

Mineração de Textos

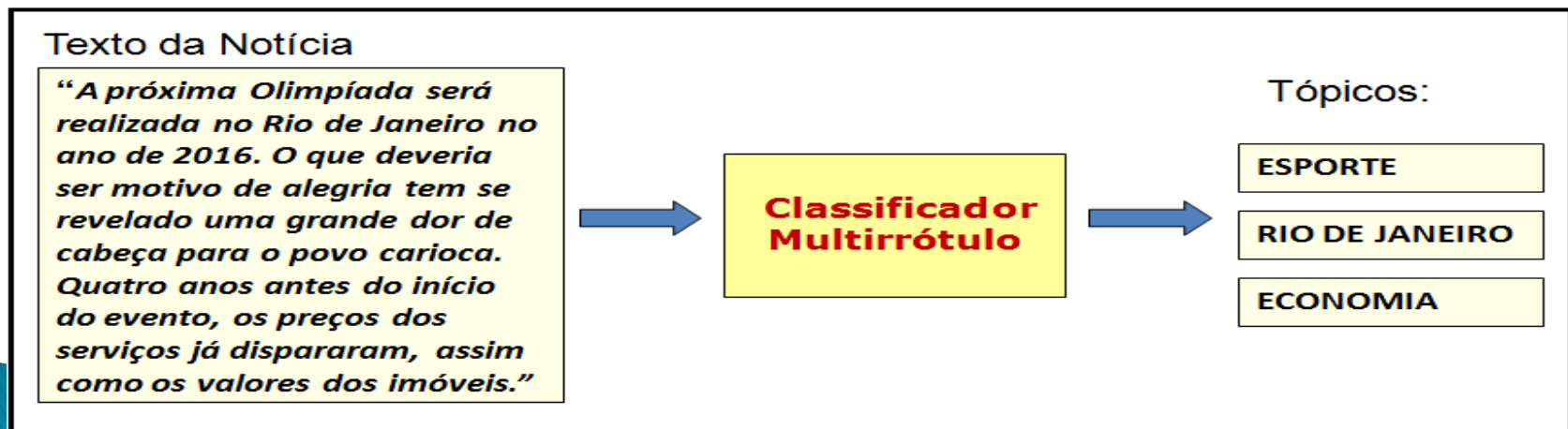
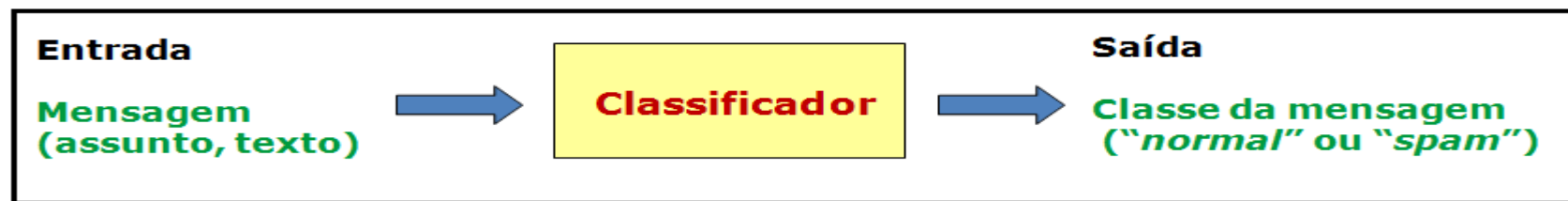
Jones Granatyr



Introdução

- ▣ Diversas formas de armazenamento
 - Livros
 - Jornais
 - Revistas
 - Páginas web
 - Blogs
 - Redes sociais
 - E-mails
 - Arquivos PDF
 - Arquivos XML
 - Arquivos JSON
- ▣ Geralmente não possuem um “esquema” para descrever sua estrutura
- ▣ Texto livre x Texto formatado

Classificação



Agrupamento (clustering)

- ▶ IBGE na descoberta de bairros com nomes similares (Jardim América e Jdim América)
- ▶ Detecção de plágio

Tao.xml <pre><?xml version="1.0"?> <produto tipo="livro"> <titulo>Tao Te Ching</titulo> <autor>Lao-Tsé</autor> <assunto>Religião</assunto> </produto></pre>	Utopia.xml <pre><?xml version="1.0"?> <livro> <titulo>Utopia</titulo> <ano>1516</ano> <autor> <nome>Thomas More</nome> <pais>Inglaterra</pais> </autor> <assunto>Filosofia</assunto> <assunto>Política</assunto> </livro></pre>	C L U S T E R 1
Brasil.xml <pre><?xml version="1.0"?> <pais sigla="BR"> <nome>Brasil</nome> <populacao>196.655.014</populacao> </pais></pre>		C L U S T E R 2

Extração da Informação

Texto

“Apesar de ter sido escrito em 1516, Utopia continua sendo um dos mais interessantes livros sobre pensamento político. A obra de Thomas More descreve uma ilha imaginária onde não existe propriedade privada e todos se preocupam com o bem da coletividade. A nova edição foi lançada pela Editora XYZ e está sendo vendida por R\$ 9,90.”

