

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL INSTITUTO DE INFORMÁTICA

INF 01203 - Estruturas de Dados

TRABALHO FINAL Similaridade Textual

1 Objetivo

O objetivo do trabalho é aplicar as estruturas de dados vistas em aula para calcular a similaridade entre dois textos.

2 Especificação da Aplicação

O cálculo da similaridade textual é uma operação bastante comum e fundamental em aplicações como detecção de plágio, recuperação de informação e sistemas de recomendação.

A tarefa consiste em atribuir um escore de similaridade para um par de textos de entrada. Neste trabalho, utilizaremos a **similaridade de Jaccard**, que é calculada da seguinte forma:

$$Sim = \frac{|Texto_A \cap Texto_B|}{|Texto_A \cup Texto_B|} = \frac{|Texto_A \cap Texto_B|}{|Texto_A| + |Texto_B| - |Texto_A \cap Texto_B|}$$

onde $|Texto_A|$ é o número de palavras distintas no Texto A, $|Texto_{|B}|$ é o número de palavras distintas no Texto B e $|Texto_A \cap Texto_B|$ é o número de palavras distintas que ocorrem em ambos os textos A e B. Em outras palavras, o coeficiente de Jaccard computa a taxa de palavras compartilhadas entre dois textos. O escore resultante estará entre $\mathbf{0}$ e $\mathbf{1}$, onde $\mathbf{1}$ é a similaridade máxima e $\mathbf{0}$ indica que os textos não têm nenhuma palavra em comum.

- ★ Uma **palavra** é uma sequência de letras. Todos os demais caracteres (numeração, espaço, pontuação) deverão ser considerados como separadores de palavras.
 - ★ 0 termo "palavras distintas" significa que repetições da mesma palavra serão desprezadas.
 - ★ A **ordem** das palavras no texto é desconsiderada (*i.e.*, utilizaremos a abordagem *bag-of-words*).
- ★ A aplicação **não é case sensitive.** Sendo assim, letras maiúsculas e minúsculas devem ser consideradas iguais.
- ★ **Stopwords** são palavras consideradas sem valor semântico (artigos, preposições, *etc.*) e por isso não devem ser consideradas no cálculo da similaridade. A lista de stopwords será fornecida (arquivo texto contendo uma palavra por linha, não necessariamente em ordem alfabética).
- ★ Sua tarefa é projetar uma aplicação que calcula de forma **eficiente** o coeficiente de Jaccard para um par de textos fornecidos. A solução pode utilizar uma combinação de estruturas de dados vistas em aula (listas e árvores). É obrigatório utilizar **pelo menos uma árvore binária balanceada** (AVL, Splay ou Rubro-Negra).
- ★ A escolha das estruturas implementadas deverá ser **justificada com riqueza de detalhes** em um relatório. A justificativa deve conter as vantagens e limitações da estrutura escolhida, constituindo esta justificativa um importante item de avaliação. A escolha da estrutura deve demonstrar conhecimento teórico e prático buscando a melhor combinação que atinja os resultados satisfatoriamente.
- ★ Os textos de entrada poderão ter um **tamanho arbitrário** (desde algumas centenas de caracteres até milhões).
- ★ Seu programa deverá ser chamado **a partir da linha de comando** (passando parâmetros para a main). Há um exemplo no Moodle que mostra como fazer essa chamada. Exemplo de chamada: C:\jaccard textoA.txt textoB.txt stopwords.txt
- ★ Após o processamento, o programa deve imprimir o coeficiente de Jaccard e o tempo gasto no processamento.
- ★ Arquivos de teste serão disponibilizados no Moodle. No dia da apresentação, novos conjuntos de arquivos serão fornecidos.

As entradas e saídas da aplicação são:

Entradas:

- (i) o nome do arquivo com o primeiro texto;
- (ii) o nome do arquivo com o segundo texto; e
- (iii) o nome do arquivo com a lista de stopwords.

Saídas:

- (i) coeficiente de Jaccard e
- (ii) tempo gasto durante toda a execução do programa.

3. Exemplo

Stopwords	Texto1	Texto2	Jaccard
a	Após a morte de seu pai, o rei Jorge	Elizabeth II (nascida Elizabeth	0.17
ao	VI, a princesa Elizabeth torna-se a	Alexandra Mary; Londres, 21 de	
até	rainha do Reino Unido, dando início	abril de 1926 - Castelo de Balmoral,	
da	ao mais longo reinado britânico.	Aberdeenshire, 8 de setembro de	
de		2022), foi rainha do Reino Unido e	
do		dos Reinos da Comunidade de	
dos		Nações de 1952 até sua morte em	
e		2022.	
em	Após a morte de seu pai, o rei Jorge	Após a morte de seu pai, o rei Jorge	1.00
foi	VI, a princesa Elizabeth torna-se a	VI, a princesa Elizabeth torna-se a	
mais	rainha do Reino Unido, dando início	rainha do Reino Unido, dando início	
na	ao mais longo reinado britânico.	ao mais longo reinado britânico.	
0	O terremoto da Turquia e Síria de	Elizabeth II (nascida Elizabeth	0.00
quando	2023 ocorreu em 6 de fevereiro de	Alexandra Mary; Londres, 21 de	
se	2023, quando dois sismos atingiram	abril de 1926 - Castelo de Balmoral,	
seu	o sul e o centro da Turquia. O	Aberdeenshire, 8 de setembro de	
sua	primeiro ocorreu a oeste da cidade de	2022), foi rainha do Reino Unido e	
um	Gaziantepe, às 04h17 TRT (01h17	dos Reinos da Comunidade de	
às	UTC), causando danos generalizados	Nações de 1952 até sua morte em	
	na Turquia e na Síria.	2022.	
	O terremoto da Turquia e Síria de	Um terremoto de magnitude 7,8	0.10
	2023 ocorreu em 6 de fevereiro de	atingiu a região central da Turquia e	
	2023, quando dois sismos atingiram	o noroeste da Síria na manhã desta	
	o sul e o centro da Turquia. O	segunda-feira (6),	
	primeiro ocorreu a oeste da cidade de		
	Gaziantepe, às 04h17 TRT (01h17		
	UTC), causando danos generalizados		
	na Turquia e na Síria.		

4. Requisitos

- O trabalho deve ser feito, preferencialmente, em **duplas**. Também aceitaremos trabalhos feitos individualmente e de duplas cujos integrantes sejam de turmas diferentes.
- A linguagem de programação aceita é **C** (Não é C++ nem C#).

5. Entrega e Apresentação

• 3 de abril de 2023 apresentação em aula presencial e entrega pelo Moodle

6. Critérios de Avaliação

O trabalho deve ser realizado em duplas e deverá ser apresentado e defendido por **todos os componentes da dupla** na data prevista. Trabalhos que não forem apresentados, não serão corrigidos.

Para a avaliação serão adotados os seguintes critérios:

- funcionamento (Peso: 30%);
- organização e documentação do código (Peso: 30%); e
- projeto das estruturas utilizadas (relatório) (Peso: 40%).

Importante:

Este trabalho deverá representar a solução da dupla para o problema proposto. O plágio é terminantemente proibido e a sua detecção irá zerar a nota do trabalho.