera um desempenho "bom o suficiente".	Ajuste fino gera modelos que falham em casos essenciais.
3. Arquitetura e Tecnologia		
3.1 Adequação da arquitetura	Arquitetura Transformer padrão é suficiente.	Requer uma arquitetura para desafios específicos.
3.2 Objetivos de treino	Objetivos padrão (MLM, CLM) são suficientes.	Requer novos objetivos para o domínio.
4. Recursos e Estratégia		
4.1 Orçamento e prazo	Orçamento e tempo limitados (dias/semanas).	Orçamento alto e bastante tempo (meses). Acesso a cluster de GPUs.
4.2 Importância estratégica	Desempenho "bom o suficiente" é aceitável. Será necessário realizar mais de uma tarefa.	Dominar o nicho é um objetivo central. O modelo é um ativo estratégico chave.

Conjuntos de dados em Português

COMMON CRAWL



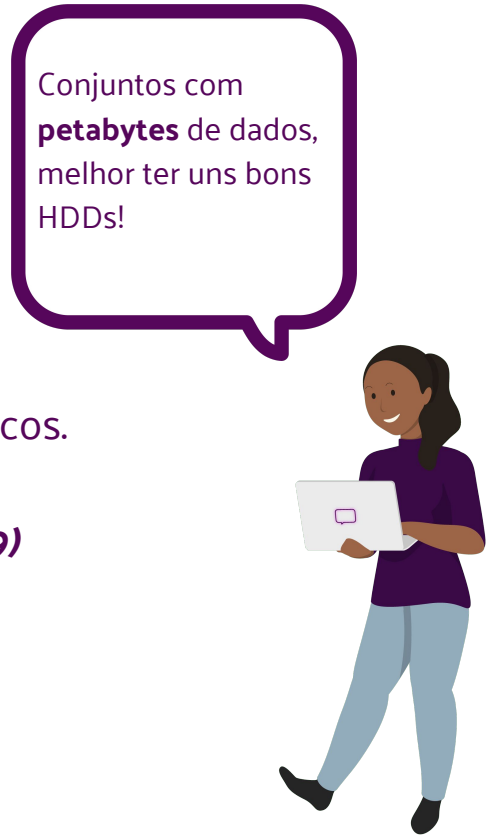
SKETCH ENGiNE

ptTenTen

Carolina

Recursos gerais

- ***Common Crawl (2008)***
  - Organização pelo ano
  - Página HTML original
- ***TenTen Corpus Family (2013)***
  - Morfossintática, gênero do discurso e classificação em tópicos.
  - Acesso mediante pagamento
- **OSCAR - *Open Super-large Crawled Aggregated coRpus (2019)***
  - Foco em línguas com poucos recursos
- **HPLT - *High Performance Language Technologies (2024)***
  - Voltado para treinamento de modelos de linguagem



Recursos para o português parte 1

- ***Linguateca (2000)***
  - Organização virtual:
    - **Corpus do Português (2006)**

    Classe de palavra, textos históricos, textos recentes, comparação de dialetos
    - **Corpus Brasileiro (2010)**
      - Textos escritos e falados, com anotações sintática e semânticas além de atributos de gênero discursivo
- ***Brazilian Portuguese Web as Corpus (brWaC) (2018)***
    - WacKy (*The Web-As-Corpus Kool Yinitiative)*
    - 134 atributos textuais diferentes

Recursos para o português parte 2

- **BlogSET-BR (2018)**
  - ***808 blogs*** brasileiros,
- **Carolina (2020)**
  - *Corpus* **aberto** com controle de fontes, gêneros textuais, data de publicação e metadados associados
- **Aroeira (2024)**
  - Corpus **curado** para treinamento de grandes modelos de língua
- **LegalPT (2024)**
  - Agregador de *datasets* jurídicos disponíveis publicamente em Português
- **GigaVerbo (2025)**
  - **145 milhões de documentos** concatenando diversos *datasets* do Português

Curadoria de dados

São tantos *datasets*... Qual devo usar? O que eles oferecem?

