Veremos hoje

- Abordagem probabilística para ORI
- Princípio de ranking probabilístico
- Modelos: BIM, BM25
- Premissas desses modelos

Sumário

Retorno de relevância da última aula

- Aula anterior: em retorno de relevância: o usuário marca documentos como relevantes/não-relevantes

Retorno de relevância da última aula

- Aula anterior: em retorno de relevância: o usuário marca documentos como relevantes/não-relevantes
- Dados documentos relevantes e não-relevantes, computamos pesos para termos não pertencentes à consulta que indicam o quão provável eles ocorrerão em documentos relevantes

Retorno de relevância da última aula

- Aula anterior: em retorno de relevância: o usuário marca documentos como relevantes/não-relevantes
- Dados documentos relevantes e não-relevantes, computamos pesos para termos não pertencentes à consulta que indicam o quão provável eles ocorrerão em documentos relevantes
- Hoje: veremos uma abordagem probabilística para retorno de relevância um modelo probabilístico para ORI

Abordagem probabilística para recuperação

- Dada uma necessidade de usuário (representada como uma consulta) e uma coleção de documentos (transformada em representações de documentos), um sistema determinar o quão bem documentos satisfazem a consulta

Abordagem probabilística para recuperação

- Dada uma necessidade de usuário (representada como uma consulta) e uma coleção de documentos (transformada em representações de documentos), um sistema determinar o quão bem documentos satisfazem a consulta
  - Um sistema ORI tem entendimento incerto da consulta do usuário e faz uma tentativa incerta sobre se um documento satisfaz uma consulta

Abordagem probabilística para recuperação

- Dada uma necessidade de usuário (representada como uma consulta) e uma coleção de documentos (transformada em representações de documentos), um sistema determinar o quão bem documentos satisfazem a consulta
  - Um sistema ORI tem entendimento incerto da consulta do usuário e faz uma tentativa incerta sobre se um documento satisfaz uma consulta
- Teoria de probabilidades provê fundação para trabalhar sob incertezas

Abordagem probabilística para recuperação

- Dada uma necessidade de usuário (representada como uma consulta) e uma coleção de documentos (transformada em representações de documentos), um sistema determinar o quão bem documentos satisfazem a consulta
  - Um sistema ORI tem entendimento incerto da consulta do usuário e faz uma tentativa incerta sobre se um documento satisfaz uma consulta
- Teoria de probabilidades provê fundação para trabalhar sob incertezas
  - Modelos probabilísticos exploram tal fundação para estimar a probabilidade de um documento ser relevante para uma consulta

Modelos probabilísticos

- Modelo clássico de recuperação probabilística

Modelos probabilísticos

- Modelo clássico de recuperação probabilística
  - Princípio de ranking probabilístico

Modelos probabilísticos

- Modelo clássico de recuperação probabilística
  - Princípio de ranking probabilístico
    - Modelo de independência binária (BIM), BestMatch25 (Okapi)

Modelos probabilísticos

- Modelo clássico de recuperação probabilística
  - Princípio de ranking probabilístico
    - Modelo de independência binária (BIM), BestMatch25 (Okapi)
- Redes bayesianas para recuperação de textos

Modelos probabilísticos

- Modelo clássico de recuperação probabilística
  - Princípio de ranking probabilístico
    - Modelo de independência binária (BIM), BestMatch25 (Okapi)
- Redes bayesianas para recuperação de textos
- Abordagem de modelo de linguagem para ORI

Modelos probabilísticos

- Modelo clássico de recuperação probabilística
  - Princípio de ranking probabilístico
    - Modelo de independência binária (BIM), BestMatch25 (Okapi)
- Redes bayesianas para recuperação de textos
- Abordagem de modelo de linguagem para ORI
- Métodos probabilísticos são antigos, porém ainda muito pesquisados em ORI

Exercício: modelo probabilístico vs. outros modelos

- Modelo booleano

Exercício: modelo probabilístico vs. outros modelos

- Modelo booleano
  - Modelos probabilísticos suportam ranking e portanto são melhores que o booleano simples

Exercício: modelo probabilístico vs. outros modelos

- Modelo booleano
  - Modelos probabilísticos suportam ranking e portanto são melhores que o booleano simples
- Modelo de espaço vetorial

Exercício: modelo probabilístico vs. outros modelos

- Modelo booleano
  - Modelos probabilísticos suportam ranking e portanto são melhores que o booleano simples
- Modelo de espaço vetorial
  - Modelo de espaço vetorial também suporta ranking

Exercício: modelo probabilístico vs. outros modelos

- Modelo booleano
  - Modelos probabilísticos suportam ranking e portanto são melhores que o booleano simples
- Modelo de espaço vetorial
  - Modelo de espaço vetorial também suporta ranking
  - Porque queremos uma alternativa para o modelo vetorial?

Modelo Probabilístico vs. vetorial

- Modelo vetorial: ranking de documentos de acordo com similaridade à consulta

Modelo Probabilístico vs. vetorial

- Modelo vetorial: ranking de documentos de acordo com similaridade à consulta
- Noção de similaridade não traduz diretamente para a pergunta "quais documentos são bons para retornar para o usuário?"

Modelo Probabilístico vs. vetorial

- Modelo vetorial: ranking de documentos de acordo com similaridade à consulta
- Noção de similaridade não traduz diretamente para a pergunta "quais documentos são bons para retornar para o usuário?"
- Documento mais similar pode ser muito relevante ou completamente irrelevante

Modelo Probabilístico vs. vetorial

- Modelo vetorial: ranking de documentos de acordo com similaridade à consulta
- Noção de similaridade não traduz diretamente para a pergunta "quais documentos são bons para retornar para o usuário?"
- Documento mais similar pode ser muito relevante ou completamente irrelevante
- Teoria de probabilidades é uma formalização mais limpa do que queremos com um sistema de ORI: encontrar documentos relevantes para o usuário

Sumário

Teoria de probabilidades

- Para eventos A e B

Teoria de probabilidades

- Para eventos A e B
  - Probabilidade conjunta $P(A \cap B)$ de ambos eventos acontecerem

Teoria de probabilidades

- Para eventos A e B
  - Probabilidade conjunta $P(A \cap B)$ de ambos eventos acontecerem
  - Probabilidade condicional $P(A|B)$ do evento A acontecer dado que o evento B já aconteceu

