GSI024 - Organização e recuperação de informação

Prof. Dr. Rodrigo Sanches Miani (FACOM/UFU)

Última atualização - Junho/2022

Pré-processamento de documentos

Agenda

“Pré-processamento de documentos”

Ideia geral

Análise léxica

Eliminação de stopwords

Stemming

Seleção de palavras-chave

QP-2

QP-2

- Marcado para o dia 23/06, próxima quinta-feira;
- E aí? Todos prontos?
- O máximo que posso fazer é adiar para o dia 30/06 (todos os outros QPs seriam deslocados também…)

Aula passada

Ideia fundamental

- A partir de uma consulta do usuário, existe um conjunto de documentos que contém exatamente os documentos relevantes (**resposta ideal**) e nenhum outro;

- Dada uma descrição desse **conjunto resposta ideal**, poderíamos recuperar os documentos relevantes;

- Quais são essas propriedades dessa descrição?
  - Resposta: não sabemos! Tudo que sabemos é que existem termos de indexação para caracterizar tais propriedades.

Definição

Seja R um conjunto de documentos inicialmente estimado como relevante para o usuário para a consulta q. Seja $\overline{R}$ o complemento de R (o conjunto de documentos não relevantes). A similaridade sim(d_j,q) entre o documento d_j e a consulta q é definida por:

$$sim(d_j, q) = \frac{P(R|\vec{d}_j, q)}{P(\overline{R}|\vec{d}_j, q)}$$

Estimar as probabilidades do conjunto de documentos relevantes

Para lidar com valores pequenos de ri, é conveniente somar 0,5 a cada um dos termos da fórmula anterior:

$$sim(d_j, q) \sim \sum_{k_i[q,d_j]} \log \left(\frac{(r_i + 0{,}5)(N - n_i - R + r_i + 0{,}5)}{(R - r_i + 0{,}5)(n_i - r_i + 0{,}5)} \right)$$

Essa fórmula é conhecida como equação **Robertson-Spark Jones** e é considerada a equação de ranqueamento clássica para o modelo probabilístico. Comporta-se bem para estimativas particulares como R = r_i.

Ajuste para ($R = r_i = 0$)

Para evitar o comportamento anômalo mostrado anteriormente, podemos eliminar o fator n_i do numerador da equação anterior, conforme sugerido por Robertson e Walker (1997):

$$sim(d_j, q) \sim \sum_{k_i \in q \wedge k_i \in d_j} \log\left(\frac{N + 0{,}5}{n_i + 0{,}5}\right)$$

Dessa forma, um termo que ocorre em todos os documentos (ni = N) produz um peso igual a zero (log(1)=0) e não existem mais pesos negativos.

Comparação entre os modelos clássicos

1. Modelo booleano é considerado o mais fraco entre os modelos clássicos;
2. O maior problema do modelo booleano é a falta de casamento parcial entre a consulta e os documentos;
3. Existe controvérsia quanto ao modelo probabilístico ser melhor do que o vetorial:
   - Experimentos realizados por Croft indicam que o modelo probabilístico fornece melhor qualidade de recuperação;
   - Outros experimentos conduzidos por Salton e Buckley contestam esses resultados.
4. Com coleções genéricas, o modelo vetorial fornece um modelo de RI razoável e robusto para fins de comparação.

Pré-processamento

Ideia básica

Motivação

Meu nome é Walter Hartwell White, moro na Alameda Riacho Negro, 308 Cidade de Albuquerque, no Novo México, CEP 87104. Para as entidades de imposição da lei, isso não é uma confissão. Estou falando para minha família agora.

Skyler... você é o amor da minha vida.

Espero que saiba disso. Walter Júnior... você é o meu rapazão.

Hav...Haverá algumas...

coisas... coisas... que vocês descobrirão sobre mim logo… em poucos dias. Só quero que vocês saibam que não... não importa o que isso pareça, eu só tenho vocês no meu coração. Adeus.

(continua...)

Motivação

- Suponha que vocês estão criando uma ferramenta de busca de informação sobre séries.
- A ferramenta irá associar termos (palavras-chave encontradas nos diálogos) a episódios de uma série (documentos).
- Algumas funcionalidades: "em qual episódio de Breaking Bad, Jesse proferiu certa piada?" "quais séries (e em quais episódios) tratam sobre tênis?" e por aí vai...

Motivação

- Dado o tamanho da base de dados, seria importante pensar em uma forma de pré-processar os termos presentes em um documento.

