jornadas FCCN

Amália IA

Treinada com dados do Arquivo.pt

Pedro Gomes

FCCN serviços digitais fct | fct Fundação para a Ciência e a Tecnologia | U.PORTO | U.PORTO FEP FACULDADE DE ECONOMIA UNIVERSIDADE DO PORTO

Amália IA

O AMALIA (Assistente Multimodal Automático de Linguagem com Inteligência Artificial) é o primeiro grande modelo de linguagem (LLM) desenvolvido especificamente com foco no português europeu e na cultura de Portugal.

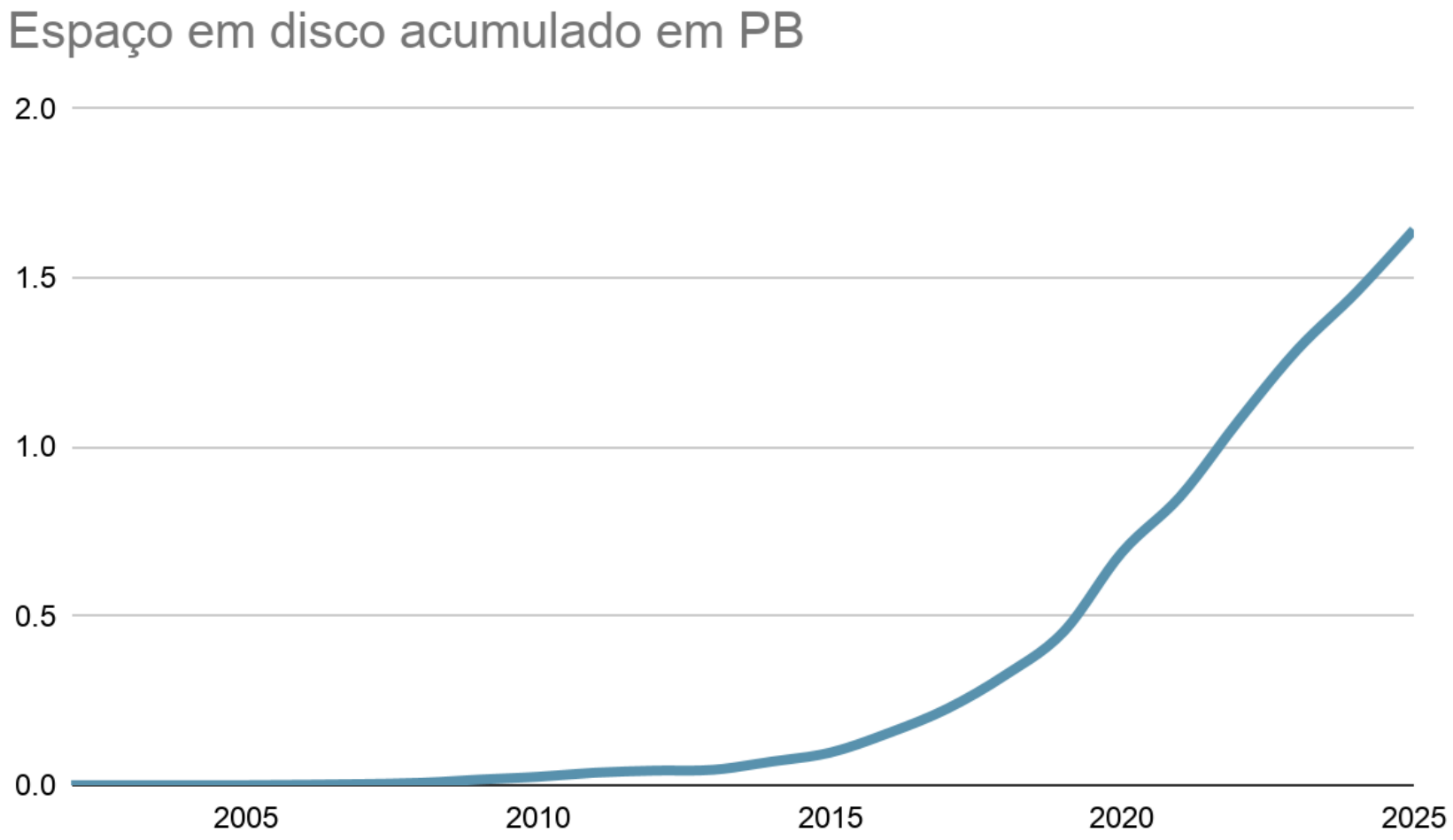
Preservação da web

arquivo.pt/numeros

>**23,000,000,000** de ficheiros recolhidos (+30% em relação a 2024)

