Recuperação da Informação

Esdras Lins Bispo Jr.
bispojr@ufg.br

Inteligência Artificial
Bacharelado em Ciência da Computação

24 de outubro de 2016

UFG
Campus Jataí

Plano de Aula

1. Pensamento
2. Recuperação da Informação - RI
   - Matriz de incidência termo-documento

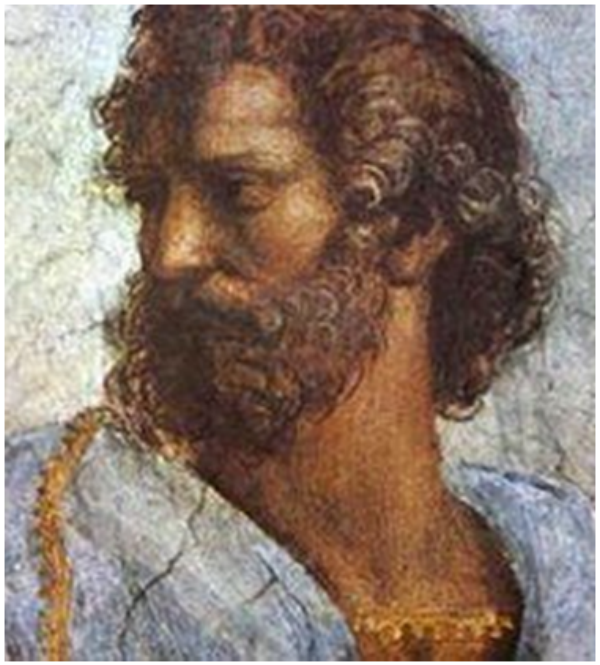
UFG
Campus Jataí

Sumário

Pensamento

UFG
Campus Jataí

Pensamento



Frase

Nós somos aquilo que fazemos repetidamente. Excelência, então, não é um modo de agir, mas um hábito.

Quem?

Aristóteles (384 a.C. - 322 a.C.)
Filósofo e lógico grego.

UFG
Campus Jataí

Sumário

UFG
Campus Jataí

Introdução à RI e Busca Web

O que é RI

Recuperação da Informação (*Information Retrieval*) é a atividade de encontrar material (normalmente documentos) de natureza não-estruturada (normalmente textos) que satisfaz uma necessidade de informação a partir de grandes coleções (normalmente armazenadas em computadores).

Introdução à RI e Busca Web

O que é RI

Recuperação da Informação (*Information Retrieval*) é a atividade de encontrar material (normalmente documentos) de natureza não-estruturada (normalmente textos) que satisfaz uma necessidade de informação a partir de grandes coleções (normalmente armazenadas em computadores).

Aplicações

Associamos RI diretamente à busca web, mas existem outras aplicações:

- Busca por emails;
- Busca por arquivos no seu PC;
- Recuperação de informações legais.

RI *versus* Banco de Dados

Dados estruturados

Tendem a referir informações através de tabelas:

Empregado	Gerente	Salário
Smith	Jones	R$ 50.000,00
Chang	Smith	R$ 60.000,00
Ivy	Smith	R$ 50.000,00

UFG
Campus Jataí

RI *versus* Banco de Dados

Dados estruturados

Tendem a referir informações através de tabelas:

Empregado	Gerente	Salário
Smith	Jones	R$ 50.000,00
Chang	Smith	R$ 60.000,00
Ivy	Smith	R$ 50.000,00

Características...

Normalmente é permitido realizar consultas exatas (através de texto)

UFG
Campus Jataí

RI *versus* Banco de Dados

Dados estruturados

Tendem a referir informações através de tabelas:

Empregado	Gerente	Salário
Smith	Jones	R$ 50.000,00
Chang	Smith	R$ 60.000,00
Ivy	Smith	R$ 50.000,00

Características...

Normalmente é permitido realizar consultas exatas (através de texto)

Exemplo: Salário $<$ 60000 AND Gerente $=$ Smith

UFG
Campus Jataí

RI *versus* Banco de Dados

Dados não-estruturados

- Normalmente refere-se a textos livres;
- Permite consultas por palavras-chave (incluindo operadores);
- Modelo clássico de busca por documentos de texto.