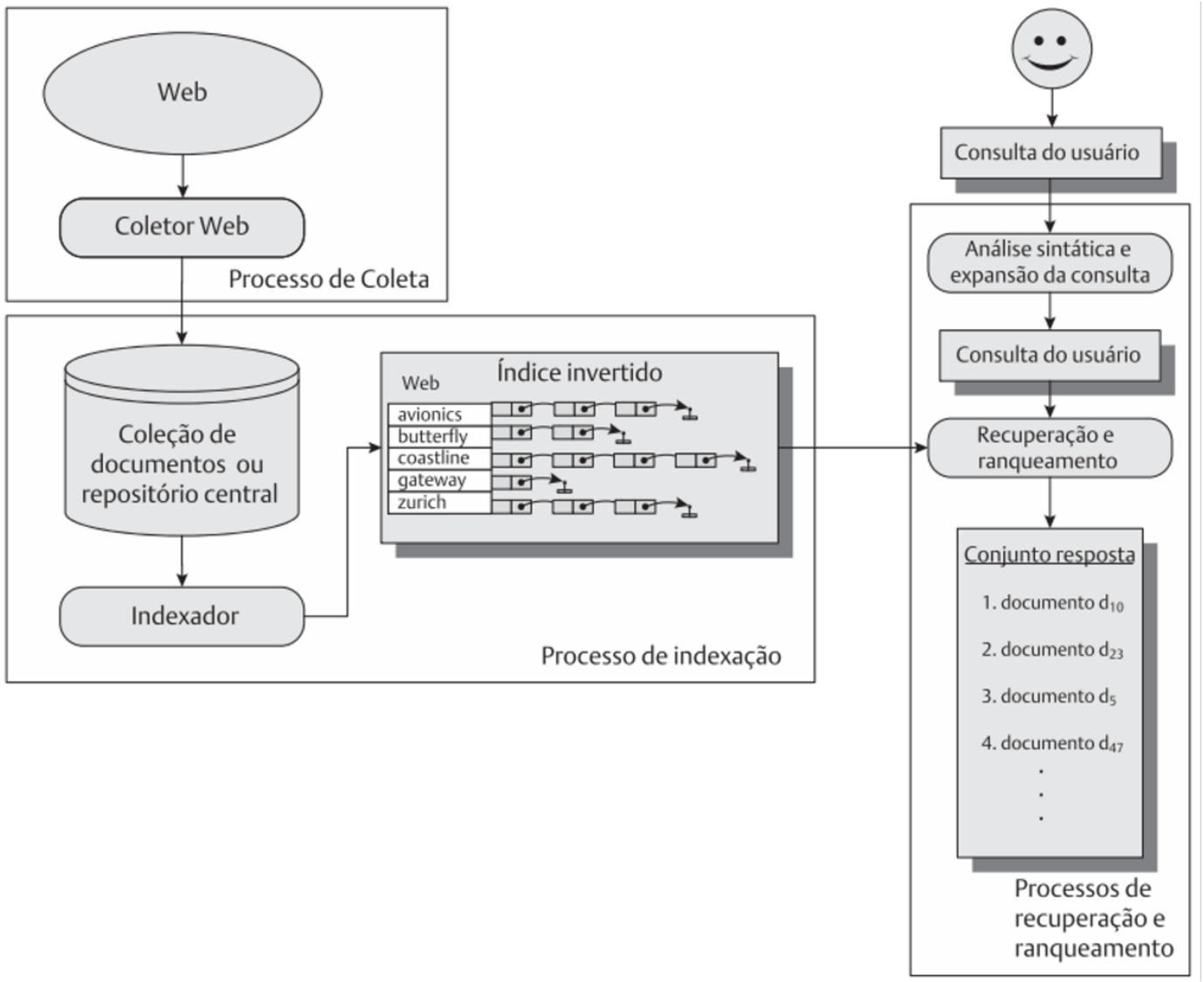
- Existem regras/sugestões para isso?

Introdução

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

- Pode ser dividido em quatro operações (ou transformações) textuais:
  1. Análise léxica do texto;
  2. Eliminação de stopwords;
  3. Stemming das palavras;
  4. Seleção de termos ou palavras-chave;

Pré-processamento x Sistema de RI



Análise léxica

Análise léxica

- Processo de conversão de uma sequência de caracteres em uma sequência de palavras;

- Ou seja, como identificar palavras em um texto?

- Usando somente espaços??

Análise léxica

1. Dígitos

2. Hífens

3. Marcas de pontuação

4. Caixa das palavras (maiúsculas e minúsculas)

Análise léxica - 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 indexação;
- Procedimentos específicos podem ser empregados para normalizar datas e números.

