# *Análisis y Recuperación de Información*

1er Cuatrimestre 2017

**Página Web**

http://www.exa.unicen.edu.ar/catedras/ayrdatos/

**Prof. Dra. Daniela Godoy**

ISISTAN Research Institute
UNICEN University
Tandil, Bs. As., Argentina
http://www.exa.unicen.edu.ar/~dgodoy
dgodoy@exa.unicen.edu.ar

# Recuperación de Información

- Definición:

Recuperación de Información (Information Retrieval, IR) es la disciplina que tiene por objetivo el desarrollo de sistemas que almacenen grandes cantidades de documentos (desestructurados) de tal forma de permitir una eficiente recuperación de aquellos documentos relevantes a las necesidades de información de sus usuarios

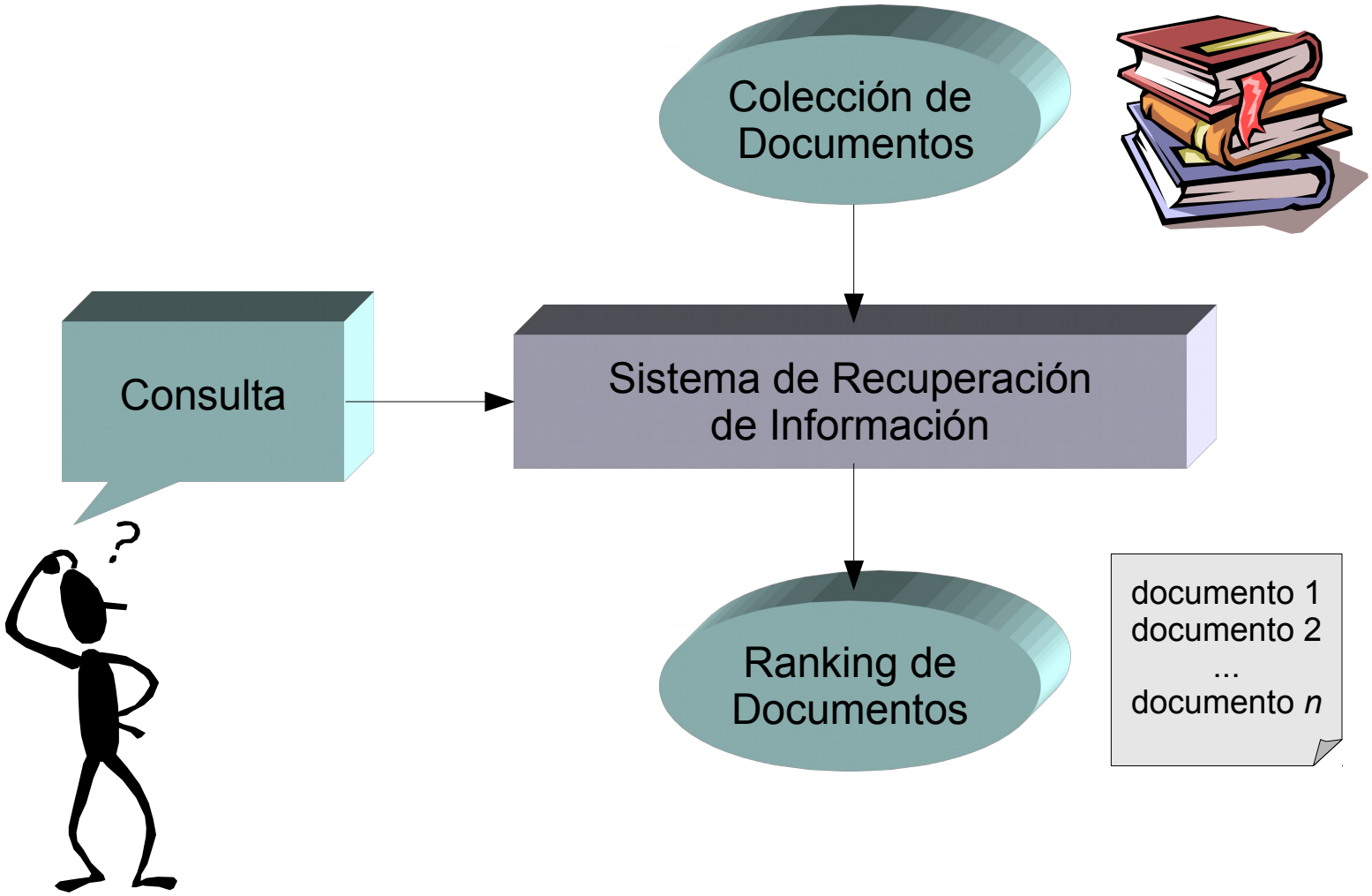
# Recuperación de Información

- Dado:
  - Un conjuntos de documentos escritos en lenguaje natural
  - Una consulta del usuario expresada textualmente
- Objetivo:
  - Encontrar un conjunto de documentos que sean *relevantes* a la consulta

