Universidade Federal de Santa Maria
1960
Frederico Westphalen

Pré-processamento

Tópicos Especiais em Recuperação de Informações

Profª. Solange Pertile

24/08/15

Fontes:
Prof. Viviane Moreira (UFRGS)
Prof. Jairo de Souza (UFJF)

Relembrando...

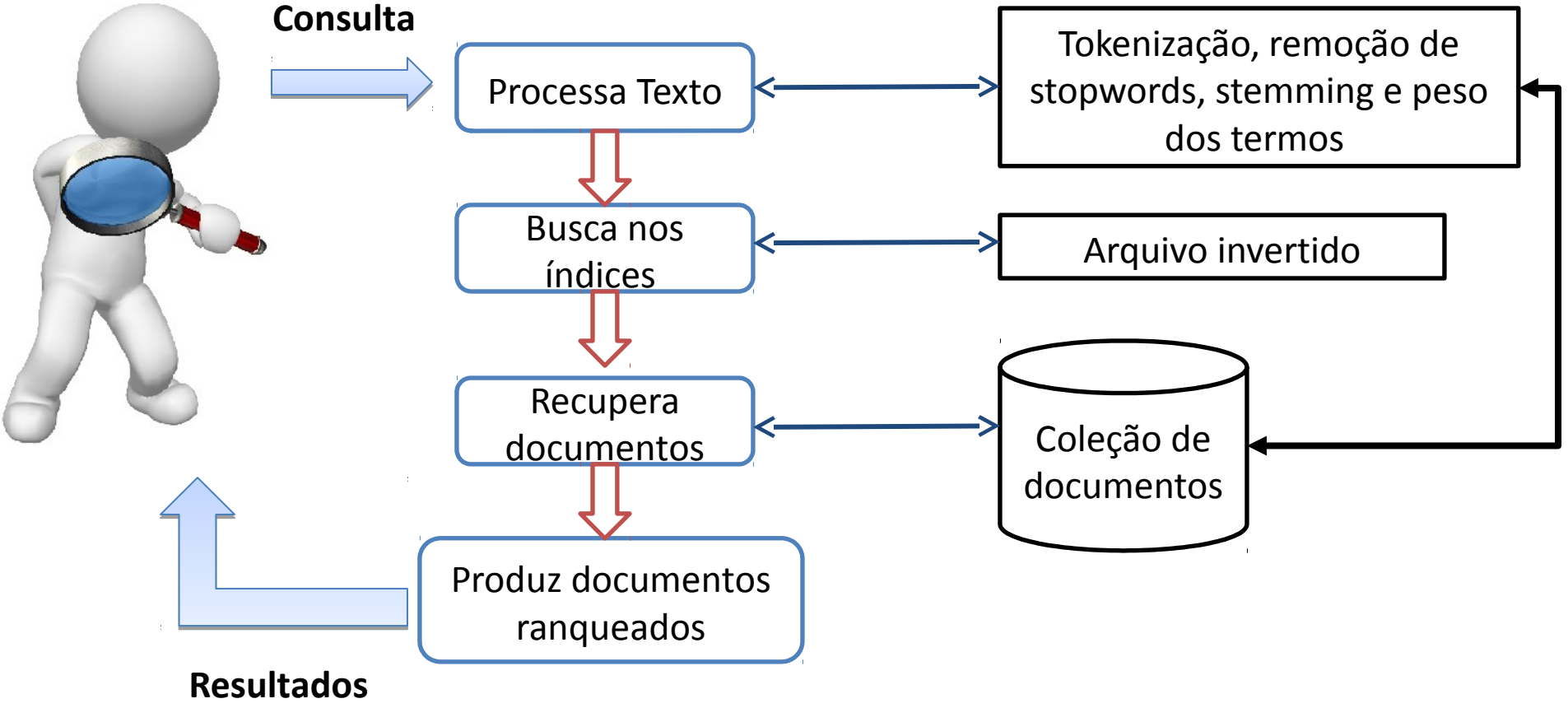
- ✓ O objetivo da área de estudo conhecida como Recuperação de Informação
- ✓Prover aos usuários o acesso fácil às informações de seu interesse
- ✓A diferença entre os objetivos iniciais da área e como esses objetivos mudaram com o advento da Web
- ✓Breve histórico

Relembrando...