1.7PB preservados (+12% em relação a 2024)

15 novas coleções - incluindo das eleições



Arquivo.pt Collections

arquivo.pt/collections

List of the collections preserved by Arquivo.pt (publicly available)

Collection ID	Description	Collection Start date	Collection End date	Total files (Number of URLs collected)	Total seeds (Preserved sites)	Total volume of WARCS files (TB)	Collection in Production? (URL, Page and Image indexed)
AWP1	1st complete crawl of the Portuguese web, mainly from the .PT domain, in 2008.	2008-02-12	2008-03-06	56,046,288	154,787	1.60	VERDADEIRO
AWP2	2nd complete crawl of the Portuguese web, mainly from the .PT domain, in 2008.	2008-03-11	2008-05-30	48,718,404	-	1.60	VERDADEIRO
AWP3	3rd complete crawl of the Portuguese web, mainly from the .PT domain, in 2008.	2008-10-21	2008-12-10	51,863,006	193,294	2.00	VERDADEIRO
AWP4	4th complete crawl of the Portuguese web, mainly from the .PT domain, in 2009.	2009-05-01	2009-05-31	68,776,707	366,880	2.50	VERDADEIRO
AWP5	5th complete crawl of the Portuguese web, mainly from the .PT domain, in 2009.	2009-10-01	2009-10-31	119,135,566	373,323	3.80	VERDADEIRO
AWP6	6th complete crawl of the Portuguese web, mainly from the .PT domain, in 2009.	2009-12-01	2009-12-31	118,810,364	340,018	3.50	VERDADEIRO
AWP7	7th complete crawl of the Portuguese web, mainly from the .PT domain, in 2010.	2010-05-01	2010-05-31	87,988,812	389,957	2.90	VERDADEIRO
AWP8	Incremental crawl of the Portuguese web, mainly from the .PT domain, in 2010. The AWP8 crawl is incremental because it was performed using DeDuplicator (http://landsbokasafn.GitHub.io/DeDuplicator/) taking the content of AWP7 as baseline. Thus, the files that remained unchanged from the AWP7 complete crawl were not archived (duplicated) on the AWP8 incremental crawl.	2010-08-01	2010-08-31	75,771,317	411,562	1.90	VERDADEIRO
AWP9	Incremental crawl of the Portuguese web, mainly from the .PT domain, in 2011. The AWP9 crawl is incremental because it was performed using DeDuplicator (http://landsbokasafn.GitHub.io/DeDu	2011-01-20	2011-03-22	81,114,575	473,588	2.10	VERDADEIRO

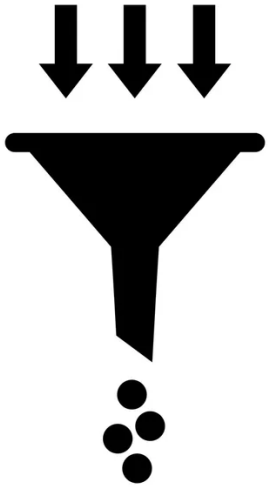
Data Pipeline

- Filtragem de URL
- Scraping
- Pós-scraping
- Filtragem por idioma
- Filtros heurísticos
- Deduplicação

Filtragem de URL

Após ler os ficheiros WARC, utilizamos dois tipos de filtragem de URL:

