Text Mining

Departamento de Engenharia Informática (DEI/ISEP)
Fátima Rodrigues
mfc@isep.ipp.pt

Text Mining

- É o processo de extração de conhecimento interessante, útil, novo e relevante a partir de dados não estruturados ou semiestruturados: **textos**

- 80% da informação nas empresas é não estruturada, tais como documentos, manuais, livros, mensagens de correio eletrónico, apresentações, Web, notícias, etc.
  - ↪ Daí a relevância do Text Mining

- A metodologia Text Mining é análoga à do processo KDD

Etapas do Processo de Text Mining

Recuperação de Informação

Localização e recuperação dos textos relevantes de acordo com os objetivos de descoberta

Extração de Informação

Identificação dos itens (características, palavras, frases, documentos) relevantes

Mineração

Aplicação de um ou vários algoritmos de mineração para identificar padrões e relacionamentos entre **palavras**, **frases**, ou até entre vários **documentos**

Interpretação

Interpretação e aplicação do conhecimento extraído

Qual a sua Importância ?

***Text Mining* auxilia:**

- na **pesquisa de informações específicas**, agilizando processos com uso de inteligência
- na **análise qualitativa e quantitativa** de grandes volumes de textos, ajudando a compreender melhor o conhecimento disponível em documentos
- a **encontrar conhecimento novo, implícito em textos**, apresentando resultados mais afinados com as reais necessidades de pessoas e organizações

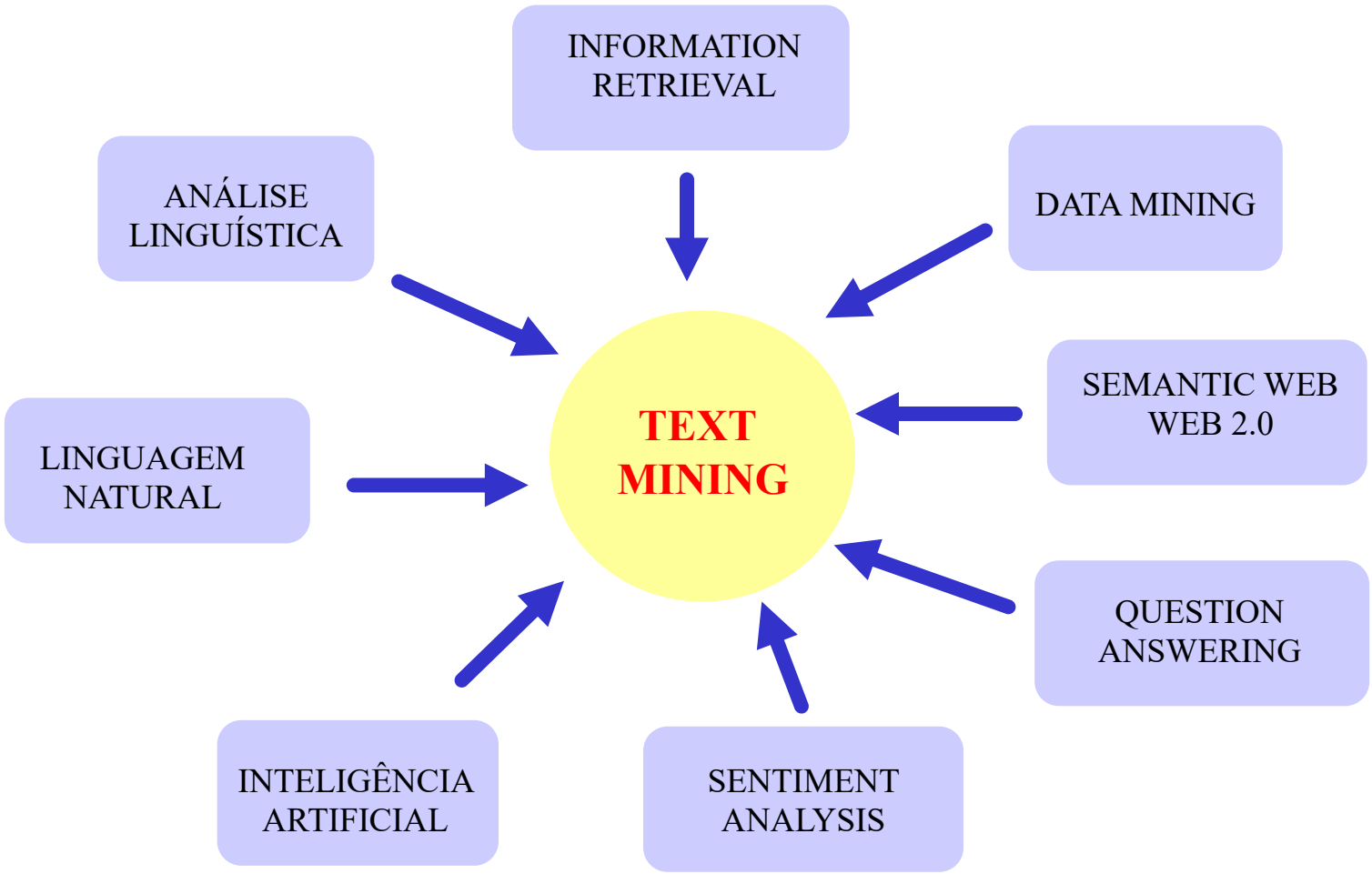
Que tipos de textos ?

- E-mails
- Textos livres resultantes de pesquisas
- Arquivos electrónicos (txt, doc, ...)
- Páginas Web
- Campos textuais (memos) em Bases de Dados
- Documentos electrónicos digitalizados
- Livrarias digitais

Aplicações mais comuns de Text Mining

- Gestão de correio eletrónico
- Gestão de documentos
- *Automatizar Help desk*
- Resumir documentos
- Motor de pesquisa
- Análise de sentimento de avaliações textuais
- Agrupar documentos...