Teoria de probabilidades

- Para eventos A e B
  - Probabilidade conjunta $P(A \cap B)$ de ambos eventos acontecerem
  - Probabilidade condicional $P(A|B)$ do evento A acontecer dado que o evento B já aconteceu
- Regra da cadeia descreve relação entre probabilidades conjuntas e probabilidades condicionais:

$$P(AB) = P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

Teoria de probabilidades

- Para eventos A e B
  - Probabilidade conjunta $P(A \cap B)$ de ambos eventos acontecerem
  - Probabilidade condicional $P(A|B)$ do evento A acontecer dado que o evento B já aconteceu
- Regra da cadeia descreve relação entre probabilidades conjuntas e probabilidades condicionais:

$$P(AB) = P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

- Similar ao complemento de um evento $P(\overline{A})$:

$$P(\overline{A}B) = P(B|\overline{A})P(\overline{A})$$

Teoria de probabilidades

- Para eventos A e B
  - Probabilidade conjunta $P(A \cap B)$ de ambos eventos acontecerem
  - Probabilidade condicional $P(A|B)$ do evento A acontecer dado que o evento B já aconteceu
- Regra da cadeia descreve relação entre probabilidades conjuntas e probabilidades condicionais:

$$P(AB) = P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

- Similar ao complemento de um evento $P(\overline{A})$:

$$P(\overline{A}B) = P(B|\overline{A})P(\overline{A})$$

- Regra da partição: se B pode ser dividido em um conjunto exaustivo de subcasos disjuntos, então $P(B)$ é a soma das probabilidades dos subcasos. Um caso especial dessa regra é:

$$P(B) = P(AB) + P(\overline{A}B)$$

Teoria de probabilidades

Teoria de probabilidades

Regra de Bayes para inverter probabilidades condicionais:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \left[\frac{P(B|A)}{\sum_{X\in\{A,\overline{A}\}} P(B|X)P(X)}\right] P(A)$$

Teoria de probabilidades

Regra de Bayes para inverter probabilidades condicionais:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \left[\frac{P(B|A)}{\sum_{X\in\{A,\overline{A}\}} P(B|X)P(X)}\right] P(A)$$

Pode ser imaginado como um jeito de atualizar as probabilidades:

Teoria de probabilidades

Regra de Bayes para inverter probabilidades condicionais:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \left[\frac{P(B|A)}{\sum_{X\in\{A,\overline{A}\}} P(B|X)P(X)}\right] P(A)$$

Pode ser imaginado como um jeito de atualizar as probabilidades:

- Iniciar com probabilidade a priori $P(A)$ (estimativa inicial da probabilidade do evento A na ausência de outras informações)

Teoria de probabilidades

Regra de Bayes para inverter probabilidades condicionais:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \left[\frac{P(B|A)}{\sum_{X\in\{A,\overline{A}\}} P(B|X)P(X)}\right] P(A)$$

Pode ser imaginado como um jeito de atualizar as probabilidades:

- Iniciar com probabilidade a priori $P(A)$ (estimativa inicial da probabilidade do evento A na ausência de outras informações)
- Obter probabilidade a posteriori $P(A|B)$ após observer a evidência B, baseada na possibilidade de B acontecer quando o evento A acontecer ou não acontecer $\overline{A}$

Sumário

O problema de ordenação de documentos (ranking)

- Cenário de recuperação em ranking: dada uma coleção de documentos, o usuário faz consulta e sistema retorna lista ordenada de documentos.

O problema de ordenação de documentos (ranking)

- Cenário de recuperação em ranking: dada uma coleção de documentos, o usuário faz consulta e sistema retorna lista ordenada de documentos.
- Usar a noção binária de relevância: $R_{d,q}$ é uma variável aleatória binária e

O problema de ordenação de documentos (ranking)

- Cenário de recuperação em ranking: dada uma coleção de documentos, o usuário faz consulta e sistema retorna lista ordenada de documentos.
- Usar a noção binária de relevância: $R_{d,q}$ é uma variável aleatória binária e
  - $R_{d,q} = 1$ se documento d é relevante em relação à consulta q

O problema de ordenação de documentos (ranking)

- Cenário de recuperação em ranking: dada uma coleção de documentos, o usuário faz consulta e sistema retorna lista ordenada de documentos.
- Usar a noção binária de relevância: $R_{d,q}$ é uma variável aleatória binária e
  - $R_{d,q} = 1$ se documento d é relevante em relação à consulta q
  - $R_{d,q} = 0$ caso contrário

O problema de ordenação de documentos (ranking)

- Cenário de recuperação em ranking: dada uma coleção de documentos, o usuário faz consulta e sistema retorna lista ordenada de documentos.
- Usar a noção binária de relevância: $R_{d,q}$ é uma variável aleatória binária e
  - $R_{d,q} = 1$ se documento d é relevante em relação à consulta q
  - $R_{d,q} = 0$ caso contrário
- Ranking probabilístico ordena documentos de acordo com:

$$P(R = 1|d, q)$$

O problema de ordenação de documentos (ranking)

- Cenário de recuperação em ranking: dada uma coleção de documentos, o usuário faz consulta e sistema retorna lista ordenada de documentos.
- Usar a noção binária de relevância: $R_{d,q}$ é uma variável aleatória binária e
  - $R_{d,q} = 1$ se documento d é relevante em relação à consulta q
  - $R_{d,q} = 0$ caso contrário
- Ranking probabilístico ordena documentos de acordo com:

$$P(R = 1|d, q)$$

- Assumir que a relevância de cada documento é independente da relevância de outros documentos

Princípio de Ranking Probabilístico (PRP)

- PRP resumido

Princípio de Ranking Probabilístico (PRP)

- PRP resumido
  - Se os documentos recuperados para uma consulta são rankeados de forma decrescente de acordo com a probabilidade de relevância, então a efetividade do ranking será a melhor possível

Princípio de Ranking Probabilístico (PRP)

- PRP resumido
  - Se os documentos recuperados para uma consulta são rankeados de forma decrescente de acordo com a probabilidade de relevância, então a efetividade do ranking será a melhor possível
- PRP completo

Princípio de Ranking Probabilístico (PRP)

- PRP resumido
  - Se os documentos recuperados para uma consulta são rankeados de forma decrescente de acordo com a probabilidade de relevância, então a efetividade do ranking será a melhor possível
- PRP completo
  - Se a resposta do sistema de ORI para cada consulta é um ranking de documentos [...] em ordem de probabilidade decrescente de relevância para a consulta, onde as probabilidades são estimadas melhores possível em relação aos dados disponíveis para o sistema para esse fim, a efetividade do sistema para o seu usuário será a melhor que é possível obter com base nesses dados

Modelo de Independência Binária (BIM)

- Tradicionalmente usado com o PRP

Premissas:

Modelo de Independência Binária (BIM)

- Tradicionalmente usado com o PRP

Premissas:

- 'Binário' (equivalente a Booleano): documentos e consultas representados como vetores binários de incidência de termos

Modelo de Independência Binária (BIM)

- Tradicionalmente usado com o PRP

Premissas:

- 'Binário' (equivalente a Booleano): documentos e consultas representados como vetores binários de incidência de termos
  - Um documento d é representado pelo vetor $\vec{x} = (x_1, \ldots, x_M)$, onde $x_t = 1$ se termo t aparece em d e, caso contrário, $x_t = 0$

Modelo de Independência Binária (BIM)

- Tradicionalmente usado com o PRP

Premissas:

- 'Binário' (equivalente a Booleano): documentos e consultas representados como vetores binários de incidência de termos
  - Um documento d é representado pelo vetor $\vec{x} = (x_1, \ldots, x_M)$, onde $x_t = 1$ se termo t aparece em d e, caso contrário, $x_t = 0$
- 'Independência': nenhuma associação entre termos, que é a premissa 'ingênua' de modelos Naive Bayes

Matriz de incidência binária

	Antônio e Cleópatra	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	
piedade	1	0	1	1	1	1	
pior	1	0	1	1	1	0	
...							

