

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.

Extração das opiniões a partir das transcrições

- Foram realizados testes com variações de um prompt, e então escolhemos um prompt específico que foi usado para realizar as extrações finais.
- Para a extração, criamos um GPT customizado utilizando o ChatGPT, para o qual passamos os textos das transcrições, e recuperamos um json contendo as opiniões extraídas.
- No momento, já realizamos a extração das 206 transcrições usando o ChatGPT.

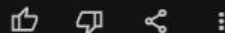
Problemas com o Gemini

- O plano original era realizar as extrações com o ChatGPT e Gemini Pro, no entanto tivemos problemas para o Gemini realizar a tarefa, o que não ocorreu com o GPT. Por tal motivo, ainda não extraímos com o Gemini.
- Principais problemas detectados: Má formatação no json, interrupção na geração das respostas, e por vezes o Gemini se recusou a realizar a tarefa.

Problemas com o Gemini

a participau00e7\u00e3o dos atingidos e fortalecer o processo de reparau00e7\u00e3o."}], [{"nome": "V\u00edtor Eduardo de Almeida Saback", "cargo": "Secret\u00e1rio de Geologia, Mineral\u00e7\u00e3o e Transformau00e7\u00e3o Mineral do Minist\u00e9rio de Minas e Energia (MME)", "opin\u00edoes": ["Refor\u00e7a o compromisso do Minist\u00e9rio de Minas e Energia com a seguran\u00e7a, as comunidades e a sustentabilidade na minera\u00e7\u00e3o.", "Destaca a import\u00e2ncia do Plano Nacional de Atingidos por Barragens (PNAB) e o di\u00e1logo com o MAB.", "Menciona a necessidade de reestruturar a Ag\u00eancia Nacional de Mineral\u00e7\u00e3o (ANM) para garantir a seguran\u00e7a das barragens.", "Reafirma o compromisso do Minist\u00e9rio em buscar um acordo justo e r\u00e1pido para Mariana, e destaca os esfor\u00e7os do Ministro Alexandre Silveira nesse sentido."]}, [{"nome": "Isabella Karen Ara\u00fajo Sim\u00f5es", "cargo": "Defensora P\u00fablica da Uni\u00e3o", "opin\u00edoes": ["Aborda a quest\u00e3o da cl\u00e1usula de confidencialidade e defende a mitiga\u00e7\u00e3o para maior acesso dos atingidos \u00e0s negocia\u00e7\u00f5es.", "Questiona a quem

Fontes ▼



Comparação com o GPT

```
{
  "assunto": "Desafios e o potencial do turismo religioso no Estado de Santa Catarina",
  "envolvidos": [
    {
      "nome": "Rodrigo Coelho",
      "cargo": "Deputado e Presidente da Comissão de Turismo",
      "opiniones": [
        "O turismo religioso movimenta cerca de 15 bilhões de reais por ano no Brasil.",
        "Santa Catarina foi o quarto Estado que mais faturou em viagens ligadas ao turismo",
        "O turismo religioso envolve destinos como Nova Trento, Brusque, Joaçaba, Camboriú",
        "Destacou a importância de investimentos em infraestrutura turística, como pavimen
      ]
    },
    {
      "nome": "Fabio Pinheiro",
      "cargo": "Secretário Nacional de Desenvolvimento e Competitividade do Turismo",
      "opiniones": [
        "O turismo religioso é um forte indutor de renda para os municípios.",
        "É importante mostrar ao turista as intervenções que acontecem no interior dos est
        "Há necessidade de promoção do turismo religioso através de campanhas publicitária
      ]
    }
  ]
}
```



Pipeline de avaliação

- Para avaliar as extrações, criamos um pipeline de avaliação, que compara as opiniões extraídas com as opiniões presentes no dataset ouro. O pipeline é composto pelos seguintes passos:
 1. Mapeamento dos envolvidos: As pessoas envolvidas na resposta ouro e na resposta extraída são listadas e enviadas para um LLM, que mapeia os que aparecem nas duas respostas.
 2. Mapeamento de opiniões: Após o mapeamento, iteramos por cada pessoa, e listamos as opiniões presentes na resposta ouro e as opiniões extraídas, e novamente enviamos para um LLM associar as opiniões similares.
 3. Cálculo de métricas: Ao fim, contamos quantas opiniões foram corretamente identificadas e calculamos a precisão e recall


Mapeamento dos envolvidos

Envolvidos da resposta ouro:

Nome: Daniel Cara. Cargo: Professor de Educação da Univ
Nome: Catarina de Almeida. Cargo: Professora da Faculda
Nome: Pastor Henrique Vieira. Cargo: Deputado (Psol-RJ)
Nome: Josevanda Franco. Cargo: Presidente da União Naci
Nome: Ana Paula Lima. Cargo: Deputada (PT-SC)

Envolvidos da resposta extraída:

Nome: Ana Paula Lima. Cargo: Deputada, Presidente da Se
Nome: Henrique Vieira. Cargo: Deputado, coautor do requ
Nome: Yann Evanovick. Cargo: Coordenador do Grupo de Tr
Nome: Cybele Amado de Oliveira. Cargo: Diretora de Form
Nome: Josevanda Franco. Cargo: Presidente da União Naci
Nome: Isabel Seixas. Cargo: Diretora do Sistema Único d
Nome: Romano Costa. Cargo: Representante da Secretaria
Nome: Daniel Cara. Cargo: Professor da Faculdade de Edu
Nome: Bruno Renato Nascimento Teixeira. Cargo: Ouvidor
Nome: Ariel de Castro Alves. Cargo: Presidente do Conse



```
{  
  'Daniel Cara': 'Daniel Cara',  
  'Catarina de Almeida': 'Não identificado',  
  'Pastor Henrique Vieira': 'Henrique Vieira',  
  'Josevanda Franco': 'Josevanda Franco',  
  'Ana Paula Lima': 'Ana Paula Lima'  
}
```


Mapeamento de opiniões

Opiniões do Josevanda Franco:

Resposta ouro

1. Soluções como a militarização não funcionam porque a violência escolar reflete o que acontece na sociedade.

Resposta extraída

1. Enfatizou que a violência nas escolas reflete a violência na sociedade.
2. Defendeu a necessidade de políticas públicas articuladas e intersetoriais.
3. Alertou contra a transformação das escolas em ambientes militarizados e defendeu o foco em soluções pedagógicas e socioemocionais.

Mapeamento: {

"1": "1, 3"

}

Cálculo de métricas

- Nós contamos a quantidade de opiniões da resposta ouro que foram mapeadas (map_gold), e também a quantidade de opiniões da resposta extraída que foram mapeadas (map_ext)
- Também anotamos o total de opiniões da resposta ouro (total_gold) e da resposta extraída (total_ext)
- O recall é calculado como $\text{map_gold}/\text{total_gold}$, e é a métrica principal analisada
- A precisão é calculada como $\text{map_ext}/\text{total_ext}$, e no momento é usada apenas para acompanharmos a quantidade de opiniões geradas.

Próximos passos

- Extrair as opiniões com o Gemini Pro
- Gerar as métricas para os todos os casos
- Escrever o relatório