Onde se enquadra o Text Mining ?



Níveis de Processamento de Texto

As técnicas de processamento de texto podem ocorrer ao nível:

- Palavra
- Frase
- Coleções de Documentos

Processamento de Documentos ao nível das Palavras

Representação através de Palavras

- Um documento é descrito através de um **conjunto representativo de palavras-chave**
- As palavras-chave constituem **pontos de acesso** ao documento
- Quando uma **palavra** é identificada como importante ela é ***mapeada*** para um **termo** que caracteriza o documento
- Diversas palavras podem ser mapeadas para um único **termo**

Propriedades das Palavras

Dificuldades

- **Palavras sinónimas**: diferentes formas o mesmo significado, ex. cantor, vocalista, refeição, almoço, …
- **Palavras homónimas**: mesma forma, significados diferentes, ex. canto (ângulo), canto (verbo cantar)
- **Polissemia**: mesma forma, significado relacionado, ex. banco alimentar, banco urgências, banco sangue
- Proximidade entre palavras é difícil medir com precisão, ex: distância entre *data mining* e *data analysis*

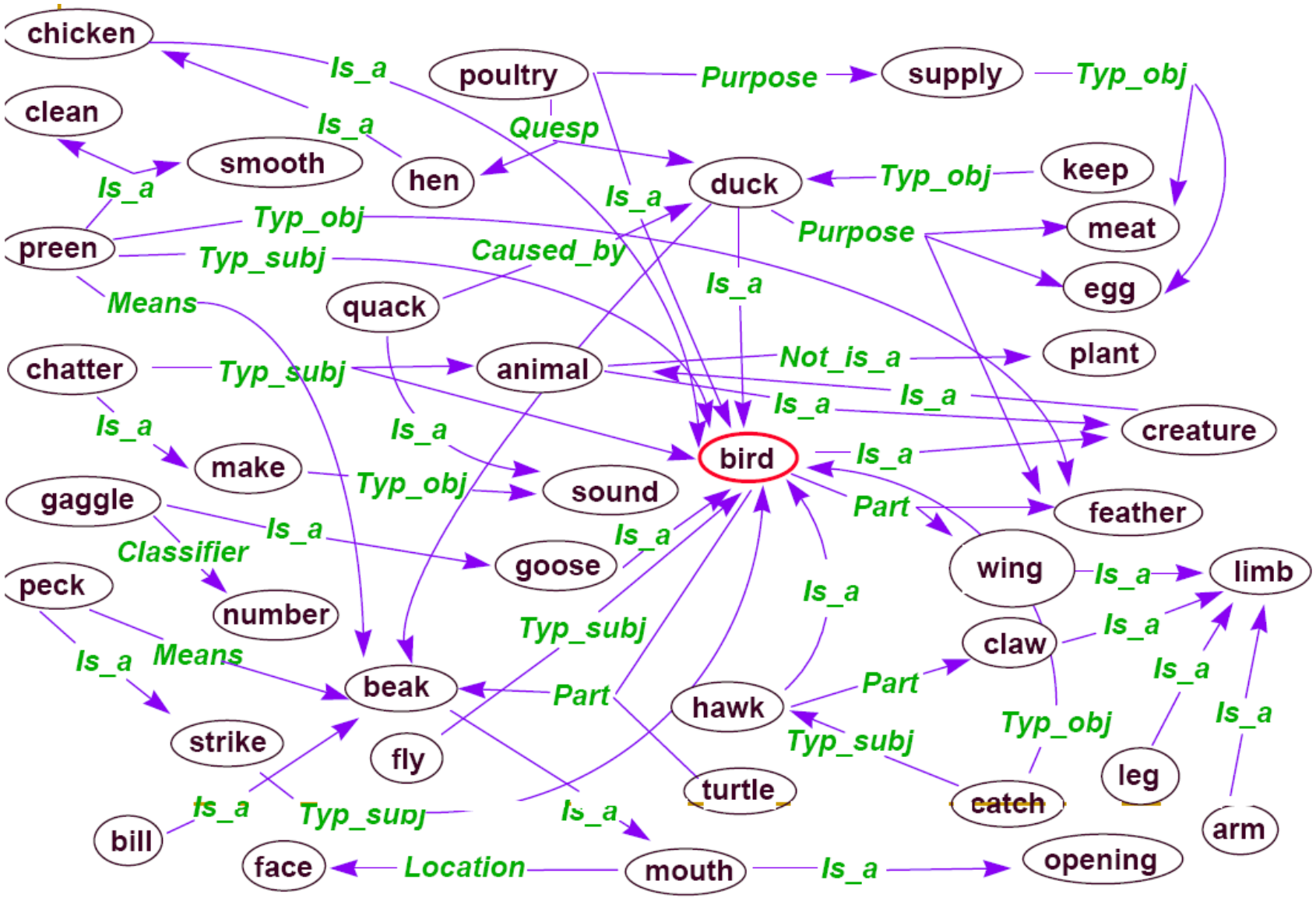
Taxonomias/*Thesaurus*

- Thesaurus é uma ferramenta similar a um dicionário, só que, ao invés de informar o significado das palavras, tem como principal função **devolver sinónimos de palavras**, informa o relacionamento entre palavras, tais como relações hierárquicas entre palavras:
  - Palavras do geral-para-específico
  - Palavras do específico-para-geral

- O Thesaurus mais desenvolvido é o WordNet que existe em várias línguas: inglês, alemão, espanhol, italiano, francês, …(EuroWordNet) → http://www.illc.uva.nl/EuroWordNet/

- WordNet contém 4 bases de dados: nomes, verbos, adjetivos e advérbios
- Cada base de dados contém vários sinónimos para cada entrada:
  - musician, instrumentalist, player
  - person, individual, someone
  - life form, organism, being