UFG
Campus Jataí

Pressupostos básicos em RI

Coleção

Um conjunto de documentos
(assumimos ser estático, neste momento).

Pressupostos básicos em RI

Coleção

Um conjunto de documentos
(assumimos ser estático, neste momento).

Objetivo

Recuperar documentos com informação que é relevante para **as necessidades de informação** do usuário e ajudá-lo a completar uma **tarefa**.

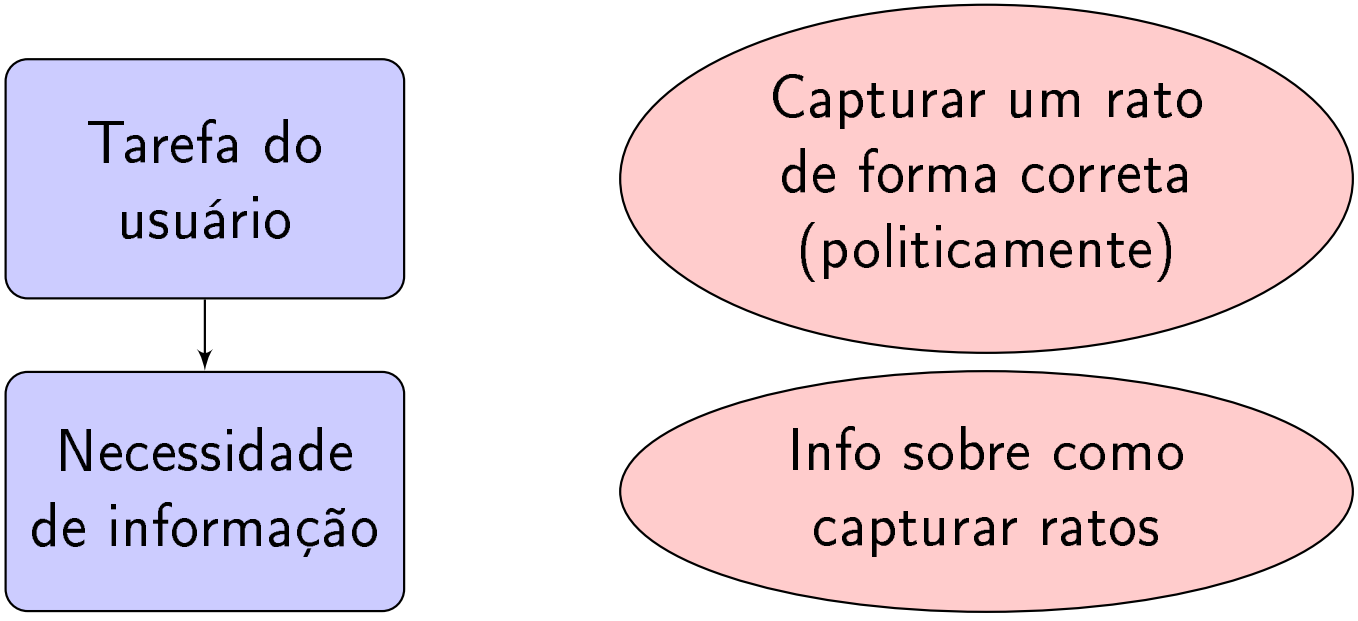
UFG
Campus Jataí

Teste

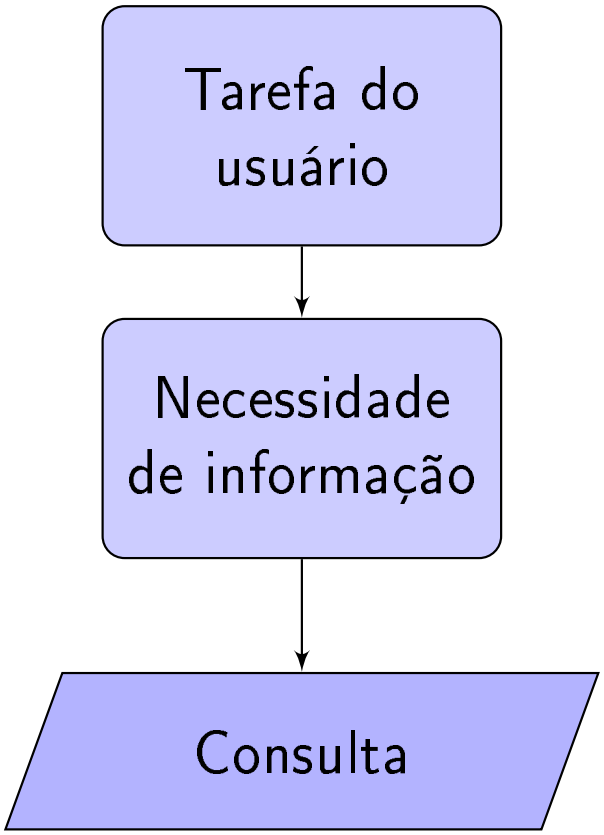
Tarefa do usuário

Capturar um rato de forma correta (politicamente)

Teste


Tarefa do usuário
Necessidade de informação
Capturar um rato de forma correta (politicamente)
Info sobre como capturar ratos

Teste

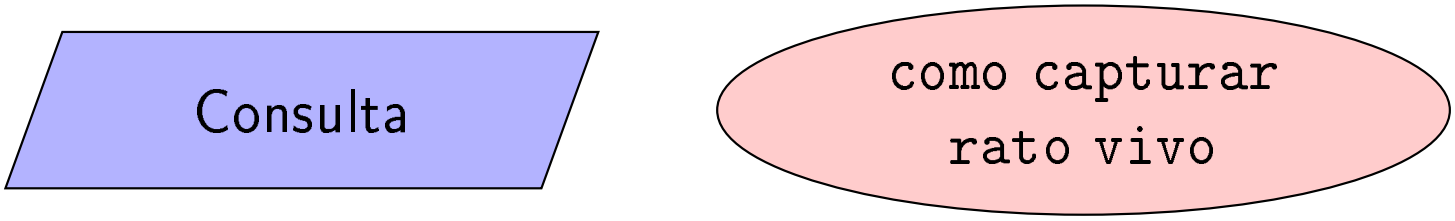


Capturar um rato de forma correta (politicamente)

