Universidade Federal de Santa Maria
1960
Frederico Westphalen

Modelo Booleano

Tópicos Especiais em Recuperação de Informações

Profª. Solange Pertile

18/09/15

Aula adaptada: Profa. Viviane Pereira (UFRGS)

Relembrando...

Componentes de um sistema RI

Consulta

Processa Texto

Busca nos índices

Recupera documentos

Produz documentos ranqueados

Resultados

Pré-processamento

Tokenization, remoção de stopwords, stemming e peso dos termos

Arquivo invertido

Coleção de documentos

Indexação

Adaptado aulas Prof. Viviane Moreira (UFRGS)

Relembrando...

✓**Operação de Indexação** - envolve a criação de estruturas de dados associados aos documentos de uma coleção. Uma estrutura de dados bastante utilizada são as listas invertidas de termos/documentos.

✓**Operação de Consulta** - envolve a especificação de um conjunto de termos, que representa a necessidade de informação do usuário.

✓**Pesquisa e Ordenação** - envolve o processo de recuperação de documentos de acordo com a consulta do usuário e sua ordenação através de um grau de similaridade entre o documento e a consulta.

Relembrando...

✓**Operação de Indexação** - envolve a criação de estruturas de dados associados aos documentos de uma coleção. Uma estrutura de dados bastante utilizada são as listas invertidas de termos/documentos.

Lista Invertida????

Relembrando...

- **Lista Invertida** *(do inglês inverted list ou inverted index):*
  - É uma estrutura de dados que mapeia termos às suas ocorrências em um documento ou conjunto de documentos, armazenados em um banco de dados.
  - É uma estratégia de indexação que permite a realização de buscas precisas e rápidas, em troca de maior dificuldade no ato de inserção e atualização de documentos.
  - É a mais popular estratégia de sistemas para obtenção de dados, usada em larga escala em sistemas de gerenciamento de bancos de dados (como o **Adabas**) e serviços de busca (como o **Google**).

Universidade Federal de Santa Maria
1960
Frederico Westphalen

Relembrando...

- ✓ **Porque o nome lista invertida????**

Relembrando...

- ✓ **Porque o nome lista invertida????**
  - ✓ Inverte a hierarquia da informação:
    - ✓ ao invés de uma lista de documentos contendo termos, é obtida uma lista de termos, referenciando documentos (através de um identificador único, como uma chave primária).
  - ✓ Outras informações podem ser armazenadas
    - ✓ **Exemplo**: a posição do termo no documento que é útil para uso de algoritmos que calculem a relevância dos resultados utilizando a proximidade de palavras.
  - ✓ Visam trazer resultados de forma **rápida** e **eficiente**.

Relembrando...

- ✓ **Exemplo de lista invertida**

Documentos

1: "Sei que sou"
2: "Sou o que sei“
3: "Sou aquilo que sou“

Lista invertida

"sei" : {1, 2}
"que" : {1, 2, 3}
"sou" : {1, 2, 3}
"o" : {2}
“aquilo" : {3}

Classificação

- ✓ Para calcular uma classificação, o sistema de RI usualmente adota um modelo para representar os documentos e a consulta do usuário.
- ✓ Os três modelos considerados clássicos são:
  - ✓ o modelo booleano,
  - ✓ o modelo vetorial e modelo probabilístico, e
  - ✓ o modelo semântico.

Modelos

- ✓ Para cada modelo, veremos:
  - ✓ A representação do documento
  - ✓ A representação da consulta
  - ✓ A função de busca

Modelo Booleano

- **Representação dos documentos**
  - Dado o conjunto de termos representativos para o corpus em questão (Vocabulário do Sistema)
    - V = {t1, t2,...,tn}
  - Os documentos são representados como conjunto de termos de indexação, sendo tais conjuntos representados como vetores de pesos binários de tamanho *n*
    - Cada posição no vetor corresponde a um termo usado na indexação dos documentos da base
    - Cada valor indica apenas se determinado termo está ou não presente no documento

Modelo Booleano

✓ Representação dos documento

Matriz incidência de termos por documentos

Docs / Termos	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6
Antônio	1	0	1	1	0	1
Brutus	1	0	1	0	1	1
César	1	0	1	0	1	0
Calpurnia	0	0	0	1	0	1
Cleópatra	0	0	0	0	1	0
misericórdia	1	1	0	0	0	0