Análise léxica - Hifens

- 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 hifens como parte integral delas;
  - Guarda-chuva, B-52...
- Adote uma regra geral, mas tome cuidado com as exceções...

Análise léxica – 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 – Caixas das palavras

- O fato das letras estarem em maiúsculo ou minúsculo normalmente não é importante para a identificação de termos de índice;

- O analisador léxico normalmente converte todo o texto para maiúsculas ou minúsculas;

- Em alguns casos a semântica pode ficar comprometida:
  - Banco e banco.

Eliminação de stopwords

Eliminação 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...)

Eliminação de stopwords

- Eliminação de stopwords proporciona a redução do tamanho da estrutura de indexação;
- Apesar dos benefícios, a eliminação de stopwords pode reduzir a revocação:
  - Procurar a frase “ser ou não ser”;
  - A eliminação de stopwords deixaria somente o termo “ser”;
  - Essa é a razão para a adoção de um índice textual completo por algumas máquinas de busca na Web.
- Existem listas que contemplam stopwords de determinados idiomas..
  - https://gist.github.com/alopes/5358189

Eliminação de stopwords

stopwords.txt

```
de
a
o
que
e
do
da
em
um
para
é
com
não
uma
os
no
se
na
por
mais
as
dos
como
mas
foi
ao
ele
```

Stemming de palavras

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.

Stemming

- Stem (radical) é a porção de uma palavra que resta após a remoção de afixos (prefixos e sufixos);
- Exemplos
  - Connect = Connected, connecting, connection, connections...
  - Automa = automata, automático, automação...

- Acredita-se que os stems sejam úteis na melhoria de performance da recuperação, porque eles reduzem as variantes da mesma palavra raiz para um conceito comum.
  - Também reduz o tamanho da estrutura de indexação!

Stemming

- Existe muita controvérsia na literatura sobre os benefícios do stemming na performance da recuperação;
- Em determinadas línguas o stemming pode ser difícil de se realizar, exigindo buscas em tabelas externas e algoritmos específicos;
- Muitas máquinas de busca não adotam algoritmos de stemming.
  - Algoritmo de Porter para a língua inglesa.
  - RSLP para a língua portuguesa (sufixo) - http://www.inf.ufrgs.br/~viviane/rslp/

Seleção de termos ou palavras-chave

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.

Seleção de palavras-chave – Representação parcial

- Na área de biblioteconomia, a seleção de termos de índice é usualmente feita por um especialista, usando uma taxonomia e um vocabulário controlado;
  - Exemplo: códigos e termos encontrados em livros;

- Uma abordagem alternativa consiste em selecionar candidatos a termos de índice automaticamente
  - Seleção de grupos de substantivos.

Seleção de palavras-chave – Representação parcial

- Uma sentença em um texto em linguagem natural é normalmente composta de substantivos, pronomes, artigos, verbos, adjetivos, advérbios e conectivos;

- A maior parte da semântica é transportada pelos substantivos:
  - Eliminação sistemática de verbos, adjetivos, advérbios, conectivos, artigos e pronomes;
  - Agrupar substantivos próximos (ex: ciência da computação, redes de computadores...)

Exercício

Motivação

Meu nome é Walter Hartwell White, moro na Alameda Riacho Negro, 308 Cidade de Albuquerque, no Novo México, CEP 87104. Para as entidades de imposição da lei, isso não é uma confissão. Estou falando para minha família agora.

Skyler... você é o amor da minha vida.

Espero que saiba disso. Walter Júnior... você é o meu rapazão.

Hav...Haverá algumas...

coisas... coisas... que vocês descobrirão sobre mim logo… em poucos dias. Só quero que vocês saibam que não... não importa o que isso pareça, eu só tenho vocês no meu coração. Adeus.

(continua...)

Comentários

No decorrer da aula vimos…

- Como conduzir o processo de pré-processamento de documentos;
- O referido processo é feito após a coleta dos documentos e antes da criação dos termos de indexação;
- Existem diversas técnicas usadas para pré-processar documentos;
- É muito comum usar bibliotecas prontas + ajustes feitos pelo próprio programador.

Estudos

- Recuperação de Informação: Conceitos e Tecnologia das Máquinas de Busca
  - Capítulo 5.6
- https://medium.com/turing-talks/uma-análise-de-dom-casmurro-com-nltk-343d72dd47a7
- https://medium.com/turing-talks/introdução-ao-processamento-de-linguagem-natural-com-baco-exu-do-blues-17cbb7404258

Próximas aulas

- Avaliação da recuperação da informação.