Introdução à Recuperação de Informações

https://github.com/fccoelho/curso-IRI

IRI 3: Dicionários e Recuperação Tolerante

Flávio Codeço Coelho

Escola de Matemática Aplicada, Fundação Getúlio Vargas

Sumário

Distinguindo entre Tipo e token

- Token – Uma instância de uma palavra ou termo ocorrendo em um documento
- Tipo – Uma classe de equivalência de tokens
- *In June, the dog likes to chase the cat in the barn.*
- 12 tokens, 9 tipos de palavras

Problemas na tokenização

- Quais são os delimitadores? Espaços? Apóstrofes? Hífen?
- Para cada um destes: às vezes eles delimitam, às vezes não.
- Muitas línguas não possuem espaços! (P.ex., Chinês)
- Não Há espaços em palavras compostas em Holandês, Alemão e Sueco (*Lebensversicherungsgesellschaftsangestellter*)

Problemas com classes de equivalência

- Um termo é uma classe de equivalência de tokens.
- Como definir Classes de equivalência?
- Números: (3/20/91 vs. 20/3/91)
- Capitalização
- Truncagem, Truncador de Porter
- Análise Morfológica : inflecional vs. derivacional
- Problemas de classes de equivalências em outras línguas
  - Morfologias mais complexas do que o inglês
  - Finlandês: Um único verbo pode ter 12000 formas diferentes
  - Acentos, tremas, etc.

Principais conclusões de hoje

Principais conclusões de hoje

- Recuperação Tolerante: O que fazer se não há correspondência exata entre o termo de consulta e os termos do documento.

Principais conclusões de hoje

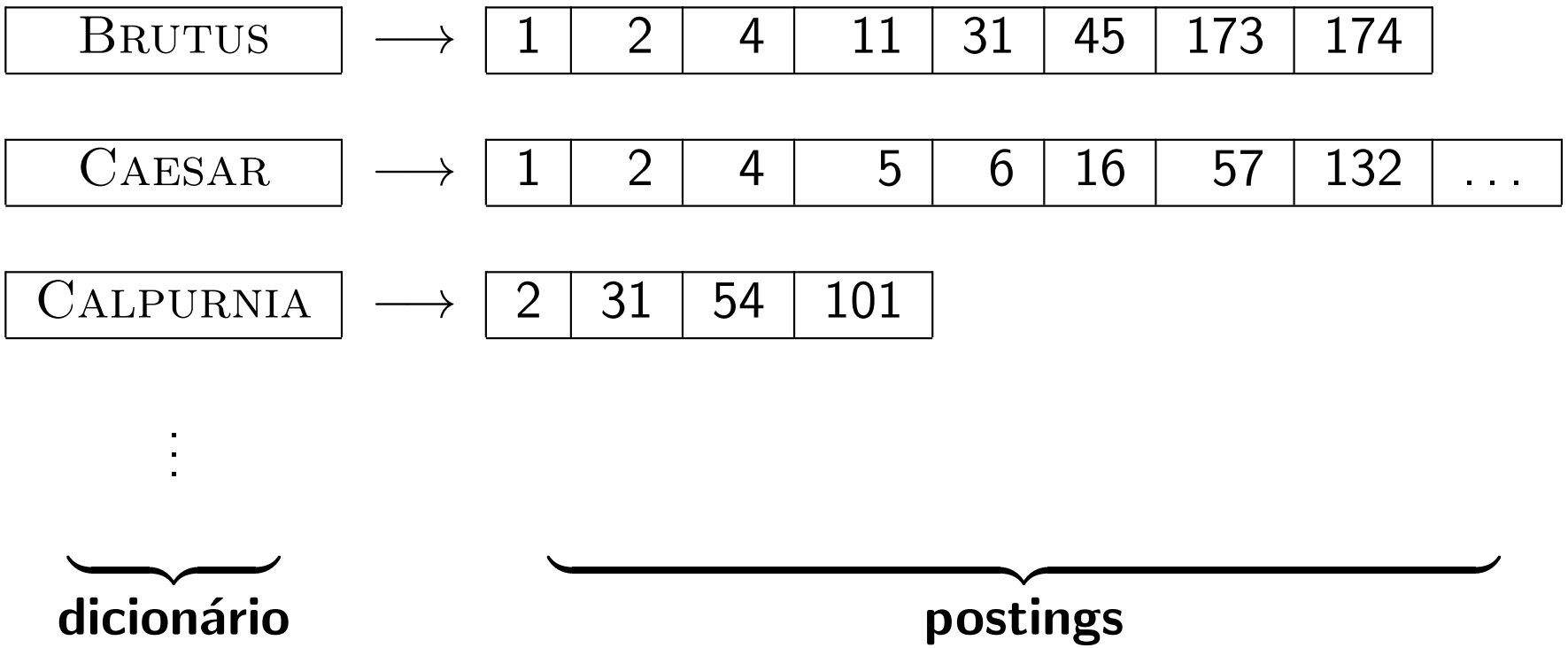
- Recuperação Tolerante: O que fazer se não há correspondência exata entre o termo de consulta e os termos do documento.
- Consultas coringas

Principais conclusões de hoje

- Recuperação Tolerante: O que fazer se não há correspondência exata entre o termo de consulta e os termos do documento.
- Consultas coringas
- Correção ortográficas

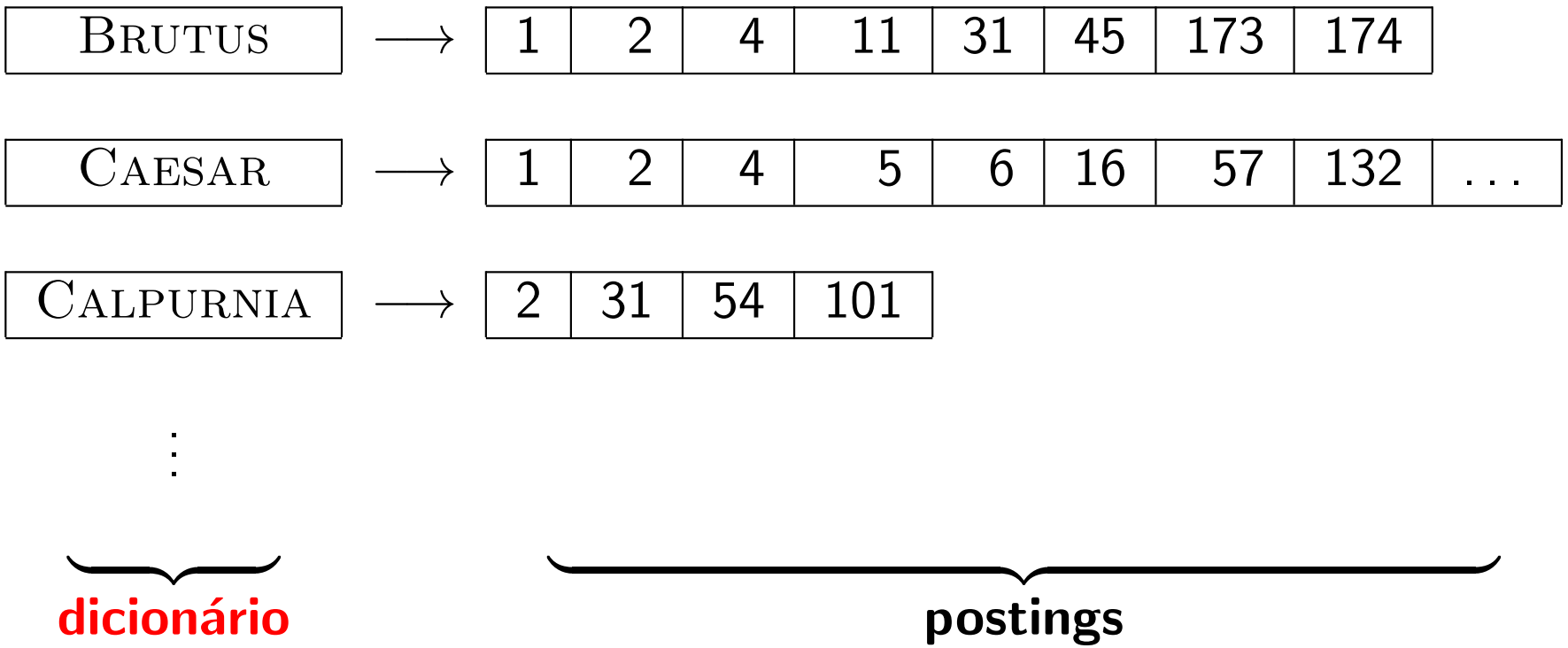
Índice invertido

Para cada termo t, armazenamos uma lista de documentos que contém t.



Índice invertido

Para cada termo t, armazenamos uma lista de documentos que contém t.



Dicionários

- O dicionário é a estrutura de dados que é usada para armazenar o vocabulário de termos.

Dicionários

- O dicionário é a estrutura de dados que é usada para armazenar o vocabulário de termos.
- vocabulário de termos: os dados

Dicionários

- O dicionário é a estrutura de dados que é usada para armazenar o vocabulário de termos.
- vocabulário de termos: os dados
- Dicionário: A estrutura de dados para armazenamento do vocabulário

Dicionário como uma matriz com entradas de tamanho fixo

- Para cada termo, precisamos armazenar um par de ítens:

Dicionário como uma matriz com entradas de tamanho fixo

- Para cada termo, precisamos armazenar um par de ítens:
  - Frequência de documentos

Dicionário como uma matriz com entradas de tamanho fixo

- Para cada termo, precisamos armazenar um par de ítens:
  - Frequência de documentos
  - Ponteiro para a lista de postings

Dicionário como uma matriz com entradas de tamanho fixo

- Para cada termo, precisamos armazenar um par de ítens:
  - Frequência de documentos
  - Ponteiro para a lista de postings
  - ...

Dicionário como uma matriz com entradas de tamanho fixo

- Para cada termo, precisamos armazenar um par de ítens:
  - Frequência de documentos
  - Ponteiro para a lista de postings
  - ...
- Assuma por ora que podemos armazenar esta informação em uma entrada de tamanho fixo.

Dicionário como uma matriz com entradas de tamanho fixo

- Para cada termo, precisamos armazenar um par de ítens:
  - Frequência de documentos
  - Ponteiro para a lista de postings
  - ...
- Assuma por ora que podemos armazenar esta informação em uma entrada de tamanho fixo.
- Assuma que armazenamos estas entradas em uma matriz.

Dicionário como uma matriz com entradas de tamanho fixo

termo	documento frequência	ponteiro para lista de postings
a	656.265	$\longrightarrow$
aachen	65	$\longrightarrow$
...	...	...
zulu	221	$\longrightarrow$
espaço necessário:	20 bytes	4 bytes	4 bytes

como acessamos um termo de consulta q_i nesta matriz em tempo de consulta? Ou seja: Que estrutura de dados usamos para localizar a entrada (linha) na matriz onde q_i está armazenado?

Estruturas de dados para acessar os termos

- Duas Classes principais de estruturas de dados: hashes e árvores

Estruturas de dados para acessar os termos

- Duas Classes principais de estruturas de dados: hashes e árvores
- Alguns sistemas de RI usam hashes, outros usam árvores.

Estruturas de dados para acessar os termos

- Duas Classes principais de estruturas de dados: hashes e árvores
- Alguns sistemas de RI usam hashes, outros usam árvores.
- Critérios de escolha:

Estruturas de dados para acessar os termos

- Duas Classes principais de estruturas de dados: hashes e árvores
- Alguns sistemas de RI usam hashes, outros usam árvores.
- Critérios de escolha:
  - Existe um número fixo de termos ou ele crescerá indefinidamente?

Estruturas de dados para acessar os termos

- Duas Classes principais de estruturas de dados: hashes e árvores
- Alguns sistemas de RI usam hashes, outros usam árvores.
- Critérios de escolha:
  - Existe um número fixo de termos ou ele crescerá indefinidamente?
  - Quais as frequências relativas com que as várias chaves serão acessadas?

Estruturas de dados para acessar os termos

- Duas Classes principais de estruturas de dados: hashes e árvores
- Alguns sistemas de RI usam hashes, outros usam árvores.
- Critérios de escolha:
  - Existe um número fixo de termos ou ele crescerá indefinidamente?
  - Quais as frequências relativas com que as várias chaves serão acessadas?
  - Quantos termos teremos?

Hashes

- Cada termo do vocabulário é "hasheado" para um inteiro.