Curadoria de dados

Construção de conjunto de dados é uma etapa **essencial** no desenvolvimento de tecnologias de PLN

- Permite treinar, avaliar, estudar
- **Não basta só acumular grandes volumes de texto**

A curadoria é um **mecanismo de controle** da qualidade dos dados

- assegura a representatividade dos diferentes registros do português
- reduz a presença de ruídos e vieses, garantindo que os textos respeitem ética e metodologia
- torna o conjunto de dados **confiável** para pesquisa e desenvolvimento em PLN

Curadoria

Qualidade

Diversidade

Proveniência

Qualidade dos dados

O que é um texto de **qualidade**? (Albalak et al., 2024)

-> Escrito por um ser humano

-> Passou por um processo de seleção

Dimensões da seleção: relevância, eficiência, e segurança

- **Relevância**: esse texto é adequado para o modelo? (idioma, semântica)
- **Eficiência**: fazer o treinamento ser mais eficaz (deduplicar conteúdo)
- **Segurança**: dados seguros -> modelo seguro (privacidade, conteúdo indesejado)

Custo da Qualidade dos dados

Qualidade pode ser medida de forma **intrínseca** e **extrínseca**
Intrínseca: nos próprios dados
Extrínseca: no modelo treinado com os dados

Medidas intrínsecas:

- Usar modelos referência
- Usar heurísticas (número de palavras únicas, número de sinais de pontuação
- Usar listas de bloqueios (urls, lista de palavras, lista de autores)

Problemas

- Modelos de referência tem vieses
- Heurísticas não são flexíveis
- Listas de bloqueios podem diminuir representatividade (não analisam semântica)

C**ontrolar qualidade é mais um guia de boas práticas, tem que saber quando elas se aplicam**

Diversidade dos dados

Diversidade não é só um problema de modelos multilinguísticos!

Língua Portuguesa = conjunto de manifestações linguísticas distintas

Variação sobre várias dimensões Wolfram (2006)

- distribuição geográfica
- grupo social
- período histórico

Variações em **registro**, **gênero**, e **estilo** (Biber; Finegan (1994) e Biber; Conrad (2009))

Diversidade dos dados

- **Registro:** variação segundo a situação em que a linguagem é produzida
  - registro da conversação pessoal, da comunicação em fóruns virtuais, da escrita em livros didáticos

- **Gênero:** variação segundo o tipo de mensagem e as convenções associadas à como aquele tipo de mensagem é construído
  - entrada de enciclopédia, carta, manual, receita culinária

- **Estilo:** variação segundo preferências estéticas de quem produz, ou consome a linguagem, podendo ser mais alinhada coletivamente ou individualmente
  - estilo da terceira fase do romantismo, estilo pessoal de Machado de Assis

Diversidade dos dados

- Pessoas adequam o uso da língua a depender da situação
    - Conversa no bar com amigos <> conversa no tribunal (registro)
    - *e-mail* para o RH <> receita de bolo (gênero)
    - poema árcade <> poema cubista (estilo)

- Ajuste fino pode não ser suficiente para o modelo capturar as diferenças intralinguísticas

    É necessário que o pré-treinamento, ou continuação desse, **inclua um nível de diversidade linguística apropriado.**

- Isso permitirá que o modelo aprenda a representar os elementos das várias sublínguas que compõem o idioma a ser dominado.

Proveniência dos dados

Proveniência: processo de rastreamento dos dados, isto é, de suas origens e histórico de processamento (Crespo et al., 2023)

- De onde os dados vieram?
- Esse dado passou por algum tipo de processamento antes? Qual? Por que?
- Quem tem permissão de utilizar esses dados? E para o quê? (licença de uso)
- Como fica a licença do modelo ao ser treinado com dados sob licença X?
- Há dados sensíveis? (CPF, número de telefone, e-mail)

Corpus Carolina

Talvez um exemplo ajude ....

Corpus Carolina

- O corpus Carolina (Córpus Geral do Português Brasileiro com Informações de Procedência e Tipologia) é um corpus **aberto.**
- Corpus **dual**, construído para treinamento de modelos de linguagem e pesquisa linguística.
- Foi desenvolvido por uma **equipe multidisciplinar** de **Linguistas** e **Cientistas da Computação**.

- **Tentativa** de conjugar aspectos discutidos até aqui
  - Empreitada difícil, pois restrições e objetivos competem
  - (ex: diversidade linguística versus licença aberta).
- O corpus Carolina tem se tornado um recurso útil por si só.
- Serve como protótipo para a construção de futuros recursos semelhantes.

Corpus Carolina

- Faz parte do desafio de NLP do C4AI (Center for Artificial Intelligence).
- Está em construção contínua desde setembro de 2020 (versão 1.0 - Ada)
- A versão atual, 2.0 (Bea):
  - mais de 2 milhões de textos
  - cerca de 1,3 bilhões de tokens.
- Todos os textos do corpus foram extraídos da web e são posteriores a 1970.

