Universidade de Brasília
Departamento de Ciência da Computação

Projeto 1
Estruturas de Dados, Turma E, 1/2014
Prof. Díbio

Dados dois ou mais documentos, programas em uma linguagem, 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 arquivos de textos 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, contudo uma filtragem deverá ser feita para avaliar os tipos de palavras a serem comparadas no caso de um programa.

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,
$D_1 . D_2$ produto interno entre dois vetores,
e $|| D_1 || * || D_2 ||$ multiplicação entre dois módulos dos vetores.

- magnitude, ou número de palavras no documento:

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

Uma maneira possível para calcular a semelhança entre programas seria:

1. ler arquivos dos documentos (doc1.txt, doc2.txt, …,doc12.txt);
2. contar ocorrências de cada símbolo de pontuação;
3. contar ocorrências de cada palavra;
4. montar **listas encadeadas** de símbolos, e palavras em separado,
   e.g. [“,”, “.”, ...]; [“banana”, “manga”, …];
5. calcular frequências dos elementos das **listas**, e.g. [[“,”, 2], [“manga”, 12], ...];
6. ordenar as **listas** pelas frequências, e.g. [[“ ”, 1200], [“ ”, 200], ...];
7. calcular os ângulos Θi entre as **listas** de cada dos arquivos de textos.

Escreva um programa em linguagem C, o qual deverá ler dois (2) arquivos de texto por vez fornecidos pelo usuário (do total de 12 textos fornecidos), calcular a semelhança entre os documentos pelo método aqui fornecido, mostrar na tela os elementos e frequências calculadas, bem como o valor de semelhança entre os dois documentos. **Um requisito obrigatório é a implementação de listas encadeadas como estruturas de dados a serem usadas nos passos 4, 5, 6, e 7 acima.** 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).

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.

Inovação/Criatividade:**

- Seria possível/interessante alterar ou incrementar a métrica sugerida para comparar os documentos?