- ✓ O problema de RI está ligado basicamente a:
  - ✓ Como extrair as informações dos documentos?
  - ✓ Como utilizar tais informações para decidir sobre a sua relevância?
- ✓ Sistema de RI pode ser dividido em seis módulos:
  1. Obtenção e coleção de documentos;
  2. Indexação dos documentos;
  3. Consulta do usuário;
  4. Recuperação de documentos;
  5. Ranqueamento dos documentos;
  6. Apresentação para o usuário.

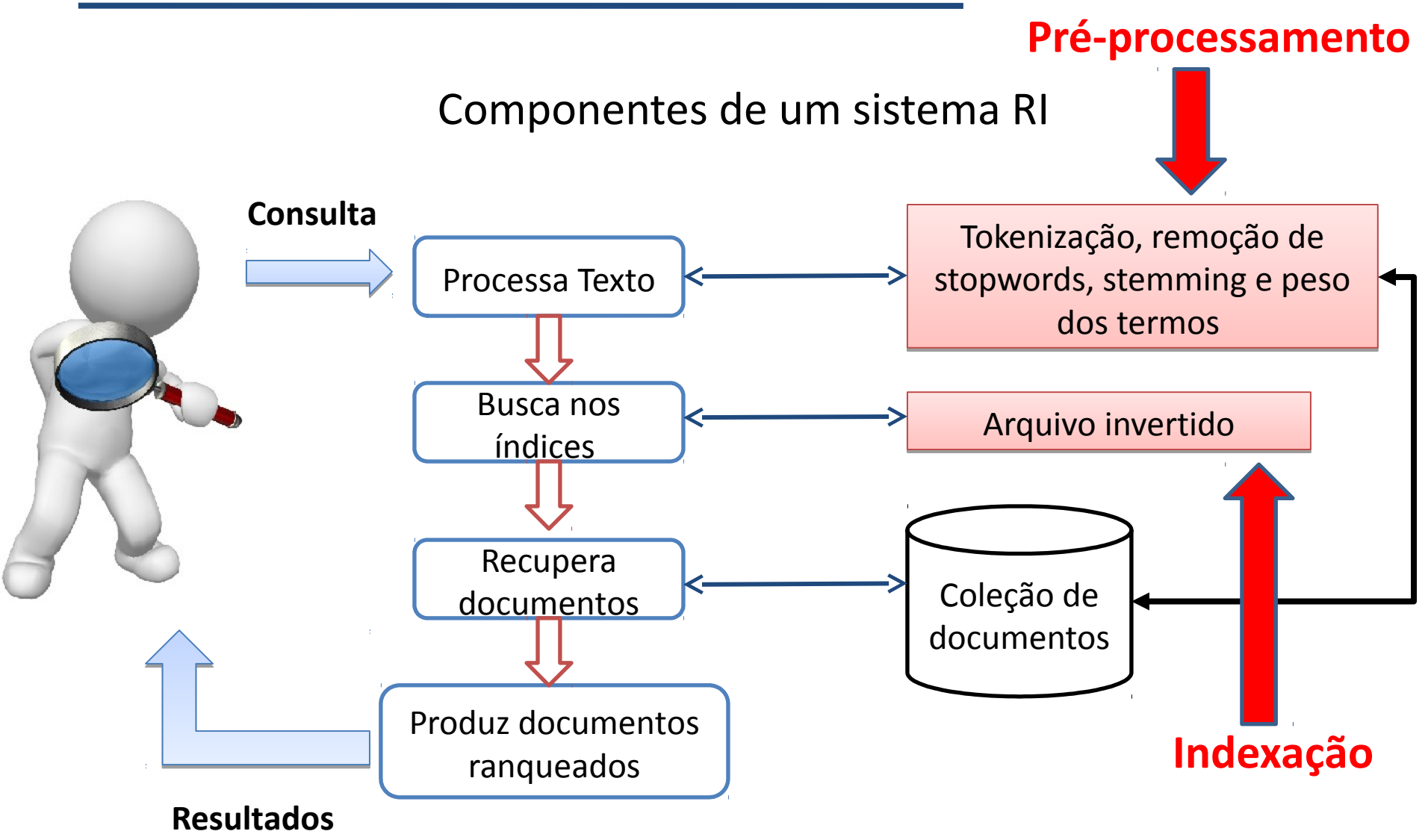
Relembrando...

