Sistemas de Recuperação de Informação

`https://github.com/fccoelho/curso-IRI`

IRI 11: Recuperação de Informação Probabilística

Flávio Codeço Coelho

Escola de Matemática Aplicada, Fundação Getúlio Vargas

Sumário da Aula

Revisão de relevância: Ideia básica

- O usuário faz uma consulta simples, curta.
- O buscador retorna um conjunto de documentos.
- O usuário marca alguns documentos como relevantes outros não.
- Buscador computa nova representação da informação requerida – deve ser melhor que a consulta inicial.
- Buscador executa nova consulta e retorna resultados.
- Novos resultados apresentação melhor revocação (espera-se).

Rocchio

X

Rocchio

Rocchio

Rocchio

X NR

Rocchio

Rocchio

NR

Rocchio

NR

Rocchio

NR

Rocchio

NR

Rocchio

$\vec{q}_{opt}$ X

Tipos de expansão de consulta

- Tesauro manual (mantido por editores, p.ex., PubMed)
- Tesauro derivado automaticamente (p.ex., baseado em estatísticas de co-ocorrência)
- Consultas equivalentes baseadas na mineração do histórico de consultas)

Expansão de Consulta em Buscadores

- Fonte principal de expansões de consulta em buscadores: logs de consulta
- Exemplo 1: Depois de consultar por [herbal], usuários frequentemente buscam por [remédio herbal].
  - → "remédio herbal" é uma expansão em potencial para "herbal" ou "erva".
- Exemplo 2: Usuários buscando por [fotos de flores] frequentemente clicam na URL photobucket.com/flor. Usuários buscando por [desenhos de flor] frequentemente clicam na mesma URL.
  - → "desenhos de flor" e "fotos de flor" São potencialmente extensões uma da outra.

ão de Hoje,

Conclusão de Hoje

- Abordagem probabilistica a RI
- Principio de Rankeamento de probabilidade
- Modelos: BIM, BM25
- Pressupostos destes modelos

Feedback de Relevância

Feedback de Relevância

- No feedback de relevância, o usuário marca documentos como relevantes ou irrelevantes

Feedback de Relevância

- No feedback de relevância, o usuário marca documentos como relevantes ou irrelevantes
- Dados alguns documentos conhecidos como relevantes e irrelevantes, computamos pesos para termos que não constam da consulta e que indicam quão provável é a sua ocorrência em documentos relevantes.

Feedback de Relevância

- No feedback de relevância, o usuário marca documentos como relevantes ou irrelevantes
- Dados alguns documentos conhecidos como relevantes e irrelevantes, computamos pesos para termos que não constam da consulta e que indicam quão provável é a sua ocorrência em documentos relevantes.
- Hoje: desenvolver uma abordagem probabilística para relevância e também um modelo probabilístico genérico para RI

Abordagem Probabilística à Recuperação

Abordagem Probabilística à Recuperação

- Da uma necessidade informacional de um usuário (representada como uma consulta) e uma coleção de documentos (transformados em representações de documentos), um sistema deve determinar quão bem os documentos satisfazem a consulta

Abordagem Probabilística à Recuperação

- Da uma necessidade informacional de um usuário (representada como uma consulta) e uma coleção de documentos (transformados em representações de documentos), um sistema deve determinar quão bem os documentos satisfazem a consulta
  - Um sistema de RI tem uma compreensão incerta da consulta do usuário, e pode "chutar" se um documento satisfaz à consulta.

Abordagem Probabilística à Recuperação

- Da uma necessidade informacional de um usuário (representada como uma consulta) e uma coleção de documentos (transformados em representações de documentos), um sistema deve determinar quão bem os documentos satisfazem a consulta
  - Um sistema de RI tem uma compreensão incerta da consulta do usuário, e pode "chutar" se um documento satisfaz à consulta.
- A teoria da Probabilidade provê os fundamentos para tal raciocínio sob incerteza

Abordagem Probabilística à Recuperação

- Da uma necessidade informacional de um usuário (representada como uma consulta) e uma coleção de documentos (transformados em representações de documentos), um sistema deve determinar quão bem os documentos satisfazem a consulta
  - Um sistema de RI tem uma compreensão incerta da consulta do usuário, e pode "chutar" se um documento satisfaz à consulta.
- A teoria da Probabilidade provê os fundamentos para tal raciocínio sob incerteza
  - Modelos Probabilísticos exploram estes fundamentos para estimar quão provável é a relevância de um documento para uma consulta

Modelos de RI Probabilísticos – visão geral

Modelos de RI Probabilísticos – visão geral

- Modelo clássico de recuperação Probabilística

Modelos de RI Probabilísticos – visão geral

- Modelo clássico de recuperação Probabilística
    - Princípio de rankeamento de probabilidade

Modelos de RI Probabilísticos – visão geral

- Modelo clássico de recuperação Probabilística
  - Princípio de rankeamento de probabilidade
    - Modelo de independência binária, BestMatch25 (Okapi)

Modelos de RI Probabilísticos – visão geral

- Modelo clássico de recuperação Probabilística
    - Princípio de rankeamento de probabilidade
        - Modelo de independência binária, BestMatch25 (Okapi)
- Redes Bayesianas para recuperação de texto

Modelos de RI Probabilísticos – visão geral

- Modelo clássico de recuperação Probabilística
  - Princípio de rankeamento de probabilidade
    - Modelo de independência binária, BestMatch25 (Okapi)
- Redes Bayesianas para recuperação de texto
- Abordagem de modelo de linguagem para RI

Modelos de RI Probabilísticos – visão geral

- Modelo clássico de recuperação Probabilística
  - Princípio de rankeamento de probabilidade
    - Modelo de independência binária, BestMatch25 (Okapi)
- Redes Bayesianas para recuperação de texto
- Abordagem de modelo de linguagem para RI
  - Importante, será discutido adiante

Modelos de RI Probabilísticos – visão geral

- Modelo clássico de recuperação Probabilística
  - Princípio de rankeamento de probabilidade
    - Modelo de independência binária, BestMatch25 (Okapi)
- Redes Bayesianas para recuperação de texto
- Abordagem de modelo de linguagem para RI
  - Importante, será discutido adiante
- Métodos probabilísticos estão entre os mais antigos, mas são um tema quente em RI

Exercício: Modelo Probabilístico vs. outros modelos

Exercício: Modelo Probabilístico vs. outros modelos

- Modelo booleano

Exercício: Modelo Probabilístico vs. outros modelos

- Modelo booleano
  - Modelos probabilísticos suportam rankeamento e portanto são melhores que o modelo booleano simples.