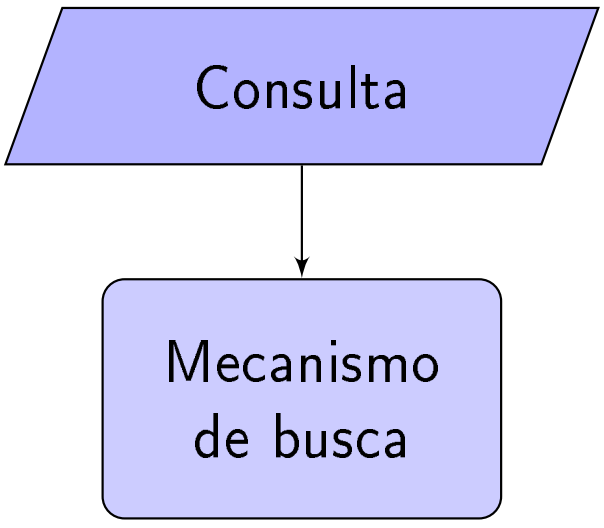
Info sobre como capturar ratos

como capturar rato vivo

Teste

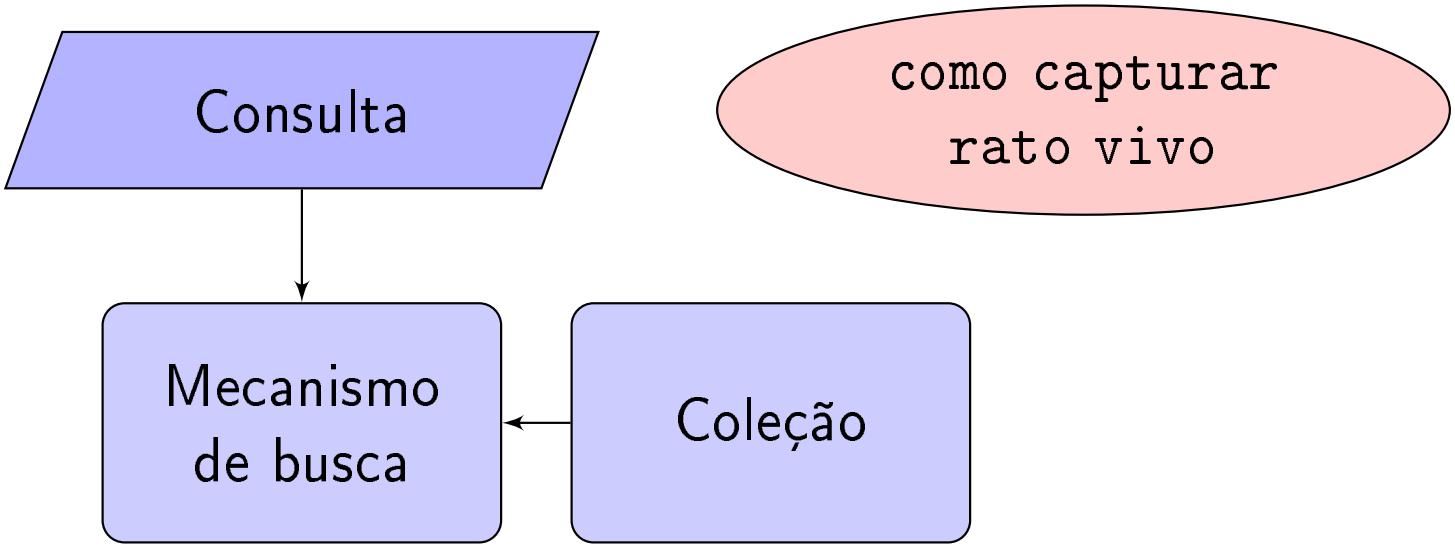


Teste

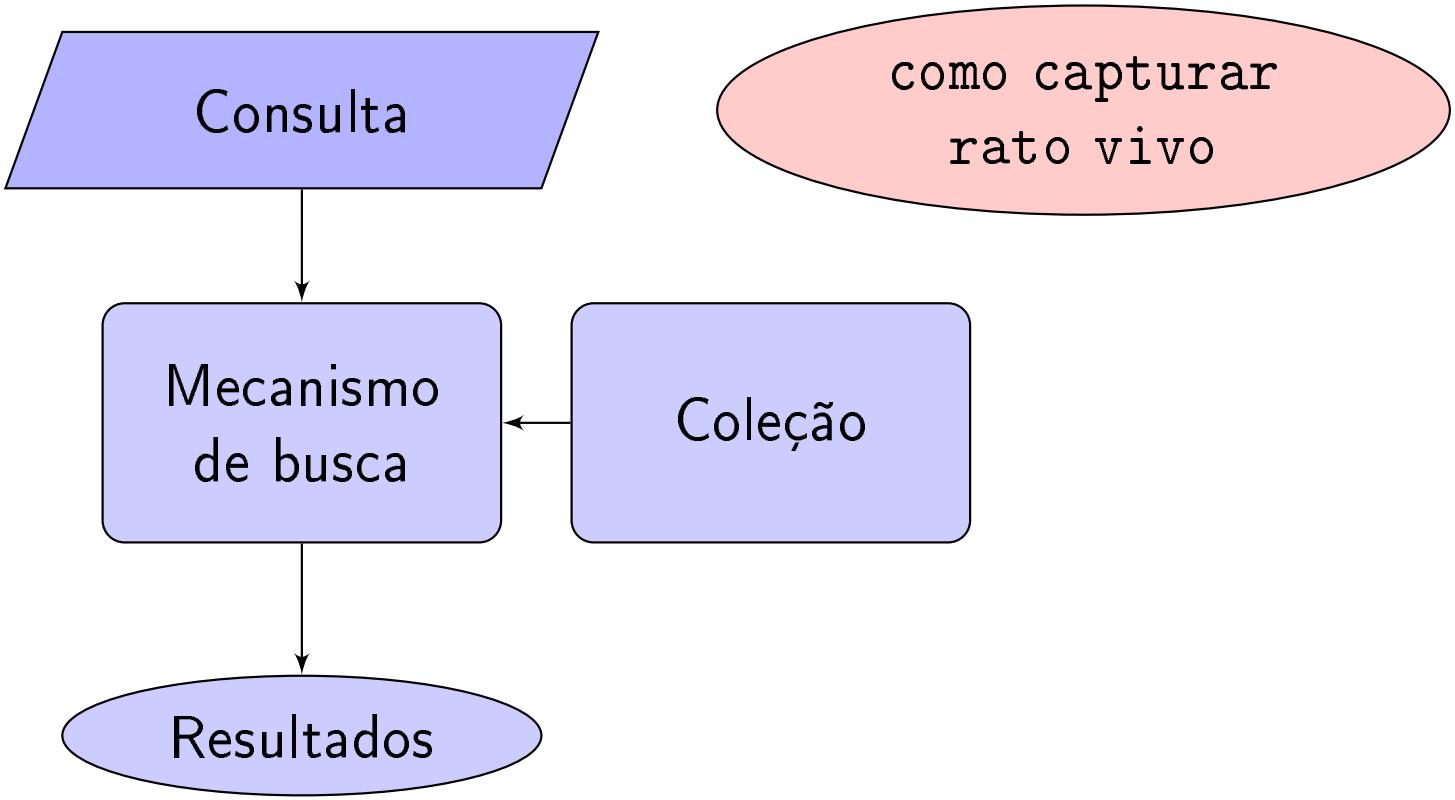


como capturar rato vivo

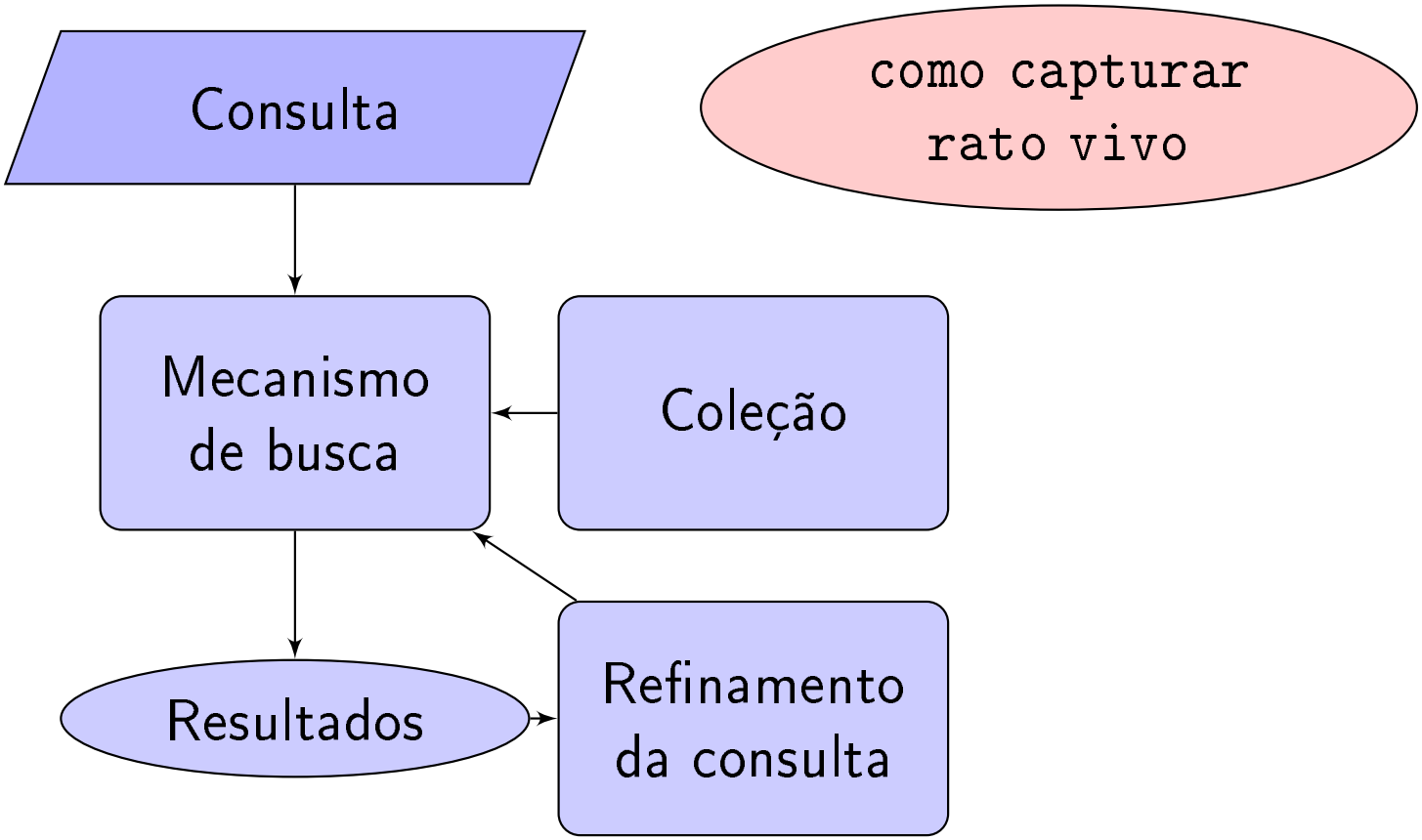
Teste


Consulta
como capturar
rato vivo
Mecanismo
de busca
Coleção

Teste


Consulta
como capturar
rato vivo
Mecanismo
de busca
Coleção
Resultados

Teste


Consulta
como capturar rato vivo
Mecanismo de busca
Coleção
Resultados
Refinamento da consulta

Dados não-estruturados em 1620

Obras de Shakespeare

- Quais peças de Shakespeare contêm as palavras 'Brutus' e 'Caesar', mas não 'Capurnia'?

Dados não-estruturados em 1620

Obras de Shakespeare

- Quais peças de Shakespeare contêm as palavras 'Brutus' e 'Caesar', mas não 'Capurnia'?
