Dataset para avaliação de sumarização de longos documentos em português

Leandro Carísio Fernandes
Guilherme Zeferino Rodrigues Dobins

Descrição do Projeto

- Criar um dataset que permita o teste de sistemas de sumarização de transcrições de audiências públicas e/ou debates da Câmara dos Deputados (CD).

- O dataset terá:
  - transcrição de uma audiência pública (cada transcrição pode ter dezenas de páginas => há casos de + de 200 páginas)
  - lista de metadados (quem disse o quê) mais relevantes da audiência

- Tarefa: dada uma transcrição, extrair quem disse o quê de relevante.

Metodologia

Taquigrafia da CD disponibiliza a transcrição das audiências.

A Agência de Notícias da CD publica matérias de várias audiências. São textos curtos e que já possuem as informações mais importantes (do ponto de vista da Agência de Notícias) do que ocorreu nas audiências.

=> Para construir o dataset, podemos antes mapear (transcrição, notícia) e, em vez de extrair os metadados das transcrições, extrair das notícias.

Mapeamento

Seleção manual das matérias/transcrições de audiências públicas.

	A	B	C
	ID	MATERIA	TRANSCRICAO
1	1	https://www.camara.leg.br/noticias/1052740-jornal	https://escriba.camara.leg.br/escriba-servicosweb/html/72444
2	2	https://www.camara.leg.br/noticias/992544-ministr	https://escriba.camara.leg.br/escriba-servicosweb/html/69425
3	3	https://www.camara.leg.br/noticias/1053182-minist	https://escriba.camara.leg.br/escriba-servicosweb/html/72463
4	4	https://www.camara.leg.br/noticias/1015002-anatel	https://escriba.camara.leg.br/escriba-servicosweb/html/70622
5	5	https://www.camara.leg.br/noticias/1050239-deputa	https://escriba.camara.leg.br/escriba-servicosweb/html/72158
6	6	https://www.camara.leg.br/noticias/995646-socio-c	https://escriba.camara.leg.br/escriba-servicosweb/html/69710
7	7	https://www.camara.leg.br/noticias/1053231-deputa	https://escriba.camara.leg.br/escriba-servicosweb/html/72431
8	8	https://www.camara.leg.br/noticias/993729-banco-	https://escriba.camara.leg.br/escriba-servicosweb/html/69586
9	9	https://www.camara.leg.br/noticias/1059408-debate	https://escriba.camara.leg.br/escriba-servicosweb/html/72151
10	10	https://www.camara.leg.br/noticias/1057055-goverr	https://escriba.camara.leg.br/escriba-servicosweb/html/72537
11	11	https://www.camara.leg.br/noticias/1026668-ESPEC	https://escriba.camara.leg.br/escriba-servicosweb/html/71477
12	12	https://www.camara.leg.br/noticias/1027933-MULHI	https://escriba.camara.leg.br/escriba-servicosweb/html/71558
13	13	https://www.camara.leg.br/noticias/1028000-AUTOI	https://escriba.camara.leg.br/escriba-servicosweb/html/71535
14	14	https://www.camara.leg.br/noticias/1028499-VITIMA	https://escriba.camara.leg.br/escriba-servicosweb/html/71543
15	15	https://www.camara.leg.br/noticias/1015019-deputa	https://escriba.camara.leg.br/escriba-servicosweb/html/70949
16	16	https://www.camara.leg.br/noticias/1023136-minist	https://escriba.camara.leg.br/escriba-servicosweb/html/71423
17	17	https://www.camara.leg.br/noticias/1023563-em-se	https://escriba.camara.leg.br/escriba-servicosweb/html/70835
18	18	https://www.camara.leg.br/noticias/1024999-presid	https://escriba.camara.leg.br/escriba-servicosweb/html/71358
19	19	https://www.camara.leg.br/noticias/1026435-CAMAI	https://escriba.camara.leg.br/escriba-servicosweb/html/71361

Metodologia

Nessa caso, a **construção do dataset** para sumarização de longos documentos passa pela sumarização de pequenos documentos (notícias). GPT-4o + correção manual:


```
{
  "assunto": "Substituição do saque-aniversário do FGTS por empréstimo consignado com juros mais baixos",
  "envolvidos": [
    {
      "nome": "Carlos Augusto Simões",
      "cargo": "Secretário de Proteção ao Trabalhador do Ministério do Trabalho e Emprego",
      "opinioes": [
        "O novo consignado em estudo poderá ter taxas semelhantes às oferecidas pelas operações de antecipação de saques.",
        "\"Nós vamos apresentar uma taxa que seja a mais próxima possível da antecipação. Com a vantagem de o trabalhador poder
        contar com o sonho de ter uma moradia no futuro. Hoje, esse sonho está em risco.\"",
        "66,3% dos trabalhadores que têm contas ativas no FGTS possuem saldo de até quatro salários mínimos, ou R$ 5.648,00.
        Quase metade deles está no saque-aniversário."
      ]
    },
    {
      "nome": "Capitão Alberto Neto",
      "cargo": "Deputado (PL-AM)",
      "opinioes": [
        "Vai requerer do governo informações sobre os cálculos atuariais do FGTS que evidenciem a insustentabilidade do fundo
        com a manutenção do saque-aniversário.",
        "O saque-aniversário tem sido importante para atender diversas necessidades do trabalhador, mas pode passar por
        aperfeiçoamentos."
      ]
    },
```

Métricas

- Depois do dataset construído, a ideia é testar a sumarização de longos documentos e comparar com os metadados extraídos das notícias.
  - Para diminuir os custos dos testes, vou usar o ChatGPT e o Gemini Pro em vez de usar o GPT via API.

- Métrica: recall. Verificar a % dos dados que o modelo informou que estão nos metadados das matérias. Como a saída do modelo é textual, a comparação pode ser feita perguntando à um LLM.

Resultados

Resultados esperados:

- (1) Criação do dataset
- (2) Prompts para extração dos metadados das transcrições com ChatGPT e Gemini
- (3) Comparação dos resultados obtidos em (2) com os metadados contidos no dataset

Cronograma

Lista de atividades a serem feitas antes de cada entrega:

- 06 de junho - entrega I - Plano de Trabalho e Dataset
- 13 de junho - entrega II - Testes com prompts para sumarizar as transcrições e fazer a avaliação
- 20 de junho - entrega III - Testes com prompts para sumarizar as transcrições e fazer a avaliação
- 27 de junho - entrega final - Dataset + avaliação + texto