Exercício: Modelo Probabilístico vs. outros modelos

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

Exercício: Modelo Probabilístico vs. outros modelos

- Modelo booleano
  - Modelos probabilísticos suportam rankeamento e portanto são melhores que o modelo booleano simples.
- Modelo de espaço vetorial
  - O Modelo de espaço vetorial também suporta rankeamento.

Exercício: Modelo Probabilístico vs. outros modelos

- Modelo booleano
  - Modelos probabilísticos suportam rankeamento e portanto são melhores que o modelo booleano simples.
- Modelo de espaço vetorial
  - O Modelo de espaço vetorial também suporta rankeamento.
  - Porque buscar uma alternativa ao modelo de espaço vetorial?

Modelo Probabilístico vs. Espaço Vetorial

Modelo Probabilístico vs. Espaço Vetorial

- Modelo de espaço vetorial: rankeia documentos de acordo com similaridade com a consulta.

Modelo Probabilístico vs. Espaço Vetorial

- Modelo de espaço vetorial: rankeia documentos de acordo com similaridade com a consulta.
- A noção de similaridade não se traduz diretamente em relevância

Modelo Probabilístico vs. Espaço Vetorial

- Modelo de espaço vetorial: rankeia documentos de acordo com similaridade com a consulta.
- A noção de similaridade não se traduz diretamente em relevância
- O documento de maior similaridade pode ser altamente relevante ou completamente irrelevante.

Modelo Probabilístico vs. Espaço Vetorial

- Modelo de espaço vetorial: rankeia documentos de acordo com similaridade com a consulta.
- A noção de similaridade não se traduz diretamente em relevância
- O documento de maior similaridade pode ser altamente relevante ou completamente irrelevante.
- A teoria da probabilidade é uma formalização mais elegante do que desejamos de um sistema de RI: Retornar documentos relevantes ao usuário.

Probabilidade Básica

Probabilidade Básica

- Para eventos A e B

Probabilidade Básica

- Para eventos A e B
  - A probabilidade conjunta $P(A \cap B)$

Probabilidade Básica

- Para eventos A e B
  - A probabilidade conjunta $P(A \cap B)$
  - A probabilidade condicional $P(A|B)$ do evento A ocorrer dado que o evento B também tenha ocorrido.

Probabilidade Básica

- Para eventos A e B
  - A probabilidade conjunta $P(A \cap B)$
  - A probabilidade condicional $P(A|B)$ do evento A ocorrer dado que o evento B também tenha ocorrido.
- Regra da cadeia: relação fundamental entre probabilidade conjunta e condicional:

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

Probabilidade Básica

- Para eventos A e B
  - A probabilidade conjunta $P(A \cap B)$
  - A probabilidade condicional $P(A|B)$ do evento A ocorrer dado que o evento B também tenha ocorrido.
- Regra da cadeia: relação fundamental entre probabilidade conjunta e condicional:

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

- Similarly for the complement of an event $P(\overline{A})$:

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

Probabilidade Básica

- Para eventos A e B
  - A probabilidade conjunta $P(A \cap B)$
  - A probabilidade condicional $P(A|B)$ do evento A ocorrer dado que o evento B também tenha ocorrido.
- Regra da cadeia: relação fundamental entre probabilidade conjunta e condicional:

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

- Similarly for the complement of an event $P(\overline{A})$:

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

- Regra da Probabilidade total: Se B pode ser dividido em uma partição de subconjuntos, então $P(B)$ é a soma das probabilidades dos conjuntos . Um caso especial desta regra é:

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

Probabilidade Básica

Probabilidade Básica

Regra de Bayes para inverter probabilidades condicionais:

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

Probabilidade Básica

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 vista como uma forma de atualizar probabilidades:

Probabilidade Básica

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 vista como uma forma de atualizar probabilidades:

- Começa com a probabilidade a priori $P(A)$ (estimativa inicial de quão provável é um evento A na ausência de outra informação)

Probabilidade Básica

Regra de Bayes para inverter probabilidades condicionais:

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

Pode ser vista como uma forma de atualizar probabilidades:

- Começa com a probabilidade a priori $P(A)$ (estimativa inicial de quão provável é um evento A na ausência de outra informação)
- A probabilidade posterior $P(A|B)$ depois de considerarmos a evidência B, baseada na verossimilhança de B ocorrer nos dois casos em que A ocorre ou não

Probabilidade Básica

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 vista como uma forma de atualizar probabilidades:

- Começa com a probabilidade a priori $P(A)$ (estimativa inicial de quão provável é um evento A na ausência de outra informação)
- A probabilidade posterior $P(A|B)$ depois de considerarmos a evidência B, baseada na verossimilhança de B ocorrer nos dois casos em que A ocorre ou não

Odds (chance) de um evento nos dá um tipo de multiplicador para como as probabilidades variam:

$$\text{Odds:} \qquad O(A) = \frac{P(A)}{P(\overline{A})} = \frac{P(A)}{1 - P(A)}$$

O Problema do Rankeamento de Documentos

O Problema do Rankeamento de Documentos

- Recuperação Rankeada : Dada uma coleção de documentos, o usuário realiza uma consulta que retorna uma lista ordenada de documentos

O Problema do Rankeamento de Documentos

- Recuperação Rankeada : Dada uma coleção de documentos, o usuário realiza uma consulta que retorna uma lista ordenada de documentos
- Assumindo noção binária de relevância: $R_{d,q}$ é uma variável aleatória dicotômica, tal que

O Problema do Rankeamento de Documentos

- Recuperação Rankeada : Dada uma coleção de documentos, o usuário realiza uma consulta que retorna uma lista ordenada de documentos
- Assumindo noção binária de relevância: $R_{d,q}$ é uma variável aleatória dicotômica, tal que
  - $R_{d,q} = 1$ se o documento d é relevante com respeito à consulta q

O Problema do Rankeamento de Documentos

- Recuperação Rankeada : Dada uma coleção de documentos, o usuário realiza uma consulta que retorna uma lista ordenada de documentos
- Assumindo noção binária de relevância: $R_{d,q}$ é uma variável aleatória dicotômica, tal que
  - $R_{d,q} = 1$ se o documento d é relevante com respeito à consulta q
  - $R_{d,q} = 0$ caso contrário

O Problema do Rankeamento de Documentos

- Recuperação Rankeada : Dada uma coleção de documentos, o usuário realiza uma consulta que retorna uma lista ordenada de documentos
- Assumindo noção binária de relevância: $R_{d,q}$ é uma variável aleatória dicotômica, tal que
  - $R_{d,q} = 1$ se o documento d é relevante com respeito à consulta q
  - $R_{d,q} = 0$ caso contrário
- O rankeamento probabilístico ordena os documentos em ordem decrescente de relevância estimada com respeito à consulta: $P(R = 1|d, q)$