Hashes

- Cada termo do vocabulário é "hasheado" para um inteiro.
- Busca-se evitar colisões

Hashes

- Cada termo do vocabulário é "hasheado" para um inteiro.
- Busca-se evitar colisões
- No momento da consulta, faz-se o seguinte: Hasheia o termo de consulta, resolve as colisões, Localiza entrada em uma matriz de elementos com tamanho constante

Hashes

- Cada termo do vocabulário é "hasheado" para um inteiro.
- Busca-se evitar colisões
- No momento da consulta, faz-se o seguinte: Hasheia o termo de consulta, resolve as colisões, Localiza entrada em uma matriz de elementos com tamanho constante
- Prós: Busca em um hash é mais rápida do que em uma árvore.

Hashes

- Cada termo do vocabulário é "hasheado" para um inteiro.
- Busca-se evitar colisões
- No momento da consulta, faz-se o seguinte: Hasheia o termo de consulta, resolve as colisões, Localiza entrada em uma matriz de elementos com tamanho constante
- Prós: Busca em um hash é mais rápida do que em uma árvore.
  - Tempo de consulta é constante.

Hashes

- Cada termo do vocabulário é "hasheado" para um inteiro.
- Busca-se evitar colisões
- No momento da consulta, faz-se o seguinte: Hasheia o termo de consulta, resolve as colisões, Localiza entrada em uma matriz de elementos com tamanho constante
- Prós: Busca em um hash é mais rápida do que em uma árvore.
  - Tempo de consulta é constante.
- Contras

Hashes

- Cada termo do vocabulário é "hasheado" para um inteiro.
- Busca-se evitar colisões
- No momento da consulta, faz-se o seguinte: Hasheia o termo de consulta, resolve as colisões, Localiza entrada em uma matriz de elementos com tamanho constante
- Prós: Busca em um hash é mais rápida do que em uma árvore.
  - Tempo de consulta é constante.
- Contras
  - Não há forma de encontrar pequenas variações (*resume* vs. *résumé*)

Hashes

- Cada termo do vocabulário é "hasheado" para um inteiro.
- Busca-se evitar colisões
- No momento da consulta, faz-se o seguinte: Hasheia o termo de consulta, resolve as colisões, Localiza entrada em uma matriz de elementos com tamanho constante
- Prós: Busca em um hash é mais rápida do que em uma árvore.
  - Tempo de consulta é constante.
- Contras
  - Não há forma de encontrar pequenas variações (*resume* vs. *résumé*)
  - Não permite busca de prefixos (Todos os termos que começam com *automat*)

Hashes

- Cada termo do vocabulário é "hasheado" para um inteiro.
- Busca-se evitar colisões
- No momento da consulta, faz-se o seguinte: Hasheia o termo de consulta, resolve as colisões, Localiza entrada em uma matriz de elementos com tamanho constante
- Prós: Busca em um hash é mais rápida do que em uma árvore.
  - Tempo de consulta é constante.
- Contras
  - Não há forma de encontrar pequenas variações (*resume* vs. *résumé*)
  - Não permite busca de prefixos (Todos os termos que começam com *automat*)
  - é necessário "rehashear" tudo periodicamente se o vocabulário continua crescendo.

Árvores

- Árvores resolvem o problema do prefixo (Encontrar todos os termos começando com *automat*).

Árvores

- Árvores resolvem o problema do prefixo (Encontrar todos os termos começando com *automat*).
- Árvore mais simples: árvore binária

Árvores

- Árvores resolvem o problema do prefixo (Encontrar todos os termos começando com *automat*).
- Árvore mais simples: árvore binária
- Busca é ligeiramente mais lenta que em hashes: $O(\log M)$, onde M é o tamanho do vocabulário.

Árvores

- Árvores resolvem o problema do prefixo (Encontrar todos os termos começando com *automat*).
- Árvore mais simples: árvore binária
- Busca é ligeiramente mais lenta que em hashes: $O(\log M)$, onde M é o tamanho do vocabulário.
- $O(\log M)$ vale apenas para árvores balanceadas.

Árvores

- Árvores resolvem o problema do prefixo (Encontrar todos os termos começando com *automat*).
- Árvore mais simples: árvore binária
- Busca é ligeiramente mais lenta que em hashes: $O(\log M)$, onde M é o tamanho do vocabulário.
- $O(\log M)$ vale apenas para árvores balanceadas.
- Rebalancear árvores binárias é caro.

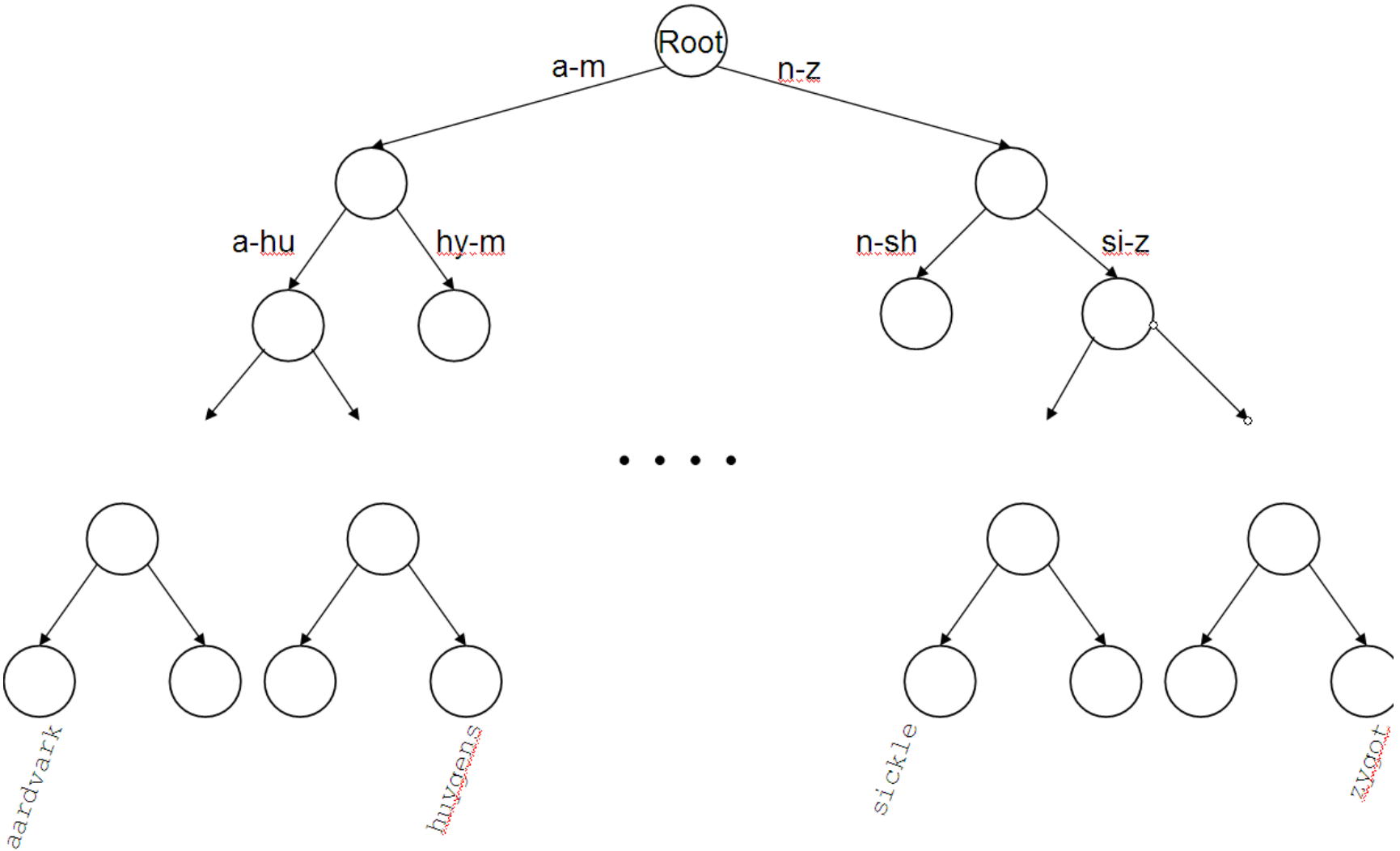
Árvores

- Árvores resolvem o problema do prefixo (Encontrar todos os termos começando com *automat*).
- Árvore mais simples: árvore binária
- Busca é ligeiramente mais lenta que em hashes: $O(\log M)$, onde M é o tamanho do vocabulário.
- $O(\log M)$ vale apenas para árvores balanceadas.
- Rebalancear árvores binárias é caro.
- Arvores-B resolvem o problema do balanceamento.

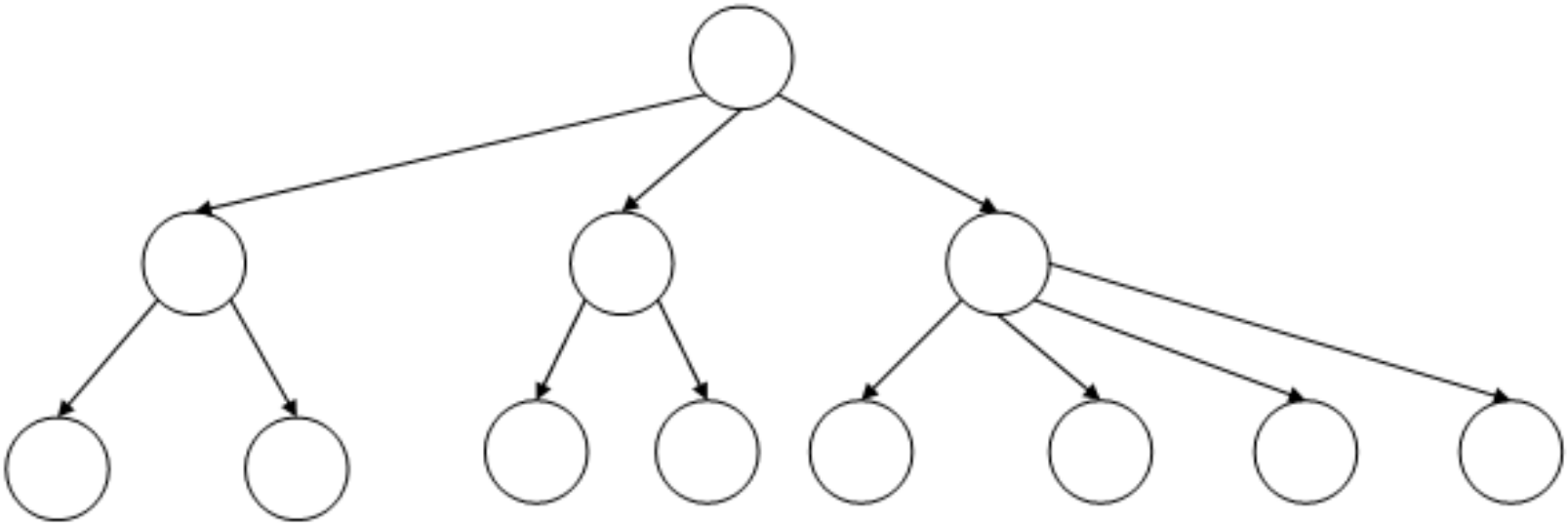
Árvores

- Árvores resolvem o problema do prefixo (Encontrar todos os termos começando com *automat*).
- Árvore mais simples: árvore binária
- Busca é ligeiramente mais lenta que em hashes: $O(\log M)$, onde M é o tamanho do vocabulário.
- $O(\log M)$ vale apenas para árvores balanceadas.
- Rebalancear árvores binárias é caro.
- Arvores-B resolvem o problema do balanceamento.
- Definição de árvore-B: cada nó interno tem um número de filhos no intervalo $[a, b]$ onde a, b são inteiros positivos apropriados , p.ex., $[2, 4]$.