indica que o termo não está presente no documento

indica que o termo está presente no documento

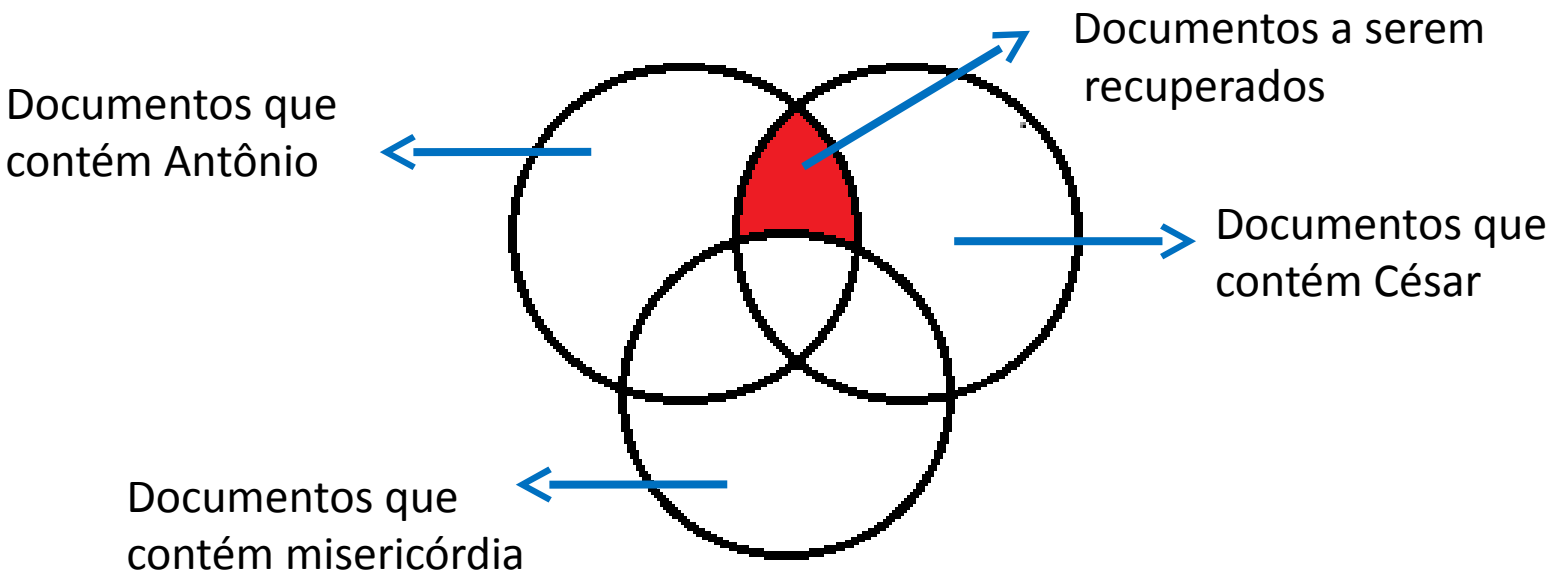
Modelo Booleano

- ✓ **Representação da Consulta**
  - ✓ Baseado na teoria de conjuntos e álgebra booleana
  - ✓ Os termos da consulta são combinados utilizando operadores booleanos como "E", "OU", "NÃO"
  - ✓ Documentos são conjuntos de palavras
  - ✓ É exato: o documento satisfaz ou não o critério da busca
  - ✓ Modelo comercial mais usado por 3 décadas

Modelo Booleano

✓ **Representação da Consulta**

Antônio **AND** César **AND NOT** misericórdia



Modelo Booleano

- ✓ **Representação dos documento**

Antônio **AND** César **AND NOT** misericórdia

Docs / Termos	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6
Antônio	1	0	1	1	0	1
Brutus	1	0	1	0	1	1
César	1	0	1	0	1	0
Calpurnia	0	0	0	1	0	1
Cleópatra	0	0	0	0	1	0
misericórdia	1	1	0	0	0	0

Modelo Booleano

✓ **Representação dos documento**

Antônio **AND** César **AND NOT** misericórdia

101010 and 101101 and 001111 = 001000

Termos \ Docs	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6
Antônio	1	0	1	1	0	1
Brutus	1	0	1	0	1	1
César	1	0	1	0	1	0
Calpurnia	0	0	0	1	0	1
Cleópatra	0	0	0	0	1	0
misericórdia	1	1	0	0	0	0

Modelo Booleano

Qual documento se encontra a consulta
Antônio **AND** César **AND NOT** misericórdia
????

101010 and 101101 and 001111 = 001000

Termos \ Docs	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6
Antônio	1	0	1	1	0	1
Brutus	1	0	1	0	1	1
César	1	0	1	0	1	0
Calpurnia	0	0	0	1	0	1
Cleópatra	0	0	0	0	1	0
misericórdia	1	1	0	0	0	0

Modelo Booleano

Qual documento se encontra a consulta
Antônio **AND** César **AND NOT** misericórdia
????

101010 and 101101 and 001111 = 001000

Doc3

Docs / Termos	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6
Antônio	1	0	1	1	0	1
Brutus	1	0	1	0	1	1
César	1	0	1	0	1	0
Calpurnia	0	0	0	1	0	1
Cleópatra	0	0	0	0	1	0
misericórdia	1	1	0	0	0	0

Modelo Booleano

- ✓ **Representação dos documento**

Quais os documentos retornados para
(misericórdia **OR** Brutus) **AND NOT** Culpurnia

Termos \ Docs	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6
Antônio	1	0	1	1	0	1
Brutus	1	0	1	0	1	1
César	1	0	1	0	1	0
Calpurnia	0	0	0	1	0	1
Cleópatra	0	0	0	0	1	0
misericórdia	1	1	0	0	0	0

Modelo Booleano

✓ **Representação dos documento**

Quais os documentos retornados para
(misericórdia **OR** Brutus) **AND NOT** Culpurnia
(110000 OR 101011) AND 111010 = 111010