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

Modelo de Independência Binária

Para fazer uma boa recuperação probabilística é necessário estimar como os termos em documentos contribuem para a relevância

- Encontrar estatísticas que afetam o julgamento sobre a relevância de documentos

Modelo de Independência Binária

Para fazer uma boa recuperação probabilística é necessário estimar como os termos em documentos contribuem para a relevância

- Encontrar estatísticas que afetam o julgamento sobre a relevância de documentos
  - Exemplos: frequência de termos, frequência de documentos, comprimento de documentos

Modelo de Independência Binária

Para fazer uma boa recuperação probabilística é necessário estimar como os termos em documentos contribuem para a relevância

- Encontrar estatísticas que afetam o julgamento sobre a relevância de documentos
  - Exemplos: frequência de termos, frequência de documentos, comprimento de documentos
- Combinar essas estatísticas para estimar a probabilidade de relevância $P(R|d,q)$ de um documento d

Modelo de Independência Binária

$P(R|d,q)$ é modelada usando vetores de incidência de termos como $P(R|\vec{x},\vec{q})$

$$P(R=1|\vec{x},\vec{q}) = \frac{P(\vec{x}|R=1,\vec{q})P(R=1|\vec{q})}{P(\vec{x}|\vec{q})}$$

$$P(R=0|\vec{x},\vec{q}) = \frac{P(\vec{x}|R=0,\vec{q})P(R=0|\vec{q})}{P(\vec{x}|\vec{q})}$$

- $P(\vec{x}|R=1,\vec{q})$ e $P(\vec{x}|R=0,\vec{q})$: probabilidade que se um documento recuperado é relevante (ou não relevante $R=0$), então a representação do documento é $\vec{x}$

Modelo de Independência Binária

$P(R|d,q)$ é modelada usando vetores de incidência de termos como $P(R|\vec{x},\vec{q})$

$$P(R=1|\vec{x},\vec{q}) = \frac{P(\vec{x}|R=1,\vec{q})P(R=1|\vec{q})}{P(\vec{x}|\vec{q})}$$

$$P(R=0|\vec{x},\vec{q}) = \frac{P(\vec{x}|R=0,\vec{q})P(R=0|\vec{q})}{P(\vec{x}|\vec{q})}$$

- $P(\vec{x}|R=1,\vec{q})$ e $P(\vec{x}|R=0,\vec{q})$: probabilidade que se um documento recuperado é relevante (ou não relevante $R=0$), então a representação do documento é $\vec{x}$
- Usar estatísticas sobre a coleção de documentos para estimar essas probabilidades

Modelo de Independência Binária

$$P(R=1|\vec{x},\vec{q}) = \frac{P(\vec{x}|R=1,\vec{q})P(R=1|\vec{q})}{P(\vec{x}|\vec{q})}$$

$$P(R=0|\vec{x},\vec{q}) = \frac{P(\vec{x}|R=0,\vec{q})P(R=0|\vec{q})}{P(\vec{x}|\vec{q})}$$

- $P(R=1|\vec{q})$ e $P(R=0|\vec{q})$: probabilidade a priori de recuperar um documento relevante (ou não) para a consulta $\vec{q}$

Modelo de Independência Binária

$$P(R=1|\vec{x},\vec{q}) = \frac{P(\vec{x}|R=1,\vec{q})P(R=1|\vec{q})}{P(\vec{x}|\vec{q})}$$

$$P(R=0|\vec{x},\vec{q}) = \frac{P(\vec{x}|R=0,\vec{q})P(R=0|\vec{q})}{P(\vec{x}|\vec{q})}$$

- $P(R=1|\vec{q})$ e $P(R=0|\vec{q})$: probabilidade a priori de recuperar um documento relevante (ou não) para a consulta $\vec{q}$
- Podemos estimar $P(R=1|\vec{q})$ e $P(R=0|\vec{q})$ a partir da porcentagem de documentos relevantes na coleção

Obtendo um ranking

- Para usar $P(R = 1|\vec{x}, \vec{q})$ e $P(R = 0|\vec{x}, \vec{q})$, deve-se saber $P(\vec{x}|\vec{q})$, $P(R = 1|\vec{q})$ e $P(R = 0|\vec{q})$

Obtendo um ranking

- Para usar $P(R = 1|\vec{x}, \vec{q})$ e $P(R = 0|\vec{x}, \vec{q})$, deve-se saber $P(\vec{x}|\vec{q})$, $P(R = 1|\vec{q})$ e $P(R = 0|\vec{q})$
- É mais fácil usar as chances (*odds*) $O(R|\vec{x}, \vec{q})$

Obtendo um ranking

- Para usar $P(R=1|\vec{x},\vec{q})$ e $P(R=0|\vec{x},\vec{q})$, deve-se saber $P(\vec{x}|\vec{q})$, $P(R=1|\vec{q})$ e $P(R=0|\vec{q})$
- É mais fácil usar as chances (*odds*) $O(R|\vec{x},\vec{q})$
- Ou seja: ordenar documentos pelas suas chances de relevância

$$O(R|\vec{x},\vec{q}) = \frac{P(R=1|\vec{x},\vec{q})}{P(R=0|\vec{x},\vec{q})} = \frac{\frac{P(R=1|\vec{q})P(\vec{x}|R=1,\vec{q})}{P(\vec{x}|\vec{q})}}{\frac{P(R=0|\vec{q})P(\vec{x}|R=0,\vec{q})}{P(\vec{x}|\vec{q})}}$$

$$= \frac{P(R=1|\vec{q})}{P(R=0|\vec{q})} \cdot \frac{P(\vec{x}|R=1,\vec{q})}{P(\vec{x}|R=0,\vec{q})}$$

Obtendo um ranking

- Para usar $P(R=1|\vec{x},\vec{q})$ e $P(R=0|\vec{x},\vec{q})$, deve-se saber $P(\vec{x}|\vec{q})$, $P(R=1|\vec{q})$ e $P(R=0|\vec{q})$
- É mais fácil usar as chances (*odds*) $O(R|\vec{x},\vec{q})$
- Ou seja: ordenar documentos pelas suas chances de relevância

$$O(R|\vec{x},\vec{q}) = \frac{P(R=1|\vec{x},\vec{q})}{P(R=0|\vec{x},\vec{q})} = \frac{\frac{P(R=1|\vec{q})P(\vec{x}|R=1,\vec{q})}{P(\vec{x}|\vec{q})}}{\frac{P(R=0|\vec{q})P(\vec{x}|R=0,\vec{q})}{P(\vec{x}|\vec{q})}}$$

$$= \frac{P(R=1|\vec{q})}{P(R=0|\vec{q})} \cdot \frac{P(\vec{x}|R=1,\vec{q})}{P(\vec{x}|R=0,\vec{q})}$$