Componentes de um sistema RI



Adaptado aulas Prof. Viviane Moreira (UFRGS)

Relembrando...



Adaptado aulas Prof. Viviane Moreira (UFRGS)

Pré-processamento

✓ O pré-processamento de documentos é um importante procedimento empregado na construção de sistemas de RI

✓ Pode ser dividido em quatro operações (ou Pode ser dividido em quatro operações (ou transformações) textuais:

1. Identificação da Unidade de Indexação
2. Identificação do idioma
3. Análise léxica do texto (ou tokenização)
4. Processamento Linguístico

Identificação da Unidade de Indexação

- ✓ O que é um documento?
- ✓ Um arquivo?
- ✓ Um livro inteiro?
- ✓ Um parágrafo?
- ✓ Uma frase?
- ✓ Vários níveis de granularidade – problema para dados em XML

Análise léxica do texto (ou tokenização)

- ✓ Dada uma sequencia de caracteres e uma unidade de indexação, a tokenização separa esta sequencia em tokens (termos ou palavras)

Como identificar palavras em um texto?

Usando somente espaços??

- ✓ Geralmente esta etapa descarta alguns caracteres, como pontuação por exemplo.

Análise léxica do texto (ou tokenização)

✓ **Exemplo:**

Entrada

"*Os coronéis da política, falsos democratas, apregoam moralidade e apresentam-se como guardiões das instituições.”*

Saída

Os	*coronéis*	*da*	*política*	*falsos*	*democrata*	*apregoam*

modalidades	*e*	*apresentam*	*se*	*como*	*Guardiões*	*das*	*instituições*

Análise léxica do texto (ou tokenização)

- ✓ **Dígitos:**
  - ✓ Números, por si só, são vagos

> *1989 pode representar um ano ou o número de pessoas que ingressaram na Universidade!*

  - ✓ Usualmente números são desconsiderados como termos de índice;
  - ✓ Procedimentos específicos podem ser empregados para normalizar datas e números.

Análise léxica do texto (ou tokenização)

- **Hífens:**
  - Difícil decisão para o analisador léxico;
  - Quebrar palavras hifenizadas pode ser útil devido a inconsistência de uso;

    Estado-da-arte = Estado da arte

  - Contudo, existem palavras que incluem hífens como parte integral delas;

    Guarda-chuva

  - Adote uma regra geral, mas tome cuidado com as exceções...

Análise léxica do texto (ou tokenização)

- ✓ **Marcas de pontuação:**
  - ✓ Removidas por completo do texto;
  - ✓ O risco de não interpretar palavras com marca de pontuação é mínimo:
  - ✓ **Por exemplo:**

510 A.C. será interpretado de maneira similar ao remover a pontuação.

Análise léxica do texto (ou tokenização)

- **Caixa das Palavras**
  - O fato das letras estarem em maiúsculo ou minúsculo normalmente não é importante para a identificação de termos de índice;
  - O analisador analisador léxico normalmente normalmente converte converte todo o texto para maiúsculas ou minúsculas;
  - Em alguns casos a semântica pode ficar comprometida:
    - **Banco e banco.**

