Ciclo de Encontros: 4 x PLN

# Recuperação de Informação

Viviane Moreira
Instituto de Informática UFRGS

25/02/25

# Sumário

# Introdução

# Introdução

- Origem: Necessidade de organizar a informação → bibliotecas → mecanismos de busca
- A Recuperação de Informação **(RI)** trata de **encontrar**, a partir de grandes coleções, material (geralmente documentos) de natureza não estruturada (geralmente texto) que satisfaçam uma **necessidade de informação** (Manning et al. 2008).
- O objetivo central da RI é a busca, ou seja, a tarefa de encontrar material **relevante** a partir de uma consulta de um usuário.
- Nosso foco: dados **textuais**
- Relação com **PLN**

# A Tarefa Central da RI

- Casar a consulta de um usuário com os documentos que são **potencialmente relevantes** a ela
- Dificuldades: **incompatibilidade de vocabulário**
  - **Sinonímia**: palavras diferentes usadas para o mesmo conceito
    - Ex: “bergamota”, “tangerina” e “mexerica
  - **Polissemia**: uma mesma palavra pode apresentar sentidos distintos
    - Ex: manga
  - **Escala**
    - Google recebe cerca de **100 mil consultas por segundo** e tem cerca de **4,3 bilhões de usuários** (https://wpdevshed.com/how-many-people-use-google)

# O Conceito de Relevância

- “Um sistema de RI retorna os resultados **em ordem de relevância**” ✗
- A relevância é um **julgamento feito pelo usuário** que indica o quão bem um documento satisfaz a consulta
- Pode ser **binária** ou ter múltiplos **níveis**
- É **subjetiva**
- A ordenação é dada por uma **medida de similaridade** entre a consulta e os documentos
- A similaridade precisa basear-se **em evidências que conseguimos calcular**
- Espera-se que a medida de similaridade esteja **positivamente correlacionada** com a ideia de relevância dos usuários

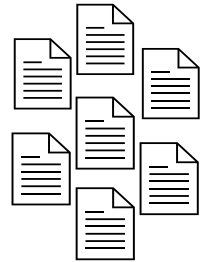
# Um Sistema Típico de RI

# Visão Geral do Processo de RI

Usuária

Consulta

Sistema de Recuperação de Informação



Documentos

Lista Ordenada

Resultados

# Visão Geral do Processo de RI

Usuária

Consulta

Sistema de Recuperação de Informação

Pré-processamento ①

Indexação ②

Índice

Documentos

offline

Resultados

Lista Ordenada

# Visão Geral do Processo de RI

Usuária

Consulta

Sistema de Recuperação de Informação

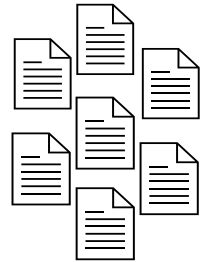
Pré-processamento ③

Pré-processamento ①

Busca e Geração do Ranking ④

Índice

Indexação ②



Documentos

Resultados

Lista Ordenada

# Pré-Processamento

- Tokenização
- Case folding (conversão para minúsculo)
- Remoção de Stopwords
- Stemming

**Trecho do livro O Tempo e o Vento – O Continente, de Érico Veríssimo, e suas versões após cada etapa do pré-processamento.**

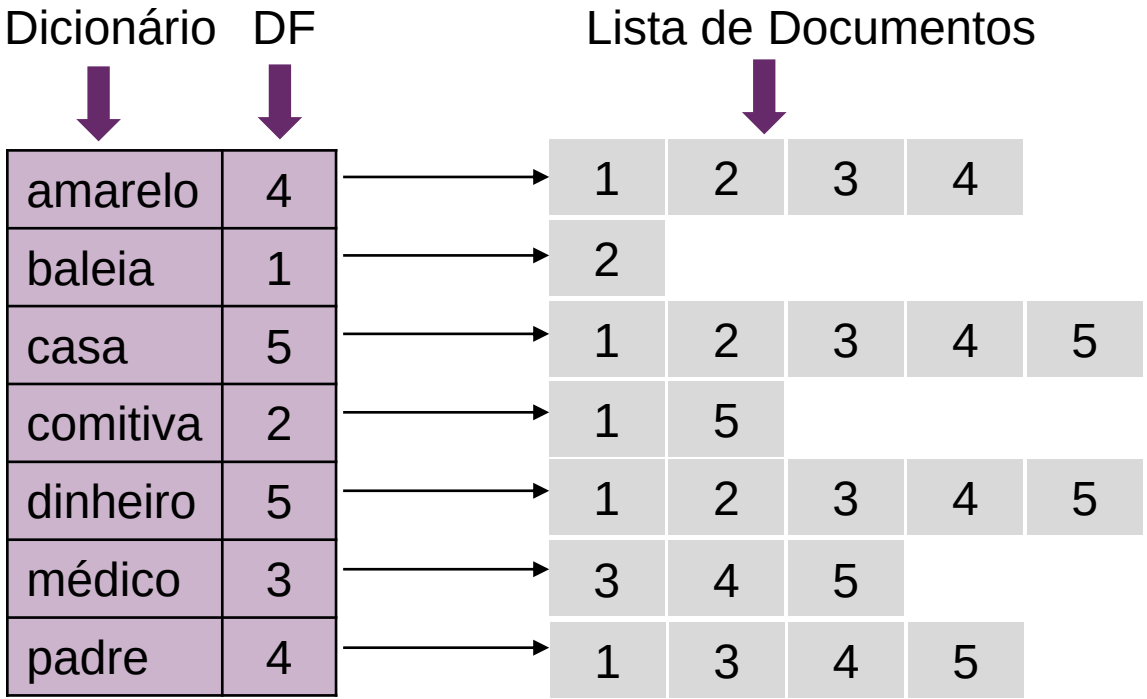
| | |
|---|---|
| Original | Quando pela primeira vez aparecera em Santa Fé, no ano em que fora assinada a paz entre farroupilhas e legalistas, causara a pior das impressões. Chegara escoteiro, montado num cavalo magro e manco, e fazendo questão de mostrar a toda a gente que tinha as guaiacas atestadas de moedas de ouro. |
| Tokenizado | Quando pela primeira vez aparecera em Santa Fé no ano em que fora assinada a paz entre farroupilhas e legalistas causara a pior das impressões Chegara escoteiro montado num cavalo magro e manco e fazendo questão de mostrar a toda a gente que tinha as guaiacas atestadas de moedas de ouro |
| Convertido para minúsculo | quando pela primeira vez aparecera em santa fé no ano em que fora assinada a paz entre farroupilhas e legalistas causara a pior das impressões chegara escoteiro montado num cavalo magro e manco e fazendo questão de mostrar a toda a gente que tinha as guaiacas atestadas de moedas de ouro |
| Após remoção de stop words | primeira vez aparecera santa fé ano assinada paz farroupilhas legalistas causara pior impressões chegara escoteiro montado cavalo magro manco fazendo questão mostrar gente guaiacas atestadas moedas ouro |
| Após stemming | primeir vez aparec sant fé ano assin paz farroupilh legal caus pior impressõ cheg escoteir mont caval magr manc faz questã mostr gent guaiac atest moed our |