# Recuperación de Información

- Vocabulario $V = \{w_1, w_2, \ldots, w_n\}$
- Consulta $q = q_1, \ldots, q_m$ donde $q_j \in V$
- Documento $d_i = w_{i1}, \ldots, w_{im}$ donde $w_{ij} \in V$
- Colección $C = \{d_1, d_2, \ldots, d_k\}$
- Conjunto de documentos relevantes $R(q) \subseteq C$
  - usualmente desconocido y dependiente del usuario
  - la consulta es una pista de qué documentos están en $R(q)$
- La tarea es calcular $R'(q) \approx R(q)$

# Recuperación de Información



# Recuperación de Información

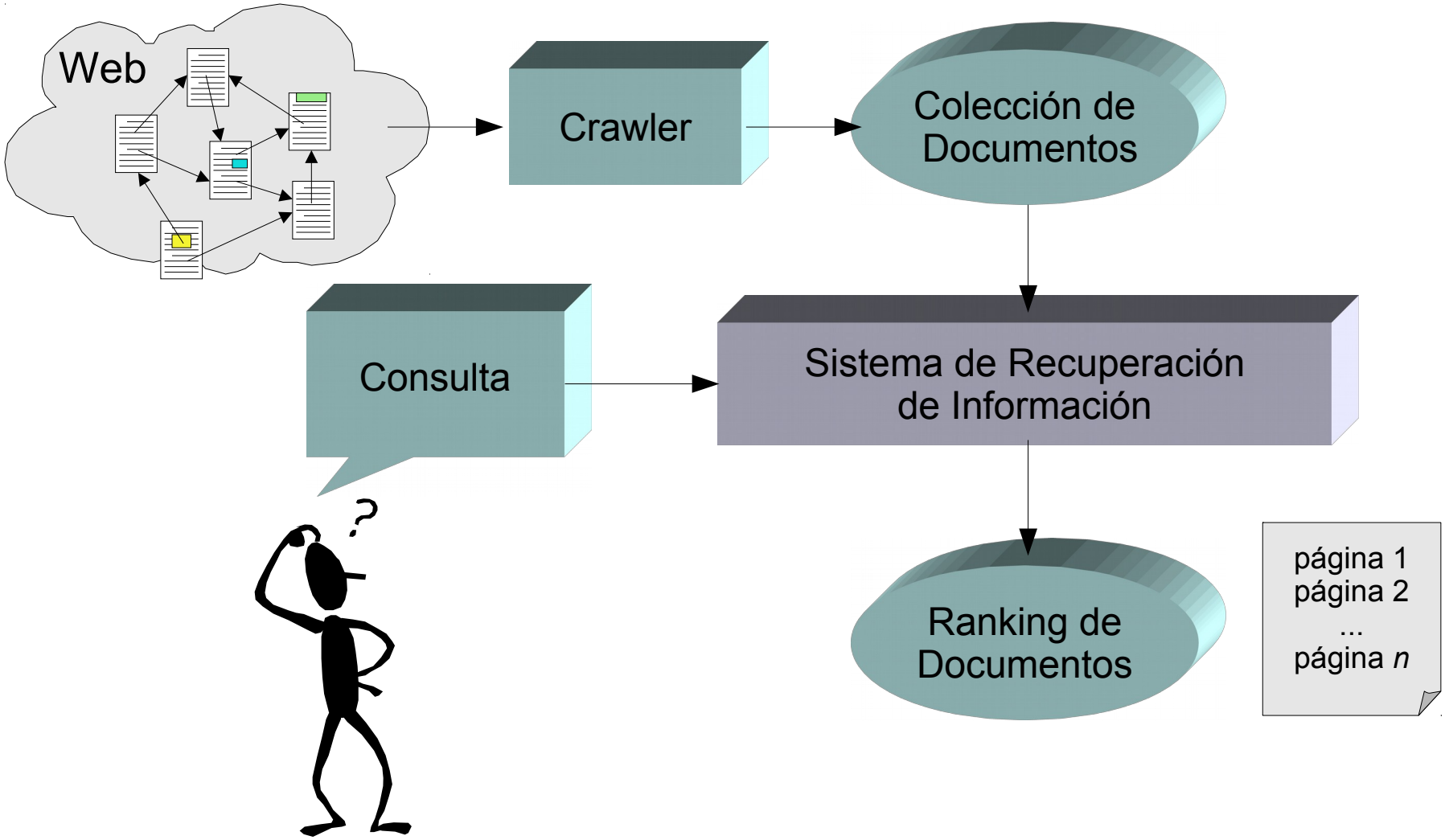
- Fuentes de Información:
  - Convencionales (ej. catálogos de librerías)
    - búsqueda por keywords, título, autor, etc.
  - Basadas en Texto (ej. Google, FAST, etc.)
    - búsqueda por keywords
  - Multimedia
    - búsqueda por apariencia visual (forma, colores, etc.)
  - Question-Answering (AskJeeves)
    - búsqueda en lenguaje natural restringido