Termos \ Docs	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6
Antônio	1	0	1	1	0	1
Brutus	1	0	1	0	1	1
César	1	0	1	0	1	0
Calpurnia	0	0	0	1	0	1
Cleópatra	0	0	0	0	1	0
misericórdia	1	1	0	0	0	0

Universidade Federal de Santa Maria
1960
Frederico Westphalen

Grandes Coleções

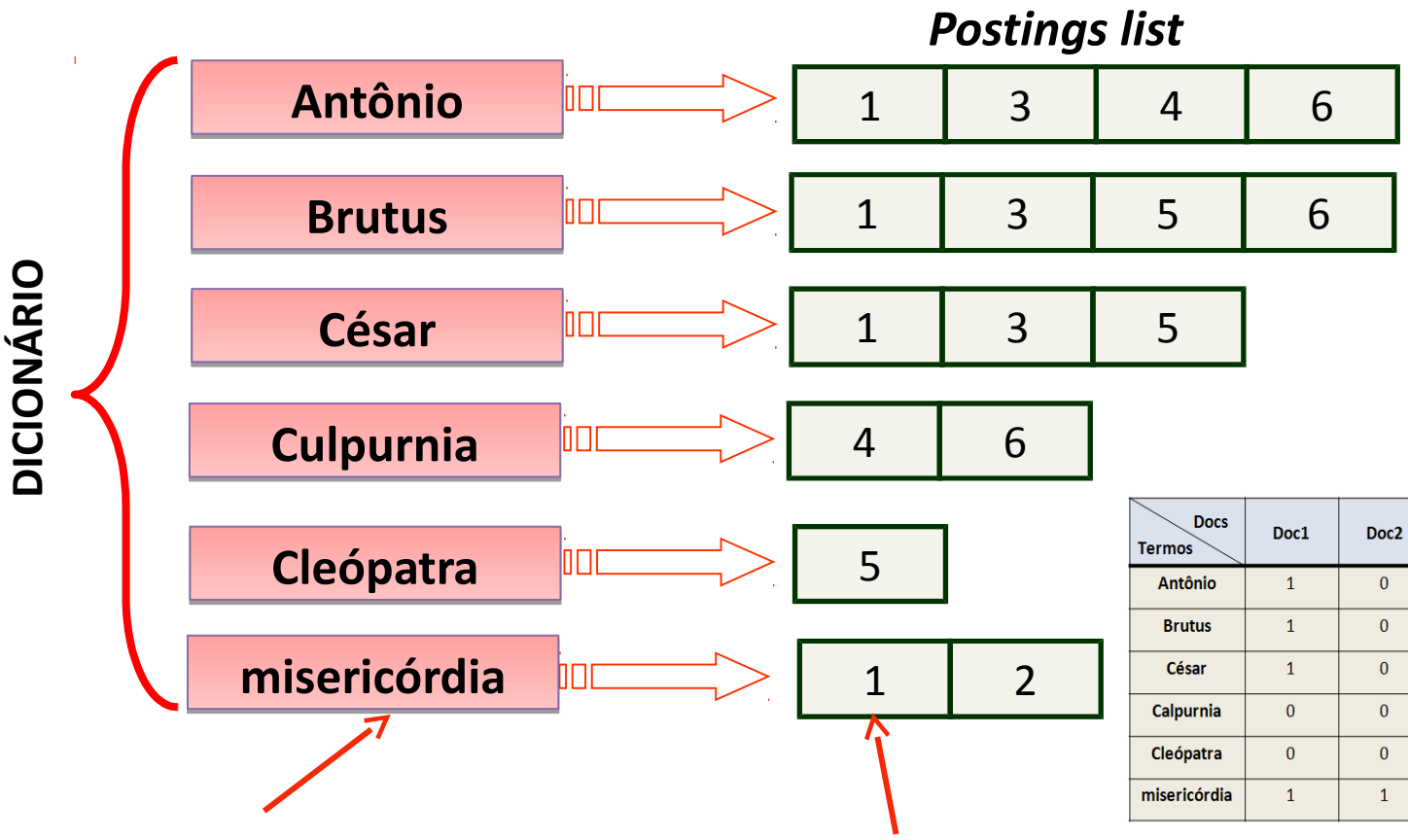
- ✓ Considere N = 1 milhão de documentos, cada um contento 1000 palavras.
- ✓ Considere 6 bytes/palavra incluindo espaço/pontuação
- ✓ 6GB de dados nos documentos.
- ✓ Suponha M = 500K temos distintos.

Dispersão Termos/Documentos

- ✓ A matriz 500K x 1M tem meio trilhão de 0's e 1's.
- ✓ Mas não tem mais de um bilhão de 1's. – A matriz é extremamente esparsa.
- ✓ Qual seria uma representação melhor? – Guardar apenas a posição dos 1.

