Pesquisa e Publicação de Informação

Operações sobre Texto

Nuno D. Mendes

Licenciatura em Sistemas e Tecnologias de Informação

4 Mai 2012

ISEGI – UNL

Operações sobre Texto

Motivação

- Nem todos termos de um documento contribuem para a representação da sua semântica

Operações sobre Texto

Motivação

- Nem todos termos de um documento contribuem para a representação da sua semântica
- A utilização de todos os termos pode introduzir ruído na tarefa de pesquisa de informação

Operações sobre Texto

Motivação

- Nem todos termos de um documento contribuem para a representação da sua semântica
- A utilização de todos os termos pode introduzir ruído na tarefa de pesquisa de informação
- No entanto, os termos a utilizar (*termos de indexação*) devem ser escolhidos criteriosamente, de modo a não ter uma representação demasiado imprecisa

Operações sobre Texto

Motivação

- Nem todos termos de um documento contribuem para a representação da sua semântica
- A utilização de todos os termos pode introduzir ruído na tarefa de pesquisa de informação
- No entanto, os termos a utilizar (*termos de indexação*) devem ser escolhidos criteriosamente, de modo a não ter uma representação demasiado imprecisa

Tipos de operações

1. **Análise lexical**: tratamento de digitos, hífens, acentuação e pontuação gráfica, capitalização de palavras

Operações sobre Texto

Motivação

- Nem todos termos de um documento contribuem para a representação da sua semântica
- A utilização de todos os termos pode introduzir ruído na tarefa de pesquisa de informação
- No entanto, os termos a utilizar (*termos de indexação*) devem ser escolhidos criteriosamente, de modo a não ter uma representação demasiado imprecisa

Tipos de operações

1. **Análise lexical**: tratamento de digitos, hífens, acentuação e pontuação gráfica, capitalização de palavras
2. **Eliminação de** *stopwords*: filtragem de termos com pouco poder discriminativo (e.g. artigos e preposições)

Operações sobre Texto

Motivação

- Nem todos termos de um documento contribuem para a representação da sua semântica
- A utilização de todos os termos pode introduzir ruído na tarefa de pesquisa de informação
- No entanto, os termos a utilizar (*termos de indexação*) devem ser escolhidos criteriosamente, de modo a não ter uma representação demasiado imprecisa

Tipos de operações

1. **Análise lexical**: tratamento de digitos, hífens, acentuação e pontuação gráfica, capitalização de palavras
2. **Eliminação de** *stopwords*: filtragem de termos com pouco poder discriminativo (e.g. artigos e preposições)
3. *Stemming*: identificação da raíz da palavra, para a qual são reduzidas todas as suas derivações

Operações sobre Texto

Motivação

- Nem todos termos de um documento contribuem para a representação da sua semântica
- A utilização de todos os termos pode introduzir ruído na tarefa de pesquisa de informação
- No entanto, os termos a utilizar (*termos de indexação*) devem ser escolhidos criteriosamente, de modo a não ter uma representação demasiado imprecisa

Tipos de operações

1. **Análise lexical**: tratamento de digitos, hífens, acentuação e pontuação gráfica, capitalização de palavras
2. **Eliminação de** *stopwords*: filtragem de termos com pouco poder discriminativo (e.g. artigos e preposições)
3. *Stemming*: identificação da raíz da palavra, para a qual são reduzidas todas as suas derivações
4. **Selecção de termos de indexação**: escolha de termos pelo seu papel sintáctico, determinação de que palavras ou conjuntos de palavras são bons termos de indexação (e.g. substantivos ou lexemas tipo "misseis balísticos")

Operações sobre Texto

Motivação

- Nem todos termos de um documento contribuem para a representação da sua semântica
- A utilização de todos os termos pode introduzir ruído na tarefa de pesquisa de informação
- No entanto, os termos a utilizar (*termos de indexação*) devem ser escolhidos criteriosamente, de modo a não ter uma representação demasiado imprecisa

