01- CONTEXTO

Como um especialista em:

* Processamento de Linguagem Natural (NLP)
* Clusterização semântica de textos curtos
* Planejamento estratégico institucional

Sua tarefa é analisar um conjunto de contribuições textuais curtas oriundas de um processo participativo de diagnóstico interno contendo forças e suas descrições, fraquezas e suas descrições, todas classificadas segundo os temas e subtemas estratégicos do PDI.

02- FORMATO DOS DADOS DE ENTRADA

Os dados possuem três colunas:

1. ID resposta (que identifica unicamente as respostas enviadas pelos participantes da consulta pública. Cada resposta é composta por uma ou mais contribuições)
2. ID contribuição (que identifica unicamente cada contribuição)
3. Unidade de vínculo;
4. Tema;
5. Subtema;
6. Tipo: “Força” ou “Fraqueza”
7. Texto: descrição + justificativa

03- OBJETIVO

Aplicar um pipeline completo de análise semântica inspirado em:

* SBERT (embeddings)
* UMAP (redução de dimensionalidade)
* HDBSCAN ou BERTopic (clusterização)
* Interpretação por LLM

E produzir uma saída estruturada adequada para uso em diagnóstico estratégico institucional.

Você deve transformar a base textual respectivas em:

- clusters semânticos consistentes;
- macrotemas estratégicos;
- consolidação SWOT;
- inteligência executiva acionável.

A arquitetura escolhida prioriza:

- qualidade semântica;

- interpretabilidade;
- estabilidade dos clusters;
- baixa intervenção manual;
- escalabilidade.

04- ETAPA 1 — PRÉ-PROCESSAMENTO (LEVE)

Execute um pré-processamento mínimo:

* Remover duplicatas exatas
* Ignorar entradas vazias ou sem sentido
* Corrigir apenas erros graves que prejudiquem entendimento
* Preservar linguagem natural (NÃO remover stopwords, NÃO aplicar stemming)

05- ETAPA 2 — REPRESENTAÇÃO SEMÂNTICA (SIMULAÇÃO DE SBERT)

* Considere que cada texto é transformado em um vetor semântico denso
* Similaridade entre textos deve ser avaliada por similaridade de significado (não por palavras iguais)
* Extraia entidades e conceitos da descrição de oportunidade ou de ameaça, tais como: tecnologias; mercados; regulações; países; concorrentes; stakeholders.

Utilize como referência conceitual:

* Similaridade por contexto
* Equivalência semântica
* Proximidade de tema

06- ETAPA 3 — REDUÇÃO DE DIMENSIONALIDADE (OPCIONAL)

* Considere que o espaço vetorial pode ser reduzido (ex: UMAP)
* Objetivo:

  * reduzir ruído
  * facilitar agrupamento
  * Preservar estrutura semântica global

07- ETAPA 4 — CLUSTERIZAÇÃO (SIMULAÇÃO DE HDBSCAN / BERTopic)

Agrupe os textos com base em densidade semântica e proximidade temática

* Regras:

  * NÃO definir previamente número de clusters

* Permitir:
  * clusters grandes (temas estruturantes)
  * clusters pequenos (temas específicos)

* Identificar e tratar:
  * duplicidades
  * textos muito genéricos
  * ruído (itens sem relação clara com outros)

* Critério principal:

* similaridade semântica global (tema)

* Critério de distância:
  * considerar implicitamente distância por similaridade de cosseno

08- ETAPA 5 — PÓS-PROCESSAMENTO DOS CLUSTERS

Para cada cluster:

* Verificar coerência interna
* Separar clusters muito heterogêneos
* Fundir clusters redundantes
* Identificar clusters:
  * muito pequenos (avaliar relevância)
  * muito genéricos (refinar ou marcar)

09- ETAPA 6 — INTERPRETAÇÃO E CONSOLIDAÇÃO

Para cada cluster, executar:

6.1 Título do cluster

* Criar um rótulo curto e representativo

6.2 Macro-evidência (OBRIGATÓRIO)

Produzir uma síntese que:

* represente o núcleo comum das contribuições
* elimine redundâncias
* seja analítica (não apenas descritiva)
* use linguagem institucional adequada ao PDI

6.3 Tipo predominante

* Definir se o cluster é:
  * Força
  * Fraqueza
* Basear-se na maioria ou no conteúdo predominante

10- ETAPA 8 — CONTROLE DE QUALIDADE

* Garantir que:
  * clusters sejam distintos entre si
  * não haja sobreposição temática significativa

* Evitar:
  * clusters genéricos demais
  * clusters redundantes

* Se necessário:
  * criar categoria "Outros / Baixa densidade temática"

11- ETAPA 9 — FORMATO DE SAÍDA (OBRIGATÓRIO)

Apresentar os resultados em tabela com as colunas:

1. ID do Cluster, no formato dos exemplos: acervo-forças-01, ead-fraquezas-05, infra-forças-24
2. Título do Cluster
3. Macro-evidência (texto consolidado)
4. Número de contribuições
5. Lista completa das contribuições originais (dados brutos que deram origem ao cluster) de acordo com o exemplo abaixo:
[acervo-2 | C00001 | Campus Bom Sucesso | Acervo Acadêmico Digital | Segurança da informação | Força] A instituição preza ela segurança da informação
[ti-32 | C02163 | Reitoria | Tecnologia da Informação | Segurança da informação e privacidade de dados | Força] POSIN e PPDP publicados

12- DIRETRIZES IMPORTANTES

* Priorizar qualidade semântica sobre velocidade
* Processar TODO o conjunto de dados
* Trabalhar com nível analítico (não superficial)
* Evitar decisões baseadas apenas em palavras-chave
* Manter rastreabilidade com os dados originais
* Limite obrigatório: no máximo 24 clusters
* Criar um cluster com título “Contribuições Inadequadas” no qual deverão ser agrupadas todas

as contribuições que não representem forças ou fraquezas, ou seja, que não representem fatores do ambiente interno.

13- RESTRIÇÕES

* NÃO agrupar por palavras iguais
* NÃO simplificar excessivamente
* NÃO perder nuances relevantes
* NÃO ignorar contribuições minoritárias relevantes

14- RESULTADO ESPERADO

Uma estrutura que permita uso direto em:

* Matriz SWOT consolidada
* Caderno de evidências
* Formulação estratégica (TOWS)

15- INSTRUÇÃO FINAL

Execute todas as etapas com profundidade analítica e consistência metodológica, simulando o comportamento de um pipeline SBERT + HDBSCAN + BERTopic, mesmo que de forma conceitual e gere um arquivo de saída em formato CSV com o formato descrito na ETAPA 9.