# Recuperação inteligente da informação
# Relatório atividade T-13

André Pacheco

Doutorado em Ciência da Computação
Programa de pós-graduação em informática
Universidade Federal do Espírito Santo

## Conteúdo



## 1 Introdução

A atividade T-13 a análise e clusterização da base de dados Narrativas [1]. Essa base, que possui dados relacionados a religião espírita, possui diversas descrições de conversas entre médiuns e espíritos. Foi utilizado o framework aLine juntamente com o k-means para clusterização do dataset.

## 2 Pré-processamento da base de dados

A base de dados disponibilizada esta no formato .docx. Neste documento, além de uma breve introdução, existe uma tabela com 980 linhas (embora exista 986 ids, o mesmo possui erro de sequência em alguns pontos do arquivo). As linhas contém um id, data, nome do médium, nome do espírito e descrição da conversa. Com objetivo de indexar essa base utilizando o aLine, cada linha foi convertida para um arquivo de extensão .txt contendo todas as informações previamente informadas. Esses arquivos foram submetidos ao aLine para indexação e clusterização.

[1] http://www.inf.ufes.br/ elias/dataSets/narrativas.docx

## 3 Resultados da clusterização

Para esta base de dados, o número de classes não é conhecido. Sendo assim, o número de clusters (k) foi variado de 2 a 5 e as métricas propostas por Salton et al. (1975) são computadas e descritas na Tabela 1. Em todos os casos, o número de iterações do k-means é igual a 300. A similaridade média (densidade) entre todas os documentos é igual a 0.1749.

**Tabela 1.** Métricas para clusterização

| | Cluster A K = 2 | Cluster B K = 3 | Cluster C K = 4 | Cluster C K = 5 |
|---|---|---|---|---|
| Tipo de indexação | $f_i^k$ | $f_i^k$ | $f_i^k$ | $f_i^k$ |
| Media de similaridade entre documentos e suas correspondentes centroides (x) | 0.1302 | 0.1403 | 0.2124 | 0.1460 |
| Media de similaridade entre as centroides e a centroide principal | 0.4959 | 0.5936 | 0.6884 | 0.4727 |
| Media de similaridade entre os pares de centroides (y) | 0.1235 | 0.0721 | 0.2723 | 0.0526 |
| Taxa y/x | 0.9485 | 0.5141 | 1.2820 | 0.3604 |

Como pode ser observado, dentre os valores de K escolhidos, o que melhor se dequa para essa base de acordo com a métricas computadas é K = 4. Para este valor, a media de similaridade entre os documentos e suas respectivas centróides é igual a 0.2124, maior do que as dos demais. Além disso, esse valor é maior, inclusive, que a media de similaridade (densidade) da base toda, que é igual a 0.1749. Especificamente para essa base, a distribuição de amostras entre os clusters é igual a 307, 252, 225, 196 para o 1ª, 2ª, 3ª e 4ª respectivamente.

Vale a pena observar que para essa base a clusterização não é tarefa tão simples, tendo em vista a essência heterogênea das conversas pisicografadas. Essa observação é confirmada pelo valor da densida média da base, já apresentada. Além disso, o número de palavras de cada psicografia varia bastante. Algumas apresentam 3 ou 4 linhas e outras apenas uma. Esses fatores ajudam a explicar a dificuldade da clusterização.

# Bibliografia

Salton, G., Wong, A., Yang, C.S., 1975. A vector space model for automatic indexing. Communications of the ACM 18, 613–620.