GUIA COMPLETO: O QUE CADA CÉLULA FAZ

Notebook: Avaliacao_QA_Shard007_FINAL_v2.ipynb

CÉLULA 0 [MARKDOWN] — Introdução

O que é: Texto de apresentação do notebook.

O que aparece: Título "Avaliação de Modelos de QA: DistilBERT vs RoBERTa" e lista das métricas que serão avaliadas.

Precisa executar? Não, é só texto.

CÉLULA 1 [CÓDIGO] — Instalação e Importação

O que faz:

1. Instala as bibliotecas necessárias (transformers, torch, pandas, etc.)
2. Importa todas as bibliotecas para uso no notebook

O que aparece: Mensagens de instalação do pip (pode demorar uns 30 segundos).

Bibliotecas importadas:

- `pandas` → manipular dados em tabelas
- `torch` → rodar os modelos de IA
- `transformers` → carregar os modelos do Hugging Face
- `matplotlib` e `seaborn` → criar gráficos
- `tqdm` → mostrar barra de progresso

CÉLULA 2 [CÓDIGO] — Carregamento dos Dados

O que faz:

1. Lê o arquivo `shard_007.csv`
2. Mostra quais colunas existem no arquivo
3. Renomeia colunas se necessário (padroniza para 'query' e 'text')

O que aparece:

```
Colunas encontradas: ['_id', 'title', 'text', 'query']
Dados carregados: 1006 linhas.
```

Importante: O arquivo CSV precisa estar na mesma pasta do notebook!

CÉLULA 3 [CÓDIGO] — Carregamento dos Modelos

O que faz:

1. Baixa o modelo DistilBERT do Hugging Face
2. Baixa o modelo RoBERTa do Hugging Face
3. Cria os "pipelines" de Question Answering

O que aparece:

```
Carregando Modelo 1 (DistilBERT)...
Carregando Modelo 2 (RoBERTa)...
Modelos carregados com sucesso!
```

Demora: Pode levar 1-2 minutos na primeira vez (baixa ~500MB).

CÉLULA 4 [CÓDIGO] — Funções Auxiliares

O que faz: Define duas funções que serão usadas depois:

1. **`get_answer(pipe, question, context)`**
   - Recebe uma pergunta e um texto
   - Retorna a resposta do modelo e o score de confiança
   - Trunca textos muito longos (>2000 caracteres) pra não dar erro
2. **`calculate_overlap(answer, context)`**
   - Calcula quantas palavras da resposta aparecem no texto original
   - Retorna um número de 0 a 1 (0% a 100%)

O que aparece:

```
Funções auxiliares 'get_answer' e 'calculate_overlap' definidas com sucesso!
```

CÉLULA 5 [CÓDIGO] — Processamento Principal (INFERÊNCIA)

O que faz:

1. Passa cada uma das 1.000 perguntas pelos dois modelos
2. Guarda: pergunta, texto, resposta M1, score M1, overlap M1, resposta M2, score M2, overlap M2
3. Cria um DataFrame com todos os resultados

O que aparece:

- Barra de progresso: `100%|██████████| 1006/1006`
- Tabela com as primeiras 5 linhas dos resultados

Demora: 5-10 minutos (processa 1.006 exemplos × 2 modelos).

⚠️ **CÉLULA MAIS DEMORADA DO NOTEBOOK!**

CÉLULA 6 [CÓDIGO] — Análise Quantitativa Inicial

O que faz:

1. Calcula o tamanho de cada resposta (em caracteres)
2. Calcula médias de score, overlap e tamanho
3. Monta uma tabela comparativa

O que aparece:

📊 Resultados Médios

Métrica	Modelo 1 (DistilBERT)	Modelo 2 (RoBERTa)
Score Médio (Confiança)	0.4131	0.3310
Overlap Médio (Fidelidade)	83.71%	78.41%
Tamanho Médio da Resposta	26.7	28.6

CÉLULA 7 [CÓDIGO] — Exportação para Excel

O que faz: Salva todos os resultados em um arquivo Excel.

O que aparece:

```
Arquivo 'analise_qualitativa.xlsx' salvo com sucesso.
```

Arquivo criado: `analise_qualitativa.xlsx` (na mesma pasta do notebook)

CÉLULA 8 [CÓDIGO] — Gráficos Visuais

O que faz: Cria 4 gráficos:

1. **Histograma de Scores** — Distribuição da confiança dos modelos
2. **Scatter Plot** — Correlação entre scores de M1 e M2
3. **Boxplot de Overlap** — Comparação da sobreposição
4. **Barras de Tamanho** — Comparação do tamanho das respostas

O que aparece: Uma figura grande com 4 gráficos empilhados, cada um com explicação embaixo.