- Poderíamos fazer um `grep all` das peças de Shakespeare para 'Brutus' e 'Caesar', e daí retirar as linhas que contêm 'Calpurnia'?

Dados não-estruturados em 1620

Obras de Shakespeare

- Quais peças de Shakespeare contêm as palavras 'Brutus' e 'Caesar', mas não 'Capurnia'?
- Poderíamos fazer um `grep all` das peças de Shakespeare para 'Brutus' e 'Caesar', e daí retirar as linhas que contêm 'Calpurnia'?
- Por que não deveríamos fazer isto?

Dados não-estruturados em 1620

Obras de Shakespeare

- Quais peças de Shakespeare contêm as palavras 'Brutus' e 'Caesar', mas não 'Capurnia'?
- Poderíamos fazer um `grep all` das peças de Shakespeare para 'Brutus' e 'Caesar', e daí retirar as linhas que contêm 'Calpurnia'?
- Por que não deveríamos fazer isto?
  - Lento (para coleções grandes);

Dados não-estruturados em 1620

Obras de Shakespeare

- Quais peças de Shakespeare contêm as palavras 'Brutus' e 'Caesar', mas não 'Capurnia'?
- Poderíamos fazer um `grep all` das peças de Shakespeare para 'Brutus' e 'Caesar', e daí retirar as linhas que contêm 'Calpurnia'?
- Por que não deveríamos fazer isto?
  - Lento (para coleções grandes);
  - NOT 'Calpurnia' não é trivial;

UFG
Campus Jataí

Dados não-estruturados em 1620

Obras de Shakespeare

- Quais peças de Shakespeare contêm as palavras 'Brutus' e 'Caesar', mas não 'Capurnia'?
- Poderíamos fazer um `grep all` das peças de Shakespeare para 'Brutus' e 'Caesar', e daí retirar as linhas que contêm 'Calpurnia'?
- Por que não deveríamos fazer isto?
  - Lento (para coleções grandes);
  - NOT 'Calpurnia' não é trivial;
  - Outras operações não são viáveis (e.g. encontrar a palavra 'Romans' próximo de 'countrymen').

Matriz incidência termo-documento

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	...
Antony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
worser	1	0	1	1	1	0	
...							

► **Figure 1.1** A term-document incidence matrix. Matrix element (t, d) is 1 if the play in column d contains the word in row t, and is 0 otherwise.

Matriz incidência termo-documento

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	...
Antony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
worser	1	0	1	1	1	0	
...							

► **Figure 1.1** A term-document incidence matrix. Matrix element (t, d) is 1 if the play in column d contains the word in row t, and is 0 otherwise.

Pergunta...

Como fazer a consulta: Brutus AND Caesar BUT NOT Calpurnia ?

Recuperação da Informação

Esdras Lins Bispo Jr.
bispojr@ufg.br

Inteligência Artificial
Bacharelado em Ciência da Computação

24 de outubro de 2016

UFG
Campus Jataí