- $O(R|\vec{q}) = \frac{P(R=1|\vec{q})}{P(R=0|\vec{q})}$ é constante para consulta $\vec{q}$ e não interfere no ranking

Obtendo um ranking

Para simplificar contas, usamos a premissa de independência condicional de Naive Bayes em que a presença de um termo em um documento é independente da presença de qualquer outro termo:

$$\frac{P(\vec{x}|R=1,\vec{q})}{P(\vec{x}|R=0,\vec{q})} = \prod_{t=1}^{M} \frac{P(x_t|R=1,\vec{q})}{P(x_t|R=0,\vec{q})}$$

Então:

$$O(R|\vec{x},\vec{q}) = O(R|\vec{q}) \cdot \prod_{t=1}^{M} \frac{P(x_t|R=1,\vec{q})}{P(x_t|R=0,\vec{q})}$$

Obtendo um ranking

Uma vez que cada x_t é 0 ou 1, podemos separar os termos:

Obtendo um ranking

Uma vez que cada x_t é 0 ou 1, podemos separar os termos:

$$O(R|\vec{x},\vec{q}) = O(R|\vec{q}) \cdot \prod_{t:x_t=1} \frac{P(x_t=1|R=1,\vec{q})}{P(x_t=1|R=0,\vec{q})} \cdot \prod_{t:x_t=0} \frac{P(x_t=0|R=1,\vec{q})}{P(x_t=0|R=0,\vec{q})}$$

Obtendo um ranking

- Seja $p_t = P(x_t = 1|R = 1, \vec{q})$ a probabilidade de um termo aparecer em um documento relevante

	doc	rel. $(R = 1)$	não-rel. $(R = 0)$
Termo presente	$x_t = 1$	p_t	u_t
Termo ausente	$x_t = 0$	$1 - p_t$	$1 - u_t$

Obtendo um ranking

- Seja $p_t = P(x_t = 1|R = 1, \vec{q})$ a probabilidade de um termo aparecer em um documento relevante
- Seja $u_t = P(x_t = 1|R = 0, \vec{q})$ a probabilidade de um termo aparecer em um documento não-relevante

	doc	rel. $(R = 1)$	não-rel. $(R = 0)$
Termo presente	$x_t = 1$	p_t	u_t
Termo ausente	$x_t = 0$	$1 - p_t$	$1 - u_t$

Obtendo um ranking

- Seja $p_t = P(x_t = 1|R = 1, \vec{q})$ a probabilidade de um termo aparecer em um documento relevante
- Seja $u_t = P(x_t = 1|R = 0, \vec{q})$ a probabilidade de um termo aparecer em um documento não-relevante
- Isso pode ser mostrado na seguinte tabela:

	doc	rel. $(R = 1)$	não-rel. $(R = 0)$
Termo presente	$x_t = 1$	p_t	u_t
Termo ausente	$x_t = 0$	$1 - p_t$	$1 - u_t$

Obtendo um ranking

Premissa simplificadora adicional: termos não ocorrendo na consulta são igualmente prováveis para ocorrer em documentos relevantes e não-relevantes

- Se $q_t = 0$, então $p_t = u_t$

Agora necessitamos somente considerar termos nos produtórios que aparecem na consulta:

$$O(R|\vec{x},\vec{q}) = O(R|\vec{q}) \cdot \prod_{t:x_t=1,q_t=1} \frac{p_t}{u_t} \cdot \prod_{t:x_t=0,q_t=1} \frac{1-p_t}{1-u_t}$$

- O produtório à esquerda é sobre termos encontrados no documento ($x_t = 1$)
- O produtório à direita é sobre termos não encontrados no documento ($x_t = 0$)

Obtendo um ranking

Saindo de:

$$O(R|\vec{x},\vec{q}) = O(R|\vec{q}) \cdot \prod_{t:x_t=1,q_t=1} \frac{p_t}{u_t} \cdot \prod_{t:x_t=0,q_t=1} \frac{1-p_t}{1-u_t}$$

Podemos incluir termos de consulta encontrados no documento no produto à direita e, simultaneamente, dividir por eles no produtório à esquerda, o que resulta em:

$$O(R|\vec{x},\vec{q}) = O(R|\vec{q}) \cdot \prod_{t:x_t=1,q_t=1} \frac{p_t(1-u_t)}{u_t(1-p_t)} \cdot \prod_{t:q_t=1} \frac{1-p_t}{1-u_t}$$

Derivando uma função de ranking para termos de consultas (2)

$$O(R|\vec{x},\vec{q}) = O(R|\vec{q}) \cdot \prod_{t:x_t=1,q_t=1} \frac{p_t(1-u_t)}{u_t(1-p_t)} \cdot \prod_{t:q_t=1} \frac{1-p_t}{1-u_t}$$

- O produtório à esquerda ainda é sobre os termos de busca encontrados no documento,

Derivando uma função de ranking para termos de consultas (2)

$$O(R|\vec{x},\vec{q}) = O(R|\vec{q}) \cdot \prod_{t:x_t=1,q_t=1} \frac{p_t(1-u_t)}{u_t(1-p_t)} \cdot \prod_{t:q_t=1} \frac{1-p_t}{1-u_t}$$

- O produtório à esquerda ainda é sobre os termos de busca encontrados no documento,
- O produtório à direita é agora sobre todos os termos de consulta portanto constantes para uma consulta particular e **não interfere no ranking**.

Derivando uma função de ranking para termos de consultas (2)

$$O(R|\vec{x},\vec{q}) = O(R|\vec{q}) \cdot \prod_{t:x_t=1,q_t=1} \frac{p_t(1-u_t)}{u_t(1-p_t)} \cdot \prod_{t:q_t=1} \frac{1-p_t}{1-u_t}$$

- O produtório à esquerda ainda é sobre os termos de busca encontrados no documento,
- O produtório à direita é agora sobre todos os termos de consulta portanto constantes para uma consulta particular e **não interfere no ranking**.
- $\rightarrow$ **Para rankear documentos em relação a uma consulta só usa-se o produtório à esquerda.**

Derivando uma função de ranking para termos de consultas

$$O(R|\vec{x},\vec{q}) = O(R|\vec{q}) \cdot \prod_{t:x_t=1,q_t=1} \frac{p_t(1-u_t)}{u_t(1-p_t)} \cdot \prod_{t:q_t=1} \frac{1-p_t}{1-u_t}$$

- Ranking usa somente o produtório à esquerda
- Define-se assim o Valor de Estado de Recuperação (Retrieval Status Value – RSV):

$$RSV_d = \log \prod_{t:x_t=q_t=1} \frac{p_t(1-u_t)}{u_t(1-p_t)} = \sum_{t:x_t=q_t=1} \log \frac{p_t(1-u_t)}{u_t(1-p_t)}$$

- Usar o log para evitar multiplicações de números pequenos.