Árvore Binária



Árvore B



Consultas coringa

- mon*: Encontre todos os documentos contendo termos começados por *mon*

Consultas coringa

- mon*: Encontre todos os documentos contendo termos começados por *mon*
- Fácil com dicionários baseados em árvore B: recupera todos os termos t no intervalo: mon $\leq t <$ moo

Consultas coringa

- mon*: Encontre todos os documentos contendo termos começados por *mon*
- Fácil com dicionários baseados em árvore B: recupera todos os termos t no intervalo: mon $\leq t <$ moo
- *mon: Encontre todos os documentos contendo termos que terminam com *mon*

Consultas coringa

- mon*: Encontre todos os documentos contendo termos começados por *mon*
- Fácil com dicionários baseados em árvore B: recupera todos os termos t no intervalo: mon $\leq t <$ moo
- *mon: Encontre todos os documentos contendo termos que terminam com *mon*
  - Mantém uma árvore adicional para termos *ao contrário*

Consultas coringa

- mon*: Encontre todos os documentos contendo termos começados por *mon*
- Fácil com dicionários baseados em árvore B: recupera todos os termos t no intervalo: mon $\leq t <$ moo
- *mon: Encontre todos os documentos contendo termos que terminam com *mon*
  - Mantém uma árvore adicional para termos *ao contrário*
  - Então recupera todos os termos t no intervalo: nom $\leq t <$ non

Consultas coringa

- mon*: Encontre todos os documentos contendo termos começados por *mon*
- Fácil com dicionários baseados em árvore B: recupera todos os termos t no intervalo: mon $\leq t <$ moo
- *mon: Encontre todos os documentos contendo termos que terminam com *mon*
  - Mantém uma árvore adicional para termos *ao contrário*
  - Então recupera todos os termos t no intervalo: nom $\leq t <$ non
- Resultado: Um conjunto de termos que correspondem à consulta coringa

Consultas coringa

- mon*: Encontre todos os documentos contendo termos começados por *mon*
- Fácil com dicionários baseados em árvore B: recupera todos os termos t no intervalo: mon $\leq t <$ moo
- *mon: Encontre todos os documentos contendo termos que terminam com *mon*
  - Mantém uma árvore adicional para termos *ao contrário*
  - Então recupera todos os termos t no intervalo: nom $\leq t <$ non
- Resultado: Um conjunto de termos que correspondem à consulta coringa
- Então recupera todos os documentos que contenham estes termos

Como lidar com um * no meio de um termo

- Exemplo: m*nchen

Como lidar com um * no meio de um termo

- Exemplo: m*nchen
- Poderíamos buscar todos os termos que satisfazem m* e *nchen Na árvore B e reter a interseção dos dois conjuntos.

Como lidar com um * no meio de um termo

- Exemplo: m*nchen
- Poderíamos buscar todos os termos que satisfazem m* e *nchen Na árvore B e reter a interseção dos dois conjuntos.
- Mas sai caro

Como lidar com um * no meio de um termo

- Exemplo: m*nchen
- Poderíamos buscar todos os termos que satisfazem m* e *nchen Na árvore B e reter a interseção dos dois conjuntos.
- Mas sai caro
- Alternativa: índice permuterm

Como lidar com um * no meio de um termo

- Exemplo: m*nchen
- Poderíamos buscar todos os termos que satisfazem m* e *nchen Na árvore B e reter a interseção dos dois conjuntos.
- Mas sai caro
- Alternativa: índice permuterm
- Idéia básica: Rotaciona cada consulta coringa, de forma que o * ocorra no final.

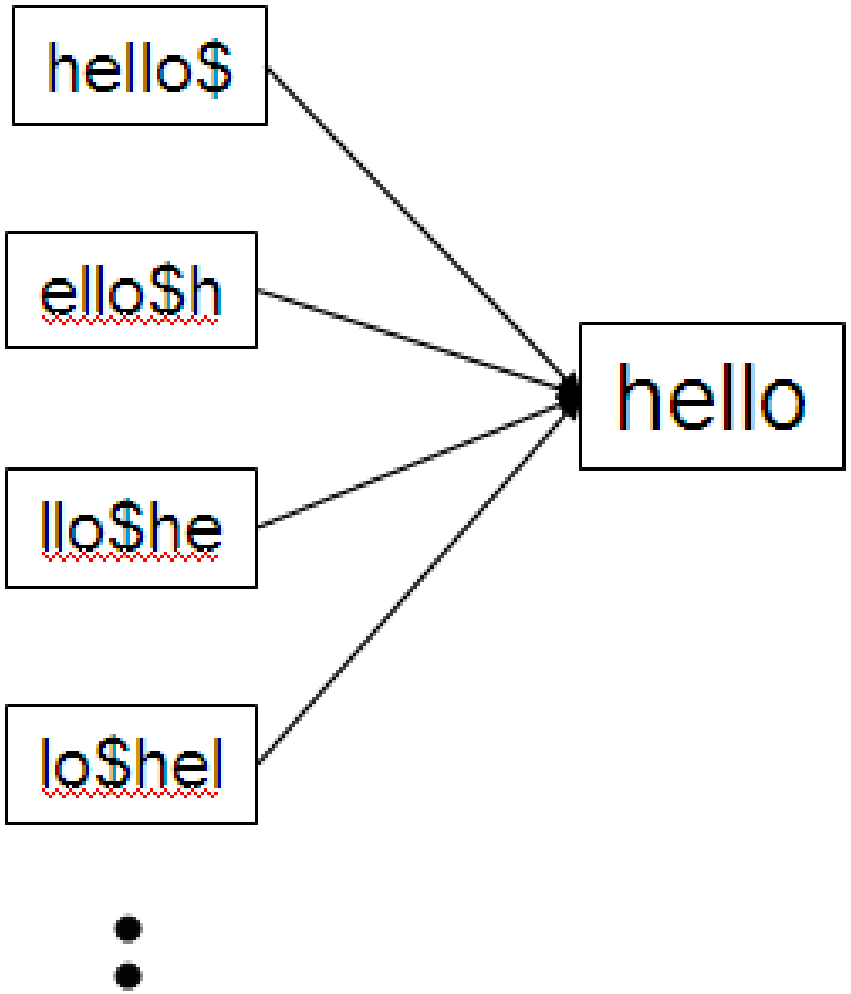
Como lidar com um * no meio de um termo

- Exemplo: m*nchen
- Poderíamos buscar todos os termos que satisfazem m* e *nchen Na árvore B e reter a interseção dos dois conjuntos.
- Mas sai caro
- Alternativa: índice permuterm
- Idéia básica: Rotaciona cada consulta coringa, de forma que o * ocorra no final.
- Armazena cada uma destas rotações no dicionário, por exemplo, em uma árvore B

Índice Permuterm

- Para o termo HELLO: adicione *hello$*, *ello$h*, *llo$he*, *lo$hel*, *o$hell*, e *$hello* à árvore B onde $ é um símbolo especial

Permuterm → mapeamento de termos



Índice Permuterm

- Para HELLO, adicionamos: *hello$*, *ello$h*, *llo$he*, *lo$hel*, e *o$hell*

Índice Permuterm

- Para HELLO, adicionamos: *hello$*, *ello$h*, *llo$he*, *lo$hel*, e *o$hell*
- Consultas

Índice Permuterm

- Para HELLO, adicionamos: *hello$*, *ello$h*, *llo$he*, *lo$hel*, e *o$hell*
- Consultas
  - Para X, acesse X$

Índice Permuterm

- Para HELLO, adicionamos: *hello$*, *ello$h*, *llo$he*, *lo$hel*, e *o$hell*
- Consultas
  - Para X, acesse X$
  - Para X*, acesse $X*

Índice Permuterm

- Para HELLO, adicionamos: *hello$*, *ello$h*, *llo$he*, *lo$hel*, e *o$hell*
- Consultas
  - Para X, acesse X$
  - Para X*, acesse $X*
  - Para *X, acesse X$*

Índice Permuterm

- Para HELLO, adicionamos: *hello$*, *ello$h*, *llo$he*, *lo$hel*, e *o$hell*
- Consultas
  - Para X, acesse X$
  - Para X*, acesse $X*
  - Para *X, acesse X$*
  - Para *X*, acesse X*

Índice Permuterm

- Para HELLO, adicionamos: *hello$*, *ello$h*, *llo$he*, *lo$hel*, e *o$hell*
- Consultas
  - Para X, acesse X$
  - Para X*, acesse $X*
  - Para *X, acesse X$*
  - Para *X*, acesse X*
  - Para X*Y, acesse Y$X*

Índice Permuterm

- Para HELLO, adicionamos: *hello$*, *ello$h*, *llo$he*, *lo$hel*, e *o$hell*
- Consultas
  - Para X, acesse X$
  - Para X*, acesse $X*
  - Para *X, acesse X$*
  - Para *X*, acesse X*
  - Para X*Y, acesse Y$X*
  - Example: Para hel*o, acesse o$hel*

Índice Permuterm

- Para HELLO, adicionamos: *hello$*, *ello$h*, *llo$he*, *lo$hel*, e *o$hell*
- Consultas
  - Para X, acesse X$
  - Para X*, acesse $X*
  - Para *X, acesse X$*
  - Para *X*, acesse X*
  - Para X*Y, acesse Y$X*
  - Example: Para hel*o, acesse o$hel*
- Um nome mais adequado para um índice Permuterm seria uma árvore permuterm tree.

Índice Permuterm

- Para HELLO, adicionamos: *hello$*, *ello$h*, *llo$he*, *lo$hel*, e *o$hell*
- Consultas
  - Para X, acesse X$
  - Para X*, acesse $X*
  - Para *X, acesse X$*
  - Para *X*, acesse X*
  - Para X*Y, acesse Y$X*
  - Example: Para hel*o, acesse o$hel*
- Um nome mais adequado para um índice Permuterm seria uma árvore permuterm tree.
- Mas índice permuterm é o nome mais comum.

Processando um acesso ao índice permuterm

- Rotacione a consulta coringa para a direita

Processando um acesso ao índice permuterm

- Rotacione a consulta coringa para a direita
- Use o acesso à árvore B como descrito anteriormente

Processando um acesso ao índice permuterm

- Rotacione a consulta coringa para a direita
- Use o acesso à árvore B como descrito anteriormente
- Problema: O Permuterm mais do que quadruplica o tamanho do dicionário quando comparado a uma árvore B regular. (observação empírica)

Correção ortográfica

Correção ortográfica

- Dois usos principais

Correção ortográfica

- Dois usos principais
  - Correção de documentos a serem indexados

Correção ortográfica

- Dois usos principais
  - Correção de documentos a serem indexados
  - Correção de consultas

Correção ortográfica

- Dois usos principais
  - Correção de documentos a serem indexados
  - Correção de consultas
- Dois métodos diferentes para correção ortográfica

Correção ortográfica

- Dois usos principais
  - Correção de documentos a serem indexados
  - Correção de consultas
- Dois métodos diferentes para correção ortográfica
- Correção de Palavra isolada

Correção ortográfica

- Dois usos principais
  - Correção de documentos a serem indexados
  - Correção de consultas
- Dois métodos diferentes para correção ortográfica
- Correção de Palavra isolada
  - Verifica cada palavra isoladamente quanto à correção ortográfica

Correção ortográfica

- Dois usos principais
  - Correção de documentos a serem indexados
  - Correção de consultas
- Dois métodos diferentes para correção ortográfica
- Correção de Palavra isolada
  - Verifica cada palavra isoladamente quanto à correção ortográfica
  - Não "pega" erros que resultam em palavras corretas, p.ex., *an asteroid that fell form the sky*

Correção ortográfica

- Dois usos principais
  - Correção de documentos a serem indexados
  - Correção de consultas
- Dois métodos diferentes para correção ortográfica
- Correção de Palavra isolada
  - Verifica cada palavra isoladamente quanto à correção ortográfica
  - Não "pega" erros que resultam em palavras corretas, p.ex., *an asteroid that fell form the sky*
- Correção ortográfica contextual

Correção ortográfica

- Dois usos principais
  - Correção de documentos a serem indexados
  - Correção de consultas
- Dois métodos diferentes para correção ortográfica
- Correção de Palavra isolada
  - Verifica cada palavra isoladamente quanto à correção ortográfica
  - Não "pega" erros que resultam em palavras corretas, p.ex., *an asteroid that fell form the sky*
- Correção ortográfica contextual
  - Olha para as palavras vizinhas

Correção ortográfica

- Dois usos principais
  - Correção de documentos a serem indexados
  - Correção de consultas
- Dois métodos diferentes para correção ortográfica
- Correção de Palavra isolada
  - Verifica cada palavra isoladamente quanto à correção ortográfica
  - Não "pega" erros que resultam em palavras corretas, p.ex., *an asteroid that fell form the sky*
- Correção ortográfica contextual
  - Olha para as palavras vizinhas
  - Pode corrigir o erro *form*/*from* acima

Corrigindo documentos

Corrigindo documentos

- Não estamos interessados em correção interativa de documentos (p.ex., MS Word) neste curso.