O Problema do Rankeamento de Documentos

- Recuperação Rankeada : Dada uma coleção de documentos, o usuário realiza uma consulta que retorna uma lista ordenada de documentos
- Assumindo noção binária de relevância: $R_{d,q}$ é uma variável aleatória dicotômica, tal que
  - $R_{d,q} = 1$ se o documento d é relevante com respeito à consulta q
  - $R_{d,q} = 0$ caso contrário
- O rankeamento probabilístico ordena os documentos em ordem decrescente de relevância estimada com respeito à consulta: $P(R = 1|d, q)$
- Assume que a relevância de cada documento é independente da relevância de outros documentos

Princípio do Rankeamento de Probabilidade (PRP)

Princípio do Rankeamento de Probabilidade (PRP)

- PRP resumidamente

Princípio do Rankeamento de Probabilidade (PRP)

- PRP resumidamente
  - Se os documentos recuperados são rankeados decrescentemente com sua probabilidade de relevância, então a efetividade do sistema será a melhor possível.

Princípio do Rankeamento de Probabilidade (PRP)

- PRP resumidamente
  - Se os documentos recuperados são rankeados decrescentemente com sua probabilidade de relevância, então a efetividade do sistema será a melhor possível.
- PRP em detalhes

Princípio do Rankeamento de Probabilidade (PRP)

- PRP resumidamente
  - Se os documentos recuperados são rankeados decrescentemente com sua probabilidade de relevância, então a efetividade do sistema será a melhor possível.
- PRP em detalhes
  - Se a resposta do sistema a cada consulta for um rankeamento dos documentos em ordem decrescente de probabilidade de relevância para a consulta, Onde as probabilidades são estimadas com o máximo de acurácia possível, utilizando toda a informação disponível

Modelo de Independência Binário (BIM)

Modelo de Independência Binário (BIM)

- Tradicionalmente usado com o PRP

Pressupostos:

Modelo de Independência Binário (BIM)

- Tradicionalmente usado com o PRP

Pressupostos:

- ‘Binário’ (equivalente ao booleano): documentos e consultas representados vetores de incidência binários

Modelo de Independência Binário (BIM)

- Tradicionalmente usado com o PRP

Pressupostos:

- 'Binário' (equivalente ao booleano): documentos e consultas representados vetores de incidência binários
  - P.Ex., documento d representado pelo vetor $\vec{x} = (x_1, \ldots, x_M)$, onde $x_t = 1$ se o termo t ocorre em d e $x_t = 0$ em caso contrário.

Modelo de Independência Binário (BIM)

- Tradicionalmente usado com o PRP

Pressupostos:

- 'Binário' (equivalente ao booleano): documentos e consultas representados vetores de incidência binários
    - P.Ex., documento d representado pelo vetor $\vec{x} = (x_1, \ldots, x_M)$, onde $x_t = 1$ se o termo t ocorre em d e $x_t = 0$ em caso contrário.
    - Documentos diferentes podem ter a mesma representação vetorial

Modelo de Independência Binário (BIM)

- Tradicionalmente usado com o PRP

Pressupostos:

- 'Binário' (equivalente ao booleano): documentos e consultas representados vetores de incidência binários
  - P.Ex., documento d representado pelo vetor $\vec{x} = (x_1, \ldots, x_M)$, onde $x_t = 1$ se o termo t ocorre em d e $x_t = 0$ em caso contrário.
  - Documentos diferentes podem ter a mesma representação vetorial
- 'Independência': não há associação entre termos

Matriz de Incidência Binária

	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	
...							

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

Modelo de Independência Binária

Modelo de Independência Binária

Para tornar precisa uma estratégia de recuperação probabilística, precisamos estimar como os termos do documento contribuem para sua relevância

- Precisamos encontrar estatísticas mensuráveis (frequência do termo, frequência de documentos, comprimento do documento) que afetem a relevância de um documento

Modelo de Independência Binária

Para tornar precisa uma estratégia de recuperação probabilística, precisamos estimar como os termos do documento contribuem para sua relevância

- Precisamos encontrar estatísticas mensuráveis (frequência do termo, frequência de documentos, comprimento do documento) que afetem a relevância de um documento
- Combinar estas estatísticas para estima a probabilidade da relevância do documento: $P(R|d, q)$

Modelo de Independência Binária

Para tornar precisa uma estratégia de recuperação probabilística, precisamos estimar como os termos do documento contribuem para sua relevância

- Precisamos encontrar estatísticas mensuráveis (frequência do termo, frequência de documentos, comprimento do documento) que afetem a relevância de um documento
- Combinar estas estatísticas para estima a probabilidade da relevância do documento: $P(R|d, q)$
- Como fazemos isso?

Modelo de Independência Binária

Modelo de Independência Binária

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

$$
\begin{aligned}
P(R=1|\vec{x},\vec{q}) &= \frac{P(\vec{x}|R=1,\vec{q})P(R=1|\vec{q})}{P(\vec{x}|\vec{q})} \\
P(R=0|\vec{x},\vec{q}) &= \frac{P(\vec{x}|R=0,\vec{q})P(R=0|\vec{q})}{P(\vec{x}|\vec{q})}
\end{aligned}
$$

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

Modelo de Independência Binária

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

$$
\begin{aligned}
P(R=1|\vec{x},\vec{q}) &= \frac{P(\vec{x}|R=1,\vec{q})P(R=1|\vec{q})}{P(\vec{x}|\vec{q})} \\
P(R=0|\vec{x},\vec{q}) &= \frac{P(\vec{x}|R=0,\vec{q})P(R=0|\vec{q})}{P(\vec{x}|\vec{q})}
\end{aligned}
$$

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

Modelo de Independência Binária

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

$$
\begin{aligned}
P(R=1|\vec{x},\vec{q}) &= \frac{P(\vec{x}|R=1,\vec{q})P(R=1|\vec{q})}{P(\vec{x}|\vec{q})} \\
P(R=0|\vec{x},\vec{q}) &= \frac{P(\vec{x}|R=0,\vec{q})P(R=0|\vec{q})}{P(\vec{x}|\vec{q})}
\end{aligned}
$$

Modelo de Independência Binária

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

$$
\begin{aligned}
P(R=1|\vec{x},\vec{q}) &= \frac{P(\vec{x}|R=1,\vec{q})P(R=1|\vec{q})}{P(\vec{x}|\vec{q})} \\
P(R=0|\vec{x},\vec{q}) &= \frac{P(\vec{x}|R=0,\vec{q})P(R=0|\vec{q})}{P(\vec{x}|\vec{q})}
\end{aligned}
$$

