Processamento de Linguagem Natural

Aula 16: Clustering de coleções

*Objetivo: ao fim desta aula, o aluno usará e interpretará técnicas de clustering para dividir coleções em conjuntos mais simples de analisar*

|  |
| --- |
| **Exercício 1** |
| *Objetivo: identificar clusters em uma figura* |
| A figura abaixo mostra o plot de duas dimensões em um dataset.  a) Como os elementos desse dataset parecem se organizar?  b) Marque na figura quais são as “sub-estruturas” que organizam os elementos do dataset. |

|  |
| --- |
| **Exercício 2** |
| *Objetivo: usar a API do scikit-learn para encontrar clusters com K-Means* |
| Ouça atentamente à explicação do professor sobre k-means. Após, no seu notebook, analise o código que exemplifica o uso de k-means.  Após a execução:  a) Qual é o conteúdo da variável x\_?  b) Qual é o conteúdo da variável means\_?  c) Qual é o conteúdo da variável labels\_? |

|  |
| --- |
| **Exercício 3** |
| *Objetivo: interpretar clusters encontrados com K-Means* |
| Em seu notebook, há uma demonstração sobre como calcular o DF de palavras em cada cluster encontrado com k-means.  a) Analise o código fornecido. Se ficarem dúvidas, pergunte.  b) Analisando os resultados das contagens de palavras, identifique:  (i) Os pontos que são observados pela maior parte do público  (ii) Se há perfis diferentes de públicos  (iii) Se há outliers, isto é, clusters com poucas pessoas  c) Modifique o código para que a visualização passe a usar Isomap, e não mais PCA. O que mudou?  d) Modifique o código para que o processo de análise use as projeções em tópicos dos documentos, e não mais os documentos em si. O que mudou nas análises? |

|  |
| --- |
| **Exercício 4** |
| *Objetivo: usar a técnica de clusterização para encontrar segmentos de clientes* |
| Neste exercício, usaremos o dataset “amazon product reviews”. Gostaríamos de entender melhor quem são nossos clientes.  Aplicando a técnica de clusterização que aprendemos hoje, descubra:  a) Quem tipicamente compra os produtos da Amazon? Quem tipicamente os usa?  b) Para que os produtos são usados?  c) Quais são as características preferidas pelos usuários? |