# PROJETO DE DISSERTAÇÃO

- **Orientadores:** Alexandre Fortes e Ricardo Cordeiro Corrêa
- **Título:** Classificação de documentos históricos
- **Objetivo:** implementar e analisar em casos práticos uma metodologia de classificação de documentos históricos quanto à sua relevância em relação a um objetivo dado.
- **Temas envolvidos:** classificadores, processamento de texto, pesquisa histórica
- **Introdução ao tema:** a seleção de documentos históricos, ou fragmentos de documentos, relevantes é prática comum em atividades de pesquisa histórica. A relevância de um documento é um atributo dependente de diversos parâmetros, variando desde o objeto da pesquisa ao contexto em que o documento foi gerado ou obtido. De uma forma geral, a determinação de relevância de um documento é uma questão complexa. No entanto, em cenários de escopo delimitado, responder à questão da relevância torna-se mais simples. Nesses casos, coloca-se, então, a questão da seleção automatizada de documentos históricos relevantes em um conjunto de documentos relativamente extenso. Esta é o tema desta proposta de dissertação.
- **Descrição sucinta da metodologia:**
  - Determinar aplicação prática no CEDIM, delimitando escopo de pesquisa histórica a ser abordado, escopo esse que torne a avaliação de relevância de documentos individuais dependente de poucos parâmetros
  - Selecionar documentos
  - Extrair textos desses documentos
  - Aplicar métodos de processamento de texto para determinar palavras representativas do conteúdo de cada documento
  - Realizar análises comparativas dos métodos aplicados
  - Definir critérios de similaridade entre documentos com base na aplicação e nas palavras representativas
  - Implementar método iterativo de classificação de relevância dos documentos: a cada iteração, o método apresenta uma pequena lista de documentos que aparentam ser relevantes e uma lista de supostos não relevantes; especialista confirma ou refuta a sugestão de cada documentos das listas. Iterações são repetidas até que as sugestões atinjam elevado grau de acerto, quando todos os demais documentos são classificados automaticamente.