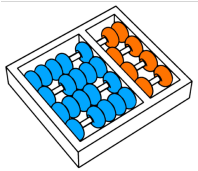
Metodologia e ferramenta para gerar datasets de conversas

Matheus Ferraroni Sanches
Orientador: Prof. Dr. Leandro A. Villas

Campinas-SP, 20 de outubro de 2021



Agenda

1. Introdução
2. Metodologia
3. Ferramenta
4. Datasets
5. Conclusão
6. Próximos passos

Agenda

1. Introdução
2. Metodologia
3. Ferramenta
4. Datasets
5. Conclusão
6. Próximos passos

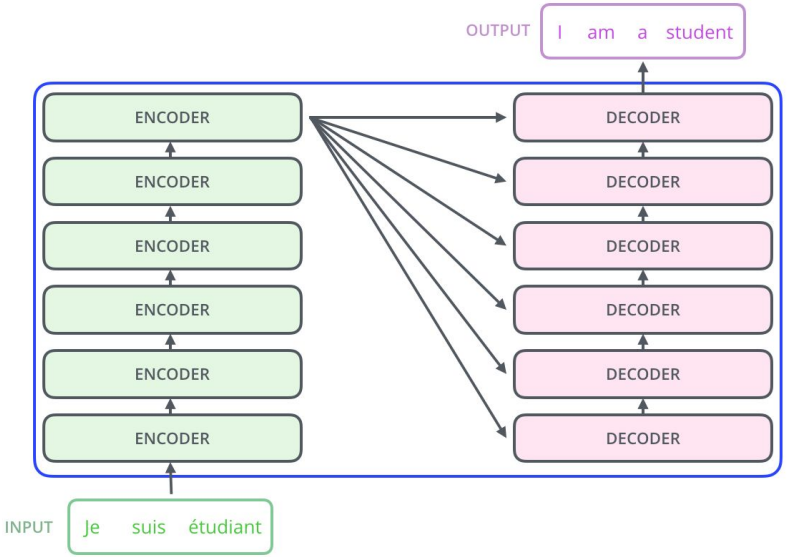
ICEIS 2022

Doutorado

Introdução

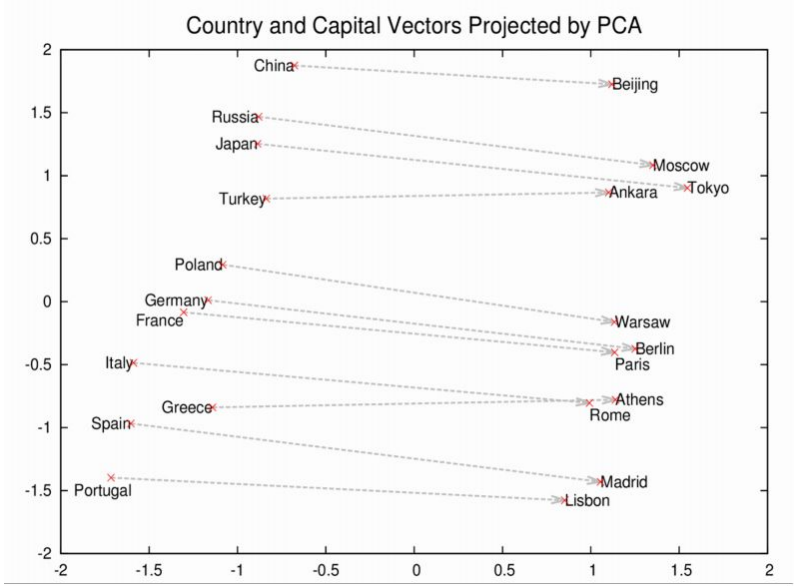
Introdução

- Modelos Sequence to Sequence
  - Bons resultados
- Dados
  - Requerimentos
    - Grande quantidade
    - Confiáveis
    - Viés
    - Idioma
  - Requisitos reduzem datasets
- Embedding



Introdução

- Modelos Sequence to Sequence
  - Bons resultados
- Dados
  - Requerimentos
    - Grande quantidade
    - Confiáveis
    - Viés
    - Idioma
  - Requisitos reduzem datasets
- Embedding



Introdução

- Falta de datasets
- PT-BR
- Mais requisitos, mais escasso
  - Idioma
  - Tamanho
  - Texto informal/formal
  - Palavras específicas
  - Conversa em turnos
    - 2 pessoas
    - 2+ pessoas
  - **Conteúdo moderado**

Introdução

- Grande dependência de modelos nos dados
- Datasets
  - BrWac, Wikipedia, OSCAR, blogs, Microsoft Research WikiQA Corpus, Yahoo
- Iniciativas para gerar dados
  - Wikipedia, Ailab, esse projeto

- **Objetivo**
  - **Criar uma metodologia para criar datasets de conversas**
    - **Conversas escritas por humanos em turnos**
    - **Diferentes respostas para a mesma pergunta/fala**
  - **Implementar ferramenta seguindo a metodologia**
  - **Caso de uso com a ferramenta**

Introdução

- Grande dependência de modelos nos dados
- Datasets
  - BrWac, Wikipedia, OSCAR, blogs, Microsoft Research WikiQA Corpus, Yahoo
- Iniciativas para gerar dados
  - Wikipedia, Ailab, esse projeto