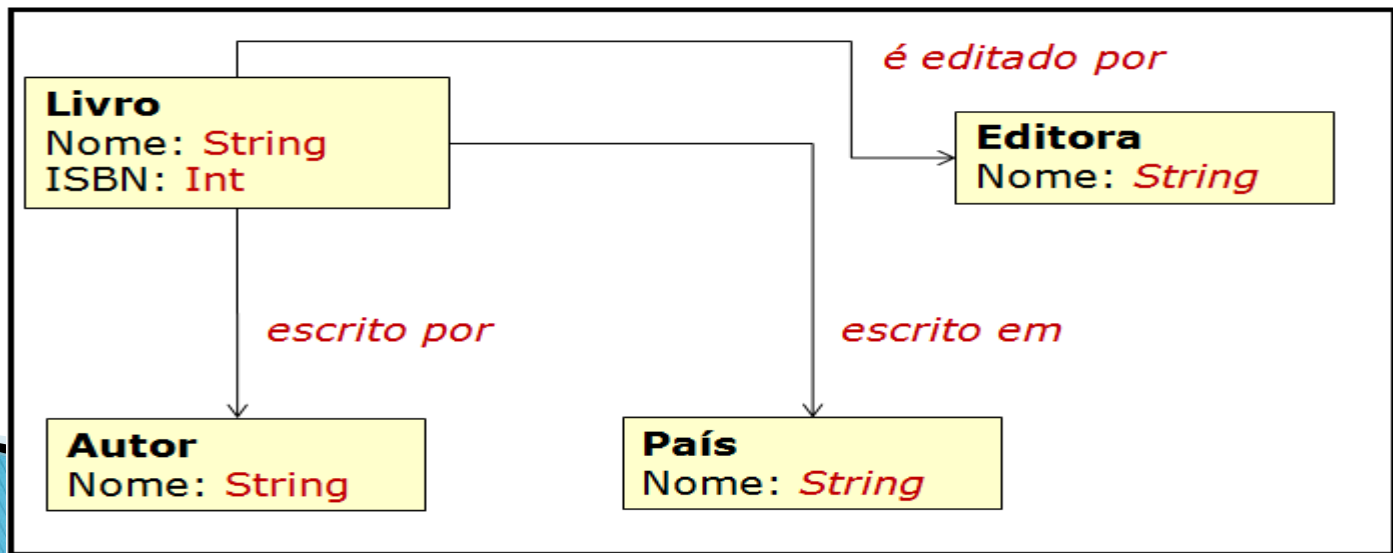


Template

Livro: “Utopia”
Ano: 1516
País: -
Autor: “Thomas More”
Editora: “XYZ”
Preço: 9,90

Extração da Informação

- ▶ Utilização de ontologias, representando conceitos e relacionamentos em um determinado domínio



Associações

► Correlação entre as palavras

“60% dos textos que contêm a palavra ‘Internacional’ também contêm a palavra ‘Grêmio’ . 3% de todos os textos contêm ambas as palavras”

Representação: {"Internacional"} \Rightarrow {"Grêmio"}

“A presença do termo ‘Pelé’ aumenta em 5 vezes a chance de ocorrência dos termos ‘Copa’ e ‘1970’”

Representação: {"Pelé"} \Rightarrow {"Copa", "1970"}

Casamento de esquemas

▶ Correspondências semânticas

S1

FUNCIONARIO	
🔑	ID_FUNC: INTEGER
◆	NOME: VARCHAR(60)
◆	ENDereco: VARCHAR(100)
◆	DAT_NASC: DATE

S2

EMPREGADO	
🔑	COD_EMP: INTEGER
◆	NOME_EMP: VARCHAR(50)
◆	LOGRADOURO: VARCHAR
◆	MUNICIPIO: VARCHAR
◆	UF: CHAR(2)
◆	CEP: INTEGER

[null]

Consulta do usuário

"encontre os livros
do Eduardo Suplicy"