Corrigindo documentos

- Não estamos interessados em correção interativa de documentos (p.ex., MS Word) neste curso.
- Em RI, Usamos a correção de documentos primariamente para documentos OCR-izados. (OCR = optical character recognition)

Corrigindo documentos

- Não estamos interessados em correção interativa de documentos (p.ex., MS Word) neste curso.
- Em RI, Usamos a correção de documentos primariamente para documentos OCR-izados. (OCR = optical character recognition)
- A filosofia geral em RI é: Não altere os documentos.

Corrigindo consultas

Corrigindo consultas

- Primeiro: correção ortográfica de palavras isoladas

Corrigindo consultas

- Primeiro: correção ortográfica de palavras isoladas
- Premissa 1: Há uma lista de palavas "corretas" a partir da qual as correções podem ser obtidas.

Corrigindo consultas

- Primeiro: correção ortográfica de palavras isoladas
- Premissa 1: Há uma lista de palavas "corretas" a partir da qual as correções podem ser obtidas.
- Premissa 2: Há uma maneira de calcular a distância entre uma palavra errada e uma correta.

Corrigindo consultas

- Primeiro: correção ortográfica de palavras isoladas
- Premissa 1: Há uma lista de palavas "corretas" a partir da qual as correções podem ser obtidas.
- Premissa 2: Há uma maneira de calcular a distância entre uma palavra errada e uma correta.
- Algoritmo simplificado: retorne a palavra "correta" com a menor distância da palavra errada.

Corrigindo consultas

- Primeiro: correção ortográfica de palavras isoladas
- Premissa 1: Há uma lista de palavas "corretas" a partir da qual as correções podem ser obtidas.
- Premissa 2: Há uma maneira de calcular a distância entre uma palavra errada e uma correta.
- Algoritmo simplificado: retorne a palavra "correta" com a menor distância da palavra errada.
- Exemplo: *informaton* $\rightarrow$ *information*

Corrigindo consultas

- Primeiro: correção ortográfica de palavras isoladas
- Premissa 1: Há uma lista de palavas "corretas" a partir da qual as correções podem ser obtidas.
- Premissa 2: Há uma maneira de calcular a distância entre uma palavra errada e uma correta.
- Algoritmo simplificado: retorne a palavra "correta" com a menor distância da palavra errada.
- Exemplo: *informaton* $\rightarrow$ *information*
- Como lista de palavras corretas, podemos usar o vocabulário de todas as palavras que ocorrem em nossa coleção.

Corrigindo consultas

- Primeiro: correção ortográfica de palavras isoladas
- Premissa 1: Há uma lista de palavas "corretas" a partir da qual as correções podem ser obtidas.
- Premissa 2: Há uma maneira de calcular a distância entre uma palavra errada e uma correta.
- Algoritmo simplificado: retorne a palavra "correta" com a menor distância da palavra errada.
- Exemplo: *informaton* $\rightarrow$ *information*
- Como lista de palavras corretas, podemos usar o vocabulário de todas as palavras que ocorrem em nossa coleção.
- Porque isto é problemático?

Alternativas ao uso do vocabulário de termos

Alternativas ao uso do vocabulário de termos

- Um dicionário padrão (Webster's, Aurélio etc.)

Alternativas ao uso do vocabulário de termos

- Um dicionário padrão (Webster's, Aurélio etc.)
- Um dicionário de um domínio específico (Para sistemas de RI especializados)

Alternativas ao uso do vocabulário de termos

- Um dicionário padrão (Webster's, Aurélio etc.)
- Um dicionário de um domínio específico (Para sistemas de RI especializados)
- O vocabulário de termos ponderado, ponderado de forma adequada

Distância entre a palavra errada e a "correta"

Distância entre a palavra errada e a "correta"

- Estudaremos várias alternativas.

Distância entre a palavra errada e a "correta"

- Estudaremos várias alternativas.
- Distância de edição e a distância de Levenshtein

Distância entre a palavra errada e a "correta"

- Estudaremos várias alternativas.
- Distância de edição e a distância de Levenshtein
- Distância de edição ponderada

Distância entre a palavra errada e a "correta"

- Estudaremos várias alternativas.
- Distância de edição e a distância de Levenshtein
- Distância de edição ponderada
- sobreposição de k-grams

Distância de edição

Distância de edição

- A distância de edição entre a string s_1 e a string s_2 é o número mínimo de operações básicas que converte s_1 em s_2.

Distância de edição

- A distância de edição entre a string s_1 e a string s_2 é o número mínimo de operações básicas que converte s_1 em s_2.
- Distância de Levenshtein: Operações válidas: inserção, deleção, e substituição

Distância de edição

- A distância de edição entre a string s_1 e a string s_2 é o número mínimo de operações básicas que converte s_1 em s_2.
- Distância de Levenshtein: Operações válidas: inserção, deleção, e substituição
- Distância de Levenshtein *dog-do*: 1

Distância de edição

- A distância de edição entre a string s_1 e a string s_2 é o número mínimo de operações básicas que converte s_1 em s_2.
- Distância de Levenshtein: Operações válidas: inserção, deleção, e substituição
- Distância de Levenshtein *dog-do*: 1
- Distância de Levenshtein *cat-cart*: 1

Distância de edição

- A distância de edição entre a string s_1 e a string s_2 é o número mínimo de operações básicas que converte s_1 em s_2.
- Distância de Levenshtein: Operações válidas: inserção, deleção, e substituição
- Distância de Levenshtein *dog-do*: 1
- Distância de Levenshtein *cat-cart*: 1
- Distância de Levenshtein *cat-cut*: 1

Distância de edição

- A distância de edição entre a string s_1 e a string s_2 é o número mínimo de operações básicas que converte s_1 em s_2.
- Distância de Levenshtein: Operações válidas: inserção, deleção, e substituição
- Distância de Levenshtein *dog-do*: 1
- Distância de Levenshtein *cat-cart*: 1
- Distância de Levenshtein *cat-cut*: 1
- Distância de Levenshtein *cat-act*: 2

Distância de edição

- A distância de edição entre a string s_1 e a string s_2 é o número mínimo de operações básicas que converte s_1 em s_2.
- Distância de Levenshtein: Operações válidas: inserção, deleção, e substituição
- Distância de Levenshtein *dog-do*: 1
- Distância de Levenshtein *cat-cart*: 1
- Distância de Levenshtein *cat-cut*: 1
- Distância de Levenshtein *cat-act*: 2
- Distância de Damerau-Levenshtein *cat-act*: 1

Distância de edição

- A distância de edição entre a string s_1 e a string s_2 é o número mínimo de operações básicas que converte s_1 em s_2.
- Distância de Levenshtein: Operações válidas: inserção, deleção, e substituição
- Distância de Levenshtein *dog-do*: 1
- Distância de Levenshtein *cat-cart*: 1
- Distância de Levenshtein *cat-cut*: 1
- Distância de Levenshtein *cat-act*: 2
- Distância de Damerau-Levenshtein *cat-act*: 1
- Distância de Damerau-Levenshtein inclui a transposição como uma quarta operação possivel.

Distância de Levenshtein: Computação

		f	a	s	t
	0	1	2	3	4
c	1	1	2	3	4
a	2	2	1	2	3
t	3	3	2	2	2
s	4	4	3	2	3

Distância de Levenshtein: Algoritmo

LevenshteinDistance(s_1, s_2)

```
 1  for i ← 0 to |s1|
 2  do m[i, 0] = i
 3  for j ← 0 to |s2|
 4  do m[0, j] = j
 5  for i ← 1 to |s1|
 6  do for j ← 1 to |s2|
 7       do if s1[i] = s2[j]
 8            then m[i, j] = min{m[i-1, j]+1, m[i, j-1]+1, m[i-1, j-1]}
 9            else  m[i, j] = min{m[i-1, j]+1, m[i, j-1]+1, m[i-1, j-1]+1}
10  return m[|s1|, |s2|]
```

Operações: inserção (custo 1), deleção (custo 1), substituição (custo 1), cópia (custo 0)

Distância de Levenshtein: Algoritmo

LevenshteinDistance(s_1, s_2)

```
 1  for i ← 0 to |s1|
 2  do m[i, 0] = i
 3  for j ← 0 to |s2|
 4  do m[0, j] = j
 5  for i ← 1 to |s1|
 6  do for j ← 1 to |s2|
 7       do if s1[i] = s2[j]
 8            then m[i, j] = min{m[i-1, j]+1, m[i, j-1]+1, m[i-1, j-1]}
 9            else  m[i, j] = min{m[i-1, j]+1, m[i, j-1]+1, m[i-1, j-1]+1}
10  return m[|s1|, |s2|]
```

Operações: inserção (custo 1), deleção (custo 1), substituição (custo 1), cópia (custo 0)

Distância de Levenshtein: Algoritmo

LevenshteinDistance(s_1, s_2)

```
 1  for i ← 0 to |s1|
 2  do m[i, 0] = i
 3  for j ← 0 to |s2|
 4  do m[0, j] = j
 5  for i ← 1 to |s1|
 6  do for j ← 1 to |s2|
 7       do if s1[i] = s2[j]
 8            then m[i, j] = min{m[i-1, j]+1, m[i, j-1]+1, m[i-1, j-1]}
 9            else  m[i, j] = min{m[i-1, j]+1, m[i, j-1]+1, m[i-1, j-1]+1}
10  return m[|s1|, |s2|]
```

Operações: inserção (custo 1), deleção (custo 1), substituição (custo 1), cópia (custo 0)

Distância de Levenshtein: Algorithm

LevenshteinDistance(s_1, s_2)

```
 1  for i ← 0 to |s1|
 2  do m[i, 0] = i
 3  for j ← 0 to |s2|
 4  do m[0, j] = j
 5  for i ← 1 to |s1|
 6  do for j ← 1 to |s2|
 7       do if s1[i] = s2[j]
 8            then m[i, j] = min{m[i-1, j]+1, m[i, j-1]+1, m[i-1, j-1]}
 9            else  m[i, j] = min{m[i-1, j]+1, m[i, j-1]+1, m[i-1, j-1]+1}
10  return m[|s1|, |s2|]
```

Operações: inserção (custo 1), deleção (custo 1), substituição (custo 1), cópia (custo 0)

Distância de Levenshtein: Algoritmo

LevenshteinDistance(s_1, s_2)

```
 1  for i ← 0 to |s1|
 2  do m[i, 0] = i
 3  for j ← 0 to |s2|
 4  do m[0, j] = j
 5  for i ← 1 to |s1|
 6  do for j ← 1 to |s2|
 7       do if s1[i] = s2[j]
 8            then m[i, j] = min{m[i-1, j]+1, m[i, j-1]+1, m[i-1, j-1]}
 9            else  m[i, j] = min{m[i-1, j]+1, m[i, j-1]+1, m[i-1, j-1]+1}
10  return m[|s1|, |s2|]
```

Operações: inserção (custo 1), deleção (custo 1), substituição (custo 1), cópia (custo 0)

Distância de Levenshtein: Exemplo

			f		a		s		t	
		0	**1**	**1**	**2**	**2**	**3**	**3**	**4**	**4**
c		**1** **1**	*1* *2*	*2* *1*	**2** **2**	3 **2**	**3** **3**	4 **3**	**4** **4**	5 **4**
a		**2** **2**	**2** 3	**2** **2**	*1* *3*	*3* *1*	*3* *2*	*4* *2*	4 **3**	5 **3**
t		**3** **3**	**3** 4	**3** **3**	3 4	**2** **2**	*2* *3*	*3* *2*	*2* *3*	*4* *2*
s		**4** **4**	**4** 5	**4** **4**	4 5	**3** **3**	**2** 4	3 **2**	*3* *3*	*3* *3*

Cada célula da matriz de Levenshtein

Custo de chegar aqui a partir do meu vizinho superior esquerdo (cópia or substituição)	Custo de chegar aqui a partir do meu vizinho superior (deleção)
Custo de chegar aqui a partir do meu vizinho esquerdo (inserção)	mínimo dos três "movimentos" possíveis; a forma mais barata de chegar aqui

Distância de Levenshtein: Exemplo

			f		a		s		t	
		0	**1**	**1**	**2**	**2**	**3**	**3**	**4**	**4**
c		**1** **1**	*1* *2*	*2* *1*	**2** **2**	3 **2**	**3** **3**	4 **3**	**4** **4**	5 **4**
a		**2** **2**	**2** 3	**2** **2**	*1* *3*	*3* *1*	*3* *2*	*4* *2*	4 **3**	5 **3**
t		**3** **3**	**3** 4	**3** **3**	3 4	**2** **2**	*2* *3*	*3* *2*	*2* *3*	*4* *2*
s		**4** **4**	**4** 5	**4** **4**	4 5	**3** **3**	**2** 4	3 **2**	*3* *3*	*3* *3*

Programação dinâmica (Cormen et al.)