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

Modelo de Independência Binária

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

$$
\begin{aligned}
P(R=1|\vec{x},\vec{q}) &= \frac{P(\vec{x}|R=1,\vec{q})P(R=1|\vec{q})}{P(\vec{x}|\vec{q})} \\
P(R=0|\vec{x},\vec{q}) &= \frac{P(\vec{x}|R=0,\vec{q})P(R=0|\vec{q})}{P(\vec{x}|\vec{q})}
\end{aligned}
$$

- $P(R=1|\vec{q})$ e $P(R=0|\vec{q})$: probabilidade *a priori* de recuperar um documento relevante ou irrelevante para uma consulta $\vec{q}$
- Estimar $P(R=1|\vec{q})$ e $P(R=0|\vec{q})$ a partir da percentagem de documentos relevantes na coleção

Modelo de Independência Binária

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

$$
\begin{aligned}
P(R=1|\vec{x},\vec{q}) &= \frac{P(\vec{x}|R=1,\vec{q})P(R=1|\vec{q})}{P(\vec{x}|\vec{q})} \\
P(R=0|\vec{x},\vec{q}) &= \frac{P(\vec{x}|R=0,\vec{q})P(R=0|\vec{q})}{P(\vec{x}|\vec{q})}
\end{aligned}
$$

- $P(R=1|\vec{q})$ e $P(R=0|\vec{q})$: probabilidade *a priori* de recuperar um documento relevante ou irrelevante para uma consulta $\vec{q}$
- Estimar $P(R=1|\vec{q})$ e $P(R=0|\vec{q})$ a partir da percentagem de documentos relevantes na coleção
- Uma vez que um documento é ou relevante ou irrelevante para uma consulta, temos que:

Modelo de Independência Binária

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

$$
\begin{aligned}
P(R=1|\vec{x},\vec{q}) &= \frac{P(\vec{x}|R=1,\vec{q})P(R=1|\vec{q})}{P(\vec{x}|\vec{q})} \\
P(R=0|\vec{x},\vec{q}) &= \frac{P(\vec{x}|R=0,\vec{q})P(R=0|\vec{q})}{P(\vec{x}|\vec{q})}
\end{aligned}
$$

- $P(R=1|\vec{q})$ e $P(R=0|\vec{q})$: probabilidade *a priori* de recuperar um documento relevante ou irrelevante para uma consulta $\vec{q}$
- Estimar $P(R=1|\vec{q})$ e $P(R=0|\vec{q})$ a partir da percentagem de documentos relevantes na coleção
- Uma vez que um documento é ou relevante ou irrelevante para uma consulta, temos que:

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

Encontrando uma função de rankeamento para termos de busca(1)

Encontrando uma função de rankeamento para termos de busca(1)

- Dada uma consulta q, rankear documentos por $P(R=1|d,q)$ é modelado no BIM como rankear por $P(R=1|\vec{x},\vec{q})$
- Mais fácil: rankear documentos por seus odds de relevância (dado o mesmo rankeamento)

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

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

- $\frac{P(R=1|\vec{q})}{P(R=0|\vec{q})}$ é constante para uma dada consulta - pode ser ignorado

Encontrando uma função de rankeamento para termos de busca (2)

Encontrando uma função de rankeamento para termos de busca (2)

É neste ponto que aceitamos o pressuposto de independência condicional do Naive Bayes segundo o qual a presença ou ausência de uma palavra em um documento é independente da presença ou ausência de qualquer outra palavra (dada a consulta):

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

logo:

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

Exercício

Exercício

Pressuposto de independência condicional do Naive Bayes: a presença ou ausência de uma palavra em um documento é independente da presença ou ausência de qualquer outra palavra (dada a consulta).

Exercício

Pressuposto de independência condicional do Naive Bayes: a presença ou ausência de uma palavra em um documento é independente da presença ou ausência de qualquer outra palavra (dada a consulta).

Porquê isto está errado? cite um bom exemplo?

Exercício

Pressuposto de independência condicional do Naive Bayes: a presença ou ausência de uma palavra em um documento é independente da presença ou ausência de qualquer outra palavra (dada a consulta).

Porquê isto está errado? cite um bom exemplo?

PRP assume que a relevância de cada documento é independente da relevância dos outros documentos.

Exercício

Pressuposto de independência condicional do Naive Bayes: a presença ou ausência de uma palavra em um documento é independente da presença ou ausência de qualquer outra palavra (dada a consulta).

Porquê isto está errado? cite um bom exemplo?

PRP assume que a relevância de cada documento é independente da relevância dos outros documentos.

Porquê isto está errado? cite um bom exemplo?

Encontrando uma função de rankeamento para termos de busca (3)

Encontrando uma função de rankeamento para termos de busca (3)

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

Encontrando uma função de rankeamento para termos de busca (3)

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

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

Encontrando uma função de rankeamento para termos de busca (4)

Encontrando uma função de rankeamento para termos de busca (4)

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

Encontrando uma função de rankeamento para termos de busca (4)

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

Encontrando uma função de rankeamento para termos de busca (4)

- 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 irrelevante

Encontrando uma função de rankeamento para termos de busca (4)

- 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 irrelevante

Encontrando uma função de rankeamento para termos de busca (4)

- 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 irrelevante
- Representando essas probabilidades em uma tabela de contingência:

	documento	relevante $(R = 1)$	irrelevante $(R = 0)$
Termo presente	$x_t = 1$	p_t	u_t
Termo ausente	$x_t = 0$	$1 - p_t$	$1 - u_t$

Encontrando uma função de rankeamento para termos de busca

Encontrando uma função de rankeamento para termos de busca

Pressuposto simplificador adicional: termos que não ocorrem na consulta são igualmente prováveis de ocorrer em documentos relevantes ou irrelevantes

- Se $q_t = 0$, então $p_t = u_t$

Agora precisamos apenas considerar termos nos produtos que aparecem na consulta:

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

Encontrando uma função de rankeamento para termos de busca

Pressuposto simplificador adicional: termos que não ocorrem na consulta são igualmente prováveis de ocorrer em documentos relevantes ou irrelevantes

- Se $q_t = 0$, então $p_t = u_t$

Agora precisamos apenas considerar termos nos produtos que aparecem na consulta:

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

- O produto da esquerda é sobre os termos de consulta encontrados no documento, e o produto da direita é sobre termos de consulta não encontrados no documento

Encontrando uma função de rankeamento para termos de busca

Encontrando uma função de rankeamento para termos de busca

Incluindo os termos de consulta encontrados no documento no produto da direita, mas ao mesmo tempo dividindo o produto da esquerda por eles, obtemos:

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

- O produto da esquerda continua a ser sobre os termos encontrados no documento, mas o da direita agora é sobre todos os termos de consulta, que é constante para uma dada consulta e pode ser ignorado.

