Aula 34: Processamento de cadeias

- Introdução
- O problema do casamento de cadeias
- Compactação: algoritmo de freqüência de caracteres

Cadeias

- Alfabeto = conjunto de elementos denominados caracteres
- Cadeia = seqüência de caracteres
  - Ex.: 0110010 é uma cadeia de alfabeto { 0 , 1 }
- Os dados (caracteres) de uma cadeia não estão estruturados entre si. Há apenas a ordem da seqüência.

Cadeias

- Cadeias aparecem em computação no processamento de palavras, mensagens, textos, códigos, etc.
- Cadeias aparecem nas aplicações de
  - tratamento computacional de dicionários
  - mensagens de correio eletrônico
  - seqüência de DNA's, em biologia computacional
  - criptografia
  - etc.
- Em geral, principal utilização de um computador pessoal: edição de textos

Problemas a serem considerados

1) Problema do casamento de cadeias

- Dadas duas cadeias de caracteres, o problema consiste em verificar se a primeira delas contém a segunda, isto é, se os caracteres que compõem a segunda aparecem sequencialmente na primeira.
  - Exemplo de aplicação: edição de textos.

2) Problema de codificação de mensagens

- Dada uma cadeia, denominada mensagem, o problema consiste em codificá-la através da atribuição de códigos a seus caracteres, de modo a minimizar o comprimento total da mensagem codificada.
  - Exemplo de aplicação: transmissão de mensagens em uma rede.

Prefixos e sufixos

- Comprimento de uma cadeia = número de caracteres da cadeia
- X = cadeia de comprimento n
  Y = cadeia de comprimento $m \leq n$
  $X = x_1\ x_2\ \ldots\ x_n$
  $Y = y_1\ y_2\ \ldots\ y_m$
- Y é subcadeia de X quando

$$x_{l+1} = y_1$$
$$\ldots$$
$$x_{l+m} = y_m$$

para algum $l \leq n - m$.

$l = 0$ => Y é prefixo de X

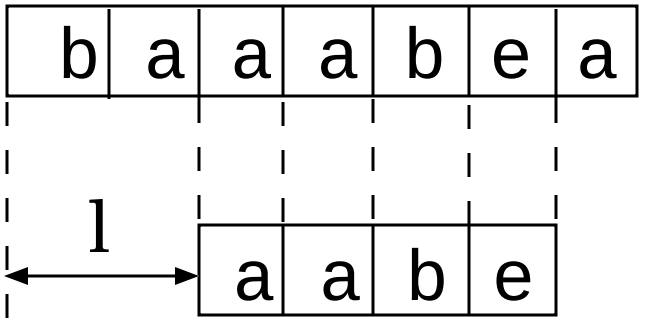
$l = n - m$ => Y é sufixo de X,

além disso, se $n \neq m$ => Y é prefixo (sufixo) próprio de X

Exemplo

X | b | a | a | a | b | e | a |

Y | a | a | b | e |

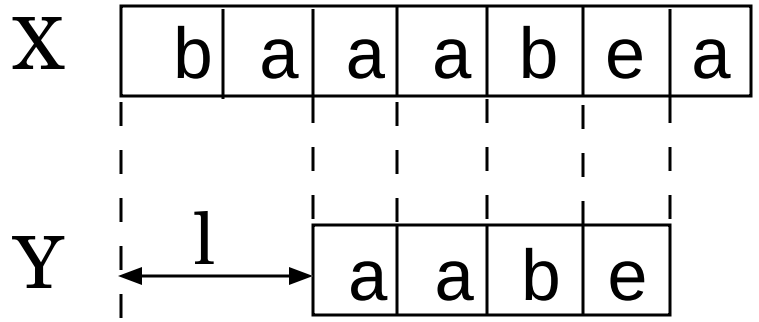


- Y é subcadeia de X, com l = 2
- A cadeia | b | a | a | é prefixo de X.
- A cadeia | e | a | é sufixo de X.

Problema do casamento de cadeias

- Dadas as cadeias X, Y verificar se Y é subcadeia de X. Em caso positivo, localizar Y em X. Diz-se então que houve um casamento de Y com X na posição l + 1.

- Exemplo:



Casamento de Y com X na posição 3

Y | a | e | b | não é subcadeia de X

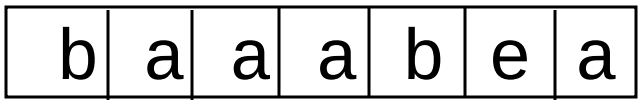
O algoritmo de força bruta

- Método: testar todas posições

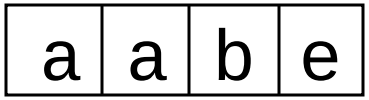
- Se Y for subcadeia de X, então Y se encontra em X, deslocado de l posições à esquerda. Então comparar Y com a subcadeia de tamanho m de X, que se inicia no caracter x_{l+1}, para l = 0, 1, ..., n - m.