- **Objetivo**
  - **Criar uma metodologia para criar datasets de conversas**
    - **Conversas escritas por humanos em turnos**
    - **Diferentes respostas para a mesma pergunta/fala**
  - **Implementar ferramenta seguindo a metodologia**
  - **Caso de uso com a ferramenta**

Metodologia

Metodologia

- Objetivo
- Arquitetura
- Entrada
  - Três etapas
- Processamento
- Saída
  - Três etapas
- Fluxo

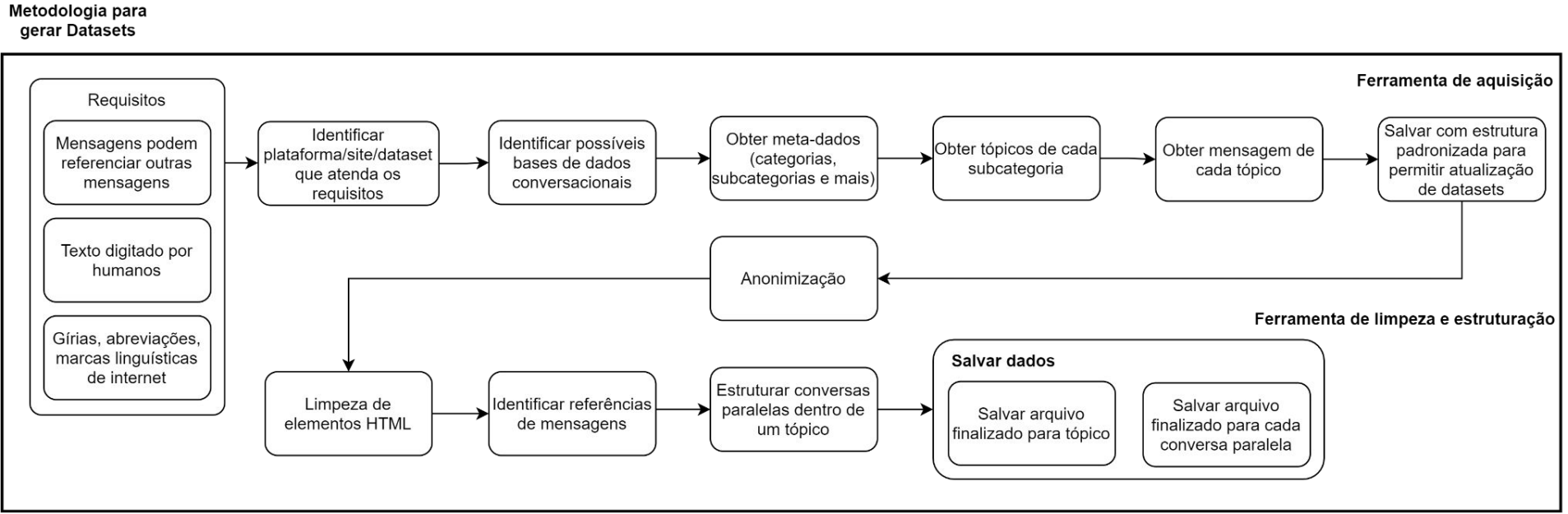
Metodologia - Objetivo

Processos descritos na metodologia

- Acessar dados
- Identificar atributos relevantes
  - Categoria, subcategoria, horários, ids
- Obter mensagens criadas por humanos
  - Marcas linguísticas próprias
- Organizar dados obtidos
  - Tratar elementos específicos
    - Marcas HTML, markdown, relações BD
  - Facilitar acesso
  - Estruturação
    - Permitir atualização do dataset
- **Dados prontos**
  - **Language Modeling**, **Q&A**, classificação, previsão de recursos

Metodologia - Arquitetura

- Três grandes partes: **Aquisição**, Anonimização, **Processamento**
  Limpeza e estruturação



Metodologia - Entrada

- Aquisição
  - Agnóstico à fonte
  - Requisitos
    - Metadados
    - Mensagens
    - Referências
- Anonimização
  - Dados gerados a partir da aquisição
- Processamento
  - Dados gerados a partir da aquisição ou anonimização
  - Formato padronizado

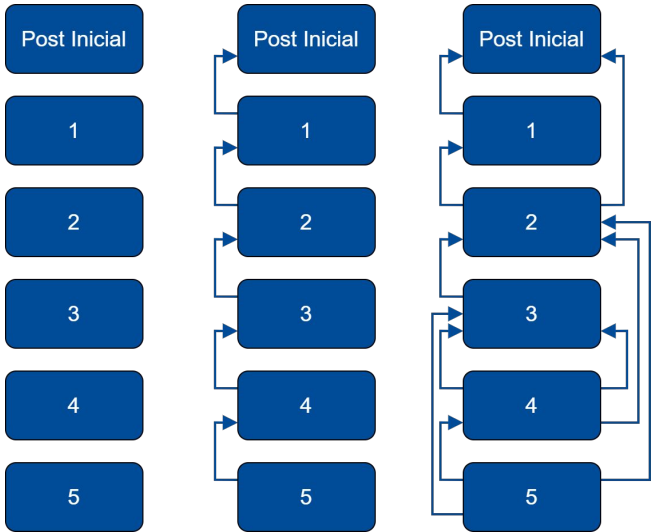
Metodologia - Anonimização

- Processo de anonimização
  - Procura por valores que podem identificar um usuário específico
    - Nome, email, id, telefone, url
    - Depende dos dados obtidos na aquisição
    - Exigência:
      - Busca apenas por chaves conhecidas
    - Opcional
      - Busca dentro das mensagens

