Universidade de Brasília
Departamento de Ciência da Computação

Projeto 1
Estruturas de Dados, Turma E, 1/2015
Prof. Díbio

Dados dois (2) documentos, quão próximos/parecidos eles são? Responder se eles são idênticos, ou não, é relativamente fácil, mas como saber o grau de modificação ou relação que um tem com o outro? Este é um problema típico ao analisarmos cadeias de DNA, identificação de autoria do documento, ou possível plágio. Neste projeto o objetivo será determinar semelhanças entre dois documentos (e.g. arquivos texto) para identificar se está ocorrendo, ou não, plágio.

Como determinar semelhanças entre documentos? A questão central é como projetar, ou usar, uma métrica (medida de distância ou similaridade) para quantificar essas semelhanças ou diferenças. Igualdade total, ou diferença total, seria simples pois bastaria comparar as sequências de caracteres na ordem. Várias métricas podem ser utilizadas, propostas, dependendo do tipo de documento, acuidade almejada, ou mesmo facilidade de aplicação. Uma métrica simples que pode ser utilizada neste projeto é baseada na frequência de palavras.

Uma palavra *w*, definida aqui como uma sequência de caracteres alfanuméricos, pode ocorrer uma, ou mais vezes em um documento. O número exato de vezes que cada palavra *w* ocorre no documento é a frequência da mesma, *D(w)*. Uma possível métrica entre dois documentos seria o produto interno dos vetores de frequência D_1 e D_2, com as seguintes relações quantitativas:

- métrica para projeção:

$$D_1 \cdot D_2 = \sum_w D_1(w) \cdot D_2(w)$$

- métrica para ângulo:

$$\theta(D_1, D_2) = \arccos\left(\frac{D_1 \cdot D_2}{\| D_1 \| * \| D_2 \|}\right)$$

$$0 \leq \theta \leq \pi/2$$

onde $\theta = 0$ significa D_1 e D_2 idênticos, e $\theta = \pi/2$ nenhuma palavra em comum.

- magnitude, ou número de palavras no documento:

$$N(D) = \| D \| = \sqrt{D \cdot D}$$

Uma maneira possível para calcular a semelhança entre dois documentos seria:

1. ler arquivos dos documentos (doc1.txt, doc2.txt, ...);
2. montar lista de palavras, e.g. [“a”, “os”, “um”, ...];
3. calcular frequências das palavras, e.g. [[“a”, 212], [“os”, 1200], ...];
4. ordenar a lista pelas frequências, e.g. [[“os”, 1200], [“a”, 200], ...];
5. calcular o ângulo Θ entre os documentos.

Escreva um programa em linguagem C, o qual deverá ler dez (10) arquivos texto fornecidos pelo usuário (doc1.txt, doc2.txt, …, doc10.txt), calcular a semelhança entre os documentos pelo método aqui fornecido, gerar um arquivo resultado compararDoc.txt com as palavras e frequências calculadas, bem como o valor de semelhança entre todos, tomados par a par, ou seja, entre o doc1.txt e todos os outros, entre doc2.txt e todos os outros, até o final. Na tela uma indicação de quais arquivos foram analisados, o resultado da métrica de comparação, e uma frase indicando se houve ou não plágio pela semelhança dos documentos (e.g. 50% de semelhança indicaria plágio). Os arquivos de documentos (doc1.txt, doc2.txt, doc3.txt, ...) serão fornecidos para teste.

O código deve ser bem documentado, de forma modular com funções para cada tarefa independente, realizado por dois (2) estudantes do curso usando *“pair programming”*, e entregue via sistema http://aprender.unb.br do curso, no prazo estipulado.

Questões importantes para refletir:

- Avaliando separadamente as funções (i.e. módulos) do programa, qual função tem maior ordem de complexidade?
- Seria possível fazer a comparação de forma mais rápida?
- Indicaria possíveis mudanças? Quais?