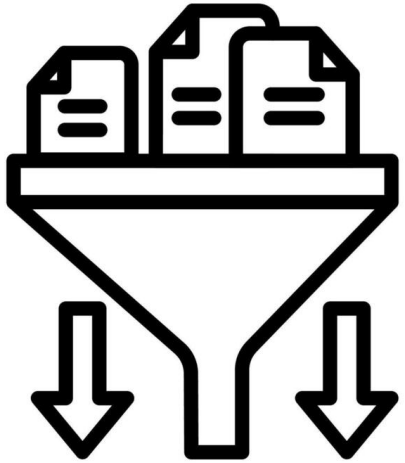
- Filtragem NSFW com blacklists
- Filtragem personalizada de URL com expressões no bloco LambdaFilter



Scraping

Utilizamos o bloco extrator Trafilatura3 para extrair dados de texto das páginas HTML.

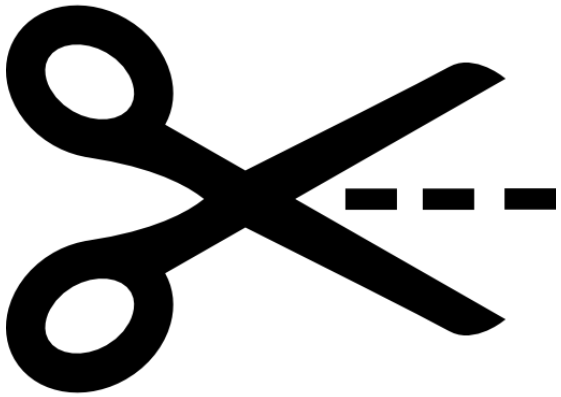
Ao privilegiar a precisão, obtém-se menos texto, mas uma extração mais correta.



Pós-scraping

Após extrair o texto, realizamos duas etapas pós-extração:

- Remover linhas duplicadas
- Remover linhas curtas (com menos de 10 caracteres alfabéticos)



Filtragem por idioma

Utilizamos um bloco de filtragem de idiomas baseado no modelo GlotLID4 para determinar o idioma dos dados de texto extraídos.

Se a probabilidade de ser português for inferior a um determinado limiar, a entrada é filtrada.