# Indexação

- Um motor de busca não tem tempo de escanear todo o texto em busca das palavras da consulta
- Consulta-se um **índice pré-construído**
- Semelhante ao índice remissivo do final de um livro

# Arquivo Invertido

- Para cada termo t: armazene uma lista de documentos que contêm



# Modelos Clássicos

# Modelos Clássicos

- Um modelo de RI especifica como **representar** os documentos, as consultas e como compará-los
- Modelos Clássicos
  - **Booleano**
  - **Vetorial**
  - **Probabilístico**
- Pressupõem a **independência dos termos** – abordagem **bag of words** (BoW)
  - Vantagem: simplificação
  - Desvantagem: perda de significado

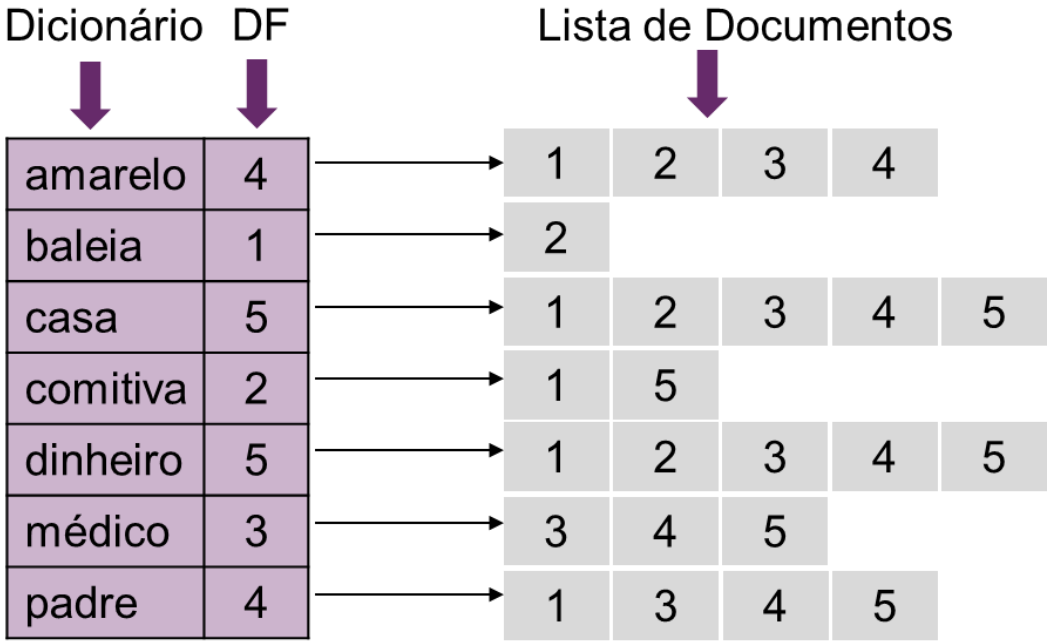
Ex: "João é mais velho do que José" **=** "José é mais velho do que João"

# Modelo Booleano

- Baseado na **teoria de conjuntos** e álgebra booleana
- Os termos da consulta são combinados utilizando operadores boleanos como “**and**”, “**or**”, “**not**”
- Documentos são **conjuntos** de palavras
- **É exato**: o documento satisfaz ou não o critério da busca
- **Não produz um ranking**
- Modelo comercial **mais usado por 3 décadas**

# Modelo Booleano

- Consultas com **AND** = **Interseção** das listas de documentos
- Consultas com **OR** = **União** das listas de documentos



# Modelo Vetorial

- Ideias básicas
  - documentos que contêm **mais vezes os termos** da consulta têm mais chance de estarem relacionados a ela e de serem relevantes (Hans Peter Luhn – anos 50)
  - os **termos** mais **raros** na coleção são mais úteis para diferenciar o conteúdo dos documentos (Karen Spärck Jones – anos 70)
- Com isso, é possível atribuir **escores** aos documentos
- Usando os escores, monta-se um **ranking**

# Escores e Pesos

- Term Frequency $tf_{t,d}$
  - Cada termo t em um documento d recebe um peso que depende do número de ocorrências de t em d
- Document Frequency $df_t$
  - Número de documentos na coleção que contêm o termo $t$
- Inverse Document Frequency $idf_t$
  - Termos raros têm mais importância

# Modelo Vetorial

- Proposto por **Gerard Salton**
- Documentos e consultas são representados como **vetores** em um espaço de ***t* dimensões**
- *t* é o **número de termos distintos** (*i.e.,* o tamanho do dicionário)
- Os termos são os eixos
- As consultas e os documentos são representados no espaço de acordo com a **força da associação** que eles têm com o termo
- Term Frequency times Inverse Document Frequency (**TF-IDF**)

$$w_{i,d_j} = TF_{i,d_j} \times IDF_t = freq_{i,d_j} \times log_{10} \frac{N}{DF_i}$$