Metodologia - Processamento

- Limpeza e estruturação
  - Limpar dados
    - Depende do tipo do dado. HTML: imagens, urls, iframe...
  - Gerar arquivos .tsv
  - Identificar referências entre mensagens
    - Criar grafo de dependências
    - Uma conversa paralela para cada caminho entre nó inicial e folhas
    - 2 ou mais usuários interagindo

Tipo do elemento	Tag inserida	Tipo do elemento	Tag inserida
Imagem externa	<image>	Conteúdo compartilhado	<shared_content> {} {} </shared_content>
Emoji	<emoji> {} </emoji>	Media	<mediaembed> {} </mediaembed>
Imagem desconhecida	<image_unknown>	Resposta	<answering> {} </answering>
Citação	<quote> {} </quote>	Spoiler	<spoiler> {} </spoiler>
Url	<url> {} {} </url>	Código	<code> {} </code>
Link	<link>	Iframe	<iframe> {} </iframe>

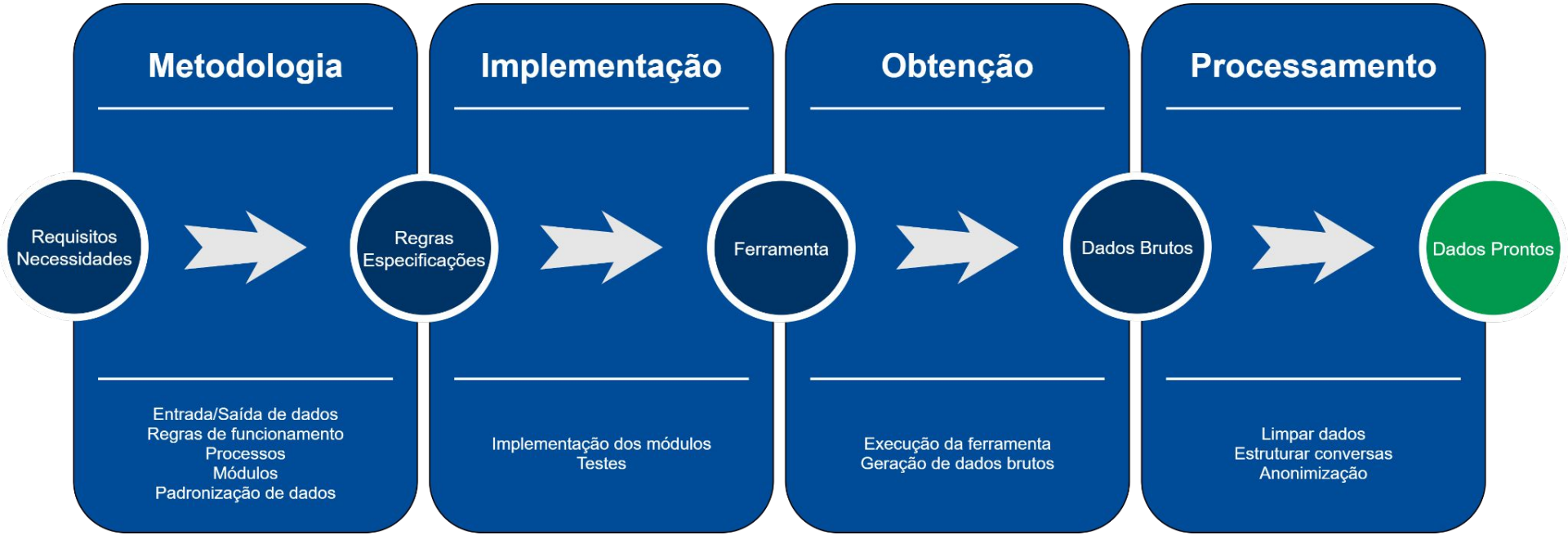


Metodologia - Saída

- Aquisição
  - Arquivos .json
- Anonimização
  - Arquivos, estrutura e dados são mantidos
    - Exceção de valores referentes a identificadores de usuários
- Processamento
  - Arquivos .tsv
  - Tópico
    - Um arquivo com todas as mensagens de um tópico
      - Ordem temporal
      - Todos usuários do tópico
    - N arquivos - Cada arquivo contendo uma conversa paralela de um tópico
      - Ordem temporal
      - 2 ou mais usuários por arquivo