Consultas com AND

Índices



Termos \ Docs	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6
Antônio	1	0	1	1	0	1
Brutus	1	0	1	0	1	1
César	1	0	1	0	1	0
Calpurnia	0	0	0	1	0	1
Cleópatra	0	0	0	0	1	0
misericórdia	1	1	0	0	0	0

Universidade Federal de Santa Maria
1960
Frederico Westphalen

Consultas com AND

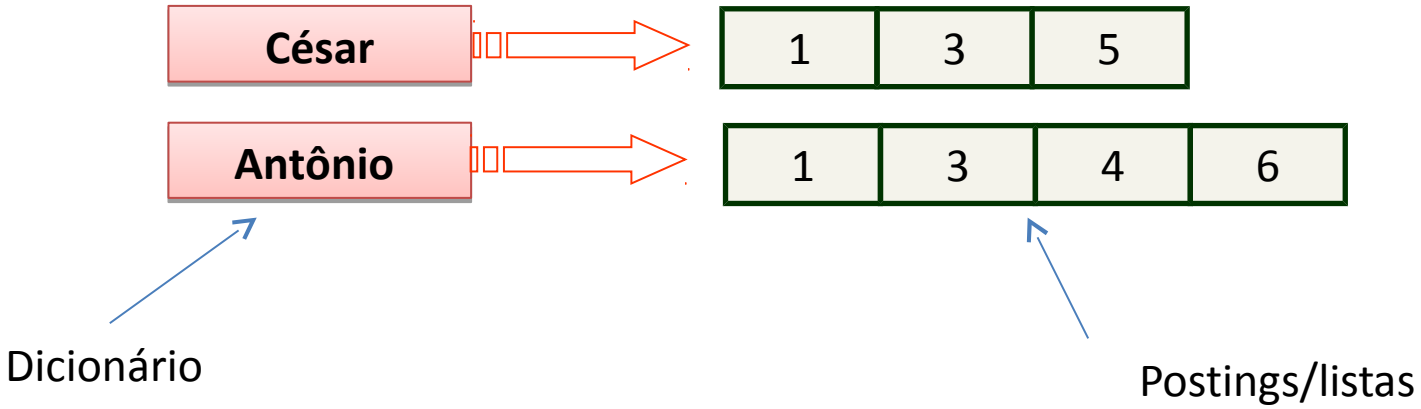
Como gerar a resposta de uma consulta a partir do índice?

Consultas com AND

- ✓ **Consultas utilizando índices**

  - ✓ **AND** = calcula-se a INTERSEÇÃO entre as listas.
  - ✓ **OR** = calcula-se a UNIÃO entre as listas.

César **AND** Antônio



Consultas com AND

1. Localizar “César” no dicionário
2. Recuperar a lista de docs que contém “César”
3. Localizar “Antônio” no dicionário
4. Recuperar a lista de docs que contém “Antônio”
5. Calcular a interseção entre as duas lista

César **AND** Antônio

Universidade Federal de Santa Maria
1960
Frederico Westphalen

Consultas com AND

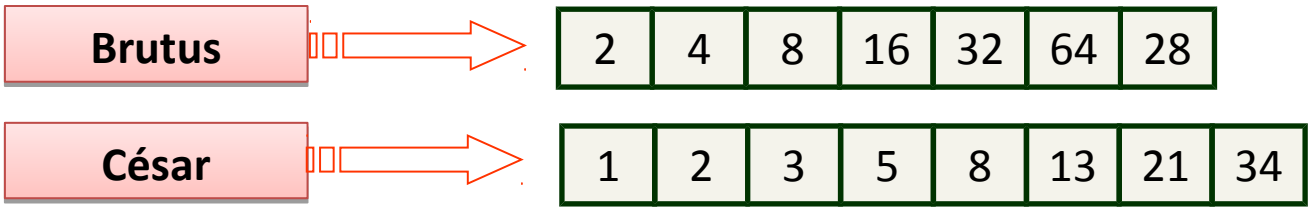
- ✓ **Interseção**
- ✓ Utiliza-se dois ponteiros para percorrer as duas listas simultaneamente
- ✓A cada passo compara-se o docID apontado por cada ponteiro
  - ✓ Se forem iguais – o doc vai para a lista de resultado e os dois ponteiros avançam
  - ✓ Se forem diferentes – o de menor valor avança
- ✓ Quando pelo menos uma das listas terminar, o processo encerra.
- ✓É fundamental que as listas estejam em ordem de docID

Consultas com AND

✓ Exemplo

Brutus **AND** César

1. Localizar “César” no dicionário
2. Recuperar a lista de docs que contém “César”
3. Localizar “Antônio” no dicionário
4. Recuperar a lista de docs que contém “Antônio”



Se as listas têm comprimento x e y, o cálculo da interseção (merge) tem complexidade O(x+y).

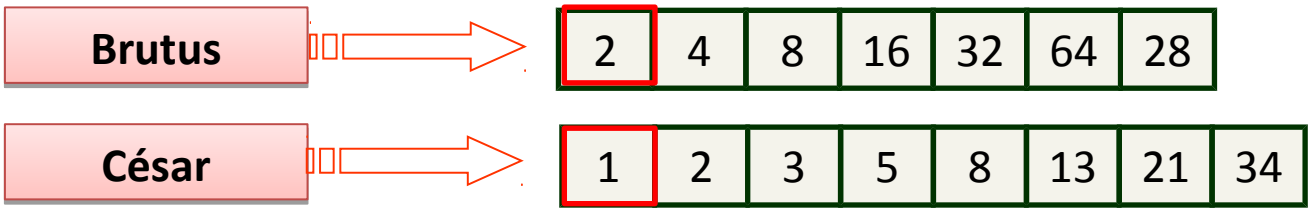
Crucial: listas de *postings* ordenadas pelo docID