WordNet – Excerto do Grafo



MWN.PT - MultiWordNet do Português

- Ontologia lexical que **inclui mais de 17200 conceitos/synsets validados manualmente** e conectados entre si pelas relações de hipo-/hiperonímia

- Os conceitos englobam mais de 21000 acepções (instâncias) e mais de 16000 lemas (tipos) das variantes Europeia e Americana do Português, encontrando-se alinhados com os conceitos equivalentes da wordnet de Princeton para o Inglês, e transitivamente com os das MultiWordNets do Italiano, Castelhano, Hebraico, Romeno e Latim

- Desenvolvido pelo *Natural Language and Speech Group,* Departamento de Informática da Faculdade de Ciências da UNL, pode ser consultado em http://mwnpt.di.fc.ul.pt/

Conjunto de Palavras Representativo

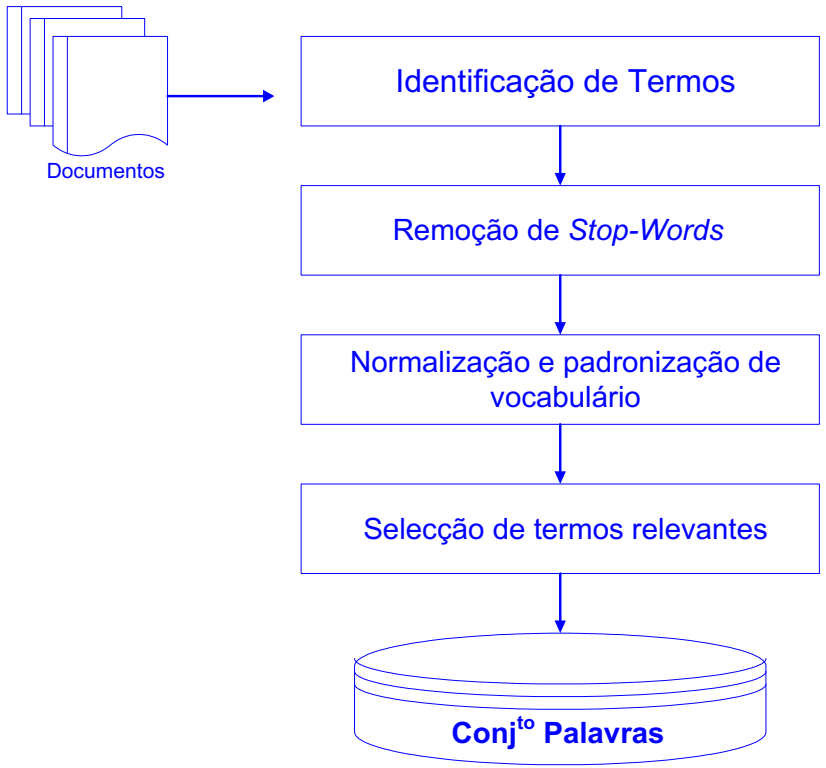
O conjunto de palavras representativo de um documento deve ser avaliado pelos factores:

- **Exaustividade** mede a quantidade de assuntos distintos que o conjunto de palavras é capaz de reconhecer → quanto **maior a exaustividade**, **maior é a abrangência**, mas **menor é a precisão**

- **Especificidade** é a capacidade dos termos do conjunto descreverem correctamente os tópicos de um documento → **quanto mais específico, menor é a abrangência, maior é a precisão**

Fases Processo Criação Conj^{to} Palavras

- Identificação de termos (simples ou compostos)
- Remoção de *stopwords* (palavras irrelevantes)
- Normalização morfológica (*stemming*)
- Seleção dos termos



Identificação de Termos

Aplicação de um *parser* (analisador léxico) que identifica as palavras presentes nos documentos, ignorando os símbolos e caracteres de controle ou de formatação:

- Identificação de nomes próprios (palavras iniciadas de forma maiúscula) ou através de um dicionário próprio
- passagem de todos os caracteres para a forma minúscula (exceto os nomes próprios)
- substituição de múltiplos espaços e tabulações por um único espaço
- padronização de datas e números
- eliminação de hífenes, sinais de pontuação, …

Identificação de Termos

Identificação de Termos Compostos

São conceitos que só podem ser descritos pela utilização de duas ou mais palavras adjacentes, ex. processo judicial

Existem basicamente duas formas de identificação:

- identificação de termos que coocorrem com frequência numa coleção de documentos, o sistema apresenta as expressões identificadas e solicita as corretas
- utilização de um dicionário de expressões que indica quais as palavras que devem ser combinadas

Identificação de Termos Válidos

... à¡±• á `>`˜ ` þÿ Na maioria das vezes os documentos devolvidos pelas ferramentas de

`>` recuperação de informação

`>` envolvem um contexto mais

amplo, fazendo com que o utilizador

tenha que minerar, ou seja, especificar ou

filtrar os documentos (o que exige tempo e conhecimento) a fim de obter a informação que ele realmente necessita `˜ ` ...

Documento original

na maioria das vezes os documentos devolvidos pelas ferramentas de recuperação de informação envolvem um contexto mais amplo fazendo com que o utilizador tenha que minerar ou seja especificar ou filtrar os documentos o que exige tempo e conhecimento a fim de obter a informação que ele realmente necessita

Documento normalizado

Remoção de "*Stopwords*"

- **Stopwords** são palavras sem valor linguístico

- Servem apenas para conectar as frases, a sua finalidade é auxiliar a estruturação da linguagem, tais como: artigos, preposições, pronomes, conjunções, alguns adjectivos e advérbios, e derivações dos verbos de ligação: ter, estar, ser e haver (ocorrem nas frases com uma frequência muito elevada).
- Representam cerca de **20%-30%** das palavras nos documentos

- Outras palavras cuja frequência na coleção de documentos é muito alta, não devem constar na estrutura do índice - não são capazes de discriminar documentos, ex. database → stopword em proceedings sobre database systems