Tipos de operações

1. **Análise lexical**: tratamento de digitos, hífens, acentuação e pontuação gráfica, capitalização de palavras
2. **Eliminação de** *stopwords*: filtragem de termos com pouco poder discriminativo (e.g. artigos e preposições)
3. *Stemming*: identificação da raíz da palavra, para a qual são reduzidas todas as suas derivações
4. **Selecção de termos de indexação**: escolha de termos pelo seu papel sintáctico, determinação de que palavras ou conjuntos de palavras são bons termos de indexação (e.g. substantivos ou lexemas tipo "misseis balísticos")
5. Construção de um Tesouro de Sinónimos (abordado na última aula)

Operações sobre Texto

Motivação

- Nem todos termos de um documento contribuem para a representação da sua semântica
- A utilização de todos os termos pode introduzir ruído na tarefa de pesquisa de informação
- No entanto, os termos a utilizar (*termos de indexação*) devem ser escolhidos criteriosamente, de modo a não ter uma representação demasiado imprecisa

Tipos de operações

1. **Análise lexical**: tratamento de digitos, hífens, acentuação e pontuação gráfica, capitalização de palavras
2. **Eliminação de** *stopwords*: filtragem de termos com pouco poder discriminativo (e.g. artigos e preposições)
3. *Stemming*: identificação da raíz da palavra, para a qual são reduzidas todas as suas derivações
4. **Selecção de termos de indexação**: escolha de termos pelo seu papel sintáctico, determinação de que palavras ou conjuntos de palavras são bons termos de indexação (e.g. substantivos ou lexemas tipo "misseis balísticos")
5. Construção de um Tesouro de Sinónimos (abordado na última aula)
6. **Compressão do texto**: reduzir o tamanho do repositório de documentos comprimindo o texto nele contido, mas permitindo acesso aleatório.

Definição

- Identifica as palavras contidas no texto, eliminando ou tratando hifenização, pontuação gráfica, capitalização e digitos

Definição

- Identifica as palavras contidas no texto, eliminando ou tratando hifenização, pontuação gráfica, capitalização e digitos
- Geralmente os números não são indexados porque são demasiado vagos, excepto se forem claramente identificados como datas ou identificadores (*e.g.* números de cartões de crédito)

Definição

- Identifica as palavras contidas no texto, eliminando ou tratando hifenização, pontuação gráfica, capitalização e digitos
- Geralmente os números não são indexados porque são demasiado vagos, excepto se forem claramente identificados como datas ou identificadores (*e.g.* números de cartões de crédito)
- O tratamento da pontuação gráfica apresenta alguns desafios. Se for simplesmente eliminada não poderá distinguir, por exemplo, `sonae.com` (um url), de `Sonaecom`, uma firma

Definição

- Identifica as palavras contidas no texto, eliminando ou tratando hifenização, pontuação gráfica, capitalização e digitos
- Geralmente os números não são indexados porque são demasiado vagos, excepto se forem claramente identificados como datas ou identificadores (*e.g.* números de cartões de crédito)
- O tratamento da pontuação gráfica apresenta alguns desafios. Se for simplesmente eliminada não poderá distinguir, por exemplo, `sonae.com` (um url), de `Sonaecom`, uma firma
- O tratamento da capitalização das palavras deve ter em atenção a informação semântica da capitalização (*e.g.* Bolsa versus bolsa, Banco versus banco)

Definição

- Identifica as palavras contidas no texto, eliminando ou tratando hifenização, pontuação gráfica, capitalização e digitos
- Geralmente os números não são indexados porque são demasiado vagos, excepto se forem claramente identificados como datas ou identificadores (*e.g.* números de cartões de crédito)
- O tratamento da pontuação gráfica apresenta alguns desafios. Se for simplesmente eliminada não poderá distinguir, por exemplo, `sonae.com` (um url), de `Sonaecom`, uma firma
- O tratamento da capitalização das palavras deve ter em atenção a informação semântica da capitalização (*e.g.* Bolsa versus bolsa, Banco versus banco)

Definição

- Identifica as palavras contidas no texto, eliminando ou tratando hifenização, pontuação gráfica, capitalização e digitos
- Geralmente os números não são indexados porque são demasiado vagos, excepto se forem claramente identificados como datas ou identificadores (*e.g.* números de cartões de crédito)
- O tratamento da pontuação gráfica apresenta alguns desafios. Se for simplesmente eliminada não poderá distinguir, por exemplo, `sonae.com` (um url), de `Sonaecom`, uma firma
- O tratamento da capitalização das palavras deve ter em atenção a informação semântica da capitalização (*e.g.* Bolsa versus bolsa, Banco versus banco)