# Similaridade entre Termos e Documentos

- A similaridade entre um documento e uma consulta é dada pelo **cosseno** entre seus vetores de pesos TF-IDF
- Cosseno = **produto escalar** normalizado
- O cosseno é máximo (1) se os vetores possuem um ângulo de 0 graus e é mínimo (0) se os vetores formarem um ângulo de 90 graus

$$cosseno(q, d_j) = \frac{\vec{q} \cdot \vec{d}_j}{|\vec{q}| \times |\vec{d}_j|} = \frac{\sum_{i=1}^{t} w_{i,q} \times w_{i,d_j}}{\sqrt{\sum_{i=1}^{t} w_{i,q}^{\ 2}} \times \sqrt{\sum_{i=1}^{t} w_{i,d_j}^{\ 2}}}$$

# Exemplo

- Coleção de documentos
  - $d1$ – O Alienista de Machado de Assis;
  - $d2$ – Vidas Secas de Graciliano Ramos;
  - $d3$ – O Continente (de O Tempo e o Vento) de Érico Veríssimo;
  - $d4$ – Capitães de Areia de Jorge Amado; e
  - $d5$ – Os Sertões de Euclides da Cunha.
- 350.260 tokens
- vocabulário de **32.594** termos (já desconsiderando-se as stop words)

$t$

| Termo | DF | IDF |
|---|---|---|
| amarelo | 4 | 0,0969 |
| baleia | 1 | 0,6990 |
| casa | 5 | 0,0000 |
| comitiva | 2 | 0,3979 |
| dinheiro | 5 | 0,0000 |
| médico | 4 | 0,0969 |
| padre | 4 | 0,0969 |

# Exemplo – Matriz de Incidência dos Termos nos Docs

| Termo | $d_1$ O Alienista | $d_2$ Vidas Secas | $d_3$ O Tempo e O Vento | $d_4$ Capitães de Areia | $d_5$ Os Sertões |
|---|---|---|---|---|---|
| amarelo | 1 | 42 | 6 | 3 | 0 |
| baleia | 0 | 86 | 0 | 0 | 0 |
| casa | 109 | 37 | 247 | 120 | 30 |
| comitiva | 4 | 0 | 0 | 0 | 4 |
| dinheiro | 7 | 9 | 33 | 43 | 3 |
| médico | 18 | 0 | 157 | 7 | 8 |
| padre | 22 | 0 | 120 | 252 | 9 |

# Exemplo – Matriz de Pesos TF-IDF

| Termo | $d_1$ O Alienista | $d_2$ Vidas Secas | $d_3$ O Tempo e O Vento | $d_4$ Capitães de Areia | $d_5$ Os Sertões |
|---|---|---|---|---|---|
| amarelo | 0,0009 | 0,0473 | 0,0024 | 0,0012 | 0,0000 |
| baleia | 0,0000 | 0,6990 | 0,0000 | 0,0000 | 0,0000 |
| casa | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 |
| comitiva | 0,0146 | 0,0000 | 0,0000 | 0,0000 | 0,0549 |
| dinheiro | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 |
| médico | 0,0160 | 0,0000 | 0,0616 | 0,0031 | 0,0258 |
| padre | 0,0196 | 0,0000 | 0,0471 | 0,0969 | 0,0291 |

# Exemplo – Matriz de Pesos TF-IDF

| Termo | $d_1$ O Alienista | $d_2$ Vidas Secas | $d_3$ O Tempo e O Vento | $d_4$ Capitães de Areia | $d_5$ Os Sertões |
|---|---|---|---|---|---|
| amarelo | 0,0009 | 0,0473 | 0,0024 | 0,0012 | 0,0000 |
| baleia | 0,0000 | 0,6990 | 0,0000 | 0,0000 | 0,0000 |
| casa | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 |
| comitiva | 0,0146 | 0,0000 | 0,0000 | 0,0000 | 0,0549 |
| dinheiro | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 |
| médico | 0,0160 | 0,0000 | 0,0616 | 0,0031 | 0,0258 |
| padre | 0,0196 | 0,0000 | 0,0471 | 0,0969 | 0,0291 |

# Exemplo

amarelo
baleia
casa
comitiva
dinheiro
médico
padre

- Consulta “comitiva médico”
- Representação: $q$ = [0; 0; 0; 0, 3979; 0; 0, 0961; 0]
- Calcula-se o **cosseno** entre a consulta e os docs

| $d_1$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ |
|---|---|---|---|---|
| 0.6156 | 0.0000 | 0.1879 | 0.0075 | 0.8765 |

- Ordena-se pelos escores (d5 > d1 > d3 > d4 > d2)
- Na prática, usa-se o índice para recuperar os docs que têm termos em comum com a consulta

# Modelos Probabilísticos

- Há **diversos** modelos probabilísticos
- O mais influente é o **Okapi BM25** (Spärck Jones et al. 2000)

$$BM25(q, d_j) = \sum_{i \in q} log \frac{(r_i + 0,5)/(R - r_i + 0,5)}{(df_i - r_i + 0,5)/(N - df_i + r_i + 0,5)} \times \frac{(k_1 + 1)tf_i}{K + tf_i} \times \frac{(k_2 + 1)qf_i}{k_2 + qf_i}$$

$R$ é o número de documentos sabidamente relevantes para a consulta $q$
$r_i$ é o número de documentos sabidamente relevantes que contêm o termo $i$
$N$ é o número de documentos da coleção; $df_i$ é o número de documentos que contêm o termo $i$
$tf_i$ é o número de ocorrências do termo $i$ no documento $j$
$qf_i$ é o número de ocorrências do termo $i$ na consulta $q$
$K$ , $k1$ e $k2$ são parâmetros que precisam ser definidos empiricamente
$b$ controla a normalização em função do tamanho do documento.

$$K = k_1((1 - b) + b \times \frac{dl}{avdl})$$

# Exemplo

$k1 = 1{,}2$, $k2 = 100$ e $b = 0{,}75$
$avdl = 276$
o tamanho de $d1$ é 161
$K = 1{,}2((1 - 0{,}75) + 0{,}75 \times 161/279) = 0{,}825$