Stopwords **são dependentes da Língua**

- **Inglês**: A, ABOUT, ABOVE, ACROSS, AFTER, AGAIN, AGAINST, ALL, ALMOST, ALONE, ALONG, ALREADY, ...
- **Português**: DE, PARA, O, A, ATÉ, TUDO, TODOS, …, EMBORA, APENAS, CONTUDO, …, PORÉM

Remoção de "*Stopwords*"

... ~~na~~ maioria ~~das~~ vezes ~~os~~ documentos devolvidos ~~pelas~~ ferramentas ~~de~~ recuperação de informação envolvem ~~um~~ contexto ~~mais~~ amplo fazendo ~~com que o~~ utilizador tenha ~~que~~ minerar ~~ou seja~~ especificar ~~ou~~ filtrar ~~os~~ documentos ~~o que~~ exige tempo ~~e~~ conhecimento ~~a fim de~~ obter ~~a~~ informação ~~que ele~~ realmente necessita ...

Documento normalizado

maioria vezes documentos devolvidos ferramentas recuperação de informação envolvem contexto amplo fazendo utilizador tenha minerar especificar filtrar documentos exige tempo conhecimento obter informação realmente necessita

Documento sem stopwords

Normalização Morfológica

As variações morfológicas são eliminadas através da identificação do **radical da palavra.** Os prefixos e os sufixos são retirados e os radicais resultantes são adicionados à estrutura de índice

As características de género, número e grau das palavras são eliminadas:

- várias palavras são mapeadas para um único termo
- **aumenta a abrangência** das consultas
- **diminui a precisão**, impossibilidade de fazer pesquisas por palavras específicas
- pode reduzir o tamanho de um índice até **50%**
- Potencialmente mais poderoso, mas menos eficiente

Normalização Morfológica

Existem duas formas possíveis de normalização morfológica:

- **stemming**

  Pode reduzir a palavra a uma outra gramaticalmente incorreta, porém ainda com valor para a análise. Os algoritmos de stemming têm um conjunto de regras para decidir como fazer os cortes

- **Lemmatization**

  Também reduz a palavra ao seu radical, retirando todas as inflexões e chegando ao *lemma*. Porém, esta redução resulta sempre numa palavra que existe na gramática. Nesta técnica, a classe gramatical da palavra é levada em consideração para fazer a redução

Original	Stemming	lemmatization
amigos	amig	amigo
amigas	amig	amigo
amizade	amizad	amizade
carreiras	carr	carreira

Normalização Morfológica

Identificação do radical das palavras

- definição de uma lista de prefixos/sufixos mais encontrados no vocabulário de uma língua
- utilização de dicionário morfológico

Exemplos:

proposal → propos

european → europ

Pode originar dois tipos de erro:

- **Overstemming**

  gramática → grama

- **Understemming**

  carreira → carr

Normalização Morfológica

O dicionário de *stemming* em inglês mais usado :

Porter Stemmer: http://www.tartarus.org/~martin/PorterStemmer/

Exemplo de regras usadas no Porter Stemmer para *stemming* de palavras:

- ATIONAL -> ATE relational -> relate
- TIONAL -> TION conditional -> condition
- ENCI -> ENCE valenci -> valence
- ANCI -> ANCE hesitanci -> hesitance
- IZER -> IZE digitizer -> digitize
- ABLI -> ABLE conformabli -> conformable
- ALLI -> AL radicalli -> radical
- ENTLI -> ENT differentli -> different
- ELI -> E vileli -> vile
- OUSLI -> OUS analogousli -> analogous

Normalização Morfológica

maioria vezes documentos devolvidos ferramentas recuperação de informação envolvem contexto amplo fazendo utilizador tenha minerar especificar filtrar documentos exige tempo conhecimento obter informação realmente necessita

Documento sem stopwords

maioria vez documento devolve ferramenta recuperação de informação envolve contexto amplo faz utilizador ter minerar especificar filtrar documento exige tempo conhecimento obter informação real necessita

Documento Normalizado

Seleção dos Termos

- Diferentes palavras-chave apresentam diferente relevância, este efeito é capturado através da atribuição de pesos numéricos a cada palavra-chave

Frequência das palavras nos textos:

- pequeno número de palavras frequentes
- elevado número de palavras com baixa frequência

Frequência do Termo: tf

- A **frequência do termo** t num documento d $\mathbf{tf_{t,d}}$ é definida como o número de vezes que t ocorre em d

- Como calcular a relevância de um documento face a uma query? Usando a frequência **tf** dos termos do documento para calcular pontuações
  - Um documento com 10 ocorrências de um termo é mais relevante do que um documento com uma ocorrência do mesmo termo
  - Mas não é 10 vezes mais relevante

- **Relevância** não aumenta proporcionalmente com a frequência do termo

Frequência dos termos num Documento

- **Termos raros** são mais informativos do que termos frequentes

- Considere-se um termo numa consulta que é raro numa coleção de documentos
  Um documento que contém este termo é muito provável que seja relevante para a consulta → **termos raros devem ter pesos altos**

- **Termos frequentes** são menos informativos do que termos raros

- Um termo frequente numa coleção de documentos não é um indicador seguro de relevância → **termos frequentes** devem ter pesos positivos, mas **pesos mais baixos do que termos raros**

Ponderação TF-IDF

- TF (*term frequency*) representa o número de vezes que o termo aparece no documento versus o número total de termos do documento

$$TF = \frac{N_t}{N_d}$$

- IDF (inverse *document frequency*) representa o número de documentos que contêm o termo no corpus

$$IDF = \log_{10}\left(\frac{N_{dc}}{N_{dt}}\right)$$

- TFI-DF é a pontuação calculada como

$$TF\text{-}IDF = TF \times IDF$$

A ponderação tf-idf:

- aumenta com o número de ocorrências do termo num documento
- aumenta com a raridade do termo na coleção de documentos

Ponderação TF-IDF: Exemplo

Considere um documento que contém 10.000 palavras em que a palavra "Palestine" aparece 300 vezes