Consultas com AND

✓ **Exemplo**

Brutus **AND** César

5. Calcular a interseção entre as duas lista



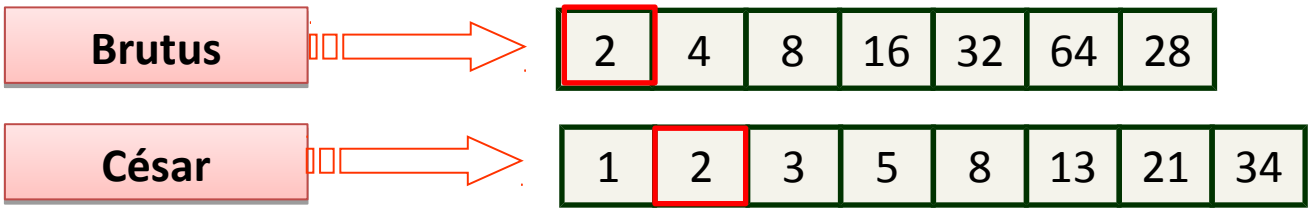
Resultado da Consulta: { }

Consultas com AND

✓ Exemplo

Brutus **AND** César

5. Calcular a interseção entre as duas lista



Resultado da Consulta: { 2}

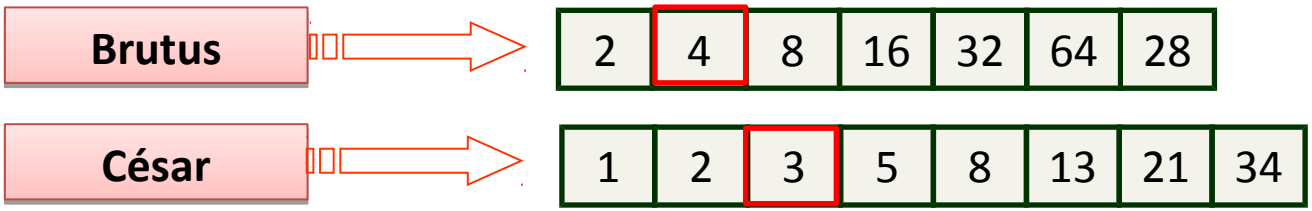
Obs: Se forem iguais – o doc vai para a lista de resultado e os dois ponteiros avançam

Consultas com AND

✓ Exemplo

Brutus **AND** César

5. Calcular a interseção entre as duas lista



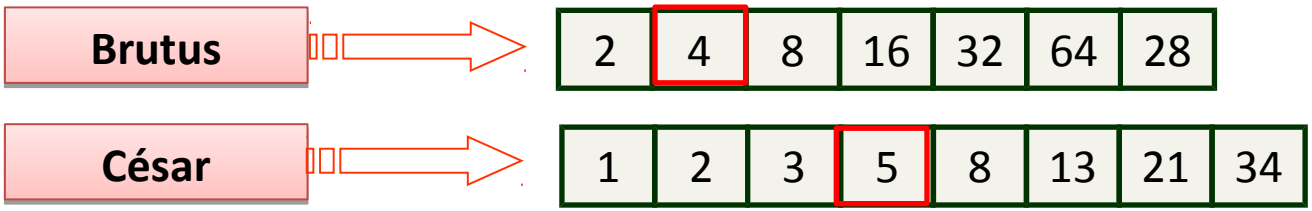
Resultado da Consulta: { 2}

Consultas com AND

✓ Exemplo

Brutus **AND** César

5. Calcular a interseção entre as duas lista



Resultado da Consulta: { 2}

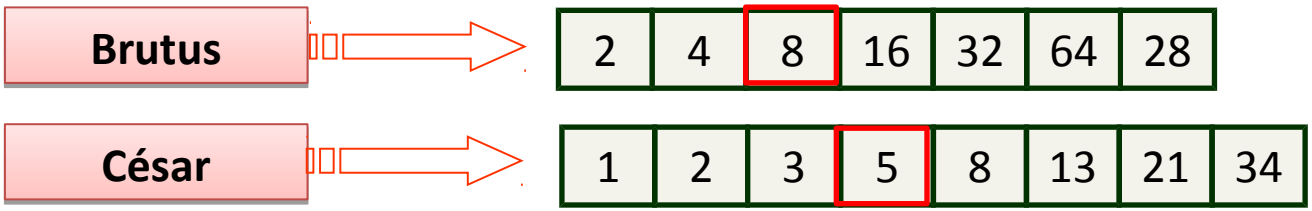
Obs: Se forem diferentes – o de menor valor avança