Programação dinâmica (Cormen et al.)

- Sub-estrutura ótima: A solução ótima para o problema, contém em si, sub-soluções, i.e., soluções ótimas para os sub-problemas.

Programação dinâmica (Cormen et al.)

- Sub-estrutura ótima: A solução ótima para o problema, contém em si, sub-soluções, i.e., soluções ótimas para os sub-problemas.
- Sobreposição de sub-soluções: As subsoluções se sobrepõem. Estas subsoluções são computadas repetidamente quando se computa a solução ótima global em um algoritmo de força bruta.

Programação dinâmica (Cormen et al.)

- Sub-estrutura ótima: A solução ótima para o problema, contém em si, sub-soluções, i.e., soluções ótimas para os sub-problemas.
- Sobreposição de sub-soluções: As subsoluções se sobrepõem. Estas subsoluções são computadas repetidamente quando se computa a solução ótima global em um algoritmo de força bruta.
- Subproblema no caso da distância de edição: Qual a distância de edição de dois prefixos

Programação dinâmica (Cormen et al.)

- Sub-estrutura ótima: A solução ótima para o problema, contém em si, sub-soluções, i.e., soluções ótimas para os sub-problemas.
- Sobreposição de sub-soluções: As subsoluções se sobrepõem. Estas subsoluções são computadas repetidamente quando se computa a solução ótima global em um algoritmo de força bruta.
- Subproblema no caso da distância de edição: Qual a distância de edição de dois prefixos
- Subsoluções sobrepostas: precisamos da maioria das distâncias entre prefixos 3 vezes – Isto corresponde a mover para a direita, diagonalmente, e para baixo.

Distância de edição ponderada

Distância de edição ponderada

- Como anteriormente, mas o peso de uma operação depende dos caracteres envolvidos.

Distância de edição ponderada

- Como anteriormente, mas o peso de uma operação depende dos caracteres envolvidos.
- Pensado para capturar erros de digitação, p.ex., a letra *m* é mais provável de ser trocada por *n* do que por *q*.

Distância de edição ponderada

- Como anteriormente, mas o peso de uma operação depende dos caracteres envolvidos.
- Pensado para capturar erros de digitação, p.ex., a letra m é mais provável de ser trocada por n do que por q.
- Logo, trocar m por n é uma distância de edição menor do que por q.

Distância de edição ponderada

- Como anteriormente, mas o peso de uma operação depende dos caracteres envolvidos.
- Pensado para capturar erros de digitação, p.ex., a letra *m* é mais provável de ser trocada por *n* do que por *q*.
- Logo, trocar *m* por *n* é uma distância de edição menor do que por *q*.
- Agora precisamos de uma matriz de pesos como entrada.

Distância de edição ponderada

- Como anteriormente, mas o peso de uma operação depende dos caracteres envolvidos.
- Pensado para capturar erros de digitação, p.ex., a letra *m* é mais provável de ser trocada por *n* do que por *q*.
- Logo, trocar *m* por *n* é uma distância de edição menor do que por *q*.
- Agora precisamos de uma matriz de pesos como entrada.
- E modificar a programação dinâmica para lidar com os pesos

Usando a distância de edição para correção ortográfica

Usando a distância de edição para correção ortográfica

- Dada uma consulta, primeiro enumera-se todas as sequências de caracteres dentro de uma distância (possivelmente ponderada) de edição prédefinida.

Usando a distância de edição para correção ortográfica

- Dada uma consulta, primeiro enumera-se todas as sequências de caracteres dentro de uma distância (possivelmente ponderada) de edição prédefinida.
- Toma-se a interseção deste conjunto com a nossa lista de palavras "corretas".

Usando a distância de edição para correção ortográfica

- Dada uma consulta, primeiro enumera-se todas as sequências de caracteres dentro de uma distância (possivelmente ponderada) de edição prédefinida.
- Toma-se a interseção deste conjunto com a nossa lista de palavras "corretas".
- Então sugere-se termos na interseção ao usuário.

Usando a distância de edição para correção ortográfica

- Dada uma consulta, primeiro enumera-se todas as sequências de caracteres dentro de uma distância (possivelmente ponderada) de edição prédefinida.
- Toma-se a interseção deste conjunto com a nossa lista de palavras "corretas".
- Então sugere-se termos na interseção ao usuário.
- $\rightarrow$ exercício em poucos slides

Exercício

1. Compute a matriz de distância de Levenshtein para as palavras OSLO – SNOW
2. Quais são as operações de Levenshtein que transformam *cat* em *catcat*?

		s	n	o	w
	0	1 1	2 2	3 3	4 4
o	1 / 1				
s	2 / 2				
l	3 / 3				
o	4 / 4				

			s		n		o		w	
		0	**1**	**1**	**2**	**2**	**3**	**3**	**4**	**4**
o	**1**	**1**	1	2						
			2	?						
s	**2**	**2**								
l	**3**	**3**								
o	**4**	**4**								

		s	n	o	w
	0	**1** **1**	**2** **2**	**3** **3**	**4** **4**
o	**1** / **1**	**1** 2 / 2 **1**			
s	**2** / **2**				
l	**3** / **3**				
o	**4** / **4**				

		s	n	o	w
	0	**1** **1**	**2** **2**	**3** **3**	**4** **4**
o	**1** / **1**	**1** 2 / 2 **1**	2 3 / 2 ?		
s	**2** / **2**				
l	**3** / **3**				
o	**4** / **4**				

		s	n	o	w
	0	**1** **1**	**2** **2**	**3** **3**	**4** **4**
o	**1** / **1**	**1** 2 / 2 **1**	**2** 3 / **2** **2**		
s	**2** / **2**				
l	**3** / **3**				
o	**4** / **4**				

		s	n	o	w
	0	**1** **1**	**2** **2**	**3** **3**	**4** **4**
o	**1** / **1**	**1** 2 / 2 **1**	**2** 3 / **2** **2**	2 4 / 3 ?	
s	**2** / **2**				
l	**3** / **3**				
o	**4** / **4**				

		s	n	o	w
	0	**1** **1**	**2** **2**	**3** **3**	**4** **4**
o	**1** / **1**	**1** 2 / 2 **1**	**2** 3 / **2** **2**	**2** 4 / 3 **2**	
s	**2** / **2**				
l	**3** / **3**				
o	**4** / **4**				

			s		n		o		w	
		0	**1**	**1**	**2**	**2**	**3**	**3**	**4**	**4**
o	**1**	**1**	**1**	2	**2**	3	**2**	4	4	5
	1		2	**1**	**2**	**2**	3	**2**	3	?
s	**2**	**2**								
l	**3**	**3**								
o	**4**	**4**								

		s	n	o	w
	0	**1** **1**	**2** **2**	**3** **3**	**4** **4**
o	**1** / **1**	**1** 2 / 2 **1**	**2** 3 / **2** **2**	**2** 4 / 3 **2**	4 5 / **3** **3**
s	**2** / **2**				
l	**3** / **3**				
o	**4** / **4**				

		s	n	o	w
	0	**1** **1**	**2** **2**	**3** **3**	**4** **4**
o	**1** / **1**	**1** 2 / 2 **1**	**2** 3 / **2** **2**	**2** 4 / 3 **2**	4 5 / **3** **3**
s	**2** / **2**	1 2 / 3 ?			
l	**3** / **3**				
o	**4** / **4**				

		s	n	o	w
	0	**1** **1**	**2** **2**	**3** **3**	**4** **4**
o	**1** / **1**	**1** 2 / 2 **1**	**2** 3 / **2** **2**	**2** 4 / 3 **2**	4 5 / **3** **3**
s	**2** / **2**	**1** 2 / 3 **1**			
l	**3** / **3**				
o	**4** / **4**				

		s	n	o	w
	0	**1** **1**	**2** **2**	**3** **3**	**4** **4**
o	**1** / **1**	**1** 2 / 2 **1**	**2** 3 / **2** **2**	**2** 4 / 3 **2**	4 5 / **3** **3**
s	**2** / **2**	**1** 2 / 3 **1**	2 3 / 2 ?		
l	**3** / **3**				
o	**4** / **4**				

		s	n	o	w
	0	**1** **1**	**2** **2**	**3** **3**	**4** **4**
o	**1** / **1**	**1** 2 / 2 **1**	**2** 3 / **2** **2**	**2** 4 / 3 **2**	4 5 / **3** **3**
s	**2** / **2**	**1** 2 / 3 **1**	**2** 3 / **2** **2**		
l	**3** / **3**				
o	**4** / **4**				

		s	n	o	w
	0	**1** **1**	**2** **2**	**3** **3**	**4** **4**
o	**1** / **1**	**1** 2 / 2 **1**	**2** 3 / **2** **2**	**2** 4 / 3 **2**	4 5 / **3** **3**
s	**2** / **2**	**1** 2 / 3 **1**	**2** 3 / **2** **2**	3 3 / 3 ?	
l	**3** / **3**				
o	**4** / **4**				

		s	n	o	w
	0	**1** **1**	**2** **2**	**3** **3**	**4** **4**
o	**1** / **1**	**1** 2 / 2 **1**	**2** 3 / **2** **2**	**2** 4 / 3 **2**	4 5 / **3** **3**
s	**2** / **2**	**1** 2 / 3 **1**	**2** 3 / **2** **2**	**3** **3** / **3** **3**	
l	**3** / **3**				
o	**4** / **4**				

		s	n	o	w
	0	**1** **1**	**2** **2**	**3** **3**	**4** **4**
o	**1** / **1**	**1** 2 / 2 **1**	**2** 3 / **2** **2**	**2** 4 / 3 **2**	4 5 / **3** **3**
s	**2** / **2**	**1** 2 / 3 **1**	**2** 3 / **2** **2**	**3** **3** / **3** **3**	3 4 / 4 ?
l	**3** / **3**				
o	**4** / **4**				

		s	n	o	w
	0	**1** **1**	**2** **2**	**3** **3**	**4** **4**
o	**1** / **1**	**1** 2 / 2 **1**	**2** 3 / **2** **2**	**2** 4 / 3 **2**	4 5 / **3** **3**
s	**2** / **2**	**1** 2 / 3 **1**	**2** 3 / **2** **2**	**3** **3** / **3** **3**	**3** 4 / 4 **3**
l	**3** / **3**				
o	**4** / **4**				

		s	n	o	w
	0	**1** **1**	**2** **2**	**3** **3**	**4** **4**
o	**1** **1**	**1** 2 2 **1**	**2** 3 2 **2**	**2** 4 3 **2**	4 5 **3** **3**
s	**2** **2**	**1** 2 3 **1**	**2** 3 2 **2**	**3** **3** **3** **3**	**3** 4 4 **3**
l	**3** **3**	3 2 4 ?			
o	**4** **4**				

			s		n		o		w	
		0	**1**	**1**	**2**	**2**	**3**	**3**	**4**	**4**
o		**1**	**1**	2	**2**	3	**2**	4	4	5
		1	2	**1**	**2**	**2**	3	**2**	**3**	**3**
s		**2**	**1**	2	**2**	3	**3**	**3**	**3**	4
		2	3	**1**	**2**	**2**	**3**	**3**	4	**3**
l		**3**	3	**2**						
		3	4	**2**						
o		**4**								
		4								

		s	n	o	w
	0	**1** **1**	**2** **2**	**3** **3**	**4** **4**
o	**1** / **1**	**1** 2 / 2 **1**	**2** 3 / **2** **2**	**2** 4 / 3 **2**	4 5 / **3** **3**
s	**2** / **2**	**1** 2 / 3 **1**	**2** 3 / **2** **2**	**3** **3** / **3** **3**	**3** 4 / 4 **3**
l	**3** / **3**	3 **2** / 4 **2**	2 3 / 3 ?		
o	**4** / **4**				