Análise léxica do texto (ou tokenização)

- ✓ Decidir o que indexar – números?
- ✓ Idiomas
  - ✓ Alemão – substantivos compostos sem espaços Computerlinguistik
  - ✓ Lebensversicherungsgesellschaftsangestellter
  - ✓ Chinês, Japonês, Coreano e Tailandês – não utilizam espaços entre as palavras
- ✓ **Possíveis soluções:**
  - ✓ Dicionários abrangentes – pega-se o maior termo do dicionário
  - ✓ Utilizar pequenas subsequências de caracteres (n-grams) em vez de palavras inteiras

Análise léxica do texto (ou tokenização)

- ✓ Separar tokens quando encontramos espaços em branco também pode causar problemas
  - Porto Alegre
  - cadeia alimentar
  - bode expiatório

Universidade Federal de Santa Maria
1960
Frederico Westphalen

Análise léxica do texto (ou tokenização)

- ✓ Como implementar
  - ✓ Usar compiler compilers (geradores de parsers)
    - ✓ JavaCC (entre outros)
    - ✓ Usam expressões regulares
  - ✓ Usar parsers prontos
    - ✓ HTMLparser
    - ✓ XMLparser

Identificação do Idioma

- ✓ Os documentos a serem indexados podem conter partes em vários idiomas (ex. um email em italiano com um attachment em francês)
- ✓ Um mesmo índice pode conter palavras em idiomas diferentes
- ✓ Este é um problema de classificação
- ✓ **Exemplo** – language guesser da Xerox

http://www.xrce.xerox.com/competencies/content-analysis/tools/guesser-ISO-8859-1.en.html

Pré-processamento linguístico

- ✓ **Uniformização**
  - ✓Lowercase, lower case e lower-case
- ✓ **Case folding** – transformar todos os caracteres para minúsculo
  - ✓Carro = carro = CaRro

Problema – dificulta a identificação de identidades nomeadas

Pré-processamento linguístico

- ✓ **Remoção de Stopwords**
  - ✓ Palavras que são muito frequentes entre os documentos de uma coleção não são boas como discriminantes;
  - ✓ Uma palavra que ocorre em 80% dos documentos de uma coleção é inútil para os propósitos de recuperação;
  - ✓ Tais palavras são frequentemente chamadas de stopwords e são normalmente removidas dos termos de índice em potencial;
  - ✓ Exemplos: artigos, preposições, conjunções (portanto, logo, pois, como...)

Pré-processamento linguístico

- ✓ **Remoção de Stopwords**
  - ✓ Além destes, verbos de ligação, advérbios e alguns adjetivos também são candidatos a stopwords
  - ✓ Estima-se que a remoção de stopwords reduz o tamanho do índice em pelo menos 40%.
  - ✓ **Problemas**
    - ✓ “Ser ou não ser, eis a questão”
    - ✓ Removendo-se as stopwords só sobra “questão”
    - ✓ Por esta razão, alguns sistemas optam por indexar as stopwords

Pré-processamento linguístico

✓ Remoção de Stopwords

~~A~~ Varig Log pediu ~~a~~ impugnação ~~dos~~ votos ~~das~~ empresas ~~de~~ leasing ~~que~~ rejeitaram ~~a~~ proposta ~~de~~ compra ~~e~~ alteração ~~do~~ plano ~~de~~ recuperação judicial ~~da~~ Varig, durante ~~a~~ assembleia ~~de~~ credores realizada nesta segunda-feira. ~~A~~ Varig Log pretende garantir ~~a~~ realização ~~do~~ leilão ~~da~~ aérea, inicialmente marcado ~~para esta~~ quarta-feira. Com ~~a~~ recusa ~~dos~~ credores, ~~a~~ Justiça poderá suspender ~~o~~ leilão ~~e~~ decretar ~~a~~ falência ~~da~~ Varig.