Derivando uma função de ranking para termos da consulta

Usar RSV_d é equivalente a rankear documentos usando log das taxas de chances para cada termo t na consulta c_t:

$$c_t = \log \frac{p_t(1-u_t)}{u_t(1-p_t)} = \log \frac{p_t}{(1-p_t)} - \log \frac{u_t}{1-u_t}$$

- A taxa de chances é a taxa de duas probabilidades: (i) a probabilidade de termos aparecerem se o documento é relevante ($p_t/(1-p_t)$), e (ii) a probabilidade do termo aparecer se o documento é não-relevante ($u_t/(1-u_t)$)

Derivando uma função de ranking para termos da consulta

Usar RSV_d é equivalente a rankear documentos usando log das taxas de chances para cada termo t na consulta c_t:

$$c_t = \log \frac{p_t(1-u_t)}{u_t(1-p_t)} = \log \frac{p_t}{(1-p_t)} - \log \frac{u_t}{1-u_t}$$

- A taxa de chances é a taxa de duas probabilidades: (i) a probabilidade de termos aparecerem se o documento é relevante ($p_t/(1-p_t)$), e (ii) a probabilidade do termo aparecer se o documento é não-relevante ($u_t/(1-u_t)$)
- $c_t = 0$: termo tem iguais chances de aparecer em docs relevantes e não-relevantes

Derivando uma função de ranking para termos da consulta

Usar RSV_d é equivalente a rankear documentos usando log das taxas de chances para cada termo t na consulta c_t:

$$c_t = \log \frac{p_t(1-u_t)}{u_t(1-p_t)} = \log \frac{p_t}{(1-p_t)} - \log \frac{u_t}{1-u_t}$$

- A taxa de chances é a taxa de duas probabilidades: (i) a probabilidade de termos aparecerem se o documento é relevante ($p_t/(1-p_t)$), e (ii) a probabilidade do termo aparecer se o documento é não-relevante ($u_t/(1-u_t)$)
- $c_t = 0$: termo tem iguais chances de aparecer em docs relevantes e não-relevantes
- c_t positivo: chances maiores de aparecer em docs relevantes

Derivando uma função de ranking para termos da consulta

Usar RSV_d é equivalente a rankear documentos usando log das taxas de chances para cada termo t na consulta c_t:

$$c_t = \log \frac{p_t(1-u_t)}{u_t(1-p_t)} = \log \frac{p_t}{(1-p_t)} - \log \frac{u_t}{1-u_t}$$

- A taxa de chances é a taxa de duas probabilidades: (i) a probabilidade de termos aparecerem se o documento é relevante ($p_t/(1-p_t)$), e (ii) a probabilidade do termo aparecer se o documento é não-relevante ($u_t/(1-u_t)$)
- $c_t = 0$: termo tem iguais chances de aparecer em docs relevantes e não-relevantes
- c_t positivo: chances maiores de aparecer em docs relevantes
- c_t negativo: chances maiores de aparecer em docs não-relevantes

Peso do termo c_t no BIM

- $c_t = \log \frac{p_t}{(1-p_t)} - \log \frac{u_t}{1-u_t}$ funciona como um peso para o termo

Peso do termo c_t no BIM

- $c_t = \log \frac{p_t}{(1-p_t)} - \log \frac{u_t}{1-u_t}$ funciona como um peso para o termo
- Valor do estado de recuperação (RSV) para documento d:

$$RSV_d = \sum_{x_t=1, q_t=1} c_t$$

Peso do termo c_t no BIM

- $c_t = \log \frac{p_t}{(1-p_t)} - \log \frac{u_t}{1-u_t}$ funciona como um peso para o termo
- Valor do estado de recuperação (RSV) para documento d:

$$RSV_d = \sum_{x_t=1, q_t=1} c_t$$

- Então BIM e o modelo vetorial são idênticos em um nível operacional . . .

Peso do termo c_t no BIM

- $c_t = \log \frac{p_t}{(1-p_t)} - \log \frac{u_t}{1-u_t}$ funciona como um peso para o termo
- Valor do estado de recuperação (RSV) para documento d:

$$RSV_d = \sum_{x_t=1, q_t=1} c_t$$

- Então BIM e o modelo vetorial são idênticos em um nível operacional . . .
- . . . exceto que os pesos de termos são diferentes e

Peso do termo c_t no BIM

- $c_t = \log \frac{p_t}{(1-p_t)} - \log \frac{u_t}{1-u_t}$ funciona como um peso para o termo
- Valor do estado de recuperação (RSV) para documento d:

$$RSV_d = \sum_{x_t=1, q_t=1} c_t$$

- Então BIM e o modelo vetorial são idênticos em um nível operacional . . .
- . . . exceto que os pesos de termos são diferentes e
- Em particular: podemos usar as mesmas estruturas de dados (índice invertido etc) para os dois modelos

Como estimar probabilidades

Para cada termo t em uma consulta, estimar c_t na coleção completa usando uma tabela de contagem de documentos na coleção, onde df_t é o número de documentos que contém o termo t:

		docs relevantes	docs não-relevantes	Total
Termo presente	$x_t = 1$	s	$\mathrm{df}_t - s$	df_t
Termo ausente	$x_t = 0$	$S - s$	$(N - \mathrm{df}_t) - (S - s)$	$N - \mathrm{df}_t$
	Total	S	$N - S$	N

$$p_t = s/S$$

$$u_t = (\mathrm{df}_t - s)/(N - S)$$

$$c_t = \log \frac{s/(S-s)}{(\mathrm{df}_t - s)/((N - \mathrm{df}_t) - (S - s))}$$

Evitando zeros

- Se qualquer dos contadores for zero, então o peso do termo não é bem-definido

Evitando zeros

- Se qualquer dos contadores for zero, então o peso do termo não é bem-definido
- Evitar zeros: somar 0.5 para cada contagem

Evitando zeros

- Se qualquer dos contadores for zero, então o peso do termo não é bem-definido
- Evitar zeros: somar 0.5 para cada contagem
- Por exemplo, use $S - s + 0.5$ na fórmula para $S - s$

Evitando zeros

- Se qualquer dos contadores for zero, então o peso do termo não é bem-definido
- Evitar zeros: somar 0.5 para cada contagem
- Por exemplo, use $S - s + 0.5$ na fórmula para $S - s$

$$c_t = \log \frac{(s+0.5)/(S-s+0.5)}{(\mathrm{df}_t-s+0.5)/((N-\mathrm{df}_t)-(S-s)+0.5)}$$

Exercício para casa

- Consulta: `Plano saúde Unimed`
- d1: `Unimed rejeita alegações sobre seu próprio plano ser ruim`
- d2: `O plano é visitar a Unimed`
- d3: `Unimed apresenta preocupações sobre reformas de plano de saúde`

Exercício para casa

- Consulta: `Plano saúde Unimed`
- d1: `Unimed rejeita alegações sobre seu próprio plano ser ruim`
- d2: `O plano é visitar a Unimed`
- d3: `Unimed apresenta preocupações sobre reformas de plano de saúde`

Estimar as probabilidades desses documentos serem relevantes à consulta. Esses são os únicos documentos na coleção. Considere os documentos $d1$ e $d3$ como relevantes e $d2$ como não relevantes.