Consultas com AND

✓ Exemplo

Brutus **AND** César

5. Calcular a interseção entre as duas lista



Resultado da Consulta: { 2}

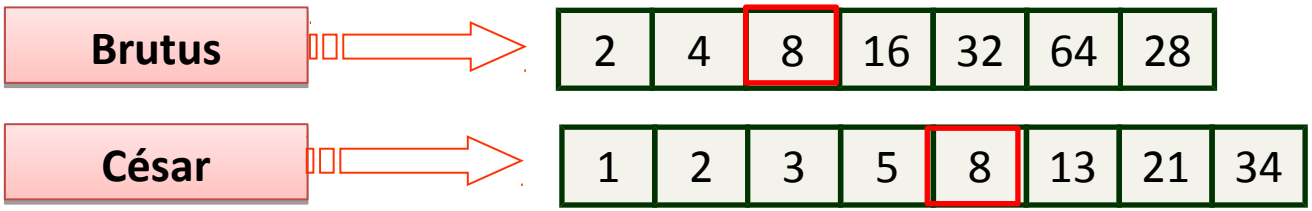
Obs: Se forem diferentes – o de menor valor avança

Consultas com AND

✓ Exemplo

Brutus **AND** César

5. Calcular a interseção entre as duas lista



Resultado da Consulta: { 2,8}

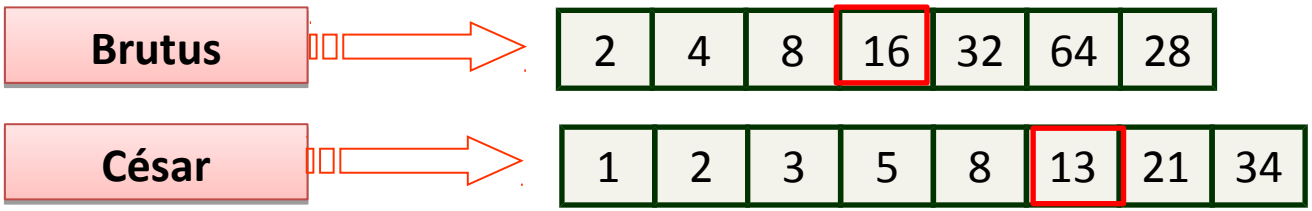
Obs: Se forem iguais – o doc vai para a lista de resultado e os dois ponteiros avançam

Consultas com AND

✓ Exemplo

Brutus **AND** César

5. Calcular a interseção entre as duas lista



Resultado da Consulta: { 2,8}

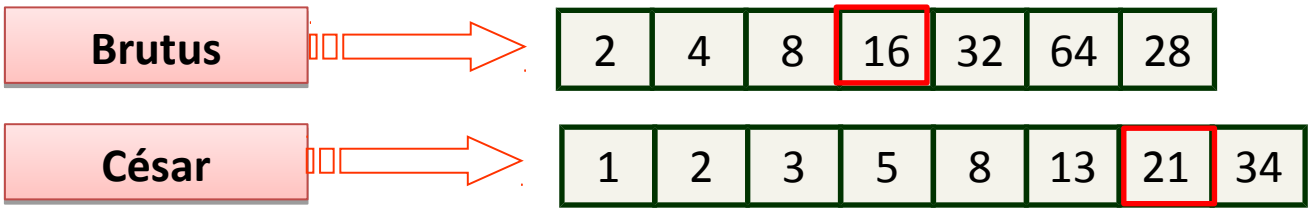
Obs: Se forem diferentes – o de menor valor avança

Consultas com AND

✓ Exemplo

Brutus **AND** César

5. Calcular a interseção entre as duas lista



Resultado da Consulta: { 2,8}

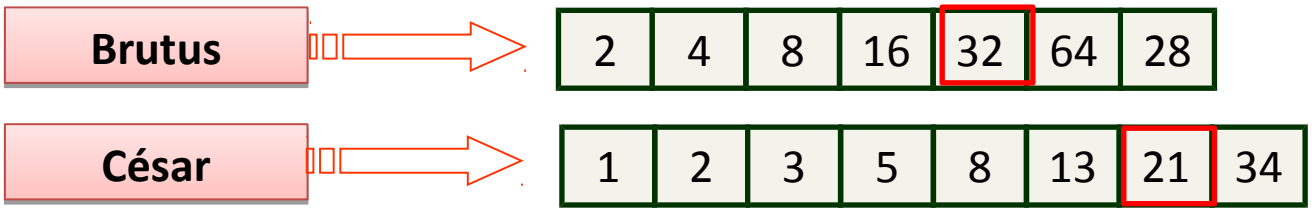
Obs: Se forem diferentes – o de menor valor avança

Consultas com AND

✓ Exemplo

Brutus **AND** César

5. Calcular a interseção entre as duas lista



Resultado da Consulta: { 2,8}

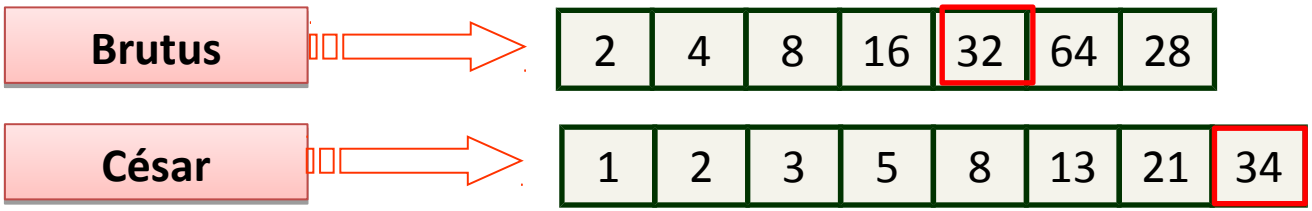
Obs: Se forem diferentes – o de menor valor avança