- Consulta "comitiva médico"

$$BM25(q, d_1) = log\frac{5 - 2 + 0{,}5}{2 + 0{,}5} \times \frac{(1{,}2 + 1)4}{0{,}825 + 4} \times \frac{(100 + 1)1}{100 + 1} +$$

$$log\frac{5 - 4 + 0{,}5}{4 + 0{,}5} \times \frac{(1{,}2 + 1)18}{0{,}825 + 18} \times \frac{(100 + 1)1}{100 + 1} =$$

comitiva

médico

$$log\frac{3{,}5}{2{,}5} \times \frac{8{,}8}{4{,}825} \times \frac{101}{101} + log\frac{1{,}5}{4{,}5} \times \frac{39{,}6}{18{,}825} \times \frac{101}{101} =$$

$$log1{,}4 \times 1{,}824 \times 1 + log0{,}33 \times 2{,}1 \times 1 =$$

$$0{,}33 \times 1{,}824 + -1{,}10866 \times 2{,}1 =$$

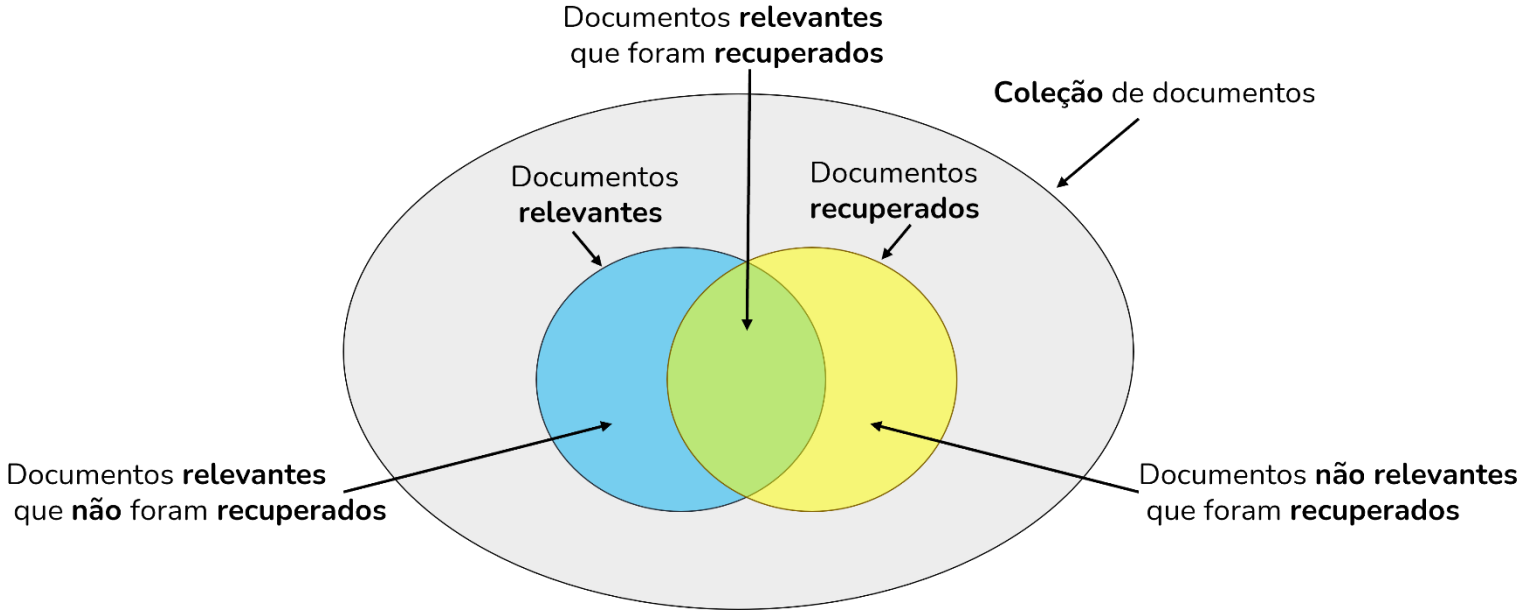
$$0{,}6137 - 2{,}311 = -1{,}6973$$

# Metodologia de Avaliação

# Paradigma de Avaliação de Qualidade

- Paradigma **Cranfield** - Cyril W. Cleverdon anos 1960
- Cálculo de métricas comparando os documentos recuperados para um conjunto de consultas e um **ground truth**
  - Baseado na **noção de relevância**, i.e., uma avaliação que diz se um documento $d_j$ é relevante a uma consulta $q$.
- Inicialmente, vamos tratar a relevância como **binária**, ou seja, o ground truth julga o documento como **relevante** (1) ou não **relevante** (0)

# Métricas Baseadas em Conjuntos




Precisão (precision): $\frac{\text{Número de relevantes recuperados}}{\text{Número total de recuperados}}$

Revocação (recall): $\frac{\text{Número de relevantes recuperados}}{\text{Número total de relevantes}}$

# Exemplo 1

- **20** documentos relevantes em toda a coleção (segundo o ground truth)
- **40** documentos recuperados pela consulta
- **10** relevantes recuperados
- Precisão = 10 ÷ 40 = 0.25 ou **25**%
- Revocação = 10 ÷ 20 = 0.5 ou **50**%

# Exemplo 2

- **20** documentos relevantes em toda a coleção (segundo o ground truth)
- **20** documentos recuperados pela consulta
- **10** relevantes recuperados
- Precisão = 10 ÷ 40 = 0.25 ou **50**%
- Revocação = 10 ÷ 20 = 0.5 ou **50**%

# Exemplo 3

- **20** documentos relevantes em toda a coleção
- **1** documento recuperado pela consulta
- **1** relevante recuperado
- Precisão = 1 ÷ 1 = 1 ou **100**%
- Revocação = 1 ÷ 20 = 0.05 ou **5**%

Conclusão: Altos índices de precisão geralmente são acompanhados por baixos níveis de revocação e vice-versa.