Encontrando uma função de rankeamento para termos de busca

Incluindo os termos de consulta encontrados no documento no produto da direita, mas ao mesmo tempo dividindo o produto da esquerda por eles, obtemos:

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

- O produto da esquerda continua a ser sobre os termos encontrados no documento, mas o da direita agora é sobre todos os termos de consulta, que é constante para uma dada consulta e pode ser ignorado.
- $\rightarrow$ A única quantidade que precisa ser estimada para rankear documentos com respeito a uma consulta, é o produto da esquerda

Encontrando uma função de rankeamento para termos de busca

Incluindo os termos de consulta encontrados no documento no produto da direita, mas ao mesmo tempo dividindo o produto da esquerda por eles, obtemos:

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

- O produto da esquerda continua a ser sobre os termos encontrados no documento, mas o da direita agora é sobre todos os termos de consulta, que é constante para uma dada consulta e pode ser ignorado.
- $\rightarrow$ A única quantidade que precisa ser estimada para rankear documentos com respeito a uma consulta, é o produto da esquerda
- Daí vem o Valor de status de Recuperação (RSV) neste modelo:

Encontrando uma função de rankeamento para termos de busca

Encontrando uma função de rankeamento para termos de busca

Equivalente: rankear documentos usando o log da razão de odds para os termos 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 razão de odds (ou razão de chances) é a razão de dois odds: (i) os odds do termo aparecer se o documento for relevante ($p_t/(1-p_t)$), e (ii) os odds do termo aparecer se o documento for irrelevante ($u_t/(1-u_t)$)

Encontrando uma função de rankeamento para termos de busca

Equivalente: rankear documentos usando o log da razão de odds para os termos 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 razão de odds (ou razão de chances) é a razão de dois odds: (i) os odds do termo aparecer se o documento for relevante ($p_t/(1-p_t)$), e (ii) os odds do termo aparecer se o documento for irrelevante ($u_t/(1-u_t)$)
- $c_t = 0$: Termo tem odds iguais de aparecer em docs relevantes e irrelevantes

Encontrando uma função de rankeamento para termos de busca

Equivalente: rankear documentos usando o log da razão de odds para os termos 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 razão de odds (ou razão de chances) é a razão de dois odds: (i) os odds do termo aparecer se o documento for relevante ($p_t/(1-p_t)$), e (ii) os odds do termo aparecer se o documento for irrelevante ($u_t/(1-u_t)$)
- $c_t = 0$: Termo tem odds iguais de aparecer em docs relevantes e irrelevantes
- c_t positivo: odds mais altos de aparecer em documentos relevantes

Encontrando uma função de rankeamento para termos de busca

Equivalente: rankear documentos usando o log da razão de odds para os termos 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 razão de odds (ou razão de chances) é a razão de dois odds: (i) os odds do termo aparecer se o documento for relevante ($p_t/(1 - p_t)$), e (ii) os odds do termo aparecer se o documento for irrelevante ($u_t/(1 - u_t)$)
- $c_t = 0$: Termo tem odds iguais de aparecer em docs relevantes e irrelevantes
- c_t positivo: odds mais altos de aparecer em documentos relevantes
- c_t negativo: odds mais altos de aparecer em documentos irrelevantes

Peso de termo c_t no BIM

Peso de termo c_t no BIM

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

Peso de termo c_t no BIM

- $c_t = \log \frac{p_t}{(1-p_t)} - \log \frac{u_t}{1-u_t}$ funciona como peso para o termo.
- RSV parap documento d: $RSV_d = \sum_{x_t=q_t=1} c_t$.

Peso de termo c_t no BIM

- $c_t = \log \frac{p_t}{(1-p_t)} - \log \frac{u_t}{1-u_t}$ funciona como peso para o termo.
- RSV parap documento d: $RSV_d = \sum_{x_t=q_t=1} c_t$.
- O BIM e o modelo de espaço vetorial são idênticos no nível operacional...

Peso de termo c_t no BIM

- $c_t = \log \frac{p_t}{(1-p_t)} - \log \frac{u_t}{1-u_t}$ funciona como peso para o termo.
- RSV parap documento d: $RSV_d = \sum_{x_t=q_t=1} c_t$.
- O BIM e o modelo de espaço vetorial são idênticos no nível operacional...
- ...exceto que os pesos dos termos são diferentes.

Peso de termo c_t no BIM

- $c_t = \log \frac{p_t}{(1-p_t)} - \log \frac{u_t}{1-u_t}$ funciona como peso para o termo.
- RSV parap documento d: $RSV_d = \sum_{x_t=q_t=1} c_t$.
- O BIM e o modelo de espaço vetorial são idênticos no nível operacional...
- ...exceto que os pesos dos termos são diferentes.
- Ou seja: podemos usar as mesmas estruturas de dados (indices invertidos, etc.) para os dois modelos.

Como Estimar Probabilidades

Como Estimar Probabilidades

para cada termo t em uma consulta, estimamos c_t em toda a coleção usando uma tabela de contingência de contagens de documentos na coleção onde Df_t é o número de documentos que contem o termo t:

	documentos	relevante	irrelevante	Total
Termo presente	$x_t = 1$	s	$Df_t - s$	Df_t
Termo ausente	$x_t = 0$	$S - s$	$(N - Df_t) - (S - s)$	$N - Df_t$
	Total	S	$N - S$	N

$$p_t = s/S$$

$$u_t = (Df_t - s)/(N - S)$$

$$c_t = K(N, Df_t, S, s) = \log \frac{s/(S-s)}{(Df_t - s)/((N - Df_t) - (S - s))}$$

Evitando zeros

Evitando zeros

- Se qualquer das contagens for zero, o peso do termo é mal-definido.

Evitando zeros

- Se qualquer das contagens for zero, o peso do termo é mal-definido.
- Estimativas de máxima verossimilhança não funcionam para eventos raros.

