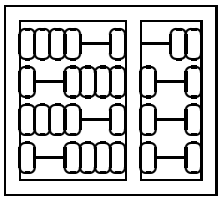

Instituto de Computação
Unicamp


# MO 906 - Introdução à Inteligência Artificial

## 2º Semestre de 2009 - Prof. Siome Goldenstein

## Clusterização - Entrega: Quinta, 15/12/2009, antes do início.

Este trabalho tem como objetivo agrupar arquivos de texto. O conjunto de dados que utilizaremos possui 18828 mensagens coleteadas de 20 newsgroups. Assumimos que os dados são não-anotados, no entanto, o nome de cada arquivo nos diz sua procedência, para que seja possível fazer a análise e comparação dos resultados.

Implemente o algoritmo de K-Means, e explore diferentes opções para escolha do vetor de características, seleção das sementes e escolha do centroide de um conjunto. Analise a variabilidade dos resultados para diferentes abordagens para cada uma destas questões. Examine o problema para 20 grupos, ou buscando números menores de grupos observando o que foi agrupado, e se há alguma razão lógica para tal.

- Grupos de até três alunos, a quantidade de trabalho desenvolvido deve ser proporcional ao tamanho do grupo.
- Não é permitido compartilhamento de resultados e funções entre grupos distintos antes da entrega do trabalho.
- Após a entrega, posso escolher um elemento de cada grupo para uma avaliação oral sobre o trabalho. O desempenho do escolhido determinará a nota do grupo inteiro.
- A qualidade do relatório é fundamental. Justifique tudo o que fizer, e acrescente o código documentado de todas as implementações realizadas.
- Os dados estão em
  `http://www.ic.unicamp.br/~siome/teaching/2009/mc906-0209/material/cluster-txt.zip`

### 1 Clusterização

Implemente o método K-Means de clusterização, e utilizê-o para separar seus dados em 20 grupos.

1. Descreva o resultado com uma matriz de pertinência (cluster X grupo).
2. Analise a sensibilidade do resultado do algoritmo para diferentes vetores de caraterísticas. Teste com e sem o uso dos cabeçalhos das mensagens.
3. Analise a sensibilidade do resultado do algoritmo para diferentes conjuntos iniciais de sementes.

### 2 Análise dos Grupos

Repita a Questão 1 para outros números de grupos. é possível achar outras estruturas hierárquicas de grupos?

### 3 Outros Algoritmos de Clusterização

Experimente a clusterização dos dados com outros algoritmos. Se implementar mostre o código, se utilizar algo de algum pacote indique exatamente o que e de onde utilizou. Compare os diferentes resultados

### 4 Bibliografia

Junto com os dados, segue também uma série de artigos relevantes para leitura suplementar. Para os que tiverem interesse, eles servem também como ponto de partida para uma busca bibliográfica mais profunda.