Recuperação de Informação: O modelo de espaço vetorial

Marcelo Keese Albertini

Faculdade de Computação - UFU

Veremos hoje

- Modelo de espaço de vetores: representação vetorial

Matriz de incidência binária

	Marco Antônio	Júlio César	A Tempestade	Hamlet	Otelo	Macbeth	...
Antônio	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
César	1	1	0	1	1	1	
Calpúrnia	0	1	0	0	0	0	
Cleópatra	1	0	0	0	0	0	
...							

Cada documento é representado como um vetor binário $\in \{0,1\}^{|V|}$.

Matriz de contagem

	Marco Antônio	Júlio César	A Tempestade	Hamlet	Otelo	Macbeth	...
Antônio	157	73	0	0	0	1	
Brutus	4	157	0	2	0	0	
César	232	227	0	2	1	0	
Calpurnia	0	10	0	0	0	0	
Cleópatra	57	0	0	0	0	0	
...							

Cada documento é representado como vetor de contagem $\in \mathbb{N}^{|V|}$.

Binário → contagem → matriz de pesos

	Marco Antônio	Júlio César	A Tempestade	Hamlet	Otelo	Macbeth	...
Antônio	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	
César	8.59	2.54	0.0	1.51	0.25	0.0	
Calpúrnia	0.0	1.54	0.0	0.0	0.0	0.0	
Cleópatra	2.85	0.0	0.0	0.0	0.0	0.0	
misericórdia	1.51	0.0	1.90	0.12	5.25	0.88	
pior	1.37	0.0	0.11	4.15	0.25	1.95	
...							

Cada documento é representado como um vetor de valores reais de pesos tf-idf $\in \mathbb{R}^{|V|}$.

Documentos na forma de vetores

- Cada documento é representado em um vetor de pesos tf-idf $\in \mathbb{R}^{|V|}$.
- Então temos um espaço vetorial com $|V|$ dimensões.
- Termos são eixos do espaço.
- Documentos são pontos ou vetores nesse espaço.
- Alto número de dimensões: dezenas de milhões de dimensões em mecanismos de busca
- Cada vetor usa muito espaço (maior parte das dimensões é zero)