Recuperação ad hoc: premissa simplificadora

$c_t = \log \frac{p_t}{(1-p_t)} - \log \frac{u_t}{1-u_t}$ funciona como um peso para o termo

Recuperação ad hoc: premissa simplificadora

$c_t = \log \frac{p_t}{(1-p_t)} - \log \frac{u_t}{1-u_t}$ funciona como um peso para o termo

- Se assumir que docs relevantes são uma pequena parte da coleção, pode-se aproximar estatísticas para documentos não-relevantes a partir da coleção inteira

Recuperação ad hoc: premissa simplificadora

$c_t = \log \frac{p_t}{(1-p_t)} - \log \frac{u_t}{1-u_t}$ funciona como um peso para o termo

- Se assumir que docs relevantes são uma pequena parte da coleção, pode-se aproximar estatísticas para documentos não-relevantes a partir da coleção inteira
- Assim, u_t (a prob. de ocorrência de termos em docs não relevantes para a consulta) é aproximado por df_t/N e

$$\log[(1-u_t)/u_t] = \log[(N-\mathrm{df}_t)/\mathrm{df}_t] \approx \log N/\mathrm{df}_t$$

Recuperação ad hoc: premissa simplificadora

$c_t = \log \frac{p_t}{(1-p_t)} - \log \frac{u_t}{1-u_t}$ funciona como um peso para o termo

- Se assumir que docs relevantes são uma pequena parte da coleção, pode-se aproximar estatísticas para documentos não-relevantes a partir da coleção inteira
- Assim, u_t (a prob. de ocorrência de termos em docs não relevantes para a consulta) é aproximado por df_t/N e

$$\log[(1-u_t)/u_t] = \log[(N-\mathrm{df}_t)/\mathrm{df}_t] \approx \log N/\mathrm{df}_t$$

- $idf_t = \log N/\mathrm{df}_t$

Recuperação ad hoc: premissa simplificadora

$c_t = \log \frac{p_t}{(1-p_t)} - \log \frac{u_t}{1-u_t}$ funciona como um peso para o termo

- Se assumir que docs relevantes são uma pequena parte da coleção, pode-se aproximar estatísticas para documentos não-relevantes a partir da coleção inteira
- Assim, u_t (a prob. de ocorrência de termos em docs não relevantes para a consulta) é aproximado por df_t/N e

$$\log[(1-u_t)/u_t] = \log[(N-\mathrm{df}_t)/\mathrm{df}_t] \approx \log N/\mathrm{df}_t$$

- $idf_t = \log N/\mathrm{df}_t$
- A aproximação anterior não pode ser facilmente extendida para documentos relevantes

Estimando probabilidades a partir de retorno de relevância

- Não há informação sobre relevância de todos os docs

Estimando probabilidades a partir de retorno de relevância

- Não há informação sobre relevância de todos os docs
- Probabilidades p_t podem ser estimadas a partir de retorno de relevância

Estimando probabilidades a partir de retorno de relevância

- Não há informação sobre relevância de todos os docs
- Probabilidades p_t podem ser estimadas a partir de retorno de relevância
  - Use a frequência de ocorrência de termos em docs reconhecidos como relevantes

Estimando probabilidades a partir de retorno de relevância

- Não há informação sobre relevância de todos os docs
- Probabilidades p_t podem ser estimadas a partir de retorno de relevância
  - Use a frequência de ocorrência de termos em docs reconhecidos como relevantes
- Essa é a base de abordagens probabilísticas para peso com retorno de relevância (método da máxima verossimilhança)

Estimativas de probabilidades em recuperação adhoc

- Recuperação ad-hoc: nenhum julgamento de relevância disponível

Estimativas de probabilidades em recuperação adhoc

- Recuperação ad-hoc: nenhum julgamento de relevância disponível
- Nesse caso: assumir que p_t é constante sobre todos os termos x_t na consulta e que $p_t = 0.5$

Estimativas de probabilidades em recuperação adhoc

- Recuperação ad-hoc: nenhum julgamento de relevância disponível
- Nesse caso: assumir que p_t é constante sobre todos os termos x_t na consulta e que $p_t = 0.5$
- Cada termo é igualmente provável de ocorrer em um doc relevante e também os fatores p_t e $(1 - p_t)$ se cancelam na expressão para RSV

Estimativas de probabilidades em recuperação adhoc

- Recuperação ad-hoc: nenhum julgamento de relevância disponível
- Nesse caso: assumir que p_t é constante sobre todos os termos x_t na consulta e que $p_t = 0.5$
- Cada termo é igualmente provável de ocorrer em um doc relevante e também os fatores p_t e $(1 - p_t)$ se cancelam na expressão para RSV
  - $c_t = \log \frac{p_t}{(1-p_t)} - \log \frac{u_t}{1-u_t} \rightarrow c_t^* = -\log \frac{u_t}{1-u_t}$

Estimativas de probabilidades em recuperação adhoc

- Recuperação ad-hoc: nenhum julgamento de relevância disponível
- Nesse caso: assumir que p_t é constante sobre todos os termos x_t na consulta e que $p_t = 0.5$
- Cada termo é igualmente provável de ocorrer em um doc relevante e também os fatores p_t e $(1 - p_t)$ se cancelam na expressão para RSV
  - $c_t = \log \frac{p_t}{(1-p_t)} - \log \frac{u_t}{1-u_t} \rightarrow c_t^* = -\log \frac{u_t}{1-u_t}$
- Combinando esse método com a aproximação anterior para u_t, o ranking de documentos é determinado por pesos IDF dos termos da consulta ocorrem em documentos

Estimativas de probabilidades em recuperação adhoc

- Recuperação ad-hoc: nenhum julgamento de relevância disponível
- Nesse caso: assumir que p_t é constante sobre todos os termos x_t na consulta e que $p_t = 0.5$
- Cada termo é igualmente provável de ocorrer em um doc relevante e também os fatores p_t e $(1 - p_t)$ se cancelam na expressão para RSV
  - $c_t = \log \frac{p_t}{(1-p_t)} - \log \frac{u_t}{1-u_t} \rightarrow c_t^* = -\log \frac{u_t}{1-u_t}$
- Combinando esse método com a aproximação anterior para u_t, o ranking de documentos é determinado por pesos IDF dos termos da consulta ocorrem em documentos
  - $c_t^* = -\log \frac{u_t}{1-u_t} \approx idf_t$