Exemplo

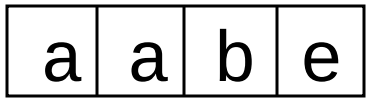
- X b a a a b e a
- Y a a b e
- X



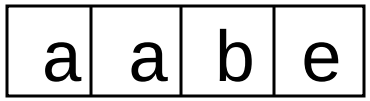
- l = 0 não



- l = 1 não



- l = 2 sim



Formulação do algoritmo

- Para tornar o método mais eficiente, utilizou-se o princípio de abandonar a verificação de uma subcadeia no meio do processo, se for detectado que um casamento não é possível.

- Algoritmo: casamento de cadeias

```
para l = 0, .., n - m faça
        i := 1
        teste := V
        enquanto i ≤ m  e teste faça
             se x[ l+i ] = y[ i ] então i := i + 1
             senão teste := F
        se teste então
             "casamento na posição l+1"
             pare
"não há casamento"
```

Complexidade

- Notação: X_k denota uma subcadeia de X, iniciada na posição k
- Pior caso do algoritmo: cada subcadeia X_{l+1}, de tamanho m, coincide com Y em todas as posições, exceto a última (não há casamento).
- Ou então: quando a condição acima se repete por todas as cadeias analisadas, exceto a última, quando é detectado um casamento (Y é sufixo de X).
- Número de subcadeias examinadas: n - m + 1
- Em cada subcadeia são realizadas m comparações.
- Logo: Número total de passos = m (n - m + 1)
- Complexidade: O (n . m)

Exercício

Sejam

X

a	b	e	b	a	e	b	a	a	b	e	a	b	a	b	b	e

Y

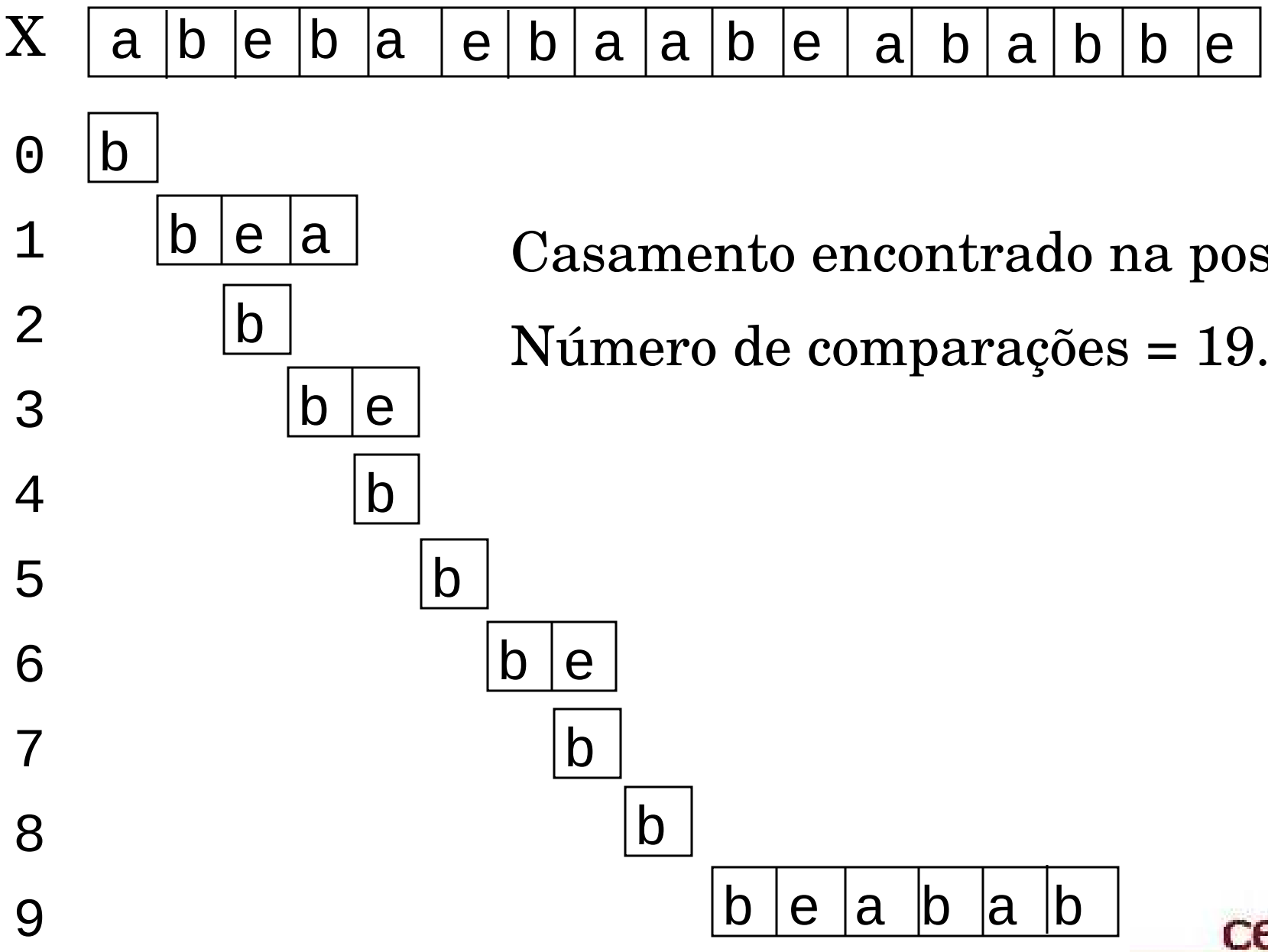
b	e	a	b	a	b

Descrever os passos efetuados pelo algoritmo de força bruta para determinar se Y é subcadeia de X.