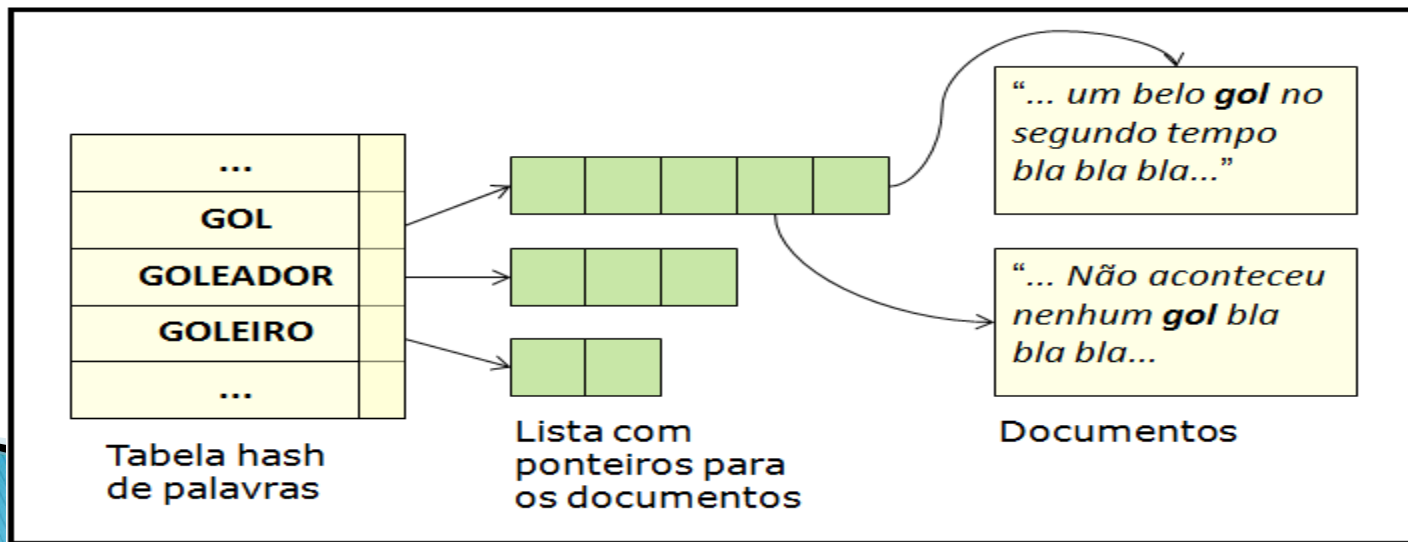
matcher

Consulta após conversão feita pelo *matcher*

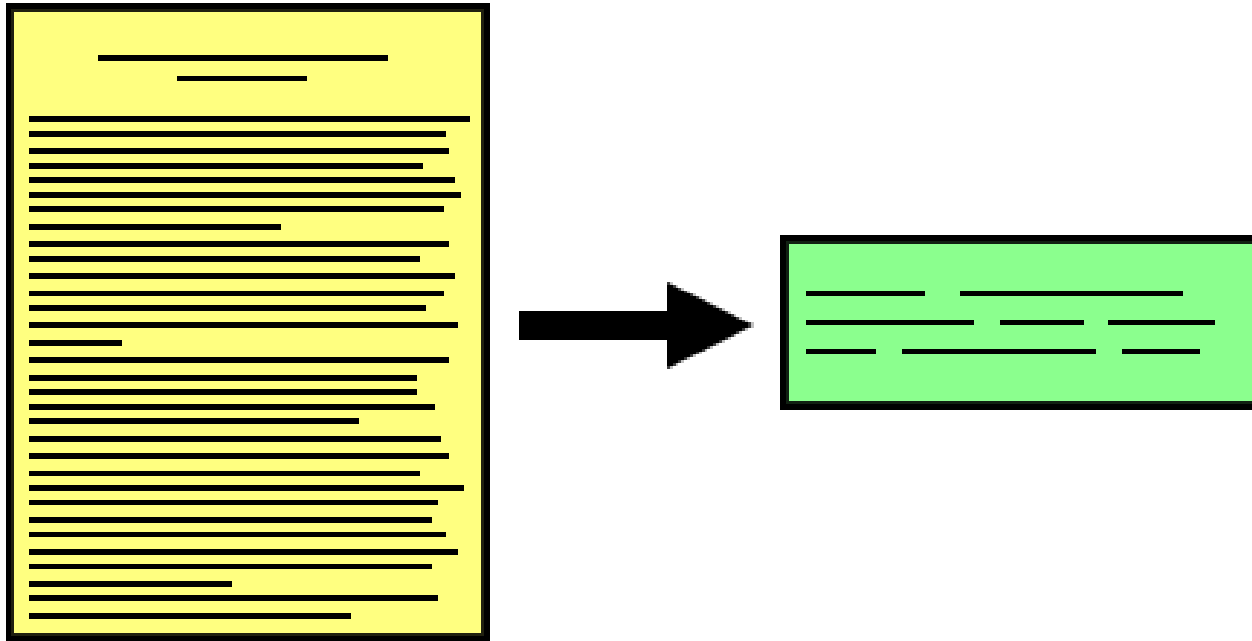
```
SELECT titulo, ano, resumo
FROM t_publicacao
WHERE autor = "Eduardo Suplicy"
AND tipo_pub = "livro"
```


Recuperação da informação

- ▶ Localizar e ranquear documentos relevantes em uma coleção
- ▶ Indexação (API Lucene)



Sumarização de documentos




Abordagens da mineração de textos

▶ Estatística

- Frequência dos termos, ignorando informações semânticas

▶ Processamento de linguagem natural

- Interpretação sintática e semântica das frases
 - Fazer o computador entender textos escritos em linguagem humana
- 

Referência

- ▶ GONÇALVES, Eduardo Corrêa. **Mineração de texto**. SQL Magazine. Rio de Janeiro, n. 105, 2012.

Conclusão