Introdução à Recuperação de Informações

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

IRI 1: Introdução

Flávio Codeço Coelho

Escola de Matemática Aplicada, Fundação Getúlio Vargas

Sumário da Aula

Definição

Recuperação de informação pode ser definida como a técnica e a arte de encontrar conteúdo em grandes coleções não (ou pouco) estruturadas de documentos (em formatos digitais) de forma a satisfazer nossas necessidades informacionais[1].

[1] adaptado de Hinrich Schütze

Definição

Recuperação de informação pode ser definida como a técnica e a arte de encontrar conteúdo em grandes coleções não (ou pouco) estruturadas de documentos (em formatos digitais) de forma a satisfazer nossas necessidades informacionais[1].

[1] adaptado de Hinrich Schütze

Definição

Recuperação de informação pode ser definida como a técnica e a arte de encontrar conteúdo em grandes coleções não (ou pouco) estruturadas de documentos (em formatos digitais) de forma a satisfazer nossas necessidades informacionais[1].

[1] adaptado de Hinrich Schütze

Definição

Recuperação de informação pode ser definida como a técnica e a arte de encontrar conteúdo em grandes coleções não (ou pouco) estruturadas de documentos (em formatos digitais) de forma a satisfazer nossas necessidades informacionais[1].

[1] adaptado de Hinrich Schütze

Definição

Recuperação de informação pode ser definida como a técnica e a arte de encontrar conteúdo em grandes coleções não (ou pouco) estruturadas de documentos (em formatos digitais) de forma a satisfazer nossas necessidades informacionais[1].

[1] adaptado de Hinrich Schütze

Definição

Recuperação de informação pode ser definida como a técnica e a arte de encontrar conteúdo em grandes coleções não (ou pouco) estruturadas de documentos (em formatos digitais) de forma a satisfazer nossas necessidades informacionais[1].

[1] adaptado de Hinrich Schütze

Definição

Recuperação de informação pode ser definida como a técnica e a arte de encontrar conteúdo em grandes coleções não (ou pouco) estruturadas de documentos (em formatos digitais) de forma a satisfazer nossas necessidades informacionais[1].

[1] adaptado de Hinrich Schütze

Mecânica do Curso

- Foco na Recuperação de informação em coleções de texto.

Mecânica do Curso

- Foco na Recuperação de informação em coleções de texto.
- Exercícios exigirão conhecimentos de programação em Python

Mecânica do Curso

- Foco na Recuperação de informação em coleções de texto.
- Exercícios exigirão conhecimentos de programação em Python
- Avaliação baseada em mini-projetos (um projeto a cada duas semanas)

Mecânica do Curso

- Foco na Recuperação de informação em coleções de texto.
- Exercícios exigirão conhecimentos de programação em Python
- Avaliação baseada em mini-projetos (um projeto a cada duas semanas)
- Projetos serão desenvolvidos em duplas rotatórias, ou seja, cada par de alunos só poderá trabalhar em um projeto.

Mecânica do Curso

- Foco na Recuperação de informação em coleções de texto.
- Exercícios exigirão conhecimentos de programação em Python
- Avaliação baseada em mini-projetos (um projeto a cada duas semanas)
- Projetos serão desenvolvidos em duplas rotatórias, ou seja, cada par de alunos só poderá trabalhar em um projeto.
- Dados e infraestrutura computacional serão fornecidos pela escola sempre que necessário

Contéudo

Este curso se restringirá à exploração e aplicação de modelos matemáticos de recuperação de informação

- Modelos Booleanos

Contéudo

Este curso se restringirá à exploração e aplicação de modelos matemáticos de recuperação de informação

- Modelos Booleanos
  - Fuzzy

Contéudo

Este curso se restringirá à exploração e aplicação de modelos matemáticos de recuperação de informação

- Modelos Booleanos
  - Fuzzy
  - Modelo Booleano extendido

Contéudo

Este curso se restringirá à exploração e aplicação de modelos matemáticos de recuperação de informação

- Modelos Booleanos
  - Fuzzy
  - Modelo Booleano extendido
- Modelos Vetoriais

Contéudo

Este curso se restringirá à exploração e aplicação de modelos matemáticos de recuperação de informação

- Modelos Booleanos
  - Fuzzy
  - Modelo Booleano extendido
- Modelos Vetoriais
  - Espaços vetoriais

Contéudo

Este curso se restringirá à exploração e aplicação de modelos matemáticos de recuperação de informação

- Modelos Booleanos
  - Fuzzy
  - Modelo Booleano extendido
- Modelos Vetoriais
  - Espaços vetoriais
  - Indexação semântica latente

Contéudo

Este curso se restringirá à exploração e aplicação de modelos matemáticos de recuperação de informação

- Modelos Booleanos
  - Fuzzy
  - Modelo Booleano extendido
- Modelos Vetoriais
  - Espaços vetoriais
  - Indexação semântica latente
  - Classificação

Contéudo

Este curso se restringirá à exploração e aplicação de modelos matemáticos de recuperação de informação

- Modelos Booleanos
  - Fuzzy
  - Modelo Booleano extendido
- Modelos Vetoriais
  - Espaços vetoriais
  - Indexação semântica latente
  - Classificação
  - Clusterização

Contéudo

Este curso se restringirá à exploração e aplicação de modelos matemáticos de recuperação de informação