Estimativas de probabilidades em recuperação adhoc

- Recuperação ad-hoc: nenhum julgamento de relevância disponível
- Nesse caso: assumir que p_t é constante sobre todos os termos x_t na consulta e que $p_t = 0.5$
- Cada termo é igualmente provável de ocorrer em um doc relevante e também os fatores p_t e $(1 - p_t)$ se cancelam na expressão para RSV
  - $c_t = \log \frac{p_t}{(1-p_t)} - \log \frac{u_t}{1-u_t} \rightarrow c_t^* = -\log \frac{u_t}{1-u_t}$
- Combinando esse método com a aproximação anterior para u_t, o ranking de documentos é determinado por pesos IDF dos termos da consulta ocorrem em documentos
  - $c_t^* = -\log \frac{u_t}{1-u_t} \approx idf_t$
  - $RSV_d^{AH} = \sum_{x_t=q_t=1} idf_t$

Estimativas de probabilidades em recuperação adhoc

- Recuperação ad-hoc: nenhum julgamento de relevância disponível
- Nesse caso: assumir que p_t é constante sobre todos os termos x_t na consulta e que $p_t = 0.5$
- Cada termo é igualmente provável de ocorrer em um doc relevante e também os fatores p_t e $(1 - p_t)$ se cancelam na expressão para RSV
  - $c_t = \log \frac{p_t}{(1-p_t)} - \log \frac{u_t}{1-u_t} \rightarrow c_t^* = -\log \frac{u_t}{1-u_t}$
- Combinando esse método com a aproximação anterior para u_t, o ranking de documentos é determinado por pesos IDF dos termos da consulta ocorrem em documentos
  - $c_t^* = -\log \frac{u_t}{1-u_t} \approx idf_t$
  - $RSV_d^{AH} = \sum_{x_t=q_t=1} idf_t$
- Para documentos curtos (títulos ou resumos) em situações mais simples, essa estimativa pode ser satisfatória

Sumário

História e resumo de premissas

- Entre os modelos mais tradicionais em ORI

História e resumo de premissas

- Entre os modelos mais tradicionais em ORI
  - Maron & Kuhns, 1960: uma vez que um sistema de ORI não pode predizer com certeza qual documento é relevante, devemos considerar probabilidades

História e resumo de premissas

- Entre os modelos mais tradicionais em ORI
  - Maron & Kuhns, 1960: uma vez que um sistema de ORI não pode predizer com certeza qual documento é relevante, devemos considerar probabilidades
- Premissas para ter aproximações razoáveis da probabilidades necessárias (no BIM):

História e resumo de premissas

- Entre os modelos mais tradicionais em ORI
  - Maron & Kuhns, 1960: uma vez que um sistema de ORI não pode predizer com certeza qual documento é relevante, devemos considerar probabilidades
- Premissas para ter aproximações razoáveis da probabilidades necessárias (no BIM):
  - Representação booleana de documentos/consultas/relevância

História e resumo de premissas

- Entre os modelos mais tradicionais em ORI
  - Maron & Kuhns, 1960: uma vez que um sistema de ORI não pode predizer com certeza qual documento é relevante, devemos considerar probabilidades
- Premissas para ter aproximações razoáveis da probabilidades necessárias (no BIM):
  - Representação booleana de documentos/consultas/relevância
  - Independência de termos

História e resumo de premissas

- Entre os modelos mais tradicionais em ORI
  - Maron & Kuhns, 1960: uma vez que um sistema de ORI não pode predizer com certeza qual documento é relevante, devemos considerar probabilidades
- Premissas para ter aproximações razoáveis da probabilidades necessárias (no BIM):
  - Representação booleana de documentos/consultas/relevância
  - Independência de termos
  - Termos fora da consulta não afetam a recuperação

História e resumo de premissas

- Entre os modelos mais tradicionais em ORI
  - Maron & Kuhns, 1960: uma vez que um sistema de ORI não pode predizer com certeza qual documento é relevante, devemos considerar probabilidades
- Premissas para ter aproximações razoáveis da probabilidades necessárias (no BIM):
  - Representação booleana de documentos/consultas/relevância
  - Independência de termos
  - Termos fora da consulta não afetam a recuperação
  - Valores da relevância de documentos são independentes

Diferenças entre espaço vetorial e BIM?

- Não são muito diferentes

Diferenças entre espaço vetorial e BIM?

- Não são muito diferentes
- Esquema de recuperação com o mesmo funcionamento básico

Diferenças entre espaço vetorial e BIM?

- Não são muito diferentes
- Esquema de recuperação com o mesmo funcionamento básico
- Para ORI probabilístico, pontua consultas não com similaridade cosseno e tf-idf em um espaço vetorial, mas por uma fórmula motivada pela Teoria de Probabilidades

Okapi BM25: visão global

- Okapi BM25 (ou BestMatch25) é um modelo probabilístico muito difundido que incorpora frequência de termos (ou seja, não é binário) e normalização de tamanho

Okapi BM25: visão global

- Okapi BM25 (ou BestMatch25) é um modelo probabilístico muito difundido que incorpora frequência de termos (ou seja, não é binário) e normalização de tamanho
- BIM foi originalmente concebido para pequenos catálogos de registros de tamanho similares e funciona bem para isso

Okapi BM25: visão global

- Okapi BM25 (ou BestMatch25) é um modelo probabilístico muito difundido que incorpora frequência de termos (ou seja, não é binário) e normalização de tamanho
- BIM foi originalmente concebido para pequenos catálogos de registros de tamanho similares e funciona bem para isso
- Em sistemas de ORI modernos, um modelo deve considerar a frequência de termos e o tamanho do documento

Okapi BM25: Ponto inicial

Okapi BM25: Ponto inicial

- A pontuação mais simples para o documento d é somente os pesos idf dos termos de consultas presentes no documento:

Okapi BM25: Ponto inicial

- A pontuação mais simples para o documento d é somente os pesos idf dos termos de consultas presentes no documento:

Okapi BM25: Ponto inicial

- A pontuação mais simples para o documento d é somente os pesos idf dos termos de consultas presentes no documento:

$$RSV_d = \sum_{t \in q} \log \frac{N}{\mathrm{df}_t}$$

Okapi BM25: peso básico

- Melhorar o termo idf [log N/df] com o uso da frequência de termos e do tamanho do documento

$$RSV_d = \sum_{t \in q} \log \left[\frac{N}{\mathrm{df}_t} \right] \cdot \frac{(k_1 + 1)\mathrm{tf}_{td}}{k_1((1 - b) + b \times (L_d / L_{\mathrm{ave}})) + \mathrm{tf}_{td}}$$