Metodologia - Fluxo

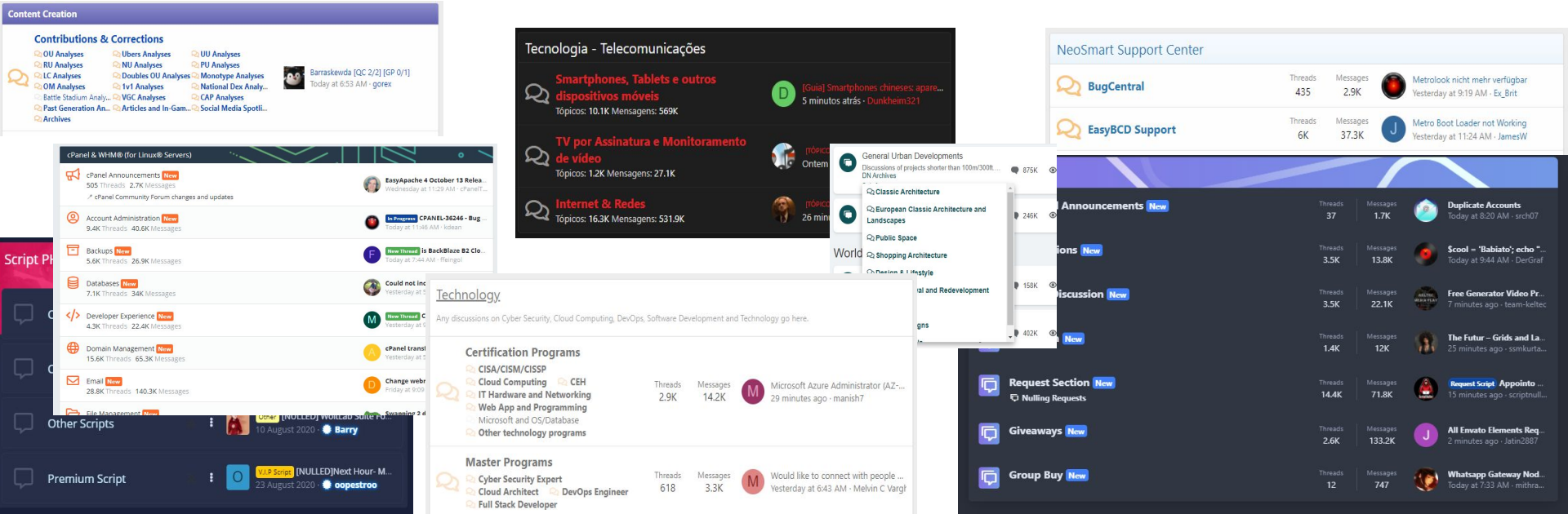
- Fluxo proposto para utilização da Metodologia~Dados prontos para uso



Ferramenta

Ferramenta

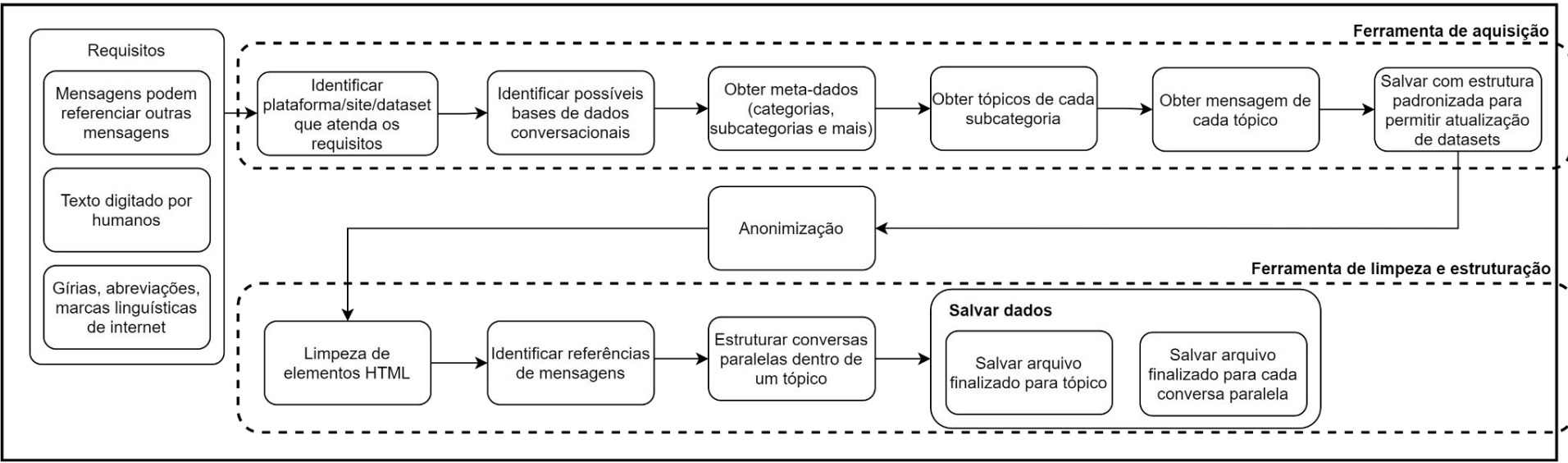
- Ferramenta baseada na metodologia proposta
  - Base de dados: Fóruns do tipo xenForo®
  - Fóruns nacionais e internacionais
  - Dados
    - Categorias, subcategorias, tópicos, mensagens, horários, usuários



Ferramenta

- Metodologia organizada em 3 módulos
  - Aquisição
  - Anonimização
  - Limpeza e estruturação