			s		n		o		w	
		0	**1**	**1**	**2**	**2**	**3**	**3**	**4**	**4**
o		**1**	**1**	2	**2**	3	**2**	4	4	5
		1	2	**1**	**2**	**2**	3	**2**	**3**	**3**
s		**2**	**1**	2	**2**	3	**3**	**3**	**3**	4
		2	3	**1**	**2**	**2**	**3**	**3**	4	**3**
l		**3**	3	**2**	**2**	3				
		3	4	**2**	3	**2**				
o		**4**								
		4								

			s		n		o		w	
		0	**1**	**1**	**2**	**2**	**3**	**3**	**4**	**4**
o		**1**	**1**	2	**2**	3	**2**	4	4	5
		1	2	**1**	**2**	**2**	3	**2**	**3**	**3**
s		**2**	**1**	2	**2**	3	**3**	**3**	**3**	4
		2	3	**1**	**2**	**2**	**3**	**3**	4	**3**
l		**3**	3	**2**	**2**	3	3	4		
		3	4	**2**	3	**2**	3	?		
o		**4**								
		4								

		s	n	o	w
	0	**1** **1**	**2** **2**	**3** **3**	**4** **4**
o	**1** / **1**	**1** 2 / 2 **1**	**2** 3 / **2** **2**	**2** 4 / 3 **2**	4 5 / **3** **3**
s	**2** / **2**	**1** 2 / 3 **1**	**2** 3 / **2** **2**	**3** **3** / **3** **3**	**3** 4 / 4 **3**
l	**3** / **3**	3 **2** / 4 **2**	**2** 3 / 3 **2**	**3** 4 / **3** **3**	
o	**4** / **4**				

			s		n		o		w	
		0	**1**	**1**	**2**	**2**	**3**	**3**	**4**	**4**
o		**1**	**1**	2	**2**	3	**2**	4	4	5
		1	2	**1**	**2**	**2**	3	**2**	**3**	**3**
s		**2**	**1**	2	**2**	3	**3**	**3**	**3**	4
		2	3	**1**	**2**	**2**	**3**	**3**	4	**3**
l		**3**	3	**2**	**2**	3	**3**	4	4	4
		3	4	**2**	3	**2**	**3**	**3**	4	?
o		**4**								
		4								

		s	n	o	w
	0	**1** **1**	**2** **2**	**3** **3**	**4** **4**
o	**1** / **1**	**1** 2 / 2 **1**	**2** 3 / **2** **2**	**2** 4 / 3 **2**	4 5 / **3** **3**
s	**2** / **2**	**1** 2 / 3 **1**	**2** 3 / **2** **2**	**3** **3** / **3** **3**	**3** 4 / 4 **3**
l	**3** / **3**	3 **2** / 4 **2**	**2** 3 / 3 **2**	**3** 4 / **3** **3**	**4** **4** / **4** **4**
o	**4** / **4**				

			s		n		o		w	
		0	**1**	**1**	**2**	**2**	**3**	**3**	**4**	**4**
o		**1**	**1**	2	**2**	3	**2**	4	4	5
		1	2	**1**	**2**	**2**	3	**2**	**3**	**3**
s		**2**	**1**	2	**2**	3	**3**	**3**	**3**	4
		2	3	**1**	**2**	**2**	**3**	**3**	4	**3**
l		**3**	3	**2**	**2**	3	**3**	4	**4**	**4**
		3	4	**2**	3	**2**	**3**	**3**	**4**	**4**
o		**4**	4	3						
		4	5	?						

		s	n	o	w
	0	**1** **1**	**2** **2**	**3** **3**	**4** **4**
o	**1** / **1**	**1** 2 / 2 **1**	**2** 3 / **2** **2**	**2** 4 / 3 **2**	4 5 / **3** **3**
s	**2** / **2**	**1** 2 / 3 **1**	**2** 3 / **2** **2**	**3** **3** / **3** **3**	**3** 4 / 4 **3**
l	**3** / **3**	3 **2** / 4 **2**	**2** 3 / 3 **2**	**3** 4 / **3** **3**	**4** **4** / **4** **4**
o	**4** / **4**	4 **3** / 5 **3**			

			s		n		o		w	
		0	**1**	**1**	**2**	**2**	**3**	**3**	**4**	**4**
o		**1**	**1**	2	**2**	3	**2**	4	4	5
		1	2	**1**	**2**	**2**	3	**2**	**3**	**3**
s		**2**	**1**	2	**2**	3	**3**	**3**	**3**	4
		2	3	**1**	**2**	**2**	**3**	**3**	4	**3**
l		**3**	3	**2**	**2**	3	**3**	4	**4**	**4**
		3	4	**2**	3	**2**	**3**	**3**	**4**	**4**
o		**4**	4	**3**	3	3				
		4	5	**3**	4	?				

		s	n	o	w
	0	**1** **1**	**2** **2**	**3** **3**	**4** **4**
o	**1** / **1**	**1** 2 / 2 **1**	**2** 3 / **2** **2**	**2** 4 / 3 **2**	4 5 / **3** **3**
s	**2** / **2**	**1** 2 / 3 **1**	**2** 3 / **2** **2**	**3** **3** / **3** **3**	**3** 4 / 4 **3**
l	**3** / **3**	3 **2** / 4 **2**	**2** 3 / 3 **2**	**3** 4 / **3** **3**	**4** **4** / **4** **4**
o	**4** / **4**	4 **3** / 5 **3**	**3** **3** / 4 **3**		

		s	n	o	w
	0	**1** **1**	**2** **2**	**3** **3**	**4** **4**
o	**1** **1**	**1** 2 / 2 **1**	**2** 3 / **2** **2**	**2** 4 / 3 **2**	4 5 / **3** **3**
s	**2** **2**	**1** 2 / 3 **1**	**2** 3 / **2** **2**	**3** **3** / **3** **3**	**3** 4 / 4 **3**
l	**3** **3**	3 **2** / 4 **2**	**2** 3 / 3 **2**	**3** 4 / **3** **3**	**4** **4** / **4** **4**
o	**4** **4**	4 **3** / 5 **3**	**3** **3** / 4 **3**	2 4 / 4 ?	

		s	n	o	w
	0	**1** 1	**2** 2	**3** 3	**4** 4
o	**1** / **1**	**1** 2 / 2 **1**	**2** 3 / **2** **2**	**2** 4 / 3 **2**	4 5 / **3** **3**
s	**2** / **2**	**1** 2 / 3 **1**	**2** 3 / **2** **2**	**3** **3** / **3** **3**	**3** 4 / 4 **3**
l	**3** / **3**	3 **2** / 4 **2**	**2** 3 / 3 **2**	**3** 4 / **3** **3**	**4** **4** / **4** **4**
o	**4** / **4**	4 **3** / 5 **3**	**3** **3** / 4 **3**	**2** 4 / 4 **2**	

		s	n	o	w
	0	**1** **1**	**2** **2**	**3** **3**	**4** **4**
o	**1** / **1**	**1** 2 / 2 **1**	**2** 3 / **2** **2**	**2** 4 / 3 **2**	4 5 / **3** **3**
s	**2** / **2**	**1** 2 / 3 **1**	**2** 3 / **2** **2**	**3** **3** / **3** **3**	**3** 4 / 4 **3**
l	**3** / **3**	3 **2** / 4 **2**	**2** 3 / 3 **2**	**3** 4 / **3** **3**	**4** **4** / **4** **4**
o	**4** / **4**	4 **3** / 5 **3**	**3** **3** / 4 **3**	**2** 4 / 4 **2**	4 5 / 3 ?

		s	n	o	w
	0	**1** **1**	**2** **2**	**3** **3**	**4** **4**
o	**1** / **1**	**1** 2 / 2 **1**	**2** 3 / **2** **2**	**2** 4 / 3 **2**	4 5 / **3** **3**
s	**2** / **2**	**1** 2 / 3 **1**	**2** 3 / **2** **2**	**3** **3** / **3** **3**	**3** 4 / 4 **3**
l	**3** / **3**	3 **2** / 4 **2**	**2** 3 / 3 **2**	**3** 4 / **3** **3**	**4** **4** / **4** **4**
o	**4** / **4**	4 **3** / 5 **3**	**3** **3** / 4 **3**	**2** 4 / 4 **2**	4 5 / **3** **3**

		s	n	o	w
	0	**1** **1**	**2** **2**	**3** **3**	**4** **4**
o	**1** / **1**	**1** 2 / 2 **1**	**2** 3 / **2** **2**	**2** 4 / 3 **2**	4 5 / **3** **3**
s	**2** / **2**	**1** 2 / 3 **1**	**2** 3 / **2** **2**	**3** **3** / **3** **3**	**3** 4 / 4 **3**
l	**3** / **3**	3 **2** / 4 **2**	**2** 3 / 3 **2**	**3** 4 / **3** **3**	**4** **4** / **4** **4**
o	**4** / **4**	4 **3** / 5 **3**	**3** **3** / 4 **3**	**2** 4 / 4 **2**	4 5 / 3 **3**

		s	n	o	w
	0	**1** **1**	**2** **2**	**3** **3**	**4** **4**
o	**1** / **1**	**1** 2 / 2 **1**	**2** 3 / **2** **2**	**2** 4 / 3 **2**	4 5 / **3** **3**
s	**2** / **2**	**1** 2 / 3 **1**	**2** 3 / **2** **2**	**3** **3** / **3** **3**	**3** 4 / 4 **3**
l	**3** / **3**	3 **2** / 4 **2**	**2** 3 / 3 **2**	**3** 4 / **3** **3**	**4** **4** / **4** **4**
o	**4** / **4**	4 **3** / 5 **3**	**3** **3** / 4 **3**	**2** 4 / 4 **2**	4 5 / **3** **3**

Como eu leio as operações de edição que transformam OSLO into SNOW?

			s		n		o		w	
		0	**1**	**1**	**2**	**2**	**3**	**3**	**4**	**4**
o	**1**	**1**	**1**	2	**2**	3	**2**	4	4	5
	1	2	**1**		**2**	**2**	3	**2**	**3**	**3**
s	**2**	**1**	2		**2**	3	**3**	**3**	**3**	4
	2	3	**1**		**2**	**2**	**3**	**3**	4	**3**
l	**3**	3	**2**		**2**	3	**3**	4	**4**	**4**
	3	4	**2**		3	**2**	**3**	**3**	**4**	**4**
o	**4**	4	**3**		**3**	**3**	**2**	4	4	5
	4	5	**3**		4	**3**	4	**2**	**3**	**3**

custo	operação	entrada	saída
1	inserção	*	w

		s	n	o	w
	0	**1** 1	**2** 2	**3** 3	**4** 4
o	**1** / **1**	**1** 2 / 2 **1**	**2** 3 / **2** **2**	**2** 4 / 3 **2**	4 5 / **3** **3**
s	**2** / **2**	**1** 2 / 3 **1**	**2** 3 / **2** **2**	**3** 3 / **3** **3**	**3** 4 / 4 **3**
l	**3** / **3**	3 **2** / 4 **2**	**2** 3 / 3 **2**	**3** 4 / **3** **3**	**4** **4** / **4** **4**
o	**4** / **4**	4 **3** / 5 **3**	**3** **3** / 4 **3**	**2** 4 / 4 **2**	4 5 / **3** **3**

custo	operação	entrada	saída
0	(cópia)	o	o
1	inserção	*	w

		s	n	o	w
	0	**1** **1**	**2** **2**	**3** **3**	**4** **4**
o	**1** / **1**	**1** 2 / 2 **1**	**2** 3 / **2** **2**	**2** 4 / 3 **2**	4 5 / **3** **3**
s	**2** / **2**	**1** 2 / 3 **1**	**2** 3 / **2** **2**	**3** **3** / **3** **3**	**3** 4 / 4 **3**
l	**3** / **3**	3 **2** / 4 **2**	**2** 3 / 3 **2**	**3** 4 / **3** **3**	**4** **4** / **4** **4**
o	**4** / **4**	4 **3** / 5 **3**	**3** **3** / 4 **3**	**2** 4 / 4 **2**	4 5 / **3** **3**

