Exercícios

1. Este exercício trata da detecção de spam. Crie um corpus de mensagens spam e outro de mensagens normais. Examine cada um deles e decida quais características parecem ser úteis para classificação: unigramas? bigramas? tamanho da mensagem, autor, horário de recebimento? Treine um algoritmo de classificação (árvore de decisão, naïve Bayes ou qualquer outro método de sua escolha) com um conjunto de treinamento e descreva a sua acurácia sobre o conjunto de teste.
2. Crie um conjunto de teste de 5 consultas, e use 3 mecanismos de busca web bem conhecidos para respondê-las. Avalie a precisão de cada um deles em 1, 3 e 10 documentos retornados.
3. Determine se os mecanisms de busca estão usando capitalização, *stemming*, sinônimo e correção ortográfica.