CÉLULA 9 [CÓDIGO] — Critérios A, B e C (Quantitativo)

O que faz:

1. **Critério A:** Calcula tamanho médio das perguntas (em palavras)
2. **Critério B:** Pega o score médio já calculado
3. **Critério C:** Pega o overlap médio já calculado

O que aparece:

📊 ANÁLISE QUANTITATIVA DETALHADA

Métrica	Modelo 1	Modelo 2
Tamanho Médio Perguntas	4.80	4.80
Score Médio (Confiança)	0.4131	0.3310
Overlap Médio (Fidelidade)	83.71%	78.41%

CÉLULA 10 [MARKDOWN] — Análise Score vs Qualidade (Texto)

O que é: Texto explicativo sobre o Critério B.

O que aparece: Análise detalhada respondendo:

- "O score médio reflete a qualidade das respostas?"
- Exemplos de quando SIM (score alto = correto)
- Exemplos de quando NÃO (score alto ≠ correto)
- Conclusão: é parcialmente confiável

Valores mostrados: Score M1 = 0.4131, Score M2 = 0.3310

CÉLULA 11 [CÓDIGO] — Estatísticas de Score

O que faz:

1. Classifica cada exemplo por faixa de score (Alto/Médio/Baixo/Muito Baixo)
2. Conta quantos exemplos em cada faixa
3. Calcula correlação entre score e overlap

O que aparece:

```
📈 DISTRIBUIÇÃO POR FAIXA DE SCORE - MODELO 1 (DistilBERT):
  Alto (≥0.90): 156 exemplos (15.5%)
  Médio (0.50-0.90): 298 exemplos (29.6%)
  Baixo (0.10-0.50): 312 exemplos (31.0%)
  Muito Baixo (<0.10): 240 exemplos (23.9%)

📊 CORRELAÇÃO SCORE × OVERLAP:
  Modelo 1 (DistilBERT): r = 0.XXXX
```

CÉLULA 12 [MARKDOWN] — Análise do Overlap (Texto)

O que é: Texto explicativo sobre o Critério C.

O que aparece: Análise detalhada do overlap:

- O que significa overlap alto (>90%) → resposta no texto
- O que significa overlap baixo (<50%) → possível alucinação
- Tabela comparando M1 vs M2
- Conclusão: DistilBERT tem mais fidelidade ao contexto

Valores mostrados: Overlap M1 = 83.71%, Overlap M2 = 78.41%

CÉLULA 13 [CÓDIGO] — Estatísticas de Overlap

O que faz:

1. Calcula média, mediana, mínimo e máximo do overlap
2. Classifica por faixa (Alto/Médio/Baixo)
3. Conta casos de possível alucinação (overlap < 50%)

O que aparece:

```
📈 ESTATÍSTICAS DE OVERLAP:
  MODELO 1 (DistilBERT):
    Média:   83.71%
    Mediana: 100.00%

📊 DISTRIBUIÇÃO POR FAIXA:
  Alto (≥90%): 626 (62.6%)
  Médio (50-90%): 297 (29.7%)
  Baixo (<50%): 77 (7.7%)

🚨 CASOS DE POSSÍVEL ALUCINAÇÃO:
  Modelo 1: 77 casos (7.7%)
  Modelo 2: 122 casos (12.2%)
```

CÉLULA 14 [CÓDIGO] — Análise Equilibrada M1 vs M2

O que faz:

1. Seleciona os 25 exemplos (10 top + 10 bottom + 5 divergentes)
2. Analisa automaticamente cada um (contexto, risco de alucinação)
3. Cria tabela comparativa

O que aparece:

- Resumo de quantos têm alto/baixo risco de alucinação
- Tabela com os 25 exemplos e análise de ambos os modelos

CÉLULA 15 [MARKDOWN] — Avaliação Qualitativa (Texto)

O que é: Texto explicativo sobre o Critério D.

O que aparece:

- Tabela dos Top 10 (maior score) com observações
- Tabela dos Bottom 10 (menor score) com observações
- Análise de padrões encontrados

CÉLULA 16 [CÓDIGO] — Preparar 25 Exemplos

O que faz:

1. Seleciona os 10 com maior score
2. Seleciona os 10 com menor score
3. Seleciona 5 onde M1 ≠ M2 (respostas diferentes)
4. Combina tudo em um DataFrame `exemplos_25`

O que aparece:

```
✅ 25 exemplos selecionados!
  - Top 10: 10
  - Bottom 10: 10
  - Divergentes: 5
```

CÉLULA 17 [CÓDIGO] — Criar Colunas de Análise

O que faz: Adiciona colunas vazias para preencher manualmente:

- `contexto_m1` / `contexto_m2` → Sim/Parcial/Não
- `correta_m1` / `correta_m2` → Sim/Parcial/Não
- `alucinacao_m1` / `alucinacao_m2` → Sim/Não
- `melhor_modelo` → M1/M2/Empate/Nenhum
- `observacoes` → texto livre

O que aparece:

```
✅ Colunas de análise criadas!
```

CÉLULA 18 [CÓDIGO] — Análise Manual: TOP 10

O que faz: Preenche a análise manual dos 10 exemplos com maior score.

Exemplos analisados:

1. who started labatt beer → John Kinder Labatt ✅
2. what nationality is lleyton hewitt → Australian ✅
3. who is the lead guitarist of metallica → Kirk Hammett ✅
4. ... (até o 10º)

O que aparece: Nada visível, só preenche os dados.

