# 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

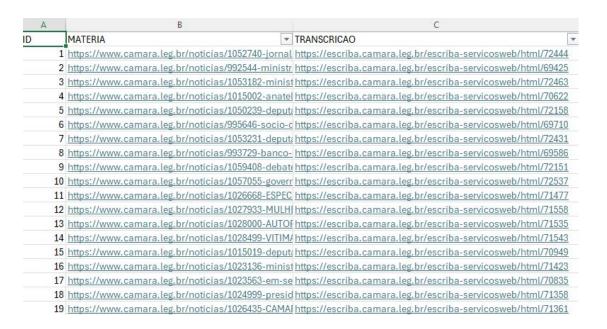
Taquigrafia da CD disponibiliza a transcrição das audiências.

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.



## 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 sague-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