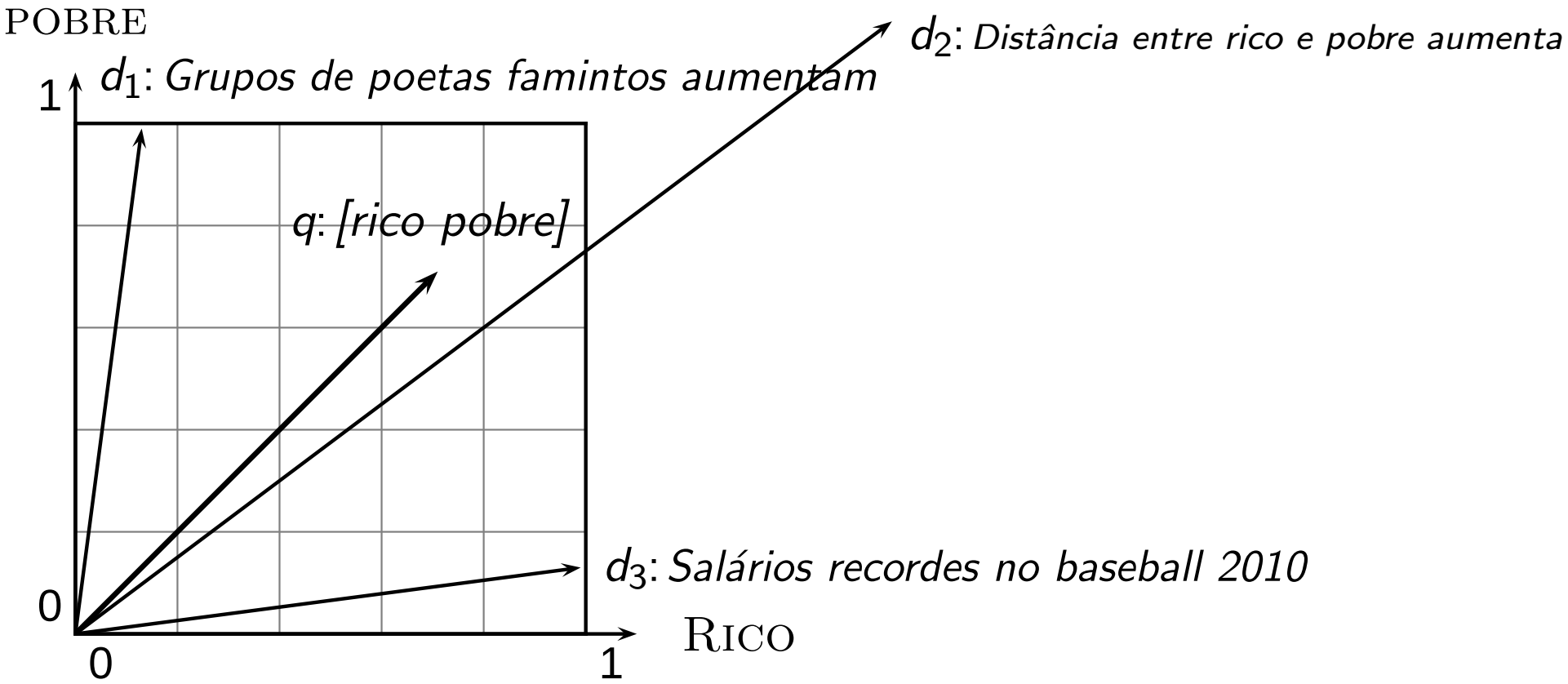
Consultas como vetores

- Ideia 1: fazer o mesmo para as consultas: representar no espaço de alta-dimensionalidade
- Ideia 2: Rankear documentos de acordo com sua proximidade à consulta
- proximidade = similaridade
- proximidade $\approx$ distância negativa
- Objetivo: estamos evitando modelo booleana e resultados tudo ou nada.
- Objetivo: rankear documentos relevantes em melhores posições que os não relevantes

Como formalizamos a similaridade no espaço vetorial

- distância (negativa) entre dois "pontos"
- (= distância entre pontos finais entre pares de vetores)
- Distância euclidiana
- Distância euclidiana é uma má ideia . . .
- . . . porque distância euclidiana é grande para vetores de diferentes comprimentos

Porque distância euclidiana é uma má ideia



A distância euclidiana de $\vec{q}$ e $\vec{d}_2$ é grande, embora a distribuição de termos na consulta q e a distribuição dos termo no documento d_2 são muito similares.

Perguntas sobre a configuração básica do espaço vetorial?