Evitando zeros

- Se qualquer das contagens for zero, o peso do termo é mal-definido.
- Estimativas de máxima verossimilhança não funcionam para eventos raros.
- Para evitar zeros: adicione 0.5 a cada contagem (Estimação por verossimilhança esperada = ELE)

Evitando zeros

- Se qualquer das contagens for zero, o peso do termo é mal-definido.
- Estimativas de máxima verossimilhança não funcionam para eventos raros.
- Para evitar zeros: adicione 0.5 a cada contagem (Estimação por verossimilhança esperada = ELE)
- Por exemplo, use $S - s + 0.5$ na fórmula para $S - s$

Exercício

Exercício

- Consulta: `Obama health plan`
- Doc1: `Obama rejects allegations about his own bad health`
- Doc2: `The plan is to visit Obama`
- Doc3: `Obama raises concerns with US health plan reforms`

Exercício

- Consulta: `Obama health plan`
- Doc1: `Obama rejects allegations about his own bad health`
- Doc2: `The plan is to visit Obama`
- Doc3: `Obama raises concerns with US health plan reforms`

Estime a probabilidade de que os documentos acima são relevantes para a consulta. Use uma tabela de contingência. Estes são os únicos três documentos na coleção

Pressuposto Simplificador

Pressuposto Simplificador

- Assumindo que documentos relevantes são uma fração bem pequena da coleção, aproxime estatísticas para documentos irrelevantes a partir de estatísticas da coleção completa

Pressuposto Simplificador

- Assumindo que documentos relevantes são uma fração bem pequena da coleção, aproxime estatísticas para documentos irrelevantes a partir de estatísticas da coleção completa
- Por conseguinte, u_t (a probabilidade de ocorrência do termo em documentos irrelevantes para uma dada consulta) é Df_t/N e

$$\log[(1-u_t)/u_t] = \log[(N-Df_t)/Df_t] \approx \log N/Df_t$$

Pressuposto Simplificador

- Assumindo que documentos relevantes são uma fração bem pequena da coleção, aproxime estatísticas para documentos irrelevantes a partir de estatísticas da coleção completa
- Por conseguinte, u_t (a probabilidade de ocorrência do termo em documentos irrelevantes para uma dada consulta) é Df_t/N e

$$\log[(1 - u_t)/u_t] = \log[(N - Df_t)/Df_t] \approx \log N/Df_t$$

- Esta equação deve parecer familiar ...

Pressuposto Simplificador

- Assumindo que documentos relevantes são uma fração bem pequena da coleção, aproxime estatísticas para documentos irrelevantes a partir de estatísticas da coleção completa
- Por conseguinte, u_t (a probabilidade de ocorrência do termo em documentos irrelevantes para uma dada consulta) é Df_t/N e

$$\log[(1-u_t)/u_t] = \log[(N-Df_t)/Df_t] \approx \log N/Df_t$$

- Esta equação deve parecer familiar ...
- A aproximação acima não pode ser facilmente estendida para documentos relevantes

Estimativas de probabilidade em feedback de relevância

Estimativas de probabilidade em feedback de relevância

- Estatísticas de documentos relevantes (p_t) no feedback de relevância pode ser estimado por meio de máxima verossimilhança ou ELE(adicionndo 0.5)

Estimativas de probabilidade em feedback de relevância

- Estatísticas de documentos relevantes (p_t) no feedback de relevância pode ser estimado por meio de máxima verossimilhança ou ELE(adicionndo 0.5)
    - Use a frequência de ocorrência de termos em documentos conhecidamente relevantes.

Estimativas de probabilidade em feedback de relevância

- Estatísticas de documentos relevantes (p_t) no feedback de relevância pode ser estimado por meio de máxima verossimilhança ou ELE(adicionndo 0.5)
  - Use a frequência de ocorrência de termos em documentos conhecidamente relevantes.
- Esta é a base das abordagens probabilísticas para feedback de relevância

Estimativas de probabilidade em feedback de relevância

- Estatísticas de documentos relevantes (p_t) no feedback de relevância pode ser estimado por meio de máxima verossimilhança ou ELE(adicionndo 0.5)
  - Use a frequência de ocorrência de termos em documentos conhecidamente relevantes.
- Esta é a base das abordagens probabilísticas para feedback de relevância
- O exercício que acabamos de fazer foi um exercício de feedback de relevância uma vez que assumimos a disponibilidade de julgamentos de relevância.

Estimativas de Probabilidade em Recuperação adhoc

Estimativas de Probabilidade em Recuperação adhoc

- Recuperação Ad-hoc: não há feedback de relevância pelo usuário

Estimativas de Probabilidade em Recuperação adhoc

- Recuperação Ad-hoc: não há feedback de relevância pelo usuário
- Neste caso: assuma que p_t é constante para todos os termos x_t na consulta e que $p_t = 0.5$

Estimativas de Probabilidade em Recuperação adhoc

- Recuperação Ad-hoc: não há feedback de relevância pelo usuário
- Neste caso: assuma que p_t é constante para todos os termos x_t na consulta e que $p_t = 0.5$
- A probabilidade de ocorrência de cada termo em um documento relevante é a mesma, e então p_t e $(1 - p_t)$ são eliminados da expressão do RSV

Estimativas de Probabilidade em Recuperação adhoc

- Recuperação Ad-hoc: não há feedback de relevância pelo usuário
- Neste caso: assuma que p_t é constante para todos os termos x_t na consulta e que $p_t = 0.5$
- A probabilidade de ocorrência de cada termo em um documento relevante é a mesma, e então p_t e $(1 - p_t)$ são eliminados da expressão do RSV
- É uma estimativa fraca, mas não conflita com a expectativa de que os termos de consulta aparecem em muitos mas não todos os documentos relevantes

Estimativas de Probabilidade em Recuperação adhoc

- Recuperação Ad-hoc: não há feedback de relevância pelo usuário
- Neste caso: assuma que p_t é constante para todos os termos x_t na consulta e que $p_t = 0.5$
- A probabilidade de ocorrência de cada termo em um documento relevante é a mesma, e então p_t e $(1 - p_t)$ são eliminados da expressão do RSV
- É uma estimativa fraca, mas não conflita com a expectativa de que os termos de consulta aparecem em muitos mas não todos os documentos relevantes
- Combinando este método com a aproximação anterior para u_t, o rankeamento de documentos é determinado simplesmente por quais termos de consulta ocorrem nos documentos ajustados por seu peso idf

Estimativas de Probabilidade em Recuperação adhoc

- Recuperação Ad-hoc: não há feedback de relevância pelo usuário
- Neste caso: assuma que p_t é constante para todos os termos x_t na consulta e que $p_t = 0.5$
- A probabilidade de ocorrência de cada termo em um documento relevante é a mesma, e então p_t e $(1 - p_t)$ são eliminados da expressão do RSV
- É uma estimativa fraca, mas não conflita com a expectativa de que os termos de consulta aparecem em muitos mas não todos os documentos relevantes
- Combinando este método com a aproximação anterior para u_t, o rankeamento de documentos é determinado simplesmente por quais termos de consulta ocorrem nos documentos ajustados por seu peso idf
- Para documento curto (títulos ou resumos) em situações de recuperação simples, esta estimativa pode ser bem satisfatória