# F-measure

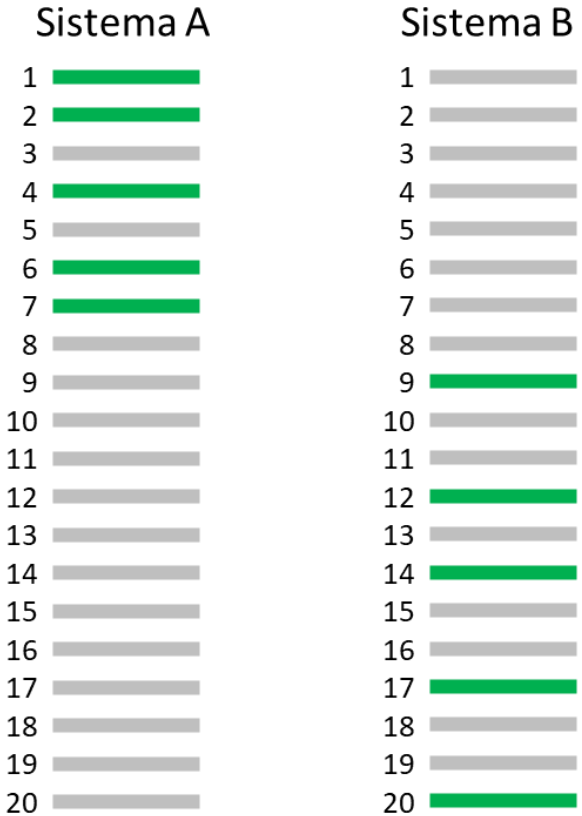
- Combina Precisão e Revocação em uma só métrica

$$F = \frac{(\beta^2 + 1) \times P \times R}{(\beta^2 \times P) + R}$$

Se β = 1, a mesma ênfase é dada a P e a R **(é a F1)** $F1 = \frac{2 \times P \times R}{P + R}$

Se β = 2, Revocação é enfatizada 2 vezes em relação à Precisão
Se β = 0.5, enfatiza a Precisão 2 vezes em relação à Revocação

# Métricas para avaliar rankings

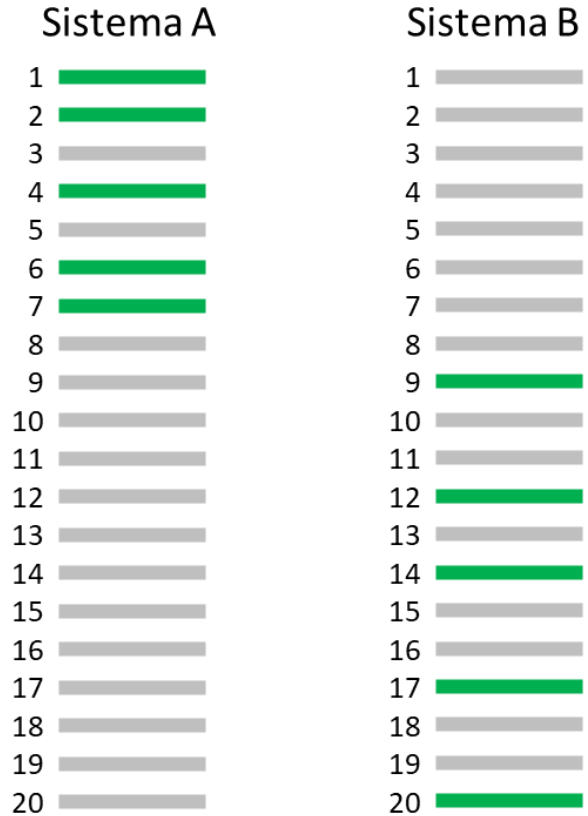



Qual sistema foi melhor? (considerando **7 relevantes**)

- Sistema A
  - Precisão = 5/20 = **0,25**
  - Revocação = 5/7 = **0,71**
  - F1 = (2*0,25*0,71)/(0,25+0,71) = **0,37**
- Sistema B
  - Precisão = 5/20 = **0,25**
  - Revocação = 5/7 = **0,71**
  - F1 = (2*0,25*0,71)/(0,25+0,71) = **0,37**

Problema: Métricas baseadas em conjuntos **não servem para avaliar rankings** !

# Precisão Média (average precision - AP) e Mean Average Precision (MAP)




Solução: calcular **quantos** documentos relevantes foram recuperados e o **quão próximos** estão do **topo** do ranking.

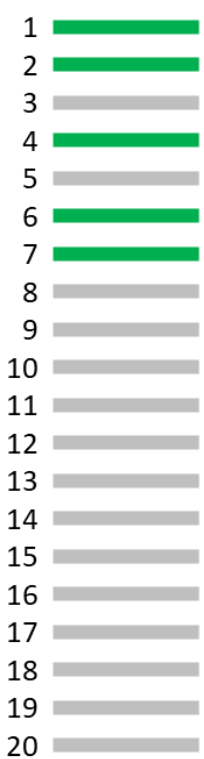
$$AP = \frac{\sum_{k=1}^{n} P(k) \times rel(k)}{\#\ relevantes}$$

$rel(k)$ = 1 se o doc na $k$-ésima posição do ranking for relevante e 0 se não for.
Os documentos relevantes que são foram recuperados serão penalizados com $P$ = 0

