

**CENTRO UNIVERSITÁRIO SALESIANO DE SÃO PAULO**

**UNIDADE LORENA**

**CURSO DE ENGENHARIA DE COMPUTAÇÃO**

**INTELIGÊNCIA ARTIFICIAL**

**PROFESSOR ME. WALMIR DUQUE**

Breno Ryan de Andrade Fernandes

Daniel Marton Barbosa

Lucas Gabriel dos Santos Moraes

Mariana Gonçalves de Freitas Ribeiro

João Vitor Ferreira Azevedo Pereira

**"Agrupamento (Clustering) /** **Inteligência Artificial Generativa e PLN"**

# Sumário

# Introdução ................................................................................................ 3

# 1. Capítulo 1 - Agrupamento (Clustering) no Python

1.1 - Business Understanding....................................................................3

1.2 - Data Understading e Data Preparation.............................................................................3

1.3 Modeling e Evaluation ............................................................................................4

**2.Capítulo 2 - Inteligência Artificial Generativa ................................................................. 5**

2.1 – Definição de texto a ser trabalhado, explicação do seu gênero textual eobjetivos de negócio...................................................................5

2.2 PLN - Remoção de Ruídos, Homogeneização Stopwords.................................................................5

2.3 - PLN - Stemming / Lemmatization..........................................................5

2.4 - - Chunk / Embedding.....................................................5

**Conclusão ............................................................................................. 28**

**Referências Bibliográficas...............................................................29 a 31**

# Introdução

Deve assinalar o sentido do trabalho, mas de nenhum modo antecipar o desenvolvimento nem a conclusão. Situar o leitor dentro do espírito do trabalho, expondo o assunto tratado no documento, exaltando a importância do assunto, o estado da arte, enfocar áreas controvertidas ou envolvidas e esclarecer a natureza e extensão da contribuição pretendida com o trabalho.

**1. Capítulo 1 - Agrupamento (Clustering) no Python**

* 1. Business Understanding

Contextualizar significa realizar uma síntese do domínio do negócio ou, mais especificamente, o Setor de Negócio escolhido pelo grupo. Por exemplo, se o tema for um Sistema Supervisório, a que Setor de Gestão esse equipamento atende ou a que Setor de Gestão o grupo pretende focar no trabalho? Uma contextualização típica deve conter aproximadamente 2 páginas (para fins de propósito acadêmico). Esses parágrafos ajudam no entendimento inicial do Sistema de Informação Computadorizado a ser desenvolvido. Nessa contextualização mencionam- se, sinteticamente:

* Quais os objetivos de negócio para o Projeto de Data Science?
* Em face aos objetivos, quais as Tarefas de Data Science que são elegíveis, a priori?
* Quais as funcionalidades (Requisitos Funcionais) e Regras de negócio? Aqui já podem ser identificados os Atores principais envolvidos. Pode ser na forma de uma Tabela!
* Quais os Requisitos Não funcionais?

**1.2 Data Understading e Data Preparation**

Para o dataset definido por você no Trabalho anterior:

• Descrever Variáveis (Features) do problema (nomes, tipos, domínio de valores).

• Dentre as Variáveis (Features), qual é a Variável Objetivo (Target) e quais as Classes? Explicar!

• Alguma preparação (ou transformação) nos dados se faz necessária para melhor atender ao Algoritmo de Machine Learning (Agrupamento)? É **necessário normalizar**? Explicar!

**De acordo com a Tarefa escolhida para Data Science e a técnica associada, podem ser necessárias conversões! Por exemplo, conversão de atributo Numérico para Nomimal (ou vice-versa), conversão de Real para Inteiro, normalização, dentre outros.**

## 1.3 Modeling e Evaluation

## • Explicar a aplicação do conceito de Elbow para a determinação do valor de K (quantidade de grupos).

## • Escolher pelo menos 3 parâmetros principais, para variar, na construção dos seus Modelos, tais como “Tipos de Inicialização (Randômica e Método K-Means++)” e “Medidas de Distância (Euclidiana e Manhattan)”, por exemplo.

## • Construir uma tabela (tal como no Lab), com sua estratégia de aplicação dos parâmetros acima – nessa tabela, também deve conter as técnicas de testes consideradas (em especial, “Percentage Split” e “Cross Validation”).

## • Para cada combinação de parâmetros, devem ser apresentados os resultados, deforma que consiga realizar as análises na Parte “Evaluation” a seguir.

## • Construa um dataset com apenas uma linha para testar seu modelo!

## • Importante: documentar, com prints, o resultado de cada experiência. As Configurações de Clusters (modelos) mais importantes para apresentação deverão ser colhidas (print) também!

## Evaluation

• Comente os resultados obtidos com base no critério “WCSSE”, ou seja, Qualidade do Modelo.

• Comente os resultados obtidos com base no critério Qualidade dos Clusters. Explicar o significado de cada cluster formado!

• Por fim, qual o melhor Modelo obtido? Justifique!

# 2. Capítulo 2 - Inteligência Artificial Generativa

## 2.1 - Definição de texto a ser trabalhado, explicação do seu gênero textual e objetivos de negócio

Para o texto definido como objeto do trabalho, explique seu gênero textual (anúncio publicitário, artigo de opinião, notícia, romance, dentre outros) e justifique sua escolha para este estudo. Defina os objetivos de negócio a serem atingidos.

### 2.2 - PLN - Tratamento do texto: Remoção de Ruídos, Homogeneização e Stopwords

Dado um texto, um tratamento inicial se faz necessário, com base nas características inatas do próprio texto - se for prosaico, será comum conjunções como “né”, por exemplo. Também podem ser necessárias remoção de pontuações, stopwords dentre outros tratamentos.

### 2.3 - PLN - Stemming / Lemmatization

Uma vez o texto tratado, deve ser escolhida e aplicada qual a estratégia de stemminge/ou lemmatization mais adequada ao seu contexto de forma a favorecer a qualidade do retorno das buscas semânticas (similaridade de vetores).

### 2.4 - Chunk / Embedding

Agora, deve ser definida a estratégia de chunk mais adequada, comou semoverlap. Da mesma forma, qual a técnica de embedding mais adequada. Tais escolhas devem favorecer a qualidade do retorno das buscas semânticas (similaridade de vetores).

# Conclusão

Deve ser fundamentada nos resultados e na discussão, contendo deduções lógicas correspondentes aos objetivos propostos. A conclusão constitui-se de uma resposta as hipóteses enunciadas. Portanto, o autor manifesta seu ponto de vista sobre os resultados obtidos. Devem surgir aqui também as dificuldades encontradas e recomendações futuras, para o aprimoramento do curso de Inteligência Artificial.

# 6. Referências Bibliográficas

Devem seguir as Normas ABNT para Monografias