História e sumário dos pressupostos

História e sumário dos pressupostos

- Dentre os modelos formais mais antigos de de RI

História e sumário dos pressupostos

- Dentre os modelos formais mais antigos de de RI
  - Maron & Kuhns, 1960: Uma vez que um sistema de RI não pode predizer com certeza qual documento é relevante, devemos lidar com probabilidades

História e sumário dos pressupostos

- Dentre os modelos formais mais antigos de de RI
  - Maron & Kuhns, 1960: Uma vez que um sistema de RI não pode predizer com certeza qual documento é relevante, devemos lidar com probabilidades
- Pressupostos para obter aproximações razoáveis das probabilidades necessárias(no BIM):

História e sumário dos pressupostos

- Dentre os modelos formais mais antigos de de RI
  - Maron & Kuhns, 1960: Uma vez que um sistema de RI não pode predizer com certeza qual documento é relevante, devemos lidar com probabilidades
- Pressupostos para obter aproximações razoáveis das probabilidades necessárias(no BIM):
  - Representação Booleana de documentos/consutas/relevância

História e sumário dos pressupostos

- Dentre os modelos formais mais antigos de de RI
  - Maron & Kuhns, 1960: Uma vez que um sistema de RI não pode predizer com certeza qual documento é relevante, devemos lidar com probabilidades
- Pressupostos para obter aproximações razoáveis das probabilidades necessárias(no BIM):
  - Representação Booleana de documentos/consutas/relevância
  - Independência de termos

História e sumário dos pressupostos

- Dentre os modelos formais mais antigos de de RI
  - Maron & Kuhns, 1960: Uma vez que um sistema de RI não pode predizer com certeza qual documento é relevante, devemos lidar com probabilidades
- Pressupostos para obter aproximações razoáveis das probabilidades necessárias(no BIM):
  - Representação Booleana de documentos/consutas/relevância
  - Independência de termos
  - Termos fora da consulta não afetam a recuperação

História e sumário dos pressupostos

- Dentre os modelos formais mais antigos de de RI
  - Maron & Kuhns, 1960: Uma vez que um sistema de RI não pode predizer com certeza qual documento é relevante, devemos lidar com probabilidades
- Pressupostos para obter aproximações razoáveis das probabilidades necessárias(no BIM):
  - Representação Booleana de documentos/consutas/relevância
  - Independência de termos
  - Termos fora da consulta não afetam a recuperação
  - Relevâncias de documentos são independentes

Quão diferentes são o modelo vetorial e o BIM?

Quão diferentes são o modelo vetorial e o BIM?

- Não são tão diferentes.

Quão diferentes são o modelo vetorial e o BIM?

- Não são tão diferentes.
- Nos dois você constroi um esquema de recuperação da mesma maneira.

Quão diferentes são o modelo vetorial e o BIM?

- Não são tão diferentes.
- Nos dois você constroi um esquema de recuperação da mesma maneira.
- Para RI probabilístico, no fim das contas, vc não pontua consultas não por similaridade (cosseno) e por tf-idf em um espaço vetorial, mas por uma fórmula ligeiramente diferente motivada pela teoria de probabilidade.

Quão diferentes são o modelo vetorial e o BIM?

- Não são tão diferentes.
- Nos dois você constroi um esquema de recuperação da mesma maneira.
- Para RI probabilístico, no fim das contas, vc não pontua consultas não por similaridade (cosseno) e por tf-idf em um espaço vetorial, mas por uma fórmula ligeiramente diferente motivada pela teoria de probabilidade.
- Em seguida: como adicionar frequencia de termos e normalização de comprimento ao modelo probabilístico.

Okapi BM25: Visão Geral

Okapi BM25: Visão Geral

- O Okapi BM25 é um modelo probabilístico que incorpora frequência dos termos (ou seja, não é binário) e normalização de comprimento.

Okapi BM25: Visão Geral

- O Okapi BM25 é um modelo probabilístico que incorpora frequência dos termos (ou seja, não é binário) e normalização de comprimento.
- O BIM foi concebido originalmente para catálogos curtos de comprimento similar, e funciona bem nestes contextos

Okapi BM25: Visão Geral

- O Okapi BM25 é um modelo probabilístico que incorpora frequência dos termos (ou seja, não é binário) e normalização de comprimento.
- O BIM foi concebido originalmente para catálogos curtos de comprimento similar, e funciona bem nestes contextos
- Para buscas de texto completo modernas, um modelo deve atentar à frequência de termos e ao comprimento do documento

Okapi BM25: Visão Geral

- O Okapi BM25 é um modelo probabilístico que incorpora frequência dos termos (ou seja, não é binário) e normalização de comprimento.
- O BIM foi concebido originalmente para catálogos curtos de comprimento similar, e funciona bem nestes contextos
- Para buscas de texto completo modernas, um modelo deve atentar à frequência de termos e ao comprimento do documento
- BestMatch25 (também conhecido como BM25 ou Okapi) é sensível a estas grandezas

Okapi BM25: Visão Geral

- O Okapi BM25 é um modelo probabilístico que incorpora frequência dos termos (ou seja, não é binário) e normalização de comprimento.
- O BIM foi concebido originalmente para catálogos curtos de comprimento similar, e funciona bem nestes contextos
- Para buscas de texto completo modernas, um modelo deve atentar à frequência de termos e ao comprimento do documento
- BestMatch25 (também conhecido como BM25 ou Okapi) é sensível a estas grandezas
- O BM25 é um dos modelos de recuperação mais robustos e amplamente utilizados

Okapi BM25

Okapi BM25

- O escore mais simples para o documento d é simplesmente o peso idf dos termos de consulta presentes no documento:

Okapi BM25

- O escore mais simples para o documento d é simplesmente o peso idf dos termos de consulta presentes no documento:

Okapi BM25

- O escore mais simples para o documento d é simplesmente o peso idf dos termos de consulta presentes no documento:

$$RSV_d = \sum_{t \in q} \log \frac{N}{Df_t}$$

Ponderação básica do Okapi BM25

Ponderação básica do Okapi BM25

- Melhora o idf do termo [log N/df] através da inclusão da frequência do termo e do comprimento do documento.

$$RSV_d = \sum_{t \in q} \log \left[\frac{N}{Df_t} \right] \cdot \frac{(k_1 + 1) tf_{td}}{k_1((1 - b) + b \times (L_d / L_{med})) + tf_{td}}$$

Ponderação básica do Okapi BM25

- Melhora o idf do termo [log N/df] através da inclusão da frequência do termo e do comprimento do documento.

$$RSV_d = \sum_{t \in q} \log \left[\frac{N}{Df_t} \right] \cdot \frac{(k_1 + 1) tf_{td}}{k_1((1 - b) + b \times (L_d / L_{med})) + tf_{td}}$$