Quantas comparações são realizadas?

Tempo: 9 minutos

Solução



Casamento encontrado na posição 10.

Número de comparações = 19.

Exercícios

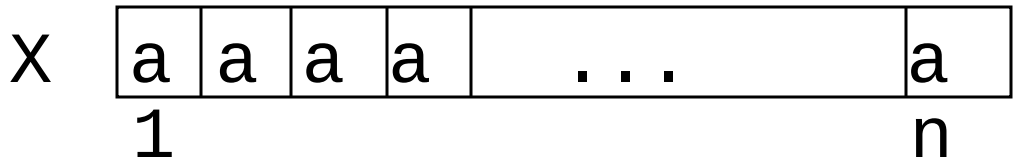
- Dê exemplo de cadeias X e Y, de comprimentos n e m, respectivamente, tais que o algoritmo de força bruta efetue um número mínimo de comparações.
  Determinar o mínimo.

- Repetir o exercício anterior, com máximo ao invés de mínimo.

Tempo: 5 minutos

Solução

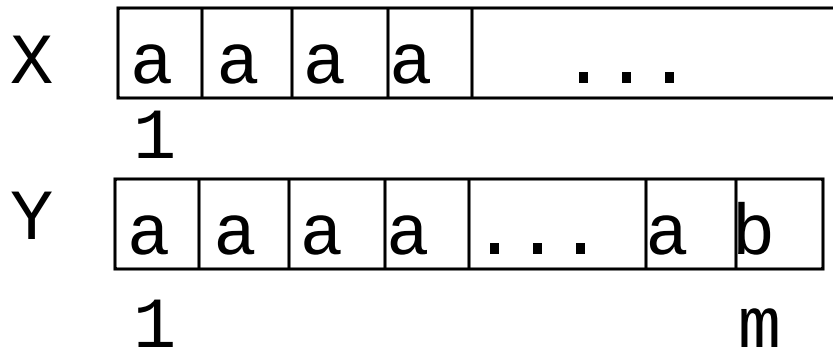
- Mínimo:



$m \leq n$

Número de comparações = m

- Máximo:



$m \leq n$

Número de comparações = m.(n - m + 1)

Eficiência

- Existem algoritmos mais eficientes?
- Sim. É possível resolver o problema do casamento de cadeias em tempo linear no tamanho da maior cadeia, isto é, O(n).
- O algoritmo de Knuth, Morris e Pratt foi o primeiro a atingir esta complexidade.

Compactação de dados

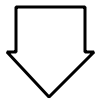
- Objetivo: Dado um arquivo de um certo tamanho n, codificá-lo, de modo que o arquivo codificado possua tamanho m < n. O objetivo é encontrar uma maneira de codificar o arquivo dado, de modo a obter a diferença n - m tão grande quanto possível.
- É necessário desenvolver algoritmos de codificação e decodificação.
- Deseja-se também que estes algoritmos sejam eficientes.
- Vantagens da compactação:
  - espaço menor para armazenamento dos dados
  - tempo de transmissão menor, em caso de redes de computadores.

Métodos a serem abordados

- Algoritmo de freqüência de caracteres
- Algoritmo de Huffman

Algoritmo de freqüência de caracteres

- Seja um arquivo composto de símbolos não numéricos.
- Para compactar este texto, determina-se a quantidade de símbolos idênticos consecutivos, existentes no texto. Substitui-se cada subcadeia de símbolos idênticos por um único símbolo, precedido da quantidade acima determinada.
- Exemplo: . . . BBBBBBBBB . . .



. . . 9B . . .

- Se a quantidade de símbolos for igual a 1, por convenção, o valor numérico será omitido.

Exemplo

- Exemplo:
  BBBEAAAAFFHHHHHCBMMALLLCDDBBBBBBBCC

  3BE4A2F5HCB2MA3LC2D7B2C

- Observação:
  se o texto possuir caracteres numéricos, pode-se utilizar um símbolo não existente no texto, que precederia cada algarismo do arquivo compactado, correspondente a um algarismo do texto original.
- Exemplo:
  BBB3333M1CCC

  3B4@3M@13C

Exercício

- Escrever o texto compactado através do método de freqüência de caracteres idênticos, correspondente ao texto seguinte:

```
BBBCCCCDEFFFBAAAAAAAAAACDDFFFFFFFCF
```

Tempo: 1 minuto

Solução

BBBCCCCDEFFFBAAAAAAAAAACDDFFFFFFFCF

3B4CDE3FB10AC2D7FCF