Pré-processamento linguístico

- ✓ **Stemming**
  - ✓ Frequentemente o usuário especifica uma palavra em uma consulta, mas apenas uma variação dela está presente em um documento relevante;
  - ✓ Plurais, gerúndios e sufixos são exemplos de variações sintáticas que evitam um casamento perfeito entre uma palavra da consulta e uma respectiva palavra no documento;
  - ✓ Substituir as palavras pelos seus respectivos stems (radicais) pode superar parcialmente esse problema.

Pré-processamento linguístico

- ✓ **Stemming**
  - ✓ O stem ou radical, é a parte que sobra da palavra após a remoção do afixo.
  - ✓ A utilidade de um stemmer é reduzir as formas variantes das palavras a um único radical.
  - ✓ O stem não precisa ser uma palavra válida, contudo ele precisa captar o significado da palavra
  - ✓ Stemming ajuda a reduzir o número de entradas no índice
  - ✓ Os termos podem ser reduzidos durante a indexação ou durante a recuperação.

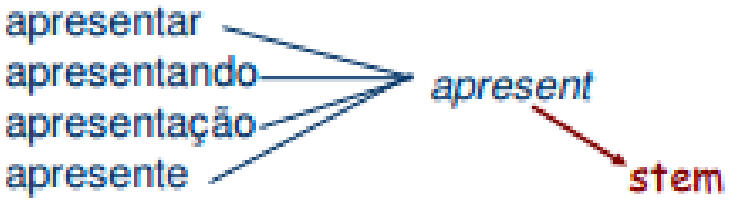
Pré-processamento linguístico

✓ Stemming

✓ Com frequência o usuário utiliza uma palavra-chave na consulta mas um documento relevante contém apenas formas variantes desta palavra.

✓ **Exemplo**:

- **Consulta**: "*como apresentar artigos científicos*"
- **doc 1**: ... apresentando um artigo científico...
- **doc 2**: ... apresentação de artigos científicos ...
- **doc 3**: ... apresente artigos científicos ...



Universidade Federal de Santa Maria
1960
Frederico Westphalen

Pré-processamento linguístico

Como fazer um stemmer??

Universidade Federal de Santa Maria
1960
Frederico Westphalen

Pré-processamento linguístico

- ✓ **Stemming**
  1. Consultas a dicionários
  2. Variedade de Sucessores
  3. N-gram
  4. Removedores de afixos
  5. Stemmer Estatístico

Pré-processamento linguístico

✓***Understemming*** – deixar de remover um sufixo, pode fazer com que palavras relacionadas não sejam combinadas.

- ✓bares → bare
- ✓adequado → adeq
- ✓adequação → adequaç

✓***Overstemming*** – remover caracteres que fazem parte do stem, pode causar que palavras não relacionadas sejam combinadas

- ✓avião → avi
- ✓bebê → beb
- ✓bebendo → beb

Pré-processamento linguístico

✓ ***Identificação de Termos Compostos***

✓ Porto Alegre, freio de mão, bode expiatório, dívida externa

✓Abordagem estatística – as palavras que compõem o termo composto devem ocorrer com certa frequência em um contexto comum – por exemplo na mesma sentença.

1. Computar as co-ocorrências entre pares de palavras
2. Se a co-ocorrência for menor do que um limiar, o par é descartado

Pré-processamento linguístico

- ***Identificação de Termos Compostos***

3. Para pares cuja co-ocorrência é maior do que o limiar, computar o valor da coesão, de acordo com a fórmula abaixo:

$$coesão(t_i, t_j) = \frac{freq\_de\_coocorr}{\sqrt{freq(t_i) \times freq(t_j)}}$$

4. Se a coesão for maior do que um dado limiar, o termo composto é retido

Universidade Federal de Santa Maria
1960
Frederico Westphalen

Pré-processamento linguístico

Quais as melhores alternativas?