Metodologia para gerar Datasets



https://github.com/MatheusFerraroni/miner-xenforo

Ferramenta

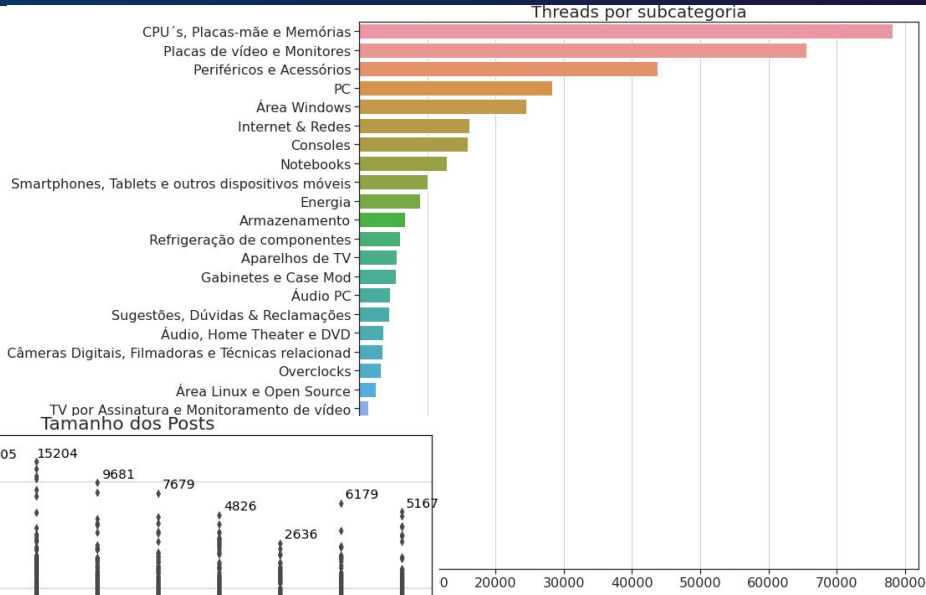
- Identificação de referências

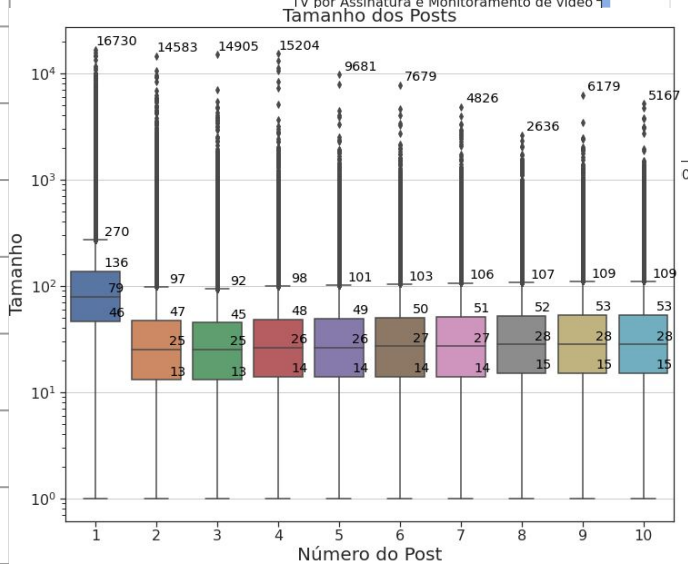
```
montar_conversas(mensagem):

    referencias = identificar_referencias(mensagem)

    for referencia in referencias:
        if referencia.data_criacao<orig.data_criacao:
            referencia['pai'] = mensagem
            montar_conversas(referencia)

    if len(referencias)==0:
        r = []
        while mensagem.pai!=None:
            r.append(mensagem)
            mensagem = mensagem.pai
        r.append(mensagem)

    return res_final
```

