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?

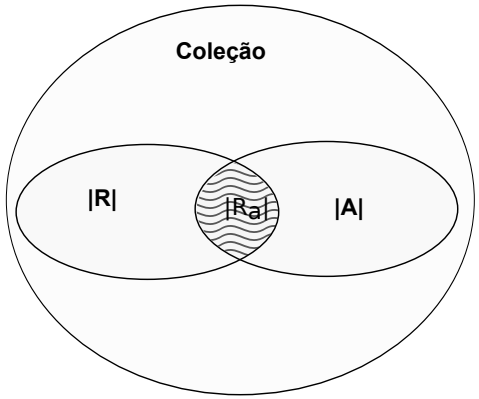
Antes de desenvolver qualquer estratégia de recuperação precisamos definir nossa meta e uma métrica de qualidade.

- 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:

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



$$Precisão = \frac{|R_a|}{|A|}$$

Problemas

- Conjunto $|R|$ em situações reais pode ser difícil ou impossível de determinar.
-