- tf_{td}: Frequência do termo no documento d

Ponderação básica do Okapi BM25

- Melhora o idf do termo [log N/df] através da inclusão da frequência do termo e do comprimento do documento.

$$RSV_d = \sum_{t \in q} \log \left[\frac{N}{Df_t} \right] \cdot \frac{(k_1 + 1)tf_{td}}{k_1((1-b) + b \times (L_d/L_{med})) + tf_{td}}$$

- tf_{td}: Frequência do termo no documento d
- L_d (L_{med}): Comprimento do documento d (Comprimento médio dos documentos na coleção inteira)

Ponderação básica do Okapi BM25

- Melhora o idf do termo [log N/df] através da inclusão da frequência do termo e do comprimento do documento.

$$RSV_d = \sum_{t \in q} \log \left[\frac{N}{Df_t} \right] \cdot \frac{(k_1 + 1) tf_{td}}{k_1((1-b) + b \times (L_d / L_{med})) + tf_{td}}$$

- tf_{td}: Frequência do termo no documento d
- L_d (L_{med}): Comprimento do documento d (Comprimento médio dos documentos na coleção inteira)
- k_1: Parâmetro de ajuste controlando a imfluência da frequencia do termo

Ponderação básica do Okapi BM25

- Melhora o idf do termo [log N/df] através da inclusão da frequência do termo e do comprimento do documento.

$$RSV_d = \sum_{t \in q} \log \left[\frac{N}{Df_t} \right] \cdot \frac{(k_1 + 1) tf_{td}}{k_1((1 - b) + b \times (L_d / L_{med})) + tf_{td}}$$

- tf_{td}: Frequência do termo no documento d
- L_d (L_{med}): Comprimento do documento d (Comprimento médio dos documentos na coleção inteira)
- k_1: Parâmetro de ajuste controlando a imfluência da frequencia do termo
- b: Parâmetro de ajuste controlando a influência do comprimento do documento

Exercício

Exercício

- Interprete a fórmula de ponderação BM25 para $k_1 = 0$
- Interprete a fórmula de ponderação BM25 para $k_1 = 1$ e $b = 0$
- Interprete a fórmula de ponderação BM25 para $k_1 \mapsto \infty$ e $b = 0$
- Interprete a fórmula de ponderação BM25 para $k_1 \mapsto \infty$ e $b = 1$

Ponderação Okapi BM25 para consultas longas

Ponderação Okapi BM25 para consultas longas

- Para consultas longas, use ponderação similar para os termos de busca

$$RSV_d = \sum_{t \in q} \left[\log \frac{N}{Df_t}\right] \cdot \frac{(k_1 + 1)tf_{td}}{k_1((1-b) + b \times (L_d/L_{med})) + tf_{td}} \cdot \frac{(k_3 + 1)tf_{tq}}{k_3 + tf_{tq}}$$

Ponderação Okapi BM25 para consultas longas

- Para consultas longas, use ponderação similar para os termos de busca

$$RSV_d = \sum_{t \in q} \left[\log \frac{N}{Df_t} \right] \cdot \frac{(k_1 + 1)tf_{td}}{k_1((1-b) + b \times (L_d/L_{med})) + tf_{td}} \cdot \frac{(k_3 + 1)tf_{tq}}{k_3 + tf_{tq}}$$

- tf_{tq}: frequência do termo na consulta q

Ponderação Okapi BM25 para consultas longas

- Para consultas longas, use ponderação similar para os termos de busca

$$RSV_d = \sum_{t \in q} \left[\log \frac{N}{Df_t} \right] \cdot \frac{(k_1 + 1) tf_{td}}{k_1((1 - b) + b \times (L_d / L_{med})) + tf_{td}} \cdot \frac{(k_3 + 1) tf_{tq}}{k_3 + tf_{tq}}$$

- tf_{tq}: frequência do termo na consulta q
- k_3: parâmetro de ajuste controlando a importância frequência do termo na consulta

Ponderação Okapi BM25 para consultas longas

- Para consultas longas, use ponderação similar para os termos de busca

$$RSV_d = \sum_{t \in q} \left[\log \frac{N}{Df_t} \right] \cdot \frac{(k_1 + 1) tf_{td}}{k_1((1-b) + b \times (L_d / L_{med})) + tf_{td}} \cdot \frac{(k_3 + 1) tf_{tq}}{k_3 + tf_{tq}}$$

- tf_{tq}: frequência do termo na consulta q
- k_3: parâmetro de ajuste controlando a importância frequência do termo na consulta
- Não há normalização de comprimento para consultas (pois a recuperação é feita com respeito a uma única consulta fixa)

Ponderação Okapi BM25 para consultas longas

- Para consultas longas, use ponderação similar para os termos de busca

$$RSV_d = \sum_{t \in q} \left[\log \frac{N}{Df_t} \right] \cdot \frac{(k_1 + 1) tf_{td}}{k_1((1-b) + b \times (L_d / L_{med})) + tf_{td}} \cdot \frac{(k_3 + 1) tf_{tq}}{k_3 + tf_{tq}}$$

- tf_{tq}: frequência do termo na consulta q
- k_3: parâmetro de ajuste controlando a importância frequência do termo na consulta
- Não há normalização de comprimento para consultas (pois a recuperação é feita com respeito a uma única consulta fixa)
- Os parâmetros de ajuste acima devem ser escolhidos de maneira a otimizar a performance em uma coleção de teste. Na ausência de tal otimização, experimentos mostram que valores razoáveis para k_1 e k_3 encontram-se entre 1.2 e 2 e $b = 0.75$

Qual Modelo de Rankeamento Devo Usar?

Qual Modelo de Rankeamento Devo Usar?

- Quero algo básico e simples $\rightarrow$ use o modelo vetorial com ponderação tf-idf.

Qual Modelo de Rankeamento Devo Usar?

- Quero algo básico e simples → use o modelo vetorial com ponderação tf-idf.
- Quero usar um modelo de rankeamento "estado-da-arte" com ótima performance → use modelos de linguagem com BM25 e parâmetros ajustados

Qual Modelo de Rankeamento Devo Usar?

- Quero algo básico e simples → use o modelo vetorial com ponderação tf-idf.
- Quero usar um modelo de rankeamento "estado-da-arte" com ótima performance → use modelos de linguagem com BM25 e parâmetros ajustados
- Algo intermediário: BM25 ou modelos de linguagem sem ajuste ou com apenas um parâmetro ajustado

Material extra

- Capítulo 11 do IIR
- Resources at `http://ifnlp.org/ir`