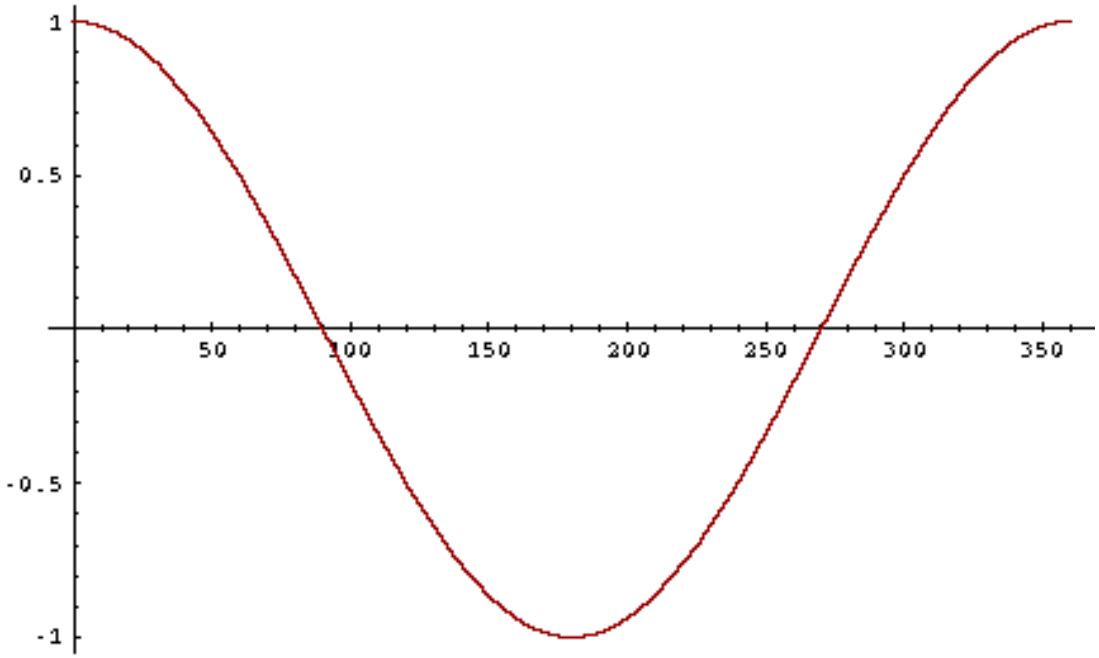
Usar ângulo em vez de distância

- Ordena documento de acordo com o ângulo em relação à consulta
- Avalie: pegue um documento d e adicione-o a si mesmo em d'.
- d e d' têm mesma informação
- O ângulo entre os dois documentos é 0, máxima similaridade
  ...
- ... mesmo que a distância euclidiana entre os dois documentos seja grande

De ângulos a cosenos

- As seguintes noções são equivalentes:
  - Ordenar documentos de acordo com o ângulo entre consulta e documento em ordem decrescente
  - Ordenar documentos de acordo com coseno(consulta, documento) em ordem crescente
- Coseno é uma função monotonicamente decrescente do ângulo para o intervalo $[0°, 180°]$

Coseno



Normalização de magnitude

- Como calcular o coseno?
- Um vetor pode ter magnitude normalizada a 1 com (norma L_2): $\vec{x} = \frac{\vec{x}}{||\vec{x}||}$
- Essa operação mapeia os vetores na unidade esférica ...
- Assim, documentos mais extensos ou curtos tem mesma informação
- Efeito nos documentos d e d' (d "dobrado") : mesmo vetor depois da normalização

Similaridade cosena entre consulta e documento

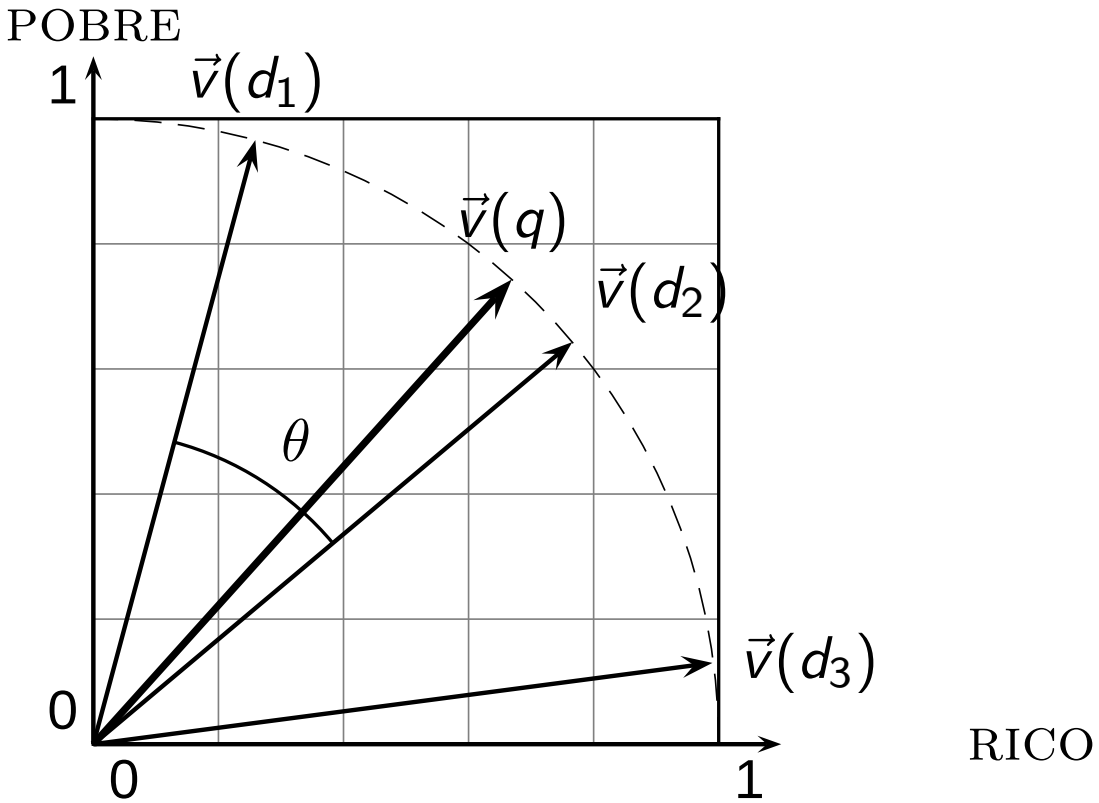
$$\cos(\vec{q}, \vec{d}) = \text{SIM}(\vec{q}, \vec{d}) = \frac{\vec{q} \cdot \vec{d}}{|\vec{q}||\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2} \sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