Frequência do termo "Palestine"

$$tf = 300/10000 = 0,03$$

Supondo que a BD contém 1000 documentos, e destes apenas em 10 documentos ocorre a palavra "Palestine"

$$idf = \log_{10} (1000/10) = 2$$

A pontuação tf-idf :

$$tf\text{-}idf = 0,03 \times 2 = 0,06 \ (6\%)$$

Cálculo de Relevância da Palavra

O cálculo da relevância das palavras pode também envolver:

- **Análise estrutural** do documento (títulos, resumos, …)
- **Posição sintáctica** da palavra (substantivos e complementos)
- **Análise semântica** baseia-se no princípio de que as partes mais relevantes de um documento já estão de alguma forma demarcadas por estruturas de formatação específicas para isso - **marcas *HTML* e *XML***

Representação do Documento

Conjunto de palavras
Saco-Palavras (*Bag-of-words)*

maioria vez documento devolve ferramenta recuperação de informação envolve contexto amplo faz utilizador ter minerar especificar filtrar documento exige tempo conhecimento obter informação real necessita

Documento Normalizado

maioria [peso]
vez [peso]
documento …
devolve
ferramenta
recuperação de informação
envolve
contexto
amplo
faz
utilizador
ter
minerar
especificar
filtrar
exige
tempo
conhecimento
obter
Informação
real
necessita ...

Perde-se toda a informação intrínseca à ordem das palavras
Limita o contexto!

Representação Documento

TRUMP MAKES BID FOR CONTROL OF RESORTS Casino owner and real estate Donald Trump has offered to acquire all Class B common shares of Resorts International Inc, a spokesman for Trump said. The estate of late Resorts chairman James M. Crosby owns 340,783 of the 752,297 Class B shares. Resorts also has about 6,432,000 Class A common shares outstanding. Each Class B share has 100 times the voting power of a Class A share, giving the Class B stock about 93 pct of Resorts' voting power.

[RESORTS:0.624] [CLASS:0.487] [TRUMP:0.367] [VOTING:0.171]
[ESTATE:0.166] [POWER:0.134] [CROSBY:0.134] [CASINO:0.119]
[DEVELOPER:0.118] [SHARES:0.117] [OWNER:0.102]
[DONALD:0.097] [COMMON:0.093] [GIVING:0.081] [OWNS:0.080]
[MAKES:0.078] [TIMES:0.075] [SHARE:0.072] [JAMES:0.070]
[REAL:0.068] [CONTROL:0.065] [ACQUIRE:0.064]
[OFFERED:0.063] [BID:0.063] [LATE:0.062] [OUTSTANDING:0.056]
[SPOKESMAN:0.049] [CHAIRMAN:0.049] [INTERNATIONAL:0.041]
[STOCK:0.035] [YORK:0.035] [PCT:0.022] [MARCH:0.011]

Conjunto de palavras
Vector elevada dimensionalidade, esparso

Seleção dos Termos

Seleção dos *n* termos mais relevantes (truncagem)

- estabelece-se um número **máximo** de características a serem utilizadas para caracterizar um documento

Filtragem baseada na **frequência de termo (term frequency)**

- consiste em eliminar todos os termos abaixo de um limiar estabelecido
- é a técnica mais simples de redução de dimensões

Outras técnicas baseiam a filtragem de termos:

- Ganho de Informação
- Informação Mútua
- Estatística χ^2 (chi-square)

Processamento de Documentos ao nível das Frases

Representação de Documentos: Nível Frase

A representação de textos através de frases tem a vantagem de permitir **identificar com maior precisão o sentido do texto**

Os textos são representados através de frases:

- com palavras contíguas frequentes
- com palavras não contíguas frequentes

A maneira mais simples de gerar frases é através da pesquisa de **n-grams frequentes**

- N-Gram é uma sequência de n palavras consecutivas:
  "Aprendizagem Máquina" é 2-gram
- Os N-grams são extraídos através de algoritmos de programação dinâmica

Análise Sintáctica de Frases

A análise sintática de frases permite diferenciar palavras e atribuir-lhes grau de importância diferente:

- Este tipo de análise é explorado pela área **Linguagem Natural** – atribui **tags** às palavras de acordo com a sua função na frase: nome, predicado, verbo, …

  Exemplo, Text Tagging:

  (*lisboa,***<np>**) (*pombalino,***<adj>**) (*rua*,**<nc>**,**<verb>**) (*larga*,**<nc>**,**<adj>**,**<v>**) (*perpendicular*,**<a_nc>**) (*sistema*,**<nc>**) (*esgoto*,**<nc>**,**<v>**) (*casa*,**<nc>**,**<v>**) (*possuir*,**<v>**) (*estrutura*,**<nc>**,**<v>**) (*gaiola*,**<nc>**) (*terreiro*,**<a_nc>**) (*paço*,**<nc>**) (*passar*,**<v>**) (*chamar*,**<v>**) (*praça*,**<nc>**) **(**comércio,**<nc>**)

Modelos de Recuperação de Informação

São modelos conceptuais ou abordagens genéricas para a recuperação de informação de natureza não-estruturada (normalmente de texto), dentro de grandes coleções que satisfaz uma query

- podem ser utilizados em qualquer tipo de documento
- basta modificar o tipo de atributo – palavras – pelo tipo de atributo adequado ao tipo de documento em questão

Taxonomia de Modelos

- Booleano
- Espaço Vectorial
- Probabilístico
- …

Modelo Booleano

- Considera os documentos como sendo conjuntos de palavras
- Manipula e descreve esses conjuntos através dos operadores booleanos:*and*, *or* e *not*
- As expressões booleanas são flexíveis permitem unir conjuntos, descrever intersecções e retirar partes de um conjunto