- Modelos Booleanos
  - Fuzzy
  - Modelo Booleano extendido
- Modelos Vetoriais
  - Espaços vetoriais
  - Indexação semântica latente
  - Classificação
  - Clusterização
- Modelos Probabilísticos

Contéudo

Este curso se restringirá à exploração e aplicação de modelos matemáticos de recuperação de informação

- Modelos Booleanos
  - Fuzzy
  - Modelo Booleano extendido
- Modelos Vetoriais
  - Espaços vetoriais
  - Indexação semântica latente
  - Classificação
  - Clusterização
- Modelos Probabilísticos
  - Redes Bayesianas

Contéudo

Este curso se restringirá à exploração e aplicação de modelos matemáticos de recuperação de informação

- Modelos Booleanos
  - Fuzzy
  - Modelo Booleano extendido
- Modelos Vetoriais
  - Espaços vetoriais
  - Indexação semântica latente
  - Classificação
  - Clusterização
- Modelos Probabilísticos
  - Redes Bayesianas
  - Graphical Models

Contéudo

Este curso se restringirá à exploração e aplicação de modelos matemáticos de recuperação de informação

- Modelos Booleanos
  - Fuzzy
  - Modelo Booleano extendido
- Modelos Vetoriais
  - Espaços vetoriais
  - Indexação semântica latente
  - Classificação
  - Clusterização
- Modelos Probabilísticos
  - Redes Bayesianas
  - Graphical Models
  - Belief Networks

Quão boa é nossa recuperação?

Antes de desenvolver qualquer estratégia de recuperação precisamos definir nossa meta e uma métrica de qualidade.

- A meta depende da necessidade informacional

Quão boa é nossa recuperação?

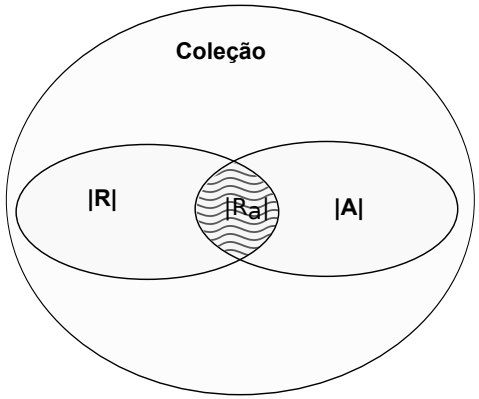
Antes de desenvolver qualquer estratégia de recuperação precisamos definir nossa meta e uma métrica de qualidade.

- A meta depende da necessidade informacional
- Existem algumas métricas classicas de qualidade

Precisão e Revocação(Recall)

Seja R um conjunto de documentos relevantes e $|R|$ o número de documentos neste conjunto. Uma requisiçã de informação I, gera um conjunto A contendo $|A|$ documentos em resposta. Seja $|R_a|$ o número de documentos da interseção entre R e A
Podemos definir revocação como:

$$Rev = \frac{|R_a|}{|R|}$$



$$Precisão = \frac{|R_a|}{|A|}$$

Na Prática

Seja $R_q = \{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123}\}$ o conjunto de documentos relevantes para uma consulta q.
Ordenando o conjunto A_q de respostas a q em ordem decrescente de relevância, temos:

Resultados ordenados

Ordem	Resultado	Precisão	Revocação
1	$\mathbf{d_{123}}$	100%	10%
2	d_{84}	50%	10%
3	$\mathbf{d_{56}}$	66%	20%
4	d_6	50%	20%
5	d_8	40%	20%
6	$\mathbf{d_9}$	50%	30%

Problemas

- Conjunto $|R|$ em situações reais pode ser difícil ou impossível de determinar.
- Revocação e Precisão são medidas correlacionadas.
- visão muito simplista sobre a qualidade da recuperação.

Média Harmônica

Como precisão e revocação são medidas correlacionadas, podemos buscar integrá-las em uma mesma medida.

Média Harmônica

$$F(j) = \frac{2}{\frac{1}{r_j} + \frac{1}{P_j}}$$

onde r_j e P_j são a revocação e a precisão do j-ésimo documento rankeado.
$F(j)$ assume valores no intervalo $[0, 1]$, sendo 0 quando nenhum documento relevante for recuperado e 1 quando todos os documentos recuperados forem relevantes.

Medida E

$E(j)$

$$E(j) = 1 - \frac{1 + b^2}{\frac{b^2}{r_j} + \frac{1}{P_j}}$$

Onde b é um parâmetro the indica a importância relativa da revocação e da precisão. Quando $b = 1$, E é o complemento da média harmônica. Quando $b < 1$, damos mais peso à precisão e quando $b > 1$ damos mais peso à revocação.

Medidas Subjetivas

Seja U um subconjunto de R que é do conhecimento do usuário. $|U|$ é o número de documentos neste conjunto. Seja $|R_k|$ o número de documentos da interseção entre A e U, e $|R_u|$ o número de documentos pertencentes a A mas não a U, i.e., $A - U$

Cobertura e Novidade

$$Cobertura = \frac{|R_k|}{|U|}$$

$$Novidade = \frac{|R_u|}{|R_u| + |R_k|}$$

Recuperação Booleana

Modelo de recuperação no qual podemod construir consultas na forma de uma expressão booleana, ou seja, os termos de busca são combinados com operadores *AND*, *OR* e *NOT*. este modelo vê cada documento como um simples conjunto de palavras.