- q_i é o peso tf-idf do termo i na consulta.
- d_i é o peso tf-idf do termo i no documento.
- $|\vec{q}|$ e $|\vec{d}|$ são as magnitudes de $\vec{q}$ e $\vec{d}$.
- Esta é a similaridade coseno entre $\vec{q}$ e $\vec{d}$ . . . . . ou, de maneira equivalente, o coseno do ângulo entre $\vec{q}$ e $\vec{d}$.

Coseno para vetores normalizados

- Para vetores normalizados, o coseno é equivalente ao produto escalar (também conhecido como produto interno).
- $\cos(\vec{q}, \vec{d}) = \vec{q} \cdot \vec{d} = \sum_i q_i \cdot d_i$
  - (se $\vec{q}$ e $\vec{d}$ são normalizados).

Similaridade de coseno ilustrada



Coseno: exemplo

O quão similares são esses livros?

ReS: Razão e Sensibilidade

OeP: Orgulho e Preconceito

MVU: Colina dos Vendavais

frequência de termos (contagem)

termo	ReS	OeP	MVU
AFEIÇÃO	115	58	20
CIÚMES	10	7	11
FOFOCA	2	0	6
VENDAVAL	0	0	38

Coseno: exemplo

frequência de termos (tf)

termo	ReS	OeP	MVU
AFEIÇÃO	115	58	20
CIÚMES	10	7	11
FOFOCA	2	0	6
VENDAVAL	0	0	38

1.0+ log da frequência

termo	ReS	OeP	MVU
AFEIÇÃO	3.06	2.76	2.30
CIÚMES	2.0	1.85	2.04
FOFOCA	1.30	0	1.78
VENDAVAL	0	0	2.58

Para simplificar este exemplo, não usaremos idf.

Se fosse usar, como seria o cálculo?

$$idf_t = log \frac{N}{df_t}$$

presença de termos (df)

termo	ReS	OeP	MVU
AFEIÇÃO	1	1	1
CIÚMES	1	1	1
FOFOCA	1	0	1
VENDAVAL	0	0	1

idf

termo	idf
AFEIÇÃO	$\log(3/3) = 0$
CIÚMES	$\log(3/3) = 0$
FOFOCA	$\log(3/2) = 0.17$
VENDAVAL	$\log(3/1) = 0.47$