Vantagens/desvantagens

1. Reduz o número de termos de indexação dos documentos

Definição

- Identifica as palavras contidas no texto, eliminando ou tratando hifenização, pontuação gráfica, capitalização e digitos
- Geralmente os números não são indexados porque são demasiado vagos, excepto se forem claramente identificados como datas ou identificadores (*e.g.* números de cartões de crédito)
- O tratamento da pontuação gráfica apresenta alguns desafios. Se for simplesmente eliminada não poderá distinguir, por exemplo, `sonae.com` (um url), de `Sonaecom`, uma firma
- O tratamento da capitalização das palavras deve ter em atenção a informação semântica da capitalização (*e.g.* Bolsa versus bolsa, Banco versus banco)

Vantagens/desvantagens

1. Reduz o número de termos de indexação dos documentos
2. Na maior parte dos casos melhora a performance do sistema de Pesquisa de Informação

Operações sobre Texto

Definição

- Identifica as palavras contidas no texto, eliminando ou tratando hifenização, pontuação gráfica, capitalização e digitos
- Geralmente os números não são indexados porque são demasiado vagos, excepto se forem claramente identificados como datas ou identificadores (*e.g.* números de cartões de crédito)
- O tratamento da pontuação gráfica apresenta alguns desafios. Se for simplesmente eliminada não poderá distinguir, por exemplo, `sonae.com` (um url), de `Sonaecom`, uma firma
- O tratamento da capitalização das palavras deve ter em atenção a informação semântica da capitalização (*e.g.* Bolsa versus bolsa, Banco versus banco)

Vantagens/desvantagens

1. Reduz o número de termos de indexação dos documentos
2. Na maior parte dos casos melhora a performance do sistema de Pesquisa de Informação
3. Pode implicar um aumento do tempo de resposta devido à necessidade de pré-processar o texto e a *query*, dependendo da complexidade das operações realizadas

Operações sobre Texto

Análise Lexical

Definição

- Identifica as palavras contidas no texto, eliminando ou tratando hifenização, pontuação gráfica, capitalização e digitos
- Geralmente os números não são indexados porque são demasiado vagos, excepto se forem claramente identificados como datas ou identificadores (*e.g.* números de cartões de crédito)
- O tratamento da pontuação gráfica apresenta alguns desafios. Se for simplesmente eliminada não poderá distinguir, por exemplo, `sonae.com` (um url), de `Sonaecom`, uma firma
- O tratamento da capitalização das palavras deve ter em atenção a informação semântica da capitalização (*e.g.* Bolsa versus bolsa, Banco versus banco)

Vantagens/desvantagens

1. Reduz o número de termos de indexação dos documentos
2. Na maior parte dos casos melhora a performance do sistema de Pesquisa de Informação
3. Pode implicar um aumento do tempo de resposta devido à necessidade de pré-processar o texto e a *query*, dependendo da complexidade das operações realizadas
4. Pode implicar perda de informação semântica sobre os documentos

Definição

- Elimina termos de fraco poder discriminativo (*e.g.* artigos, preposições, conjunções), porque um termo que aparece numa grande percentagem de documentos é inútil para determinar a relevância

Definição

- Elimina termos de fraco poder discriminativo (*e.g.* artigos, preposições, conjunções), porque um termo que aparece numa grande percentagem de documentos é inútil para determinar a relevância
- A lista de *stopwords* pode ser ainda enriquecida com alguns verbos, advérbios ou adjectivos comuns

Definição

- Elimina termos de fraco poder discriminativo (*e.g.* artigos, preposições, conjunções), porque um termo que aparece numa grande percentagem de documentos é inútil para determinar a relevância
- A lista de *stopwords* pode ser ainda enriquecida com alguns verbos, advérbios ou adjectivos comuns

Vantagens/desvantagens

1. Elimina termos inúteis para a tarefa de Pesquisa de Informação

Definição

- Elimina termos de fraco poder discriminativo (*e.g.* artigos, preposições, conjunções), porque um termo que aparece numa grande percentagem de documentos é inútil para determinar a relevância
- A lista de *stopwords* pode ser ainda enriquecida com alguns verbos, advérbios ou adjectivos comuns