Datasets

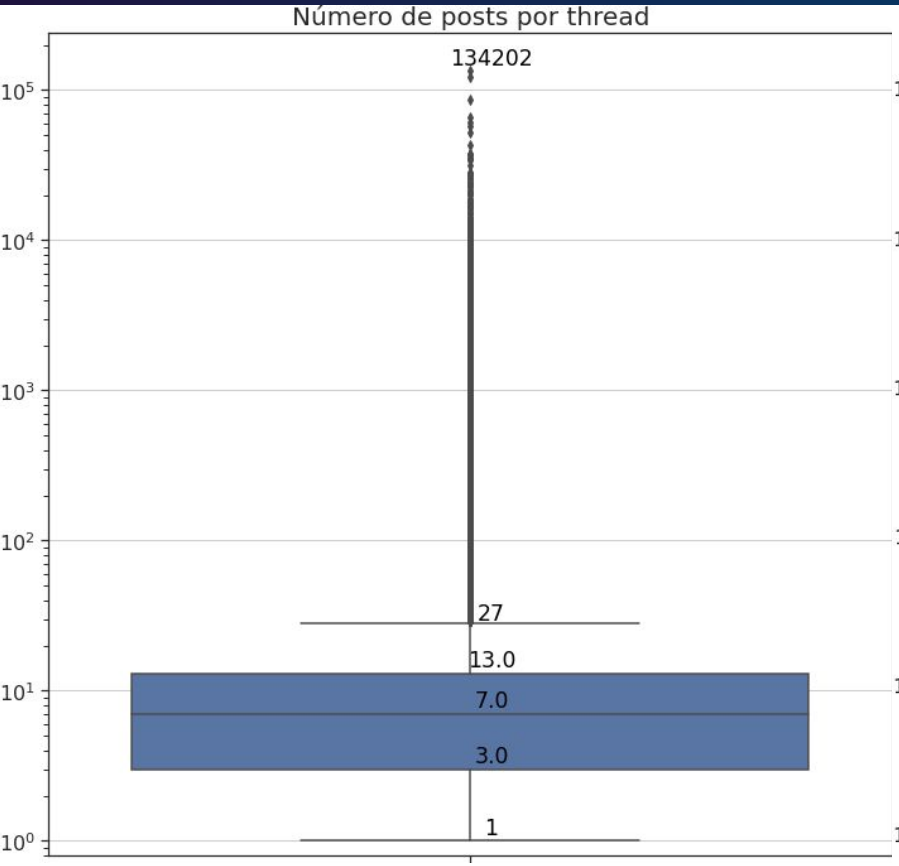
Datasets - Adrenaline

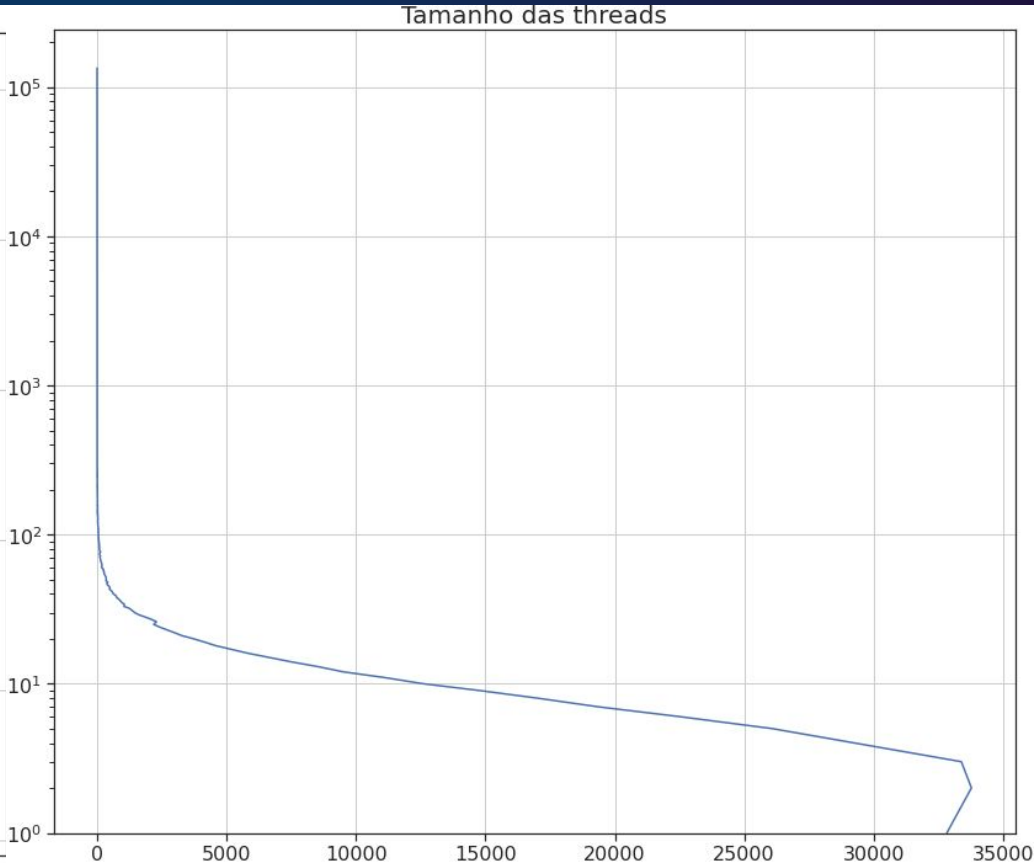
Dado	Valor
Total threads	356.763
Total posts	9.550.206
Total tokens	477.619.581
Tamanho médio thread	26,7
Média tokens post	50,0
Tamanho médio conversa	34,2
Total conversas	4.539.655
Total tokens distintos	1.701.482
Tokens médios thread	1.338
Tamanho bruto	2.0 gb
Tamanho processado	71 gb