custo	operação	entrada	saída
1	substituição	l	n
0	(cópia)	o	o
1	inserção	*	w

		s		n		o		w	
		0	**1**	**1**	**2**	**2**	**3**	**3**	**4**	**4**
o	**1**	**1**	2	**2**	3	**2**	4	4	5
	1	2	**1**	**2**	**2**	3	**2**	**3**	**3**
s	**2**	**1**	2	**2**	3	**3**	**3**	**3**	4
	2	3	**1**	**2**	**2**	**3**	**3**	4	**3**
l	**3**	3	**2**	**2**	3	**3**	4	**4**	**4**
	3	4	**2**	3	**2**	**3**	**3**	**4**	**4**
o	**4**	4	**3**	**3**	**3**	**2**	4	4	5
	4	5	**3**	4	**3**	4	**2**	**3**	**3**

custo	operação	entrada	saída
0	(cópia)	s	s
1	substituição	l	n
0	(cópia)	o	o
1	inserção	*	w

		s		n		o		w	
		0	**1**	**1**	**2**	**2**	**3**	**3**	**4**	**4**
o		**1**	**1**	2	**2**	3	**2**	4	4	5
		1	2	**1**	**2**	**2**	3	**2**	**3**	**3**
s		**2**	**1**	2	**2**	3	**3**	**3**	**3**	4
		2	3	**1**	**2**	**2**	**3**	**3**	4	**3**
l		**3**	3	**2**	**2**	3	**3**	4	**4**	**4**
		3	4	**2**	3	**2**	**3**	**3**	**4**	**4**
o		**4**	4	**3**	**3**	**3**	**2**	4	4	5
		4	5	**3**	4	**3**	4	**2**	**3**	**3**

custo	operação	entrada	saída
1	deleção	o	*
0	(cópia)	s	s
1	substituição	l	n
0	(cópia)	o	o
1	inserção	*	w

			c		a		t		c		a		t	
		0	**1**	**1**	**2**	**2**	**3**	**3**	**4**	**4**	**5**	**5**	**6**	**6**
c	**1**	**1**	**0**	2	2	3	3	4	**3**	5	5	6	6	7
	1		2	**0**	**1**	**1**	**2**	**2**	**3**	**3**	**4**	**4**	**5**	**5**
a	**2**	**2**	2	**1**	**0**	2	2	3	3	4	**3**	5	5	6
	2		3	**1**	2	**0**	**1**	**1**	**2**	**2**	**3**	**3**	**4**	**4**
t	**3**	**3**	3	**2**	2	**1**	**0**	2	2	3	3	4	**3**	5
	3		4	**2**	3	**1**	2	**0**	**1**	**1**	**2**	**2**	**3**	**3**

		c	a	t	c	a	t
	0	**1** **1**	**2** **2**	**3** **3**	**4** **4**	**5** **5**	**6** **6**
c	**1** **1**	**0** 2 2 **0**	2 3 **1** **1**	3 4 **2** **2**	**3** 5 **3** **3**	5 6 **4** **4**	6 7 **5** **5**
a	**2** **2**	2 **1** 3 **1**	**0** 2 2 **0**	2 3 **1** **1**	3 4 **2** **2**	**3** 5 **3** **3**	5 6 **4** **4**
t	**3** **3**	3 **2** 4 **2**	2 **1** 3 **1**	**0** 2 2 **0**	2 3 **1** **1**	3 4 **2** **2**	**3** 5 **3** **3**

custo	operação	entrada	saída
1	inserção	*	c
1	inserção	*	a
1	inserção	*	t
0	(cópia)	c	c
0	(cópia)	a	a
0	(cópia)	t	t

		c	a	t	c	a	t
	0	**1** **1**	**2** **2**	**3** **3**	**4** **4**	**5** **5**	**6** **6**
c	**1** **1**	**0** 2 2 **0**	2 3 **1** **1**	3 4 **2** **2**	**3** 5 **3** **3**	5 6 **4** **4**	6 7 **5** **5**
a	**2** **2**	2 **1** 3 **1**	**0** 2 2 **0**	2 3 **1** **1**	3 4 **2** **2**	**3** 5 **3** **3**	5 6 **4** **4**
t	**3** **3**	3 **2** 4 **2**	2 **1** 3 **1**	**0** 2 2 **0**	2 3 **1** **1**	3 4 **2** **2**	**3** 5 **3** **3**

custo	operação	entrada	saída
0	(cópia)	c	c
1	inserção	*	a
1	inserção	*	t
1	inserção	*	c
0	(cópia)	a	a
0	(cópia)	t	t

			c		a		t		c		a		t	
		0	**1**	**1**	**2**	**2**	**3**	**3**	**4**	**4**	**5**	**5**	**6**	**6**
c	**1**	**1**	**0**	2	2	3	3	4	**3**	5	5	6	6	7
	1		2	**0**	**1**	**1**	**2**	**2**	**3**	**3**	**4**	**4**	**5**	**5**
a	**2**	**2**	2	**1**	**0**	2	2	3	3	4	**3**	5	5	6
	2		3	**1**	2	**0**	**1**	**1**	**2**	**2**	**3**	**3**	**4**	**4**
t	**3**	**3**	3	**2**	2	**1**	**0**	2	2	3	3	4	**3**	5
	3		4	**2**	3	**1**	2	**0**	**1**	**1**	**2**	**2**	**3**	**3**

custo	operação	entrada	saída
0	(cópia)	c	c
0	(cópia)	a	a
1	inserção	*	t
1	inserção	*	c
1	inserção	*	a
0	(cópia)	t	t

			c		a		t		c		a		t	
		0	**1**	**1**	**2**	**2**	**3**	**3**	**4**	**4**	**5**	**5**	**6**	**6**
c	**1**	**1**	**0**	2	2	3	3	4	**3**	5	5	6	6	7
	1	2	**0**		**1**	**1**	**2**	**2**	**3**	**3**	**4**	**4**	**5**	**5**
a	**2**	**2**	2	**1**	**0**	2	2	3	3	4	**3**	5	5	6
	2	3	**1**		2	**0**	**1**	**1**	**2**	**2**	**3**	**3**	**4**	**4**
t	**3**	**3**	3	**2**	2	**1**	**0**	2	2	3	3	4	**3**	5
	3	4	**2**		3	**1**	2	**0**	**1**	**1**	**2**	**2**	**3**	**3**

custo	operação	entrada	saída
0	(cópia)	c	c
0	(cópia)	a	a
0	(cópia)	t	t
1	inserção	*	c
1	inserção	*	a
1	inserção	*	t

Correção ortográfica

Correção ortográfica

- Agora que podemos computar a distância de edição, como podemos usá-la para correção de palavras isoladas?

Correção ortográfica

- Agora que podemos computar a distância de edição, como podemos usá-la para correção de palavras isoladas?
- Índices de k-gramas para correção ortográfica de palavras isoladas.

Correção ortográfica

- Agora que podemos computar a distância de edição, como podemos usá-la para correção de palavras isoladas?
- Índices de k-gramas para correção ortográfica de palavras isoladas.
- Correção sensível ao contexto

Correção ortográfica

- Agora que podemos computar a distância de edição, como podemos usá-la para correção de palavras isoladas?
- Índices de k-gramas para correção ortográfica de palavras isoladas.
- Correção sensível ao contexto
- Questões gerais

Índices de k-gramas para correção ortográfica

Índices de k-gramas para correção ortográfica

- Enumera-se todos os k-gramas no termo de consulta

Índices de k-gramas para correção ortográfica

- Enumera-se todos os k-gramas no termo de consulta
- Exemplo: Indice de bigramas, Palavras errada: *bordroom*

Índices de *k*-gramas para correção ortográfica

- Enumera-se todos os *k*-gramas no termo de consulta
- Exemplo: Indice de bigramas, Palavras errada: *bordroom*
- Bigramas: *bo*, *or*, *rd*, *dr*, *ro*, *oo*, *om*

Índices de *k*-gramas para correção ortográfica

- Enumera-se todos os *k*-gramas no termo de consulta
- Exemplo: Indice de bigramas, Palavras errada: *bordroom*
- Bigramas: *bo*, *or*, *rd*, *dr*, *ro*, *oo*, *om*
- Usa-se o índice de *k*-gramas para recuperar palavras "corretas" que correspondem aos *k*-gramas do termo de consulta

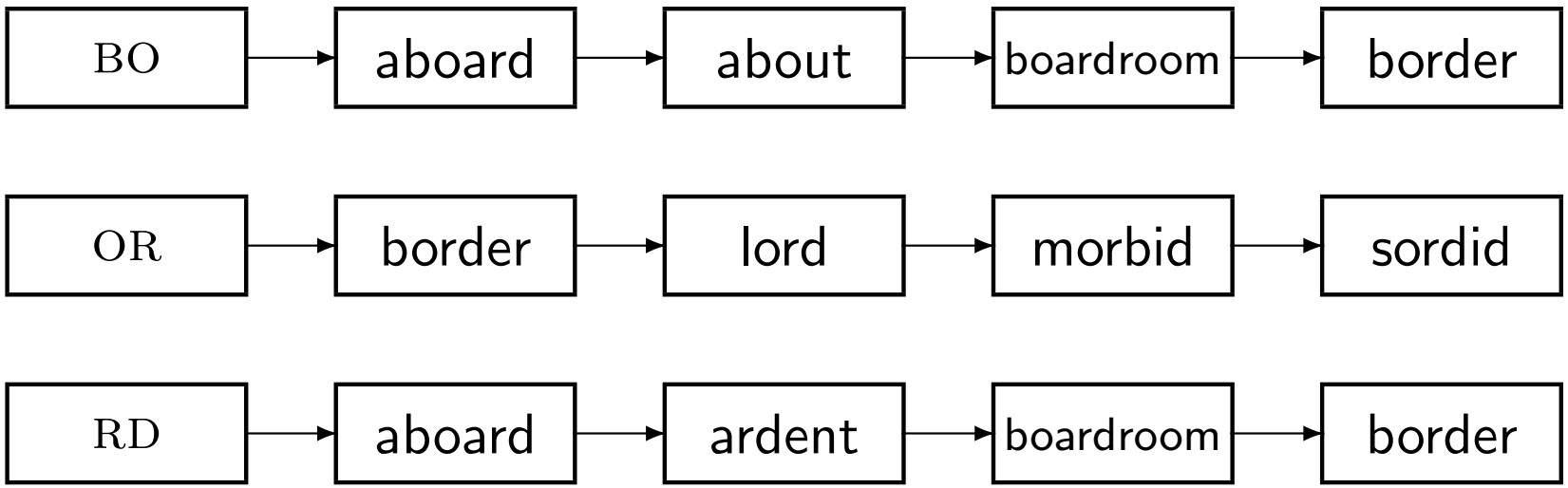
Índices de *k*-gramas para correção ortográfica

- Enumera-se todos os *k*-gramas no termo de consulta
- Exemplo: Indice de bigramas, Palavras errada: *bordroom*
- Bigramas: *bo*, *or*, *rd*, *dr*, *ro*, *oo*, *om*
- Usa-se o índice de *k*-gramas para recuperar palavras "corretas" que correspondem aos *k*-gramas do termo de consulta
- Define-se um limiar para o número de *k*-gramas correspondentes

Índices de *k*-gramas para correção ortográfica

- Enumera-se todos os *k*-gramas no termo de consulta
- Exemplo: Indice de bigramas, Palavras errada: *bordroom*
- Bigramas: *bo, or, rd, dr, ro, oo, om*
- Usa-se o índice de *k*-gramas para recuperar palavras "corretas" que correspondem aos *k*-gramas do termo de consulta
- Define-se um limiar para o número de *k*-gramas correspondentes
- P. ex., apenas termos de vocabulário que diferem por, no máximo 3 *k*-gramas

Índices de *k*-gramas para correção ortográfica: *bordroom*



Correção ortográfica sensível ao contexto

Correção ortográfica sensível ao contexto

- Nosso exemplo era: *an asteroid that fell form the sky*

Correção ortográfica sensível ao contexto

- Nosso exemplo era: *an asteroid that fell form the sky*
- Como podemos corrigir *form* aqui?

Correção ortográfica sensível ao contexto

- Nosso exemplo era: *an asteroid that fell form the sky*
- Como podemos corrigir *form* aqui?
- Uma idéia: correção ortográfica baseada em frequência

Correção ortográfica sensível ao contexto

- Nosso exemplo era: *an asteroid that fell form the sky*
- Como podemos corrigir *form* aqui?
- Uma idéia: correção ortográfica baseada em frequência
  - Recupere termos "corretos" próximos a cada termo de consulta

Correção ortográfica sensível ao contexto

- Nosso exemplo era: *an asteroid that fell form the sky*
- Como podemos corrigir *form* aqui?
- Uma idéia: correção ortográfica baseada em frequência
  - Recupere termos "corretos" próximos a cada termo de consulta
  - para *flew form munich*: *flea* para *flew*, *from* para *form*, *munch* para *munich*

Correção ortográfica sensível ao contexto

- Nosso exemplo era: *an asteroid that fell form the sky*
- Como podemos corrigir *form* aqui?
- Uma idéia: correção ortográfica baseada em frequência
  - Recupere termos "corretos" próximos a cada termo de consulta
  - para *flew form munich*: *flea* para *flew*, *from* para *form*, *munch* para *munich*
  - Agora tente todas as frases resultantes como consultas com uma palavra "fixa" de cada vez.