Vantagens/desvantagens

1. Elimina termos inúteis para a tarefa de Pesquisa de Informação
2. Reduz **substancialmente** ($\sim$ 40%) o tamanho do estrutura que indexa os termos (diminui o vocabulário)

Definição

- Elimina termos de fraco poder discriminativo (*e.g.* artigos, preposições, conjunções), porque um termo que aparece numa grande percentagem de documentos é inútil para determinar a relevância
- A lista de *stopwords* pode ser ainda enriquecida com alguns verbos, advérbios ou adjectivos comuns

Vantagens/desvantagens

1. Elimina termos inúteis para a tarefa de Pesquisa de Informação
2. Reduz **substancialmente** ($\sim 40\%$) o tamanho do estrutura que indexa os termos (diminui o vocabulário)
3. Pode dificultar drasticamente algumas tarefas de pesquisa (*e.g.* a frase "to be or not to be", eliminando as *stopwords* mais comuns fica reduzida a "be be")

Definição

- Reduz os termos à sua raíz (*e.g. ligar*, *ligando*, *ligação*, *ligações* ficam reduzidos ao mesmo termo)

Definição

- Reduz os termos à sua raíz (*e.g. ligar*, *ligando*, *ligação*, *ligações* ficam reduzidos ao mesmo termo)
- Para a língua inglesa existe o algoritmo de Porter para a eliminação de sufixos de termos

Definição

- Reduz os termos à sua raíz (*e.g. ligar*, *ligando*, *ligação*, *ligações* ficam reduzidos ao mesmo termo)
- Para a língua inglesa existe o algoritmo de Porter para a eliminação de sufixos de termos
- Existem quatro tipos de estratégias:
  - **Remoção de afixos** (prefixos, sufixos e infixos), segundo regras específicas de cada língua
  - **Pesquisa tabular**, que exige uma tabela de todas as derivações consideradas de cada palavra da língua
  - **Derivação de sucessores**, baseado na identificação dos morfemas recorrendo a princípios de linguística estrutural
  - **Método dos *n*-gramas**, que agrupa termos em classes consoante os conjuntos de *n* caracteres que partilham (digramas, trigramas, etc)

Definição

- Reduz os termos à sua raíz (*e.g. ligar*, *ligando*, *ligação*, *ligações* ficam reduzidos ao mesmo termo)
- Para a língua inglesa existe o algoritmo de Porter para a eliminação de sufixos de termos
- Existem quatro tipos de estratégias:
  - **Remoção de afixos** (prefixos, sufixos e infixos), segundo regras específicas de cada língua
  - **Pesquisa tabular**, que exige uma tabela de todas as derivações consideradas de cada palavra da língua
  - **Derivação de sucessores**, baseado na identificação dos morfemas recorrendo a princípios de linguística estrutural
  - **Método dos *n*-gramas**, que agrupa termos em classes consoante os conjuntos de *n* caracteres que partilham (digramas, trigramas, etc)

Vantagens/desvantages

1. Reduz o número de termos de indexação e, consequentemente, o vocabulário

Definição

- Reduz os termos à sua raíz (*e.g.* *ligar*, *ligando*, *ligação*, *ligações* ficam reduzidos ao mesmo termo)
- Para a língua inglesa existe o algoritmo de Porter para a eliminação de sufixos de termos
- Existem quatro tipos de estratégias:
  - **Remoção de afixos** (prefixos, sufixos e infixos), segundo regras específicas de cada língua
  - **Pesquisa tabular**, que exige uma tabela de todas as derivações consideradas de cada palavra da língua
  - **Derivação de sucessores**, baseado na identificação dos morfemas recorrendo a princípios de linguística estrutural
  - **Método dos *n*-gramas**, que agrupa termos em classes consoante os conjuntos de *n* caracteres que partilham (digramas, trigramas, etc)

Vantagens/desvantages

1. Reduz o número de termos de indexação e, consequentemente, o vocabulário
2. Pode correr o risco de coalescer termos que se querem distintos (*e.g.* *ligar*, *ligante*, *liga*)