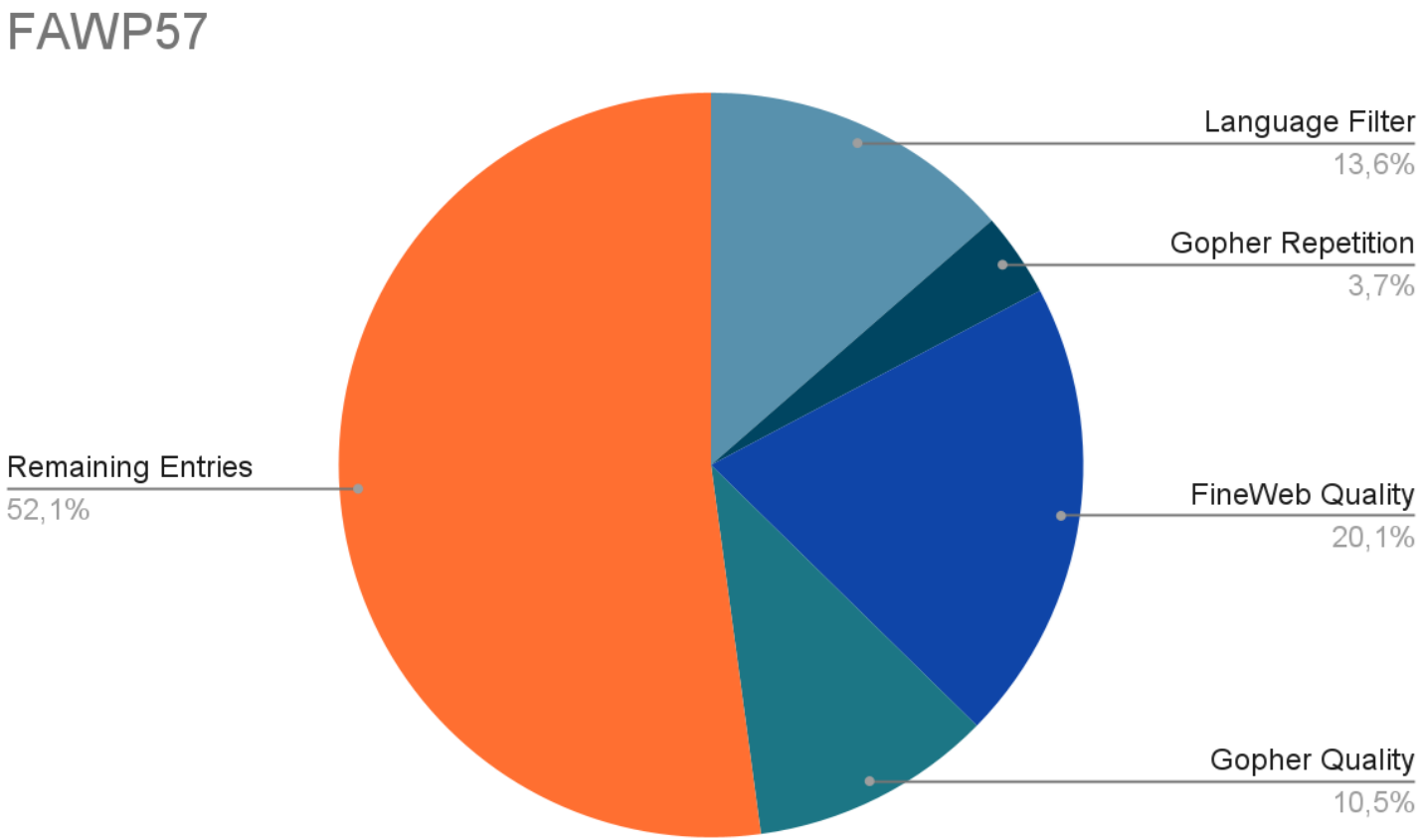
Filtros heurísticos

Estes filtros são utilizados para remover texto de baixa qualidade com base em regras heurísticas.

Existem três grupos principais de filtros:

- Filtro de Repetição Gopher;
  - Exp: ““Ficha Técnica\n**Passatempo: Veja o programa** Boa Cama Boa Mesa na SIC Notícias e concorra até dia 19 de junho.\n**Passatempo: Veja o programa** Boa Cama Boa Mesa na SIC Notícias e concorra até dia 12 de junho."”
- Filtro de Qualidade FineWeb;
- Filtro de Qualidade Gopher;

Impactos dos filtros

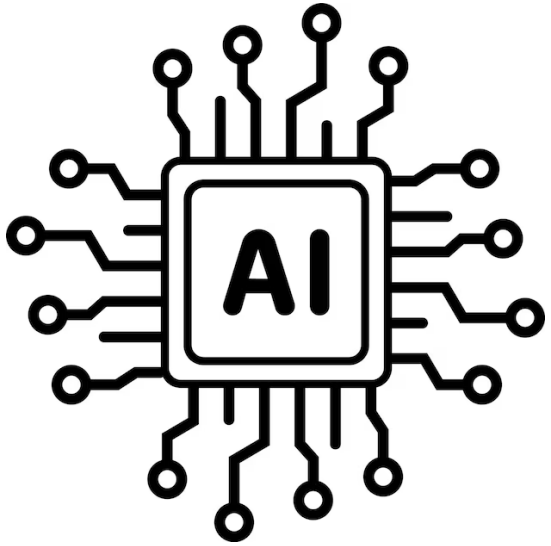


Deduplicação

Após a filtragem, é realizada uma etapa de deduplicação utilizando o método MinHash.

	Depois dos filtros anteriores	Depois da Deduplicação
FAWP 57	6 678 792	544 601

Por exemplo, na coleção FAWP57, a aplicação da deduplicação reduz o número de entradas em 92%.

AMALIA Dados Gerais

- No modelo atual estão a ser usados **5.8 B (milhares de milhões) de tokens** de dados do **Arquivo.pt**.
- O tamanho do modelo é **9 B**.
- O modelo está a correr num cluster com 4 GPUs A100.