CÉLULA 19 [CÓDIGO] — Análise Manual: BOTTOM 10

O que faz: Preenche a análise manual dos 10 exemplos com menor score.

Exemplos analisados:

1. what books did tolkien write → Parcial (múltiplas respostas)
2. what is a latte liberal → Definição abstrata, difícil
3. kyrgyz population → Sem resposta no contexto
4. ... (até o 10º)

O que aparece: Nada visível, só preenche os dados.

CÉLULA 20 [CÓDIGO] — Análise Manual: DIVERGENTES

O que faz: Preenche a análise manual dos 5 exemplos onde M1 ≠ M2.

Exemplos analisados:

1. what is the jazz guitar → M1 melhor (definição correta)
2. what river is javari located in → M1 melhor
3. youngest female chess player → Nenhum (M2 disse Bobby Fischer, que é HOMEM!)
4. who makes jupiter computer → M2 melhor
5. what are jewish holidays → M2 melhor

O que aparece: Nada visível, só preenche os dados.

CÉLULA 21 [CÓDIGO] — Estatísticas e Exportação

O que faz:

1. Conta quantas vezes cada modelo foi melhor
2. Salva tudo em Excel

O que aparece:

📊 RESUMO DA ANÁLISE QUALITATIVA:
Modelo 1 (M1) melhor: X casos
Modelo 2 (M2) melhor: Y casos
Empate: Z casos
Nenhum acertou: W casos

✅ Análise exportada para: Analise_Qualitativa_25_Exemplos_COMPLETO.xlsx

---

CÉLULA 22 [CÓDIGO] — Formatação Profissional do Excel

O que faz:

1. Lê o Excel gerado
2. Cria novo Excel com formatação bonita:
    - Cores por categoria (verde/laranja/cinza)
    - Headers congelados
    - Bordas e alinhamento
    - Múltiplas abas organizadas

O que aparece:

✅ PLANILHA REFORMATADA COM SUCESSO!
📁 Arquivo: Analise_Qualitativa_25_Exemplos_FORMATADO.xlsx

CÉLULA 23 [CÓDIGO] — Visualização Colorida das Tabelas

O que faz:

1. Lê o arquivo `Analises.xlsx`
2. Aplica cores nas tabelas:
    - Top 10 → Verde (`#E8F5E9`)
    - Bottom 10 → Laranja (`#FFF3E0`)
    - Divergentes → Cinza (`#ECEFF1`)
3. Exibe as tabelas formatadas no notebook

O que aparece: 3 tabelas coloridas e bonitas com os 25 exemplos analisados.

⚠️ **Precisa do arquivo `Analises.xlsx` na mesma pasta!**

CÉLULA 24 [MARKDOWN] — Conclusão Final

O que é: Texto com a resposta à pergunta obrigatória.

O que aparece:

- Tabela resumo das métricas
- Escolha: **DistilBERT (Modelo 1)**
- Justificativa baseada nos 25 exemplos:
    1. Top 10: M1 acertou tudo
    2. Bottom 10: M1 sinalizou incerteza corretamente
    3. Divergentes: Erros do M2 foram mais graves
    4. Menor risco de alucinação
- Arquitetura recomendada para produção

📋 RESUMO: ORDEM DE EXECUÇÃO

Célula	Tempo	O que faz
0	-	Introdução (texto)
1	30s	Instala bibliotecas
2	2s	Carrega o CSV
3	1-2min	Baixa os modelos
4	1s	Define funções
5	**5-10min**	**PROCESSA 1.000 EXEMPLOS**
6	1s	Calcula médias
7	1s	Salva Excel
8	3s	Gera gráficos
9	1s	Métricas A, B, C
10	-	Análise score (texto)
11	1s	Estatísticas de score
12	-	Análise overlap (texto)

Célula	Tempo	O que faz
13	1s	Estatísticas de overlap
14	2s	Análise equilibrada
15	-	Critério D (texto)
16	1s	Seleciona 25 exemplos
17	1s	Cria colunas
18	1s	Análise Top 10
19	1s	Análise Bottom 10
20	1s	Análise Divergentes
21	1s	Salva Excel
22	2s	Formata Excel
23	2s	Tabelas coloridas
24	-	Conclusão (texto)

Tempo total estimado: 10-15 minutos (a maior parte é a célula 5)

⚠️ ARQUIVOS NECESSÁRIOS

Para o notebook funcionar, você precisa ter na mesma pasta:

1. `shard_007.csv` — Dataset com as 1.000+ perguntas
2. `Analises.xlsx` — Planilha com análises manuais (para célula 23)

🔧 SE DER ERRO

Erro	Solução
"FileNotFoundError: shard_007.csv"	Coloque o CSV na mesma pasta
"FileNotFoundError: Analises.xlsx"	Coloque o Excel na mesma pasta
"CUDA out of memory"	Reinicie o kernel e rode de novo
"Connection error"	Verifique sua internet (precisa baixar modelos)

Erro	Solução
Demora muito na célula 5	Normal, espere 5-10 minutos