Nome:
_____Matrícula:_____
Data:__/__/____

PROJETO PRÁTICO DE APRENDIZADO DE MÁQUINA
PLN – PROCESSAMENTO DE LINGUAGEM NATURAL
APRENDIZADO NÃO-SUPERVISIONADO - AGRUPAMENTO

Exercício: Criando agrupamento de vídeos do YouTube por assuntos relacionados

Objetivo

Neste exercício, você irá:

1. Coletar **50 transcrições** de vídeos do YouTube sobre temas aleatórios (exemplo: tecnologia, política, economia, ciência, esportes, entretenimento, etc.).

2. Processar essas transcrições e representar seus conteúdos de diferentes formas (**Bag of Words OU Embeddings**).

3. Aplicar **K-Means** para agrupar os vídeos em **10 grupos** e comparar os resultados.

4. **Analisar e interpretar os grupos**, verificando se fazem sentido e qual técnica funcionou melhor.

Passo 1: Coleta de Dados

Vocês devem escolher aleatoriamente 50 vídeos do YouTube (sugestão: vídeos de 5 minutos).

Como fazer?

1. Escolha os vídeos do YouTube.

2. Use *pytube* para baixar os vídeos e *youtube_transcript_api* para obter as transcrições.

3. Salve as transcrições em arquivos .csv sendo cada transcrição em uma linha. Você terá 50 linhas de transcrições formando um dataset de dados não estruturados.

Passo 2: Representação dos Textos

Agora, precisamos transformar essas transcrições em um formato numérico que possa ser utilizado para agrupamento. Existem **três métodos principais**:

1. **Bag of Words (BoW):** Conta quantas vezes cada palavra aparece em cada transcrição, ignorando a ordem.

2. **Embeddings:** Utiliza modelos pré-treinados para converter frases em vetores que capturam o significado das palavras.

Como fazer?

Vocês podem usar as seguintes bibliotecas:

- *CountVectorizer* para *BoW*

- *SentenceTransformer* para *embeddings*

Passo 3: Aplicação do Algoritmo de Agrupamento

O agrupamento será feito com o algoritmo **K-Means**, que tenta separar os dados em grupos baseados em suas semelhanças.

Como fazer?

1. Defina o número de grupos como **10**.

2. Aplique o algoritmo sobre os dados transformados para gerar o modelo.

Passo 4: Análise dos Resultados

Depois de realizar o agrupamento, vocês devem **analisar os grupos** e verificar se fazem sentido.

Dicas para análise:

- Analise os 10 grupos gerados. Em geral, teremos 10 grupos de 5 vídeos cada.

- Listar os títulos dos vídeos em cada grupo e verificar padrões.

- Ver se vídeos com temas parecidos ficaram no mesmo grupo.

- Comparar os resultados de BoW e Embeddings: qual técnica criou os agrupamentos mais coerentes?

Entrega

Vocês devem entregar um relatório contendo:

1. Como foi feita a coleta dos vídeos e as transcrições.
2. **Comparação das três técnicas (BoW, Embeddings)** e os resultados obtidos.
3. **Lista dos 10 grupos** e os títulos dos vídeos em cada um. (CSV)
4. **É possível rotular os 10 grupos? Rotule os grupos na lista que será entregue**. (exemplo: tecnologia, ciência, esportes, entretenimento, etc.).
5. **Conclusões sobre qual método foi melhor e por quê**.
6. **Código-fonte**.