Operações sobre Texto

Stemming

Definição

- Reduz os termos à sua raíz (*e.g. ligar*, *ligando*, *ligação*, *ligações* ficam reduzidos ao mesmo termo)
- Para a língua inglesa existe o algoritmo de Porter para a eliminação de sufixos de termos
- Existem quatro tipos de estratégias:
  - **Remoção de afixos** (prefixos, sufixos e infixos), segundo regras específicas de cada língua
  - **Pesquisa tabular**, que exige uma tabela de todas as derivações consideradas de cada palavra da língua
  - **Derivação de sucessores**, baseado na identificação dos morfemas recorrendo a princípios de linguística estrutural
  - **Método dos *n*-gramas**, que agrupa termos em classes consoante os conjuntos de *n* caracteres que partilham (digramas, trigramas, etc)

Vantagens/desvantages

1. Reduz o número de termos de indexação e, consequentemente, o vocabulário
2. Pode correr o risco de coalescer termos que se querem distintos (*e.g. ligar*, *ligante*, *liga*)
3. As operações de *stemming* têm, em geral, de ser adaptadas a cada língua utilizada. Pode ser desafiante em colecções de documentos multilingues

Definição

- Consiste na selecção de lexemas (tipocamente substantivos simples ou compostos) como termos de indexação

Definição

- Consiste na selecção de lexemas (tipocamente substantivos simples ou compostos) como termos de indexação
- Estes substantivos representam *conceitos*

Definição

- Consiste na selecção de lexemas (tipocamente substantivos simples ou compostos) como termos de indexação
- Estes substantivos representam *conceitos*

Vantagens/desvantagens

1. Permitem a representação de um documentos em termos do seu conjunto de conceitos

Definição

- Consiste na selecção de lexemas (tipocamente substantivos simples ou compostos) como termos de indexação
- Estes substantivos representam *conceitos*

Vantagens/desvantagens

1. Permitem a representação de um documentos em termos do seu conjunto de conceitos
2. Diminui o número de termos de indexação

Definição

- Consiste na selecção de lexemas (tipocamente substantivos simples ou compostos) como termos de indexação
- Estes substantivos representam *conceitos*

Vantagens/desvantagens

1. Permitem a representação de um documentos em termos do seu conjunto de conceitos
2. Diminui o número de termos de indexação
3. Requer um método automático de identificação de conceitos

Definição

- Consiste na selecção de lexemas (tipocamente substantivos simples ou compostos) como termos de indexação
- Estes substantivos representam *conceitos*

Vantagens/desvantagens

1. Permitem a representação de um documentos em termos do seu conjunto de conceitos
2. Diminui o número de termos de indexação
3. Requer um método automático de identificação de conceitos
4. Como os métodos anteriores, pode introduzir imprecisões e indirecções na tarefa de pesquisa da informação

Definição

- Reduz o tamanho dos documentos, comprimindo o texto

Definição

- Reduz o tamanho dos documentos, comprimindo o texto
- Tipos de compressão de texto comuns:
  - **Codificação Huffman**, atribui uma representação binária a cada símbolo diferente no documento. Símbolos mais frequentes são codificados por um número menor de bits

Definição

- Reduz o tamanho dos documentos, comprimindo o texto
- Tipos de compressão de texto comuns:
  - **Codificação Huffman**, atribui uma representação binária a cada símbolo diferente no documento. Símbolos mais frequentes são codificados por um número menor de bits
  - **Codificação aritmética**, semelhante mas o código é construído incrementalmente

Definição

- Reduz o tamanho dos documentos, comprimindo o texto
- Tipos de compressão de texto comuns:
  - **Codificação Huffman**, atribui uma representação binária a cada símbolo diferente no documento. Símbolos mais frequentes são codificados por um número menor de bits
  - **Codificação aritmética**, semelhante mas o código é construído incrementalmente
  - **Métodos baseados em dicionários**, constroem um dicionário com sequências que vão sendo observadas, e substituem as sequências no texto por referências ao dicionário construído (*e.g.* algoritmos de compressão da família Ziv-Lempel)