Consultas com AND

✓ Exemplo

Brutus **AND** César

5. Calcular a interseção entre as duas lista



Resultado da Consulta: { 2,8}

Obs: Quando pelo menos uma das listas terminar, o processo encerra

Consultas com AND

- ✓ **Algoritmo de Interseção**

```
INTERSECT(p1, p2)
 1  answer ← ⟨ ⟩
 2  while p1 ≠ NIL and p2 ≠ NIL
 3  do if docID(p1) = docID(p2)
 4        then ADD(answer, docID(p1))
 5             p1 ← next(p1)
 6             p2 ← next(p2)
 7        else if docID(p1) < docID(p2)
 8                then p1 ← next(p1)
 9                else p2 ← next(p2)
10  return answer
```

Universidade Federal de Santa Maria
1960
Frederico Westphalen

Consultas com OR

- ✓ Mesmo procedimento descrito para AND, obtém-se as listas de *postings* para cada termo da consulta
  - ✓ Calcula-se a **união** entre as listas

Universidade Federal de Santa Maria
1960
Frederico Westphalen

Merge com NOT

Brutus AND **NOT** Caesar
Brutus OR **NOT** Caesar

Ainda é possível obter complexidade O(x+y)???

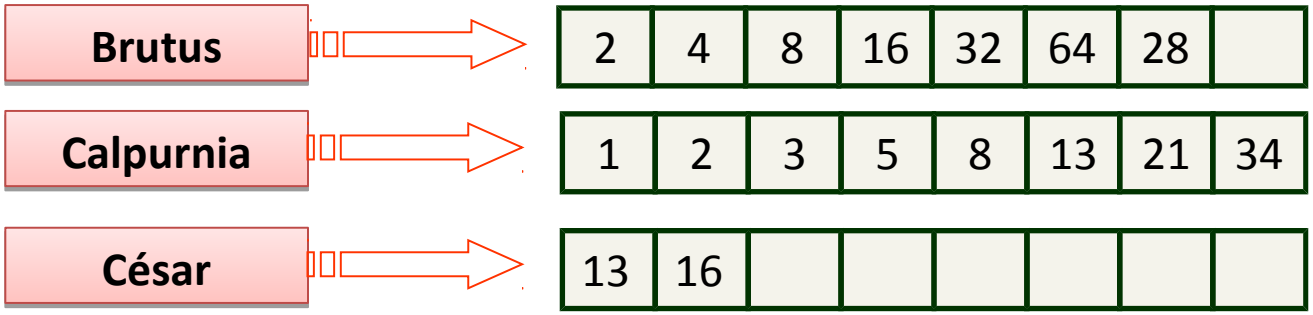
Otimização de Consultas

✓Qual a melhor ordem para o processamento de consultas?

✓ Considere uma consulta que é um AND de *n* termos.

✓Para cada um dos *t* termos, obtenha a lista de *postings* e faça o merge 2 a 2.

Consulta : Brutus AND Calpurnia AND César

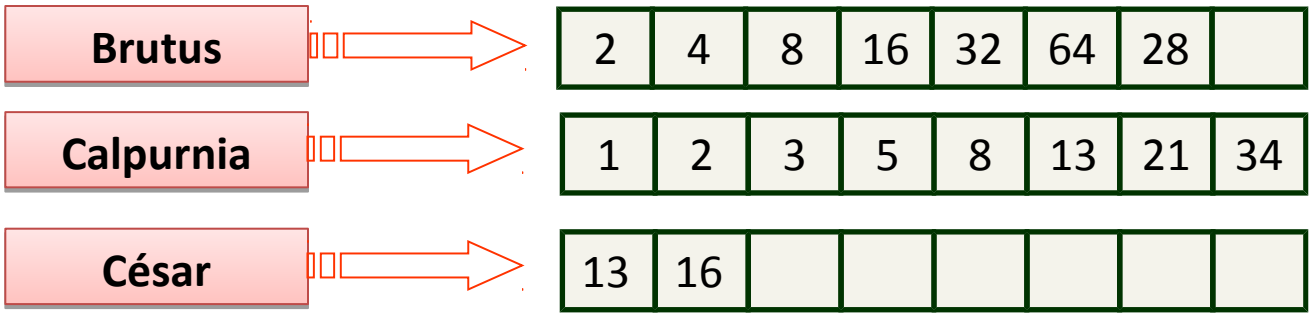


Otimização de Consultas

- ✓Processar em ordem crescente de frequência:
  - ✓Começar com o menos frequente.