Correção ortográfica sensível ao contexto

- Nosso exemplo era: *an asteroid that fell form the sky*
- Como podemos corrigir *form* aqui?
- Uma idéia: correção ortográfica baseada em frequência
  - Recupere termos "corretos" próximos a cada termo de consulta
  - para *flew form munich*: *flea* para *flew*, *from* para *form*, *munch* para *munich*
  - Agora tente todas as frases resultantes como consultas com uma palavra "fixa" de cada vez.
  - Tente a consulta *"flea form munich"*

Correção ortográfica sensível ao contexto

- Nosso exemplo era: *an asteroid that fell form the sky*
- Como podemos corrigir *form* aqui?
- Uma idéia: correção ortográfica baseada em frequência
  - Recupere termos "corretos" próximos a cada termo de consulta
  - para *flew form munich*: *flea* para *flew*, *from* para *form*, *munch* para *munich*
  - Agora tente todas as frases resultantes como consultas com uma palavra "fixa" de cada vez.
  - Tente a consulta *"flea form munich"*
  - Tente a consulta *"flew from munich"*

Correção ortográfica sensível ao contexto

- Nosso exemplo era: *an asteroid that fell form the sky*
- Como podemos corrigir *form* aqui?
- Uma idéia: correção ortográfica baseada em frequência
  - Recupere termos "corretos" próximos a cada termo de consulta
  - para *flew form munich*: *flea* para *flew*, *from* para *form*, *munch* para *munich*
  - Agora tente todas as frases resultantes como consultas com uma palavra "fixa" de cada vez.
  - Tente a consulta *"flea form munich"*
  - Tente a consulta *"flew from munich"*
  - Tente a consulta *"flew form munch"*

Correção ortográfica sensível ao contexto

- Nosso exemplo era: *an asteroid that fell form the sky*
- Como podemos corrigir *form* aqui?
- Uma idéia: correção ortográfica baseada em frequência
  - Recupere termos "corretos" próximos a cada termo de consulta
  - para *flew form munich*: *flea* para *flew*, *from* para *form*, *munch* para *munich*
  - Agora tente todas as frases resultantes como consultas com uma palavra "fixa" de cada vez.
  - Tente a consulta *"flea form munich"*
  - Tente a consulta *"flew from munich"*
  - Tente a consulta *"flew form munch"*
  - A consulta correta *"flew from munich"* apresenta a maior freqûencia.

Correção ortográfica sensível ao contexto

- Nosso exemplo era: *an asteroid that fell form the sky*
- Como podemos corrigir *form* aqui?
- Uma idéia: correção ortográfica baseada em frequência
  - Recupere termos "corretos" próximos a cada termo de consulta
  - para *flew form munich*: *flea* para *flew*, *from* para *form*, *munch* para *munich*
  - Agora tente todas as frases resultantes como consultas com uma palavra "fixa" de cada vez.
  - Tente a consulta *"flea form munich"*
  - Tente a consulta *"flew from munich"*
  - Tente a consulta *"flew form munch"*
  - A consulta correta *"flew from munich"* apresenta a maior freqûencia.
- Suponha que temos 7 alternativas para *flew*, 20 para *form* e 3 para *munich*, quantas frases "corrigidas" iremos enumerar?

Correção ortográfica sensível ao contexto

Correção ortográfica sensível ao contexto

- O algoritmo “baseado em frequência” que delineamos não é muito eficiente.

Correção ortográfica sensível ao contexto

- O algoritmo "baseado em frequência" que delineamos não é muito eficiente.
- Alternativa mais eficiente: Olhe para a "coleção" de consultas e não de documentos

Problemas comuns em correção ortográfica

Problemas comuns em correção ortográfica

- Interface com o usuário

Problemas comuns em correção ortográfica

- Interface com o usuário
  - correção automática vs. sugerida

Problemas comuns em correção ortográfica

- Interface com o usuário
  - correção automática vs. sugerida
  - *Você quis dizer...* só funciona para uma sugestão.

Problemas comuns em correção ortográfica

- Interface com o usuário
  - correção automática vs. sugerida
  - *Você quis dizer...* só funciona para uma sugestão.
  - E quando temos mais de uma correção possível?

Problemas comuns em correção ortográfica

- Interface com o usuário
  - correção automática vs. sugerida
  - *Você quis dizer...* só funciona para uma sugestão.
  - E quando temos mais de uma correção possível?
  - Toma lá da cá: UI simples vs. poderosa

Problemas comuns em correção ortográfica

- Interface com o usuário
  - correção automática vs. sugerida
  - *Você quis dizer...* só funciona para uma sugestão.
  - E quando temos mais de uma correção possível?
  - Toma lá da cá: UI simples vs. poderosa
- Custo

Problemas comuns em correção ortográfica

- Interface com o usuário
  - correção automática vs. sugerida
  - *Você quis dizer...* só funciona para uma sugestão.
  - E quando temos mais de uma correção possível?
  - Toma lá da cá: UI simples vs. poderosa
- Custo
  - Correção ortográfica é potencialmente cara.

Problemas comuns em correção ortográfica

- Interface com o usuário
  - correção automática vs. sugerida
  - *Você quis dizer...* só funciona para uma sugestão.
  - E quando temos mais de uma correção possível?
  - Toma lá da cá: UI simples vs. poderosa
- Custo
  - Correção ortográfica é potencialmente cara.
  - Devemos evitar rodar a cada consulta?

Problemas comuns em correção ortográfica

- Interface com o usuário
  - correção automática vs. sugerida
  - *Você quis dizer...* só funciona para uma sugestão.
  - E quando temos mais de uma correção possível?
  - Toma lá da cá: UI simples vs. poderosa
- Custo
  - Correção ortográfica é potencialmente cara.
  - Devemos evitar rodar a cada consulta?
  - Talvez apenas em consultas que retornam poucos documentos.

Problemas comuns em correção ortográfica

- Interface com o usuário
  - correção automática vs. sugerida
  - *Você quis dizer...* só funciona para uma sugestão.
  - E quando temos mais de uma correção possível?
  - Toma lá da cá: UI simples vs. poderosa
- Custo
  - Correção ortográfica é potencialmente cara.
  - Devemos evitar rodar a cada consulta?
  - Talvez apenas em consultas que retornam poucos documentos.
  - Mais provável: A correção ortográfica na maioria dos principais engines de busca é suficientemente eficiente para ser executada em todas as consultas.

Exercício: Entender corretor ortográfico do Peter Norvig's

```
import re, collections
def words(text): return re.findall('[a-z]+', text.lower())
def train(features):
    model = collections.defaultdict(lambda: 1)
    for f in features:
        model[f] += 1
    return model
NWORDS = train(words(file('big.txt').read()))
alphabet = 'abcdefghijklmnopqrstuvwxyz'
def edits1(word):
   splits     = [(word[:i], word[i:]) for i in range(len(word) + 1)]
   deletes    = [a + b[1:] for a, b in splits if b]
   transposes = [a + b[1] + b[0] + b[2:] for a, b in splits if len(b) gt 1]
   replaces   = [a + c + b[1:] for a, b in splits for c in alphabet if b]
   inserts    = [a + c + b     for a, b in splits for c in alphabet]
   return set(deletes + transposes + replaces + inserts)
def known_edits2(word):
    return set(e2 for e1 in edits1(word) for e2 in edits1(e1) if e2 in NWORDS)
def known(words): return set(w for w in words if w in NWORDS)
def correct(word):
    candidates = known([word]) or known(edits1(word)) or known_edits2(word) or
[word]
    return max(candidates, key=NWORDS.get)
```

Soundex

Soundex

- Soundex é a base para encontrar alternativas fonéticas (ao invés de ortográficas).

Soundex

- Soundex é a base para encontrar alternativas fonéticas (ao invés de ortográficas).
- Exemplo: *chebyshev / tchebyscheff*

Soundex

- Soundex é a base para encontrar alternativas fonéticas (ao invés de ortográficas).
- Exemplo: *chebyshev / tchebyscheff*
- Algoritmo:

Soundex

- Soundex é a base para encontrar alternativas fonéticas (ao invés de ortográficas).
- Exemplo: *chebyshev / tchebyscheff*
- Algoritmo:
  - Converte-se cada token a ser indexado em uma forma reduzida de 4 caracteres

Soundex

- Soundex é a base para encontrar alternativas fonéticas (ao invés de ortográficas).
- Exemplo: *chebyshev / tchebyscheff*
- Algoritmo:
  - Converte-se cada token a ser indexado em uma forma reduzida de 4 caracteres
  - Faz-se o mesmo com o termos de consultas

Soundex

- Soundex é a base para encontrar alternativas fonéticas (ao invés de ortográficas).
- Exemplo: *chebyshev / tchebyscheff*
- Algoritmo:
  - Converte-se cada token a ser indexado em uma forma reduzida de 4 caracteres
  - Faz-se o mesmo com o termos de consultas
  - Constroi-se um índices e busca-se as formas reduzidas

Algoritmo Soundex

1. Retém-se o primeiro caracter do termo.
2. Troca-se todas as ocorrências das seguintes letras por '0' (zero): A, E, I, O, U, H, W, Y
3. Troca-se letras por dígitos da seguinte forma:
   - B, F, P, V por 1
   - C, G, J, K, Q, S, X, Z por 2
   - D,T por 3
   - L por 4
   - M, N por 5
   - R por 6
4. Remove-se repetidamente um de cada par de digitos identicos consecutivos.
5. Remove-se todos os zeros da string resultante; completa-se a string resultante com zeros e retorna-se as primeiras 4 posições, que devem consistir de uma letra seguida por três dígitos.

Exemplo: Soundex de *HERMAN*

Exemplo: Soundex de *HERMAN*

- Retém-se o H

Exemplo: Soundex de *HERMAN*

- Retém-se o H
- *ERMAN* $\rightarrow$ *0RM0N*

Exemplo: Soundex de *HERMAN*

- Retém-se o H
- *ERMAN* $\rightarrow$ *0RM0N*
- *0RM0N* $\rightarrow$ *06505*

Exemplo: Soundex de *HERMAN*

- Retém-se o H
- *ERMAN* $\rightarrow$ *0RM0N*
- *0RM0N* $\rightarrow$ *06505*
- *06505* $\rightarrow$ *06505*

Exemplo: Soundex de *HERMAN*

- Retém-se o H
- *ERMAN* $\rightarrow$ *0RM0N*
- *0RM0N* $\rightarrow$ *06505*
- *06505* $\rightarrow$ *06505*
- *06505* $\rightarrow$ *655*

Exemplo: Soundex de *HERMAN*

- Retém-se o H
- *ERMAN* $\rightarrow$ *0RM0N*
- *0RM0N* $\rightarrow$ *06505*
- *06505* $\rightarrow$ *06505*
- *06505* $\rightarrow$ *655*
- Retorna-se *H655*

Exemplo: Soundex de *HERMAN*

- Retém-se o H
- *ERMAN* $\rightarrow$ *0RM0N*
- *0RM0N* $\rightarrow$ *06505*
- *06505* $\rightarrow$ *06505*
- *06505* $\rightarrow$ *655*
- Retorna-se *H655*
- Note: *HERMANN* irá gerar o mesmo código

Quão útil é o Soundex?

Quão útil é o Soundex?

- Não muito – Para recuperação de informação

Quão útil é o Soundex?

- Não muito – Para recuperação de informação
- Ok para tarefas que requerem "alta revocação" em outros tipos de aplicação (p. ex., Interpol)

Quão útil é o Soundex?

- Não muito – Para recuperação de informação
- Ok para tarefas que requerem "alta revocação" em outros tipos de aplicação (p. ex., Interpol)
- Zobel e Dart (1996) sugerem alternativas melhores para correspondência onética em RI.

Exercício

- Compute o código Soundex do seu último nome