Definição

- Reduz o tamanho dos documentos, comprimindo o texto
- Tipos de compressão de texto comuns:
  - **Codificação Huffman**, atribui uma representação binária a cada símbolo diferente no documento. Símbolos mais frequentes são codificados por um número menor de bits
  - **Codificação aritmética**, semelhante mas o código é construído incrementalmente
  - **Métodos baseados em dicionários**, constroem um dicionário com sequências que vão sendo observadas, e substituem as sequências no texto por referências ao dicionário construído (*e.g.* algoritmos de compressão da família Ziv-Lempel)
  - **Huffman tomando palavras como símbolos**, igual ao método de Huffman mas os símbolos em vez de caracteres são palavras

Definição

- Reduz o tamanho dos documentos, comprimindo o texto
- Tipos de compressão de texto comuns:
  - **Codificação Huffman**, atribui uma representação binária a cada símbolo diferente no documento. Símbolos mais frequentes são codificados por um número menor de bits
  - **Codificação aritmética**, semelhante mas o código é construído incrementalmente
  - **Métodos baseados em dicionários**, constroem um dicionário com sequências que vão sendo observadas, e substituem as sequências no texto por referências ao dicionário construído (*e.g.* algoritmos de compressão da família Ziv-Lempel)
  - **Huffman tomando palavras como símbolos**, igual ao método de Huffman mas os símbolos em vez de caracteres são palavras

Para a frase *For each rose, a rose is a rose*, uma representação possível é:

Símbolo	**Freq.**	**Código**
each	1/9	0000
,	1/9	0001
for	1/9	0010
is	1/9	0011
a	2/9	01
rose	1/3	1

Definição

- Reduz o tamanho dos documentos, comprimindo o texto
- Tipos de compressão de texto comuns:
  - **Codificação Huffman**, atribui uma representação binária a cada símbolo diferente no documento. Símbolos mais frequentes são codificados por um número menor de bits
  - **Codificação aritmética**, semelhante mas o código é construído incrementalmente
  - **Métodos baseados em dicionários**, constroem um dicionário com sequências que vão sendo observadas, e substituem as sequências no texto por referências ao dicionário construído (*e.g.* algoritmos de compressão da família Ziv-Lempel)
  - **Huffman tomando palavras como símbolos**, igual ao método de Huffman mas os símbolos em vez de caracteres são palavras
- Tipos de compressão especiais foram desenvolvidos para ficheiros invertidos (a ver nas próximas aulas)

Definição

- Reduz o tamanho dos documentos, comprimindo o texto
- Tipos de compressão de texto comuns:
  - **Codificação Huffman**, atribui uma representação binária a cada símbolo diferente no documento. Símbolos mais frequentes são codificados por um número menor de bits
  - **Codificação aritmética**, semelhante mas o código é construído incrementalmente
  - **Métodos baseados em dicionários**, constroem um dicionário com sequências que vão sendo observadas, e substituem as sequências no texto por referências ao dicionário construído (*e.g.* algoritmos de compressão da família Ziv-Lempel)
  - **Huffman tomando palavras como símbolos**, igual ao método de Huffman mas os símbolos em vez de caracteres são palavras
- Tipos de compressão especiais foram desenvolvidos para ficheiros invertidos (a ver nas próximas aulas)

Vantagens/desvantagens

1. Diminui grandemente o tamanho dos documentos (importante para grandes repositórios)

Definição

- Reduz o tamanho dos documentos, comprimindo o texto
- Tipos de compressão de texto comuns:
  - **Codificação Huffman**, atribui uma representação binária a cada símbolo diferente no documento. Símbolos mais frequentes são codificados por um número menor de bits
  - **Codificação aritmética**, semelhante mas o código é construído incrementalmente
  - **Métodos baseados em dicionários**, constroem um dicionário com sequências que vão sendo observadas, e substituem as sequências no texto por referências ao dicionário construído (*e.g.* algoritmos de compressão da família Ziv-Lempel)
  - **Huffman tomando palavras como símbolos**, igual ao método de Huffman mas os símbolos em vez de caracteres são palavras
- Tipos de compressão especiais foram desenvolvidos para ficheiros invertidos (a ver nas próximas aulas)

Vantagens/desvantagens

1. Diminui grandemente o tamanho dos documentos (importante para grandes repositórios)
2. Introduz a latência das operações de compressão (e descompressão, para técnicas que não permitem acesso aleatório)