Pré-processamento linguístico

- ✓O sistema de IR deve permitir diferentes alternativas
- ✓O responsável pelo sistema deve:
  - ✓ Tem um bom conhecimento da coleção
  - ✓ Saber qual o tipo de usuário
  - ✓ Saber quais consultas serão feitas
  - ✓ Conhecer o padrão de uso

Seleção de Palavras-chave

✓ Quais termos serão usados para fazer a indexação do documento?

1. Representação do texto completo – todas as palavras no texto são usadas como termos de índice;
2. Representação parcial – nem todas as palavras são usadas como termos de índice.

Bag of Words

- ✓Na maioria dos sistemas de IR a ordenação dos termos nos documentos é descartada.
- ✓Armazena-se apenas informações sobre o número de ocorrências dos termos nos documentos
- ✓Este modelo é conhecido por bag of words
- ✓<u>Vantagem</u>
  - ✓ simplificação
- ✓<u>Desvantagem</u>
  - ✓ “João é mais velho do que José” = “José é mais velho do que João”

Universidade Federal de Santa Maria
1960
Frederico Westphalen

Ponderação dos Termos (term weighting)

✓ Todos os termos de um documento são igualmente importantes?

Ponderação dos Termos (term weighting)

- ✓ Todos os termos de um documento são igualmente importantes?
- ✓ Não – alguns termos são mais importantes do que outros.
- ✓ Princípios
  - ✓ Termos que ocorram com muita frequência na coleção de documentos são menos importantes.
  - ✓ Um documento que contenha os termos da consulta mais vezes está mais relacionado à consulta – cuidado para não beneficiar documentos longos.

Ponderação dos Termos (term weighting)

- ✓ TF×IDF
  - ✓ Term Frequency times Inverse Document Frequency

$$w_{t,d} = tf_{t,d} idf_t$$

weight (peso) do termo t no documento d

$$w_{t,d} = tf_{t,d} idf_t$$

freq$_{t,d}$ = número de ocorrências do termo t no doc d

max$_l$ = número de ocorrências do termo mais frequente em d

$$w_{t,d} = tf_{t,d} idf_t$$

N = número de documentos na coleção

n$_t$ = número de documentos com o termo t

Ponderação dos Termos (term weighting)

✓E outras palavras, TF-IDF atribui ao termo t uma importância no documento d que é:

- ✓ **Alta** se *t* ocorrer muitas vezes em um número pequeno de documentos
- ✓ **Menor** se *t* ocorrer poucas vezes no documento ou muitas vezes na coleção
- ✓ **Muito baixa** se *t* ocorrer em quase todos os documentos

Ponderação dos Termos (term weighting)

Qual seria o ***idf*** de um termo que ocorre em todos os documentos?

Índices

- ✓ A maneira de evitar a necessidade de varrer os textos para resolver cada consulta é construir um índice antecipadamente.
- ✓ A forma mais comum de armazenarmos um índice é usando um arquivo invertido.
- ✓ A fim de construir o índice, várias atividades de pré-processamento podem ser necessárias.

Índices

- ✓ A maneira de evitar a necessidade de varrer os textos para resolver cada consulta é construir um índice antecipadamente.
- ✓ A forma mais comum de armazenarmos um índice é usando um arquivo invertido.
- ✓ A fim de construir o índice, várias atividades de pré-processamento podem ser necessárias.

Próximas aulas

- ✓ Estudo dos modelos clássicos de recuperação e ranqueamento de documentos:
- ✓ Modelo booleano;
- ✓ Modelo vetorial;
- ✓ Modelo probabilístico.

Exercícios

1. Construa o índice invertido e posicional passo a passo para a seguinte coleção de documentos:

Base de documentos	
Documento	**Texto**
1	Pease porridge hot, pease porridge cold
2	Pease porridge in the pot
3	Nine days cold
4	Some like it hot, some like it cold
5	Some like it in the pot
6	Nine days old