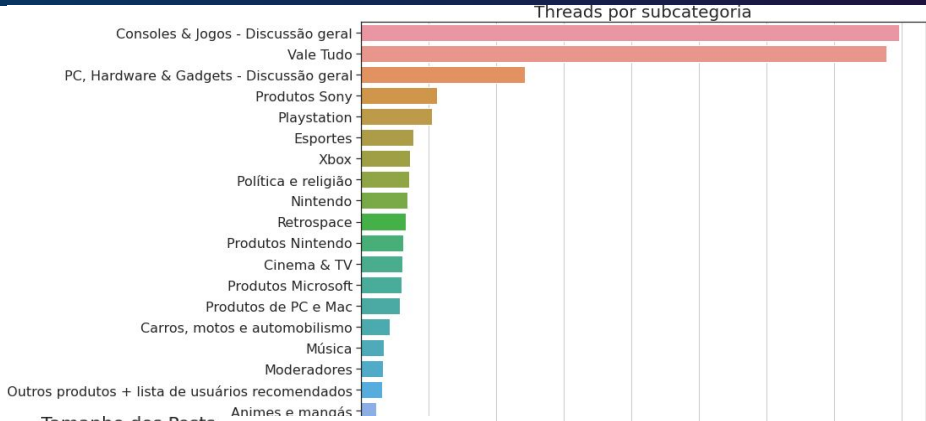
Datasets - Adrenaline

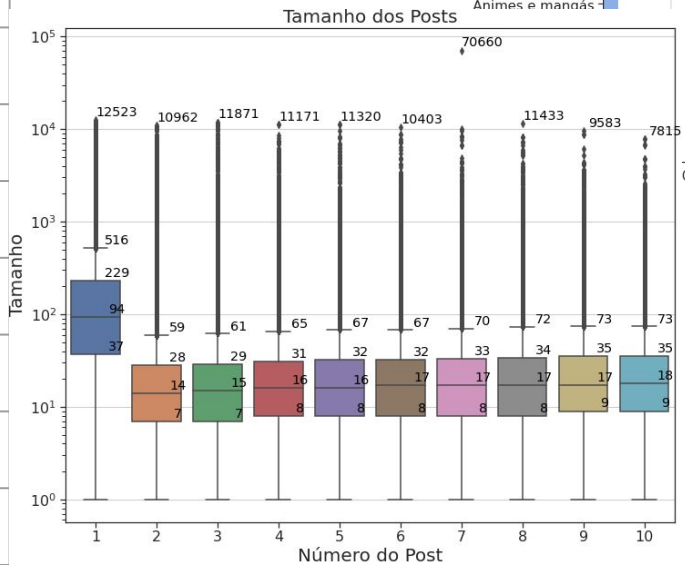




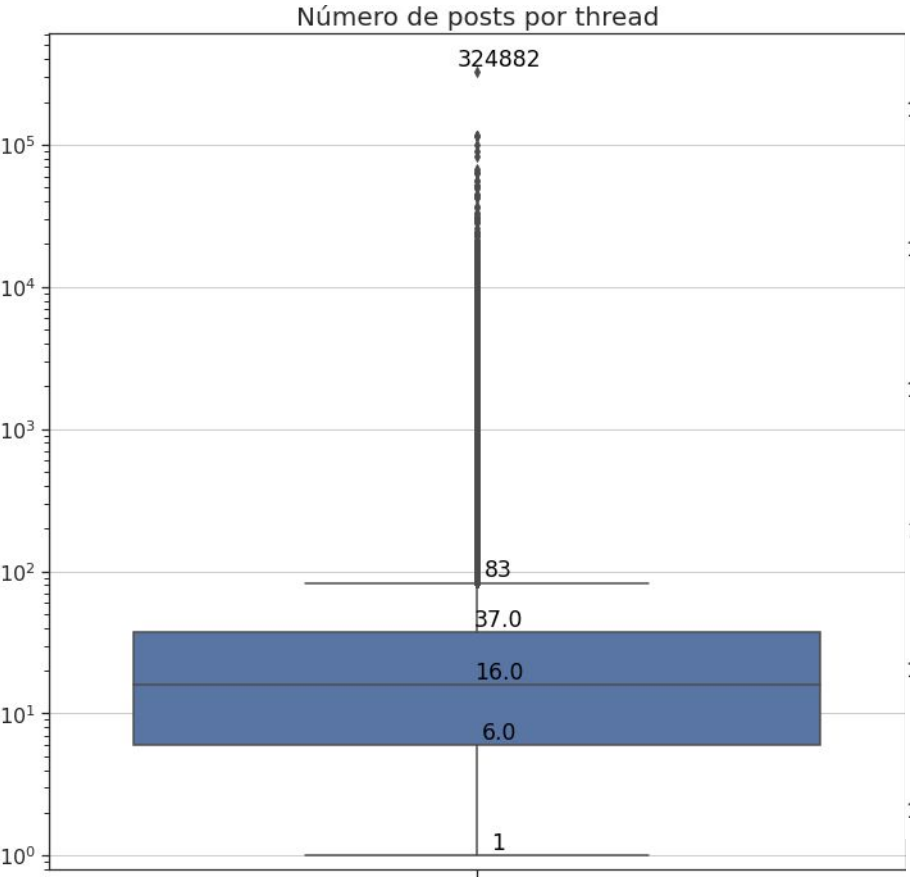
Datasets - Outerspace

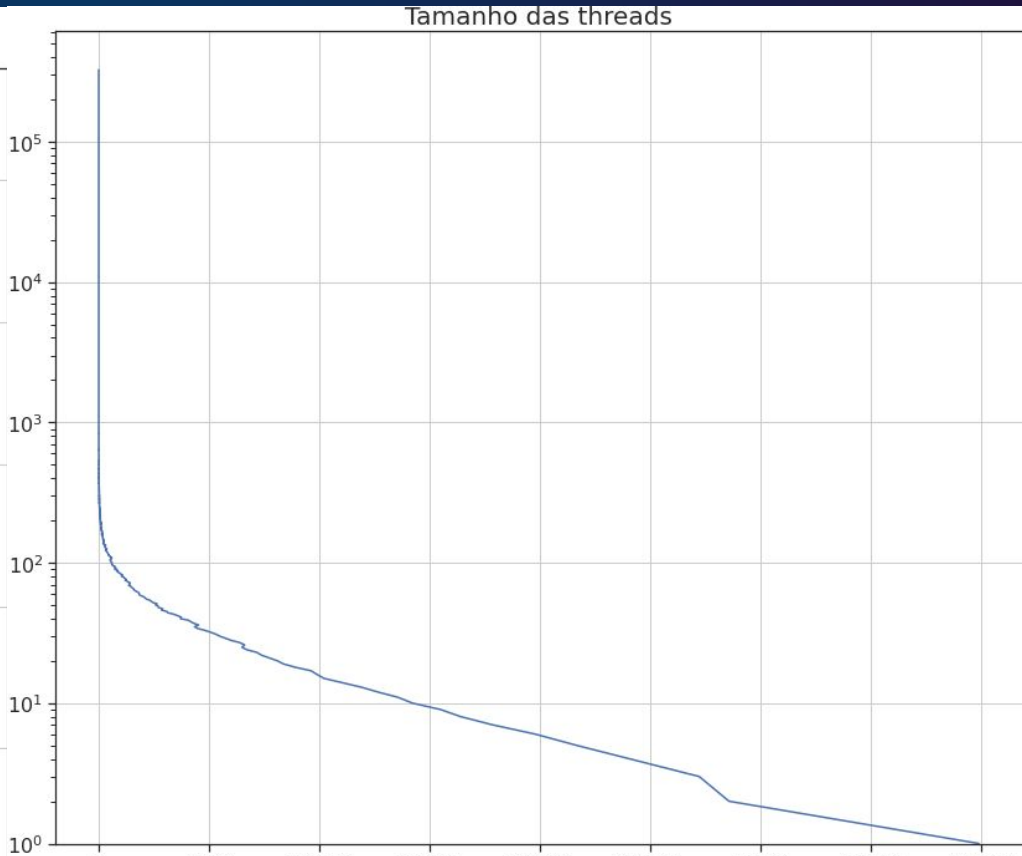
Dado	Valor
Total threads	570.105
Total posts	24.514.161
Total tokens	1.068.077.323
Tamanho médio thread	42,9
Média tokens post	43,5
Tamanho médio conversa	3,3
Total conversas	3.080.489
Total tokens distintos	2.670.306
Tokens médios thread	1.873
Tamanho bruto	4.6 gb
Tamanho processado	5.4 gb





Datasets - Outerspace





Datasets - Comparação

Dado	Adrenaline	Outerspace	Adrenaline/Outerspace
Total threads	356763	570105	0.62
Total posts	9.550.206	24514161	0.38
Total tokens	477.619.581	1.068.077.323	0.44
Tamanho médio thread	26,7	42,9	0.62
Média tokens post	50,0	43,5	1.14
Tamanho médio conversa	34,2	3,3	**10.3**
Total conversas	4.539.655	3.080.489	**1.47**
Total tokens distintos	1.701.482	2.670.306	**0.63**