Okapi BM25: peso básico

- Melhorar o termo idf [log N/df] com o uso da frequência de termos e do tamanho do documento

$$RSV_d = \sum_{t \in q} \log \left[\frac{N}{\mathrm{df}_t} \right] \cdot \frac{(k_1 + 1)\mathrm{tf}_{td}}{k_1((1 - b) + b \times (L_d / L_{\mathsf{ave}})) + \mathrm{tf}_{td}}$$

- tf_{td}: frequência de termos no documento d

Okapi BM25: peso básico

- Melhorar o termo idf [log N/df] com o uso da frequência de termos e do tamanho do documento

$$RSV_d = \sum_{t \in q} \log \left[\frac{N}{\mathrm{df}_t} \right] \cdot \frac{(k_1 + 1)\mathrm{tf}_{td}}{k_1((1-b) + b \times (L_d/L_{\mathrm{ave}})) + \mathrm{tf}_{td}}$$

- tf_{td}: frequência de termos no documento d
- L_d (L_{ave}): tam. do documento d e tam. médio do documento na coleção

Okapi BM25: peso básico

- Melhorar o termo idf [log N/df] com o uso da frequência de termos e do tamanho do documento

$$RSV_d = \sum_{t \in q} \log \left[\frac{N}{\mathrm{df}_t} \right] \cdot \frac{(k_1 + 1)\mathrm{tf}_{td}}{k_1((1-b) + b \times (L_d/L_{\mathrm{ave}})) + \mathrm{tf}_{td}}$$

- tf_{td}: frequência de termos no documento d
- L_d (L_{ave}): tam. do documento d e tam. médio do documento na coleção
- k_1: parâmetro de controle da ponderação da frequência de termos dos documentos, $k_1 \geq 0$

Okapi BM25: peso básico

- Melhorar o termo idf [log N/df] com o uso da frequência de termos e do tamanho do documento

$$RSV_d = \sum_{t \in q} \log \left[\frac{N}{\mathrm{df}_t} \right] \cdot \frac{(k_1 + 1)\mathrm{tf}_{td}}{k_1((1-b) + b \times (L_d/L_{\mathrm{ave}})) + \mathrm{tf}_{td}}$$

- tf_{td}: frequência de termos no documento d
- L_d (L_{ave}): tam. do documento d e tam. médio do documento na coleção
- k_1: parâmetro de controle da ponderação da frequência de termos dos documentos, $k_1 \geq 0$
- b: parâmetro de controle da ponderação do tamanho do documento, $0 \leq b \leq 1$

Exercício

- Interpretar fórmula de pesos BM25 para $k_1 = 0$
- Interpretar fórmula de pesos BM25 para $k_1 = 1$ e $b = 0$
- Interpretar fórmula de pesos BM25 para $k_1 \mapsto \infty$ e $b = 0$
- Interpretar fórmula de pesos BM25 para $k_1 \mapsto \infty$ e $b = 1$

Pesagem Okapi BM25 para consultas longas

- Para consultas longas, use pesagem similar para consulta de termos

Pesagem Okapi BM25 para consultas longas

- Para consultas longas, use pesagem similar para consulta de termos

$$RSV_d = \sum_{t \in q} \left[\log \frac{N}{\mathrm{df}_t} \right] \cdot \frac{(k_1 + 1)\mathrm{tf}_{td}}{k_1((1-b) + b \times (L_d/L_{\mathrm{ave}})) + \mathrm{tf}_{td}} \cdot \frac{(k_3 + 1)\mathrm{tf}_{tq}}{k_3 + \mathrm{tf}_{tq}}$$

- tf_{tq}: frequência de termos na consulta q

Pesagem Okapi BM25 para consultas longas

- Para consultas longas, use pesagem similar para consulta de termos

$$RSV_d = \sum_{t \in q} \left[\log \frac{N}{\mathrm{df}_t} \right] \cdot \frac{(k_1 + 1)\mathrm{tf}_{td}}{k_1((1-b) + b \times (L_d/L_{\mathrm{ave}})) + \mathrm{tf}_{td}} \cdot \frac{(k_3 + 1)\mathrm{tf}_{tq}}{k_3 + \mathrm{tf}_{tq}}$$

- tf_{tq}: frequência de termos na consulta q
- k_3: parâmetro de controle da escalonagem da frequência de termos da consulta, $k_3 \geq 0$

Pesagem Okapi BM25 para consultas longas

- Para consultas longas, use pesagem similar para consulta de termos

$$RSV_d = \sum_{t \in q} \left[\log \frac{N}{\mathrm{df}_t} \right] \cdot \frac{(k_1 + 1)\mathrm{tf}_{td}}{k_1((1-b) + b \times (L_d/L_{\mathrm{ave}})) + \mathrm{tf}_{td}} \cdot \frac{(k_3 + 1)\mathrm{tf}_{tq}}{k_3 + \mathrm{tf}_{tq}}$$

- tf_{tq}: frequência de termos na consulta q
- k_3: parâmetro de controle da escalonagem da frequência de termos da consulta, $k_3 \geq 0$
- Sem normalização de tamanho de consultas (porque recuperação é feito em relação a uma única consulta fixa)

Pesagem Okapi BM25 para consultas longas

- Para consultas longas, use pesagem similar para consulta de termos

$$RSV_d = \sum_{t \in q} \left[\log \frac{N}{\mathrm{df}_t} \right] \cdot \frac{(k_1 + 1)\mathrm{tf}_{td}}{k_1((1-b) + b \times (L_d/L_{\mathrm{ave}})) + \mathrm{tf}_{td}} \cdot \frac{(k_3 + 1)\mathrm{tf}_{tq}}{k_3 + \mathrm{tf}_{tq}}$$

- tf_{tq}: frequência de termos na consulta q
- k_3: parâmetro de controle da escalonagem da frequência de termos da consulta, $k_3 \geq 0$
- Sem normalização de tamanho de consultas (porque recuperação é feito em relação a uma única consulta fixa)
- A definição de parâmetros deve idealmente ser configurado para otimizar o desempenho em uma coleção de testes para uso durante o desenvolvimento. Na ausência de tal otimização, experimentos têm mostrado resultados razoáveis para k_1 e k_3 estão entre 1.2 e 2 e $b = 0.75$

Qual modelo usar?

- Se quer algo básico e simples $\rightarrow$ usar espaço vetorial com pesos TF-IDF

Qual modelo usar?

- Se quer algo básico e simples $\rightarrow$ usar espaço vetorial com pesos TF-IDF
- Se quer algo com desempenho excelente $\rightarrow$ usar modelos de linguagem ou BM25 com parâmetros bem configurados

Qual modelo usar?

- Se quer algo básico e simples $\rightarrow$ usar espaço vetorial com pesos TF-IDF
- Se quer algo com desempenho excelente $\rightarrow$ usar modelos de linguagem ou BM25 com parâmetros bem configurados
- Meio termo: BM25 ou modelos de linguagem sem ou apenas um parâmetro de configuração