$$MAP = \frac{\sum_{q=1}^{|Q|} AP_q}{|Q|}$$

# Precisão Média

Sistema A




1/1 = 1

2/2 = 1

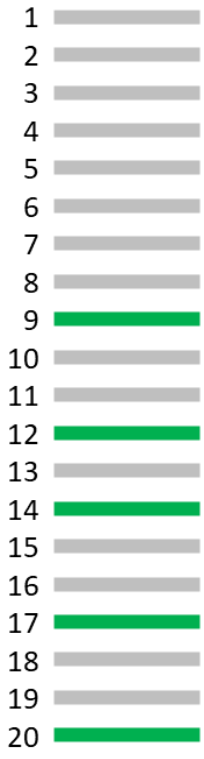
3/4 = 0,75

4/6 = 0,66

5/7 = 0,85

$$AP = \frac{1 + 1 + 0{,}75 + 0{,}66 + 0{,}85}{7} = \mathbf{0{,}61}$$

Sistema B




$$AP = \frac{0{,}11 + 0{,}16 + 0{,}21 + 0{,}24 + 0{,}25}{7} = \mathbf{0{,}14}$$

1/9 = 0,11

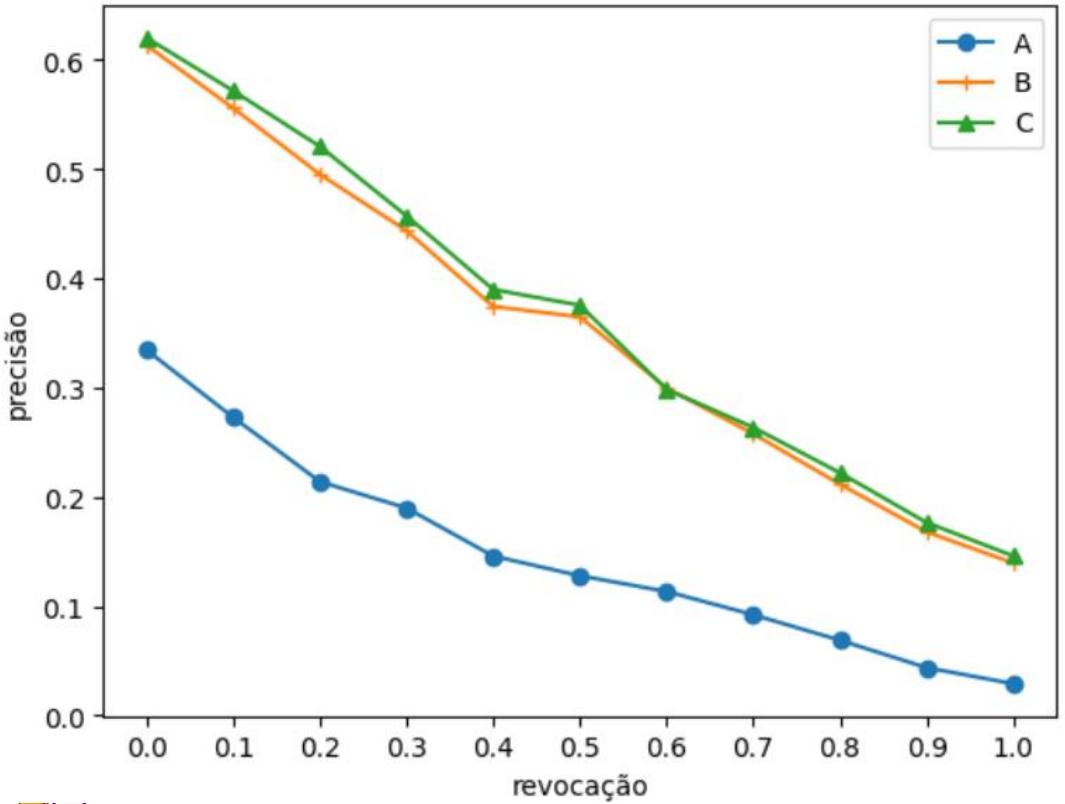
2/12 = 0,16

3/14 = 0,21

4/17 = 0,24

5/20 = 0,25

# Curvas de Revocação e Precisão




- Precisão em **11 pontos** de revocação
- **Regra de interpolação** "a precisão interpolada para um nível de revocação *j* é o maior valor de precisão para qualquer nível de revocação maior ou igual a *j*
- Curvas **monotonicamente decrescentes**
- **Quanto maior** a área sob a curva, **melhor**

# Outras métricas

- **Precisão** em diferentes pontos do ranking (**P@1, P@5, P@10**)
  - Útil para quando não se tem julgamentos de relevância completos

- **Mean Reciprocal Rank**
  - Inverso da posição do primeiro documento relevante retornado

# Ganho Acumulado Descontado Normalizado (NDCG)

- Proposto por JÄRVELIN, K.; KEKÄLÄINEN, J. Cumulated gain-based evaluation of IR techniques. ACM TOIS, v. 20, n. 4, p. 422–446, 2002
- Usado para avaliar rankings quando há **diversos níveis de relevância**

  3 = altamente relevante, 2 = moderadamente relevante,

  1 = marginalmente, relevante, 0 = não relevante
- Para entender o NDCG, primeiro temos que entender o **DCG** e o **IDCG**
- $DCG = \sum_{i=1}^{n} \frac{2^{rel_i} - 1}{log_2(i+1)}$

  $n$ é o número de documentos recuperados

  $i$ é a posição do documento no ranking

  $rel_i$ é o nível de relevância do documento
- **IDCG é o DGC ideal** (i.e., o DCG da recuperação perfeita)
- $NDCG = \frac{DCG}{IDCG}$

# Coleções de Teste

- Compostas por:
  - **Corpus** de Documentos
  - **Consultas** (chamadas de tópicos)
  - **Julgamentos de relevância** – quais documentos são relevantes para cada consulta – i.e., *ground truth*

**Alto custo de obtenção!**

# Coleções de Teste em Português

- **CHAVE** - Santos, D. e Rocha, P. (2004). The key to the first CLEF with Portuguese: Topics, questions and answers in CHAVE. CLEF, pages 821–832.
  - Notícias da **Folha de São Paulo e O Público** (~150K docs)
  - 100 consultas + julgamentos de relevância **humanos**
- **REGIS** - Oliveira, L. L, Romeu, e V. P. Moreira. "REGIS: A test collection for geoscientific documents in Portuguese." SIGIR. 2021
  - Teses, dissertações e artigos no **domínio geocientífico**
  - 34 consultas + julgamentos de relevância **humanos** (21K docs)
- **MS MARCO** - Bonifacio, L. H. et al. mMARCO: A Multilingual Version of MS MARCO Passage Ranking Dataset.
  - 8,8 milhões de documentos (passagens) e 6980 consultas do log do Bing + respostas **humanas**
  - **Traduzida** do inglês
- **Quati -** Bueno, M., de Oliveira, E. S., Nogueira, R., Lotufo, R. A., and Pereira, J. A. (2024). Quati: A Brazilian Portuguese information retrieval dataset from native speakers. arXiv preprint arXiv:2404.06976.
  - **Passagens** de mil caracteres extraídas de **páginas web** (10 milhões)
  - 200 consultas + julgamentos de relevância pelo **GPT4**

# Exemplos de Consultas

(a) CHAVE

<top>
<num> C267 </num>
<PT-title> Melhor Filme Estrangeiro </PT-title>
<PT-desc> Quais foram os filmes candidatos ao Oscar de Melhor Filme Estrangeiro? </PT-desc>
<PT-narr> Documentos relevantes devem indicar o título e nacionalidade dos filmes indicados para o Óscar na categoria de Melhor Filme em Língua Estrangeira. </PT-narr>
</top>

(b) REGIS

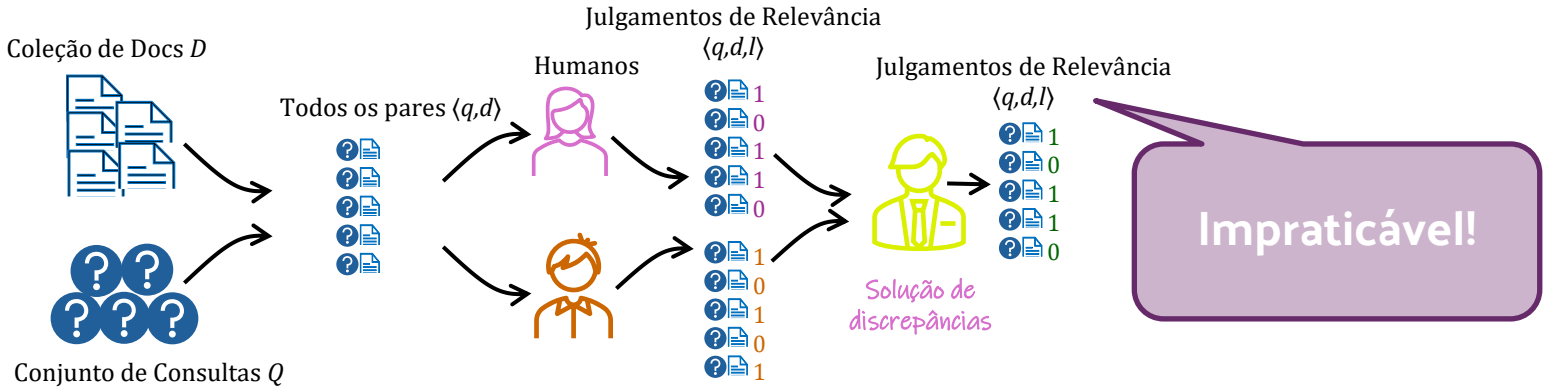
<top>
<num>Q5</num>
<title>Permeabilidade em Marlim</title>
<desc>Informações sobre o campo de Marlim, mas não dos campos de “Marlim Sul” ou de “Marlim Leste”.</desc>
<narr>Interessam documentos que contenham informação sobre a permeabilidade das rochas do campo de Marlim. Principalmente, interessam informações quantitativas (como dados de permeabilidade expressos em Darcies ou milidarcies).</narr>
</top>

(c) MS MARCO

| | |
|---|---|
| 350,732 | como limpar copos de melamina manchados |
| 1,067,511 | por que meu ligamento colateral lateral dói |
| 906,071 | o que adicionar à sua água para ser mais saudável |

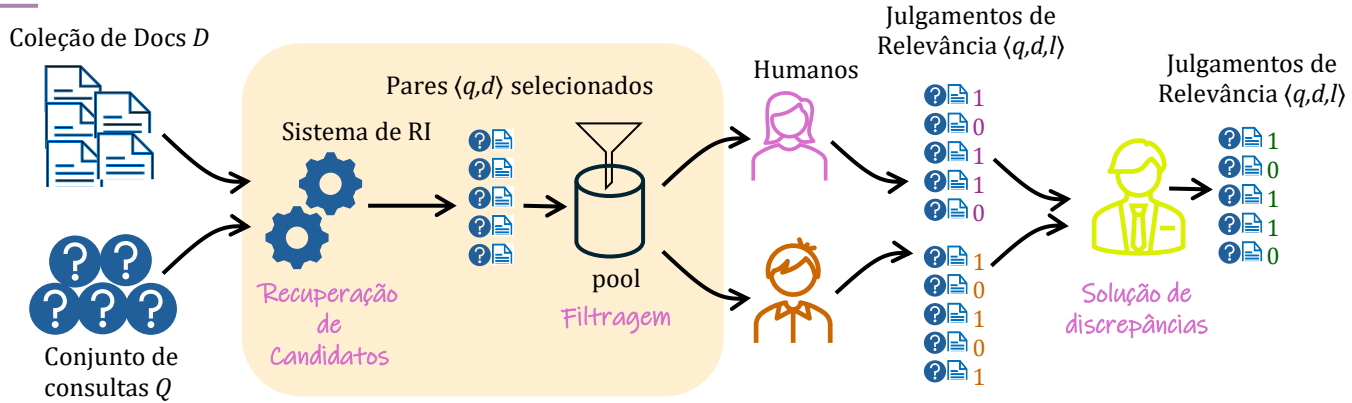
# Como se cria uma coleção de teste?

- Primeira coleção de teste desenvolvida no Cranfield College of Aeronautics (Inglaterra) no final dos anos 60
- 1400 abstracts de documentos sobre aeronáutica
- 225 consultas
- Cada documento julgado por mais de um avaliador – mais de meio milhão de julgamentos de relevância (225 × 1400 × 2 = 630 000)



# Método Pooling

Spärck Jones, Karen. 1975. Report on the need for and provision of an 'ideal' information retrieval test collection. *Computer Laboratory*




- Repositórios (**pools**) são criados a partir dos **resultados de diferentes sistemas**
- Apenas os **top n documentos** de cada sistema são adicionados ao pool
- Os avaliadores julgam **somente** documentos que estão no **pool**
- Os documentos **não julgados são considerados irrelevantes**
- Como o pool é construído com documentos retornados por vários sistemas, existe uma **probabilidade** de que todos os documentos relevantes sejam encontrados

# O quão confiável é o paradigma de avaliação?

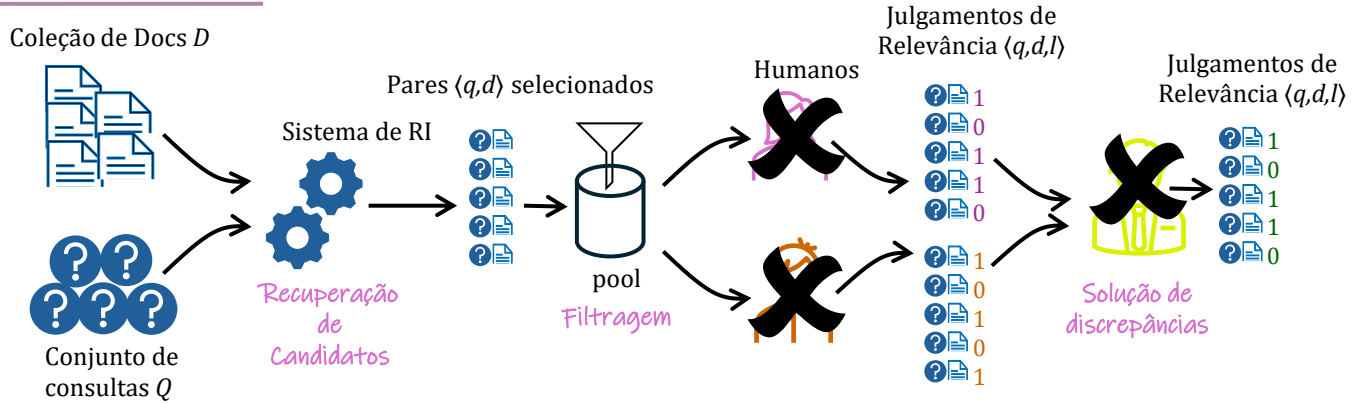
- Diversos trabalhos avaliam a confiabilidade do paradigma experimental de IR
  - Voorhees, E. M. (2000). Variations in relevance judgments and the measurement of retrieval effectiveness. Information Processing & Management, 36(5):697–716.

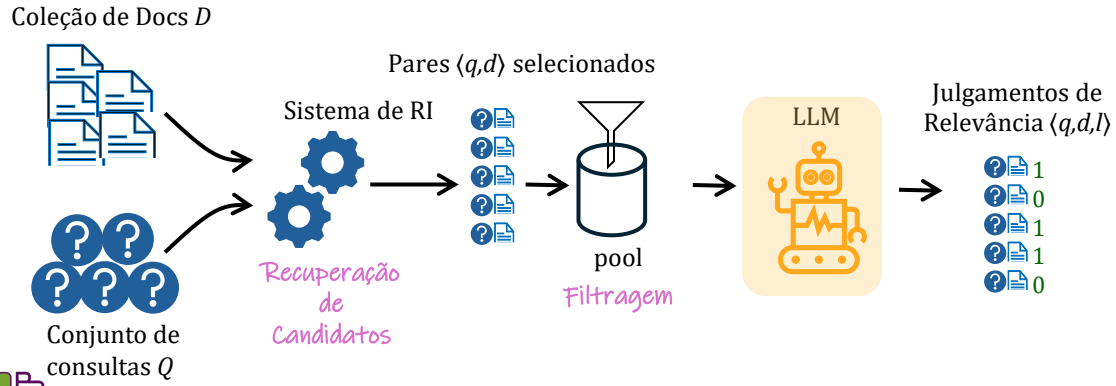
Table 5
Kendall correlation between rankings produced at different qrels sets on the TREC-6 data. Also shown is the number of swaps the correlation represents given that there are 74 systems represented in the rankings (and thus 2701 possible swaps)

| qrels pair | Correlation | Swaps |
|---|---|---|
| NIST & Waterloo | 0.8956 | 141 |
| NIST & union | 0.9556 | 60 |
| NIST & intersection | 0.8904 | 148 |
| Waterloo & union | 0.8956 | 141 |
| Waterloo & intersection | 0.9237 | 103 |
| Union & intersection | 0.8652 | 182 |
| Mean | 0.9044 | 129 |

- Conclusão – o valor do escore pode mudar com os diferentes conjuntos de julgamentos de relevância, mas o **ranking comparativo dos sistemas se mantém**

# Tendência Atual – julgamentos por LLM





Faggioli, G., Dietz, L., Clarke, C. L., Demartini, G., Hagen, M., Hauff, C., Kando, N., Kanoulas, E., Potthast, M., Stein, B., et al. (2023). Perspectives on large language models for relevance judgment. ICTIR

Thomas, P., Spielman, S., Craswell, N., and Mitra, B. (2023). Large language models can accurately predict searcher preferences. arXiv

Bueno, M., de Oliveira, E. S., Nogueira, R., Lotufo, R. A., and Pereira, J. A. (2024) Quati: A Brazilian Portuguese Information Retrieval Dataset from Native Speakers.

Soviero, B., Kuhn, D., Salle, A., and Moreira, V. P. (2024). ChatGPT goes shopping: LLMs can predict relevance in ecommerce search. ECIR

Bencke, L; Paula, F.; Santos, B. Moreira, V. P. (2024) Can we trust LLMs as relevance judges?. SBBD

# Questões em Aberto

- **Ainda precisamos de humanos** para julgar relevância?

- Quem sabe em **domínios específicos**, idiomas com **poucos recursos**...

# Considerações Finais

# Considerações Finais

- Fizemos uma **introdução à** área de **Recuperação de Informação**
- Falamos dos **Modelos Clássicos** e do **paradigma de avaliação**
- **O que ficou de fora**? (um universo inteiro)
  - Técnicas de melhoria de qualidade
  - Learning to Rank
  - Abordagens baseadas em vetores densos
  - Reranqueamento usando modelos neurais
  - RI na Web

# Obrigada!