Tokens médios thread	1.338	1.873	0.71
Tamanho bruto	2.0 gb	4.6 gb	**0.43**
Tamanho processado	71 gb	5.4 gb	**13.1**

Conclusão

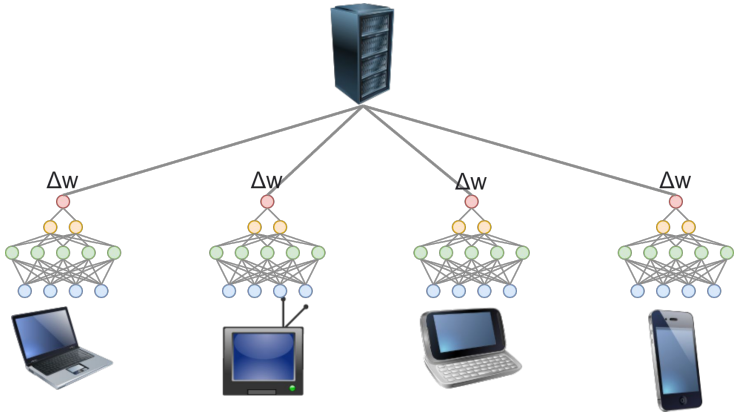
Conclusão

- Metodologia
  - Aquisição
  - Anonimização
  - Limpeza e identificação de conversas paralelas
- Ferramenta baseada na metodologia
  - Focada em fóruns XenForo
- Caso de uso com a ferramenta
  - Adrenaline
  - Outerspace

Próximos passos

Próximos passos - Modelos

- Datasets
  - Adrenaline & Outerspace
  - Limpeza BrWac
  - Estruturação Wikipédia
- GPT2 - Language Modeling
  - Novo Tokenizador
  - Small (andamento)
    - Diferentes porções dos dados
    - Diferentes configurações de treinamento
      - Congelando camadas, pesos aleatórios x modelo pré treinado
  - Base
  - Destilado
- BERT
  - Destilado
    - Tamanho: 40% de redução
    - Score: 97% do modelo original
    - Velocidade: +60%



Obrigado

ferraroni@lrc.ic.unicamp.br