Exemplo: Informações sobre onde realizar compras na Internet

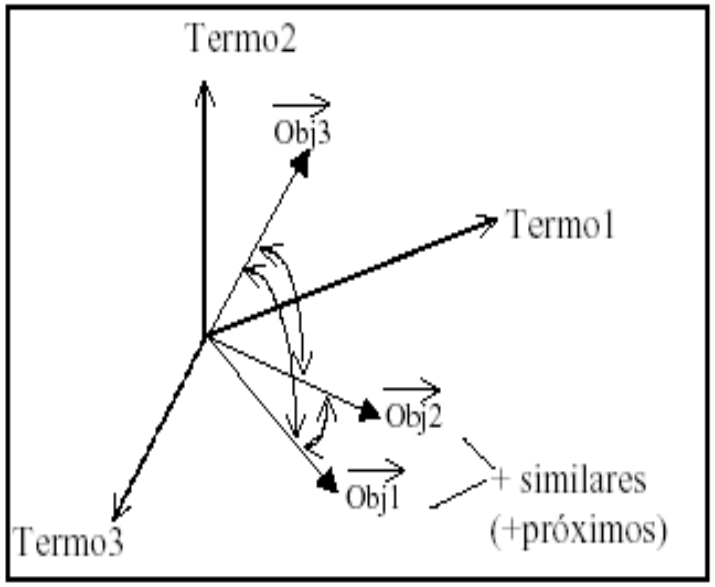
- Consulta do tipo → *"virtual and store"*
- Diminuir a abrangência para livros → *"virtual and book and store"*
- Aumentar a abrangência adicionando os tipos de itens desejados: → *"virtual and (book or cd) and store"*
- excluir alguns locais não desejados:

→*"virtual and (book or cd) and store and* ***not*** *(Amazon or Elsevier)"*

Modelo Espaço Vectorial

- Cada documento é representado por um vector de termos
- Cada termo possui um valor associado que indica o grau de importância (denominado *peso*) no documento
- O vector de termos é constituído por pares de elementos na forma:

{ (palavra_1, peso_1), (palavra_2, peso_2), ... (palavra_n, peso_n) }



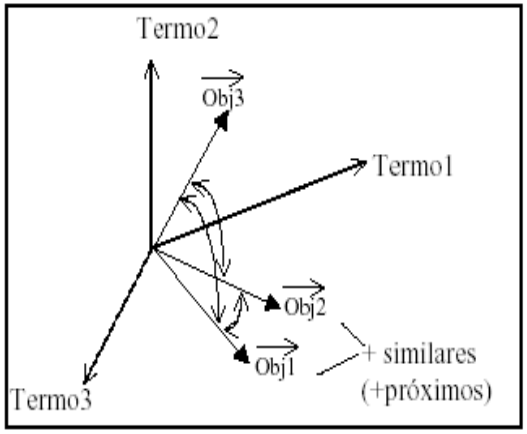
Similaridade entre Documentos

A medida mais usada para determinar a similaridade entre dois vectores é através do **co-seno do ângulo entre os vectores**

$$Sim(d_1, d_2) = \cos(\vec{v}(d_1), \vec{v}(d_2)) = \frac{\vec{v}(d_1) \bullet \vec{v}(d_2)}{\left\|\vec{v}(d_1)\right\| \left\|\vec{v}(d_2)\right\|} = \frac{\sum x_{1i} x_{2i}}{\sqrt{\sum_i x_{1i}^2} \sqrt{\sum_j x_{2j}^2}}$$

- Eficiente a calcular (somatório do produto da intersecção das palavras)
- Similaridade varia entre 0 (diferente) e 1 (o mesmo)

Modelo Espaço Vectorial



$$similaridade(Q,D) = \frac{\sum_{k=1}^{n}(w_{q,k} \times w_{d,k})}{\sqrt{\sum_{k=1}^{n}(w_{q,k})^2 \times \sum_{k=1}^{n}(w_{d,k})^2}}$$

- A distância entre um documento e o outro indica o **grau de similaridade**
- As consultas são também representadas por um vector
- Os vectores dos documentos são comparados com os vectores da consulta
- A proximidade entre os vectores é calculada através do ângulo entre os vectores

Operações Text Mining sobre Documentos

Sumarização de Documentos

Tarefa: produzir um sumário a partir de um documento original

Duas aproximações são possíveis:

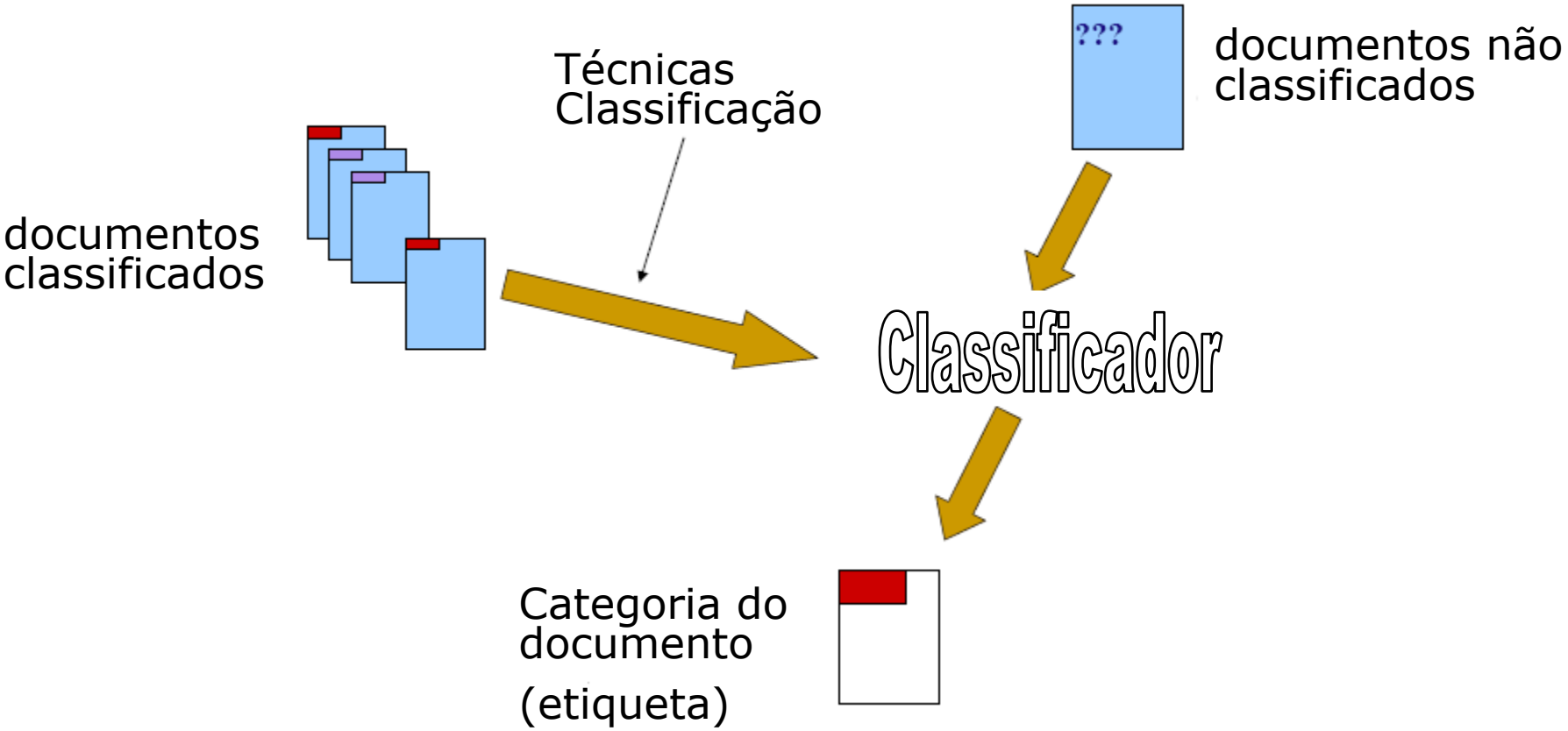
- Seleção de frases do documento, com base:
  - Peso(F) = LocalizaçãonoTexto(F) + FreqPalavras(F) + PresAdicionalnoutrasFrases()

- Significado das frases do documento
  - é realizada a análise semântica das frases
  - é criada uma rede semântica de conceitos
  - o sumário é criado a partir da rede semântica, de acordo com a restrição de tamanho imposta

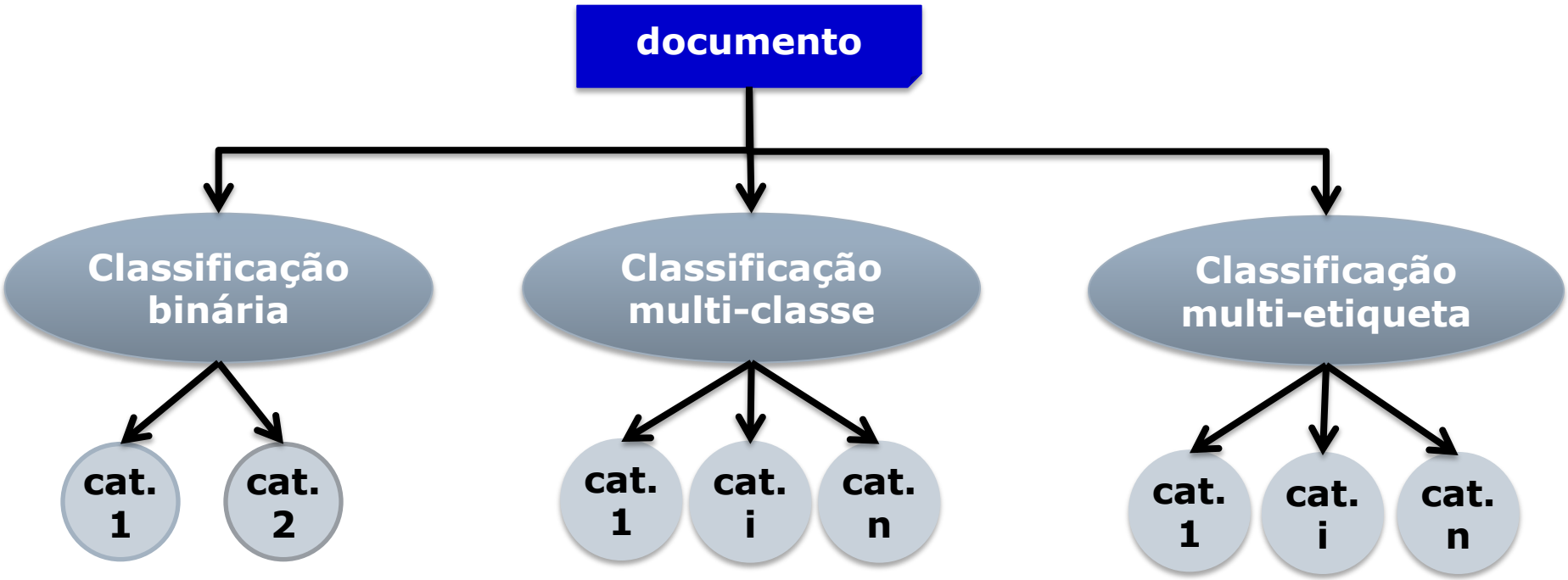
Classificação de Documentos

Dados: Um conjunto de documentos classificados em categorias de acordo com o seu conteúdo

Objectivo: construir um modelo que atribui automaticamente a categoria a documentos não classificados



Tipos de Classificação



Exemplo: classificar um conjunto de emails: *spam* ou *não spam*

Exemplo: classificar um conjunto de noticias **em apenas uma** categoria: desporto, politica, tecnologia

Exemplo: classificar um conjunto de noticias **em uma ou mais** categorias: desporto individual, desporto coletivo, politica nacional, politica internacional, tecnologia, ….

