

# Sistemas de Recomendação

## Recomendadores Content-based

*Representação dos Itens*



Pós-Graduação Lato Sensu

**Nícollas Silva**

# Modelo Trivial



# Representação dos Itens

	Título	Autor	Tamanho	Pop G+	Gênero	Descrição	Personagens
$i_1$	Guerra dos Tronos	George R. R. Martin	694	94%	Ação Drama Aventura	Guerra dos Tronos é o primeiro livro da série de fantasia épica As Crônicas de Gelo e Fogo.	Jhon Snow Daenerys T. Cersei L.
$i_2$	Fúria dos Reis	George R. R. Martin	761	93%	Ação Drama Aventura	Fúria dos Reis é o segundo livro da série de fantasia épica As Crônicas de Gelo e Fogo	Jhon Snow Daenerys T. Cersei L.

# Representação dos usuários

	Título	Autor	Tamanho	Pop G+	Gênero	Descrição	Personagens
$i_1$	Guerra dos Tronos	George R. R. Martin	694	94%	Ação Drama Aventura	Guerra dos Tronos é o primeiro livro da série de fantasia épica As Crônicas de Gelo e Fogo.	Jhon Snow Daenerys T. Cersei L.
$i_2$	Fúria dos Reis	George R. R. Martin	761	93%	Ação Drama Aventura	Fúria dos Reis é o segundo livro da série de fantasia épica As Crônicas de Gelo e Fogo	Jhon Snow Daenerys T. Cersei L.

	Título	Autor	Tamanho	Pop G+	Gênero	Descrição	Personagens
$u_1$	Guerra dos Tronos Fúria dos Reis	George R. R. Martin	727.5	93.5%	Ação Drama Aventura	... livro da série de fantasia épica As Crônicas de Gelo e Fogo ...	Jhon Snow Daenerys T. Cersei L.

# Tarefa de predição

Título	Autor	Tamanho	Pop G+	Gênero	Descrição	Personagens
Guerra dos Tronos Fúria dos Reis	George R. R. Martin	727.5	93.5%	Ação Drama Aventura	... livro da série de fantasia épica As Crônicas de Gelo e Fogo ...	Jhon Snow Daenerys T. Cersei L.

Solução simples:

- Calcular a sobreposição de *features* entre o perfil do usuário e um item
  - Coeficiente de Dice

$$sim(u_1, i_3) = \frac{2|k(u_1) \cap k(i_3)|}{|k(u_1)| + |k(i_3)|}$$

- **Tokenização**

- Como indexar uma palavra?
  - “*information retrieval*” => *information* + *retrieval* ?
  - 信息检索 => 信息 + 检索 ?

- **Normalização dos termos**

- Estratégias de *stemming* para reduzir os termos em *tokens*
  - “*universal*” => “*univers*”
  - “*university*” => “*univers*”
  - “*universe*” => “*univers*”

- **Stop-words**

- Devem descartar termos recorrentes, como preposições e artigos
  - *a, an, and, be, will, were, that, ...*
  - *“To be or not to be: that is the question”?*

- **Informatividade das palavras**

- *information* ocorre em 35% dos itens
- *retrieval* ocorre em 0.1% dos itens
- Qual delas é mais discriminativa?

# Modelo Vetorial





# Representação dos Itens

- Cada item é um vetor de *componentes* com valores  $\{0, 1\}$

	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
ANTHONY	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
...						

# Representação Binária

- Cada item é um vetor de *componentes* com valores  $\{0, 1\}$ 
  - Existe um ou mais componentes para cada item.
  - Possui uma alta dimensionalidade.
  - É uma representação altamente esparsa.

***Quão representativo é  
cada termo?***

	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
ANTHONY	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
WORSE	1	0	1	1	1	0
...						

# Term-frequency

- $TF_{td}$  : é a frequência do termo  $t$  em um documento  $d$

	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
ANTHONY	157	73	0	0	0	1
BRUTUS	4	157	0	2	0	0
CAESAR	232	227	0	2	1	0
CALPURNIA	0	10	0	0	0	0
CLEOPATRA	57	0	0	0	0	0
MERCY	2	0	3	8	5	8
WORSER	2	0	1	1	1	5
...						

- $TF_{td}$  : é a frequência do termo  $t$  em um documento  $d$ 
  - Como ficaria uma palavra popular como “*kill*”?
  - Ela seria capaz de descrever algum item?

***Quão discriminativo é  
cada termo?***

	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
ANTHONY	157	73	0	0	0	1
BRUTUS	4	157	0	2	0	0
CAESAR	232	227	0	2	1	0
CALPURNIA	0	10	0	0	0	0
CLEOPATRA	57	0	0	0	0	0
MERCY	2	0	3	8	5	8
WORSER	2	0	1	1	1	5
...						

- $TF_{td}$  : é a frequência do termo  $t$  em um documento  $d$
- $IDF_t$  : frequência inversa do documento

$$IDF_i = \log \frac{n}{n_i}$$

- $n$  : número total de itens na coleção
  - $nt$  : número de itens onde o termo  $t$  aparece
- $TF_{td} * IDF_t$  : melhor solução para penalizar os termos

# Representação TF-IDF

Representação vetorial:

	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
ANTHONY	5.25	3.18	0.0	0.0	0.0	0.35
BRUTUS	1.21	6.10	0.0	1.0	0.0	0.0
CAESAR	8.59	2.54	0.0	1.51	0.25	0.0
CALPURNIA	0.0	1.54	0.0	0.0	0.0	0.0
CLEOPATRA	2.85	0.0	0.0	0.0	0.0	0.0
MERCY	1.51	0.0	1.90	0.12	5.25	0.88
WORSER	1.37	0.0	0.11	4.15	0.25	1.95
...						