Coseno: exemplo

log da frequência

termo	ReS	OeP	MVU
AFEIÇÃO	3.06	2.76	2.30
CIÚMES	2.0	1.85	2.04
FOFOCA	1.30	0	1.78
VENDAVAL	0	0	2.58

log da frequência & normalização do coseno

termo	ReS	OeP	MVU
AFEIÇÃO	0.789	0.832	0.524
CIÚMES	0.515	0.555	0.465
FOFOCA	0.335	0.0	0.405
VENDAVAL	0.0	0.0	0.588

- $\cos(\text{ReS},\text{OeP}) \approx$
  $0.789 * 0.832 + 0.515 * 0.555 + 0.335 * 0.0 + 0.0 * 0.0 \approx 0.94.$
- $\cos(\text{ReS},\text{MVU}) \approx 0.79$
- $\cos(\text{OeP},\text{MVU}) \approx 0.69$

Componentes do peso tf-idf

Frequência de termos		Frequência em Documentos		Normalização	
n (natural)	$\mathrm{tf}_{t,d}$	n (não)	1	n (nenhum)	1
l (logaritmo)	$1+\log(\mathrm{tf}_{t,d})$	t (idf)	$\log \frac{N}{\mathrm{df}_t}$	c (cosine)	$\frac{1}{\sqrt{w_1^2+w_2^2+\ldots+w_M^2}}$
a (aumentado)	$0.5+\frac{0.5\times\mathrm{tf}_{t,d}}{\max_t(\mathrm{tf}_{t,d})}$	p (prob idf)	$\max\{0, \log \frac{N-\mathrm{df}_t}{\mathrm{df}_t}\}$	u (pivoted unique)	$1/u$
b (boolean)	$\begin{cases}1 & \text{if } \mathrm{tf}_{t,d}>0\\ 0 & \text{senão}\end{cases}$			b (byte size)	$1/CharLength^{\alpha}$, $\alpha<1$
L (log ave)	$\frac{1+\log(\mathrm{tf}_{t,d})}{1+\log(\mathrm{ave}_{t\in d}(\mathrm{tf}_{t,d}))}$				

Melhor combinação conhecida de opções de pesos

Padrão: sem peso

Exemplo tf-idf

- Frequentemente utiliza-se diferentes opções de pesos para consultas e documentos.
- Notação: ddd.qqq
- Exemplo: lnc.ltn
- documento: log tf, sem peso df, normalização coseno
- consulta: log tf, idf, sem normalização
- É ruim não colocar peso idf no documento?
- Exemplo consulta: “melhor seguro carro ”
- Exemplo documento: “melhor seguro carro auto ”

Exemplo tf-idf : lnc.ltn

Consulta: "melhor seguro carro". Documento: "carro seguro auto seguro".

palavra	consulta					documento				produto
	tf	tf-com-peso	df	idf	peso	tf	tf-com-peso	peso	norm.	
auto	0	0	5000	2.3	0	1	1	1	0.52	0
melhor	1	1	50000	1.3	1.3	0	0	0	0	0
carro	1	1	10000	2.0	2.0	1	1	1	0.52	1.04
seguro	1	1	1000	3.0	3.0	2	1.3	1.3	0.68	2.04

Colunas: tf: (sem peso) frequência de termo , tf-com-peso: log frequência de termo , df: frequência de documento , idf: frequência de documento inversa, peso: o peso final do termo na consulta ou documento, norm.: pesos de documentos depois de normalização , produto: produto do peso final da consulta e peso final do documento

$\sqrt{1^2 + 0^2 + 1^2 + 1.3^2} \approx 1.92$
$1/1.92 \approx 0.52$
$1.3/1.92 \approx 0.68$

Resultado final de similaridade entre consulta e documento:
$\sum_i w_{qi} \cdot w_{di} = 0 + 0 + 1.04 + 2.04 = 3.08$

Perguntas?

Resumo: recuperação ordenada no modelo de espaço vetorial

- Representar a consulta como um vetor tf-idf com pesos
- Representar cada documento como um vetor tf-idf com pesos
- Calcular a similaridade coseno entre o vetor consulta e vetor documento
- Rankear documentos em relação à consulta
- Exibir os K melhores resultados (e.g., $K = 10$) ao usuário