Por isso a frequência dos documentos é armazenada no dicionário

Consulta : Brutus AND Calpurnia AND César

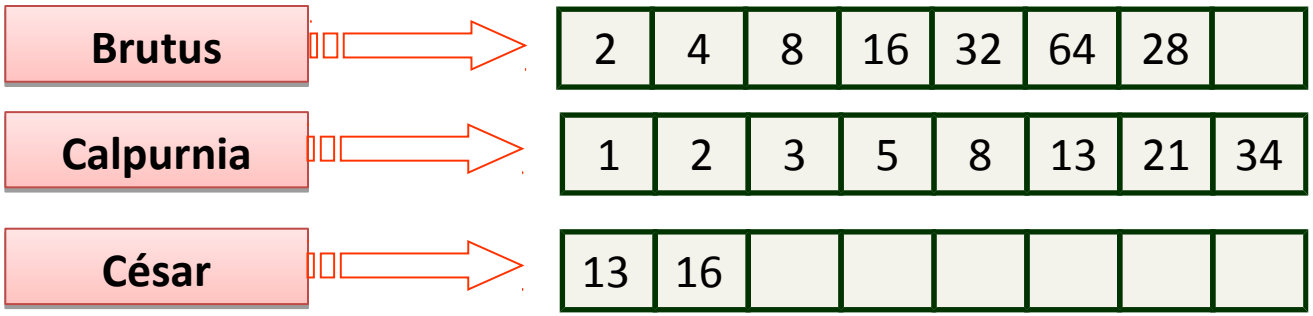


Otimização de Consultas

- ✓Processar em ordem crescente de frequência:
  - ✓Começar com o menos frequente.

Nesse caso, qual a ordem da consulta?

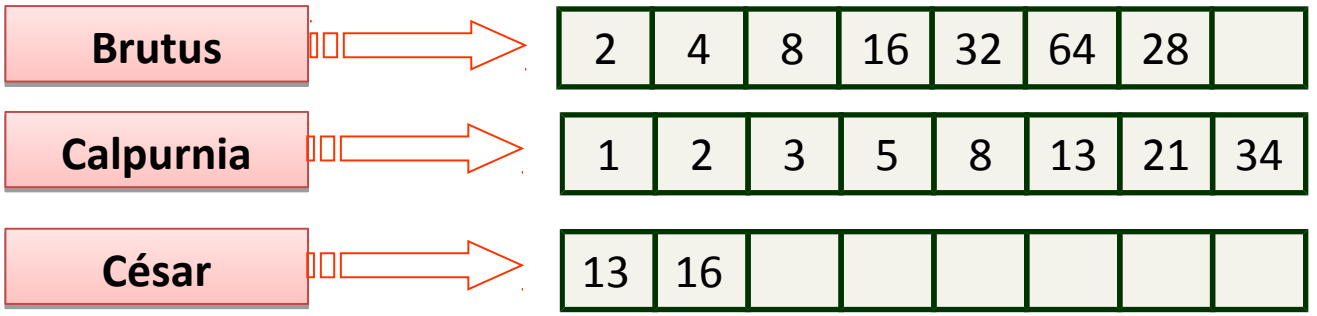
Consulta : Brutus AND Calpurnia AND César



Otimização de Consultas

- ✓Processar em ordem crescente de frequência:
  - ✓Começar com o menos frequente.

Consulta : Brutus AND Calpurnia AND César



Executar a consulta como:

(César **AND** Brutus) **AND** Calpurnia

Universidade Federal de Santa Maria
1960
Frederico Westphalen

Otimização de Consultas

(madding **OR** crowd) **AND**
(ignoble **OR** strife)

Como escolher a ordem de processamento?

Universidade Federal de Santa Maria
1960
Frederico Westphalen

Otimização de Consultas

(madding **OR** crowd) **AND**
(ignoble **OR** strife)

- ✓ Obtém-se a frequência de todos os termos.
- ✓ Estima-se o tamanho de cada OR pela soma das frequências.
- ✓ Processar os ANDs em ordem crescente de tamanho de OR.

Modelo Booleano

- **Vantagens**
  - Modelo simples baseado em teoria bem fundamentada
  - Fácil de entender e implementar em computador

- **Desvantagens**
  - Não permite casamento parcial entre consulta e documento
  - Não permite ordenação dos documentos recuperados
  - A necessidade de informação do usuário deve ser expressa em termos de uma expressão booleana
  - Nem todo usuário é capaz disso
  - Todos os termos de indexação têm o mesmo peso na descrição dos assuntos de um documento.

Próxima Aula

- ✓ Consultas booleanas são exatas.
- ✓ Muitas vezes queremos classificar os resultados
- ✓ Necessidade de medir a proximidade da consulta para cada documento.
- ✓ Precisa decidir se docs apresentadas ao usuário são

únicos, ou um grupo de documentos que abrangem vários aspectos da a consulta.

Exercícios

1) Desenhe a matriz de incidência para a coleção de documentos abaixo:

Doc 1 : breakthrough drug for schizophrenia

Doc 2 : new schizophrenia drug

Doc 3 : new approach for treatment of schizophrenia

Doc 4 : new hopes for schizophrenia patients

2) Considerando a matriz de incidência do exercício 1, quais seriam os documentos retornados para as consultas:

a) schizophrenia AND drug

b) for AND NOT (drug OR approach)

Exercícios

3) Como seria o índice invertido para os textos do Exercício 1?

4) Escreva um algoritmo para fazer a União de duas listas de postings para uma consulta x OR y.

5) Sugira uma ordem de processamento de consulta para (passos no Slide 46):

(tangerine OR trees) AND
(marmalade OR skies) AND
(kaleidoscope OR eyes)

Term	Freq
eyes	213312
kaleidoscope	87009
marmalade	107913
skies	271658
tangerine	46653
trees	316812