Corpus Carolina

O corpus Carolina foi criado com o intuito de preservar informações de proveniência e de tipologia textual. Para alcançar esse objetivo:

- Apresentamos um cabeçalho com metadados anotados em lotes de textos, que **conservam o máximo de informações** que podemos obter de nossas fontes e seguem o padrão TEI (Text Encoding Initiative) (TEI Consortium, 2021).
- **Informações** como a **licença** e **URL da fonte** são **obrigatoriamente anotadas** garantindo
  - a rastreabilidade dos textos originais,
  - conformidade com os direitos autorais de cada um.

```
<sourceDesc>
    <biblFull>
        <fileDesc>
            <titleStmt>
                <title>
                    <name>Discussão:Mutirão:Página principal</name>
                    <media mimeType="text/xml" url="https://br.wikimedia.org/wiki?curid=13" source="
                    brwikimedia-20210701-pages-meta-current.xml" />
                </title>
                <author />
                <editor role="translator" />
                <sponsor>Unknown</sponsor>
            </titleStmt>
            <extent>
                <measure unit="bytes" quantity="14471932" />
                <measure unit="tokens" quantity="1586470" />
                <measure unit="pages" quantity="-1" />
            </extent>
            <publicationStmt>
                <publisher />
                <authority>Wikimedia Foundation</authority>
                <date>2014-09-16</date>
                <availability status="free">
                    <license target="https://creativecommons.org/licenses/by-sa/3.0/deed.pt_BR">Creative Commons - Atribuição -
                    Compartilha Igual 3.0 Não Adaptada (CC BY-SA 3.0)</license>
                </availability>
```

A *tag* “<license>” contém as informações da licença da fonte e “<media>” a URL do texto.

Corpus Carolina

Também anotamos informações das duas tipologias textuais do corpus:

- **Tipologia Ampla** (ou Tipologia do Carolina): Organiza os textos dentro da estrutura do corpus.
- **Tipologia da Fonte**: Termos utilizados pelas próprias fontes para se referirem aos seus textos.

Além disso, anotamos o Domínio de cada texto, que entendemos como a esfera em que circulam.

- Exemplos: Acadêmico, Pedagógico, Comercial etc.

Esses metadados nos fornecem informações importantes acerca do registro dos textos que compõem o Carolina e podem ser usados para mensurar sua diversidade linguística.

```
            <profileDesc>
                <creation resp="user" />
                <textDesc>
                    <channel mode="w" />
                    <constitution type="single" />
                    <derivation />
                    <domain>Virtual Forum</domain>
                    <factuality />
                    <interaction />
                    <preparedness type="monitored" />
                    <purpose />
                </textDesc>
                <textClass>
                    <catRef scheme="#Source_typology" target="#DISCUSSION_VIR_W" />
                </textClass>
                <langUsage>
                    <language ident="pt-BR" />
                </langUsage>
            </profileDesc>
        </biblFull>
    </sourceDesc>
</fileDesc>
<profileDesc>
    <textClass>
        <catRef scheme="#Carolina_typology" target="#WIKIS" />
    </textClass>
</profileDesc>
```

As *tags* “<catRef>” contêm os tipos amplos e da fonte. A *tag* “<domain>” apresenta o domínio.

Exemplo de uso de informações

Tipologia da Fonte	# Documentos	menor	mediana	p80	maior
---	---	---	---	---	---
str	u32	i64	f64	f64	i64
#VOCABULARY_ENTRY_INS_W	836073	14	157.0	480.0	36707
#TWEET_VIR_W	711003	1	13.0	23.0	85
#NEWS_JOU_W	189731	3	370.0	594.0	10434
#PRODUCT_REVIEW_COM_W	160614	1	15.0	30.0	795
#DISCUSSION_VIR_W	46675	10	110.0	349.0	68254

Corpus Carolina 2.0 - 5 tipologias mais frequentes

Considerações finais

Em resumo:

- Pré-treinamento
- alguns dos principais conjuntos de dados disponíveis
- principais aspectos que alguém deve considerar ao selecionar / construir um conjunto de dados

E se em algum momento você obtiver um modelo cheio de alucinações, apesar de todo o seu esforço na seleção de dados, não perca a esperança: lembre-se que até Dom Quixote acabou são no final!

Miguel de Mello Carpi
Felipe Ribas Serras
Mariana Lourenço Sturzeneker
Mayara Feliciano Palma
Gabriela Alves Lachi
Aline Silva Costa
Maria Clara Paixão de Sousa
Cristiane Namiuti
Vanessa Martins do Monte
Marcelo Finger

Obrigada!