Classificação Bayesiana

A classificação Bayesiana de documentos baseia-se no **Teorema de Bayes**

$$P(x \mid y) = \frac{P(y \mid x)P(x)}{P(y)}$$

Na classificação Naive Bayes a melhor classe é a mais provável designada por ***máxima probabilidade à posteriori* (MAP)** c_{MAP} dada por

$$c_{MAP} = ArgMax_{c_j \in C} P(d_i \mid c_j)P(c_j)$$

$P(c_j)$ é estimada a partir do conjto de treino

$ArgMax_{c_j \in C} P(d_i \mid c_j)$ é estimado assumindo-se a independência dos termos $t_1, t_2, \ldots, t_n$ do documento – **suposição Naive**

Clustering de Documentos

- Clustering é o processo de descoberta de grupos naturais de modo não supervisionado (sem classes previamente definidas)

A operação chave na operação de clustering é a medida de similaridade usada para comparar documentos:

↪ a medida mais usada – **similaridade coseno**

Algoritmos de clustering mais usados na segmentação de documentos:

- Algoritmo K-Means
- Algoritmos de clustering hierárquico aglomerativo …

Avaliação de Pesquisas sobre BD Documentos

Avaliação dos Resultados

O resultado de pesquisas em documentos é avaliado através de métricas provenientes da área – **Bibliometria**

A eficiência e a eficácia de uma pesquisa é avaliada de acordo com a sua capacidade em recuperar **o máximo possível de documentos relevantes** ao mesmo tempo que **filtra o maior número de documentos irrelevantes**

Avaliação de Sistemas e Técnicas de Recuperação

- *Text Retrieval Conference (TREC)* – Conferência Internacional Anual sobre a análise de desempenho de SRI
  - ➥ dispõe de coleções públicas de documentos preparadas para avaliar técnicas de pesquisa

Avaliação dos Resultados

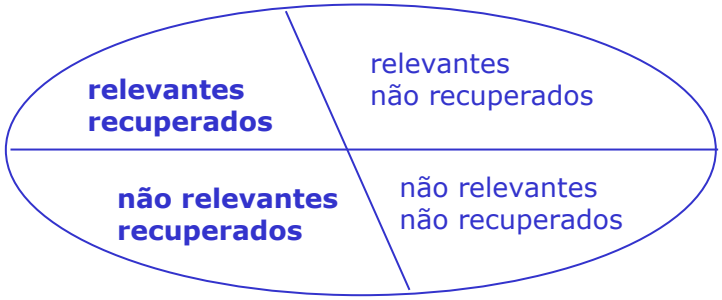
Para uma **coleção de documentos conhecida**, quando uma pesquisa dispara uma busca o conjunto de documentos é dividido em quatro segmentos lógicos:

Medidas mais comuns para a área de recuperação de informação:

- *Precision* (precisão)
- *Recall* (abrangência)
- *Fallout*

Avaliação dos Resultados

Precisão *(precision)* mede a habilidade do sistema em recuperar documentos relevantes às necessidades de informação de um utilizador



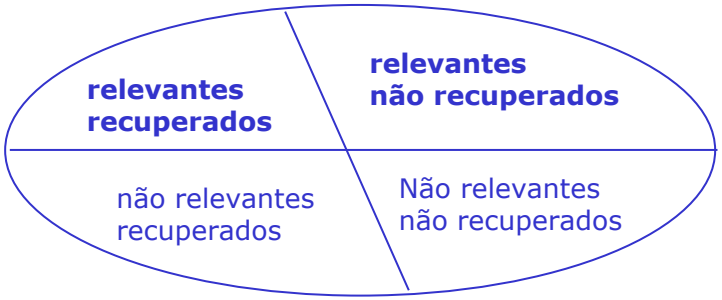
$$precision = \frac{n\_recuperados\_relevantes}{n\_total\_documentos\_recuperados}$$

$$precision = \frac{TP}{TP + FP}$$

A precisão indica o esforço (*overhead*) que o utilizador teria para analisar uma determinada busca

Avaliação dos Resultados

Abrangência *(recall)* Fração de documentos relevantes na coleção que são recuperados



$$recall = \frac{n\_recuperados\_relevantes}{n\_total\_documentos\_relevantes}$$

$$recall = \frac{TP}{TP + FN}$$

O recall indica a percentagem de informação relevante que o utilizador tem acesso numa determinada busca.

Avaliação dos Resultados

Exemplo

Base documental com 500 documentos: Sabe-se que 100 documentos são relevantes para uma consulta *q*

Dos 150 documentos recuperados na consulta *q, apenas* 25 documentos são relevantes

Precision? Recall?

	recuperados	não recuperados	
relevantes	25		100
não relevantes			
	150		500

Precision = 25/150 = 17%
Recall = 25/100 = 25%

Avaliação dos Resultados

A quantidade de informação que um sistema possui pode influenciar diretamente nas métricas precision e recall

Exemplo: Adicionando 100 novos documentos é aceitável que a quantidade de documentos irrelevantes recuperados aumente (situação possível, embora não obrigatória)

Pior caso: Sistema recupera todos os novos documentos adicionados e nenhum é relevante à consulta

	recuperados	não recuperados	
relevantes	25		100
não relevantes			
	250		**600**

Precision = 25/250 = 10% Recall = 25/100 = 25%

Avaliação dos Resultados

Fallout: mede a proporção de documentos não-relevantes recuperados relativamente a todos os documentos não-relevantes disponíveis

$$fallout = \frac{n\_documentos\_irrelevantes\_recuperados}{n\_total\_documentos\_irrelevantes}$$

$$fallout = \frac{FP}{FP + TN}$$

Fallout (inicial) = 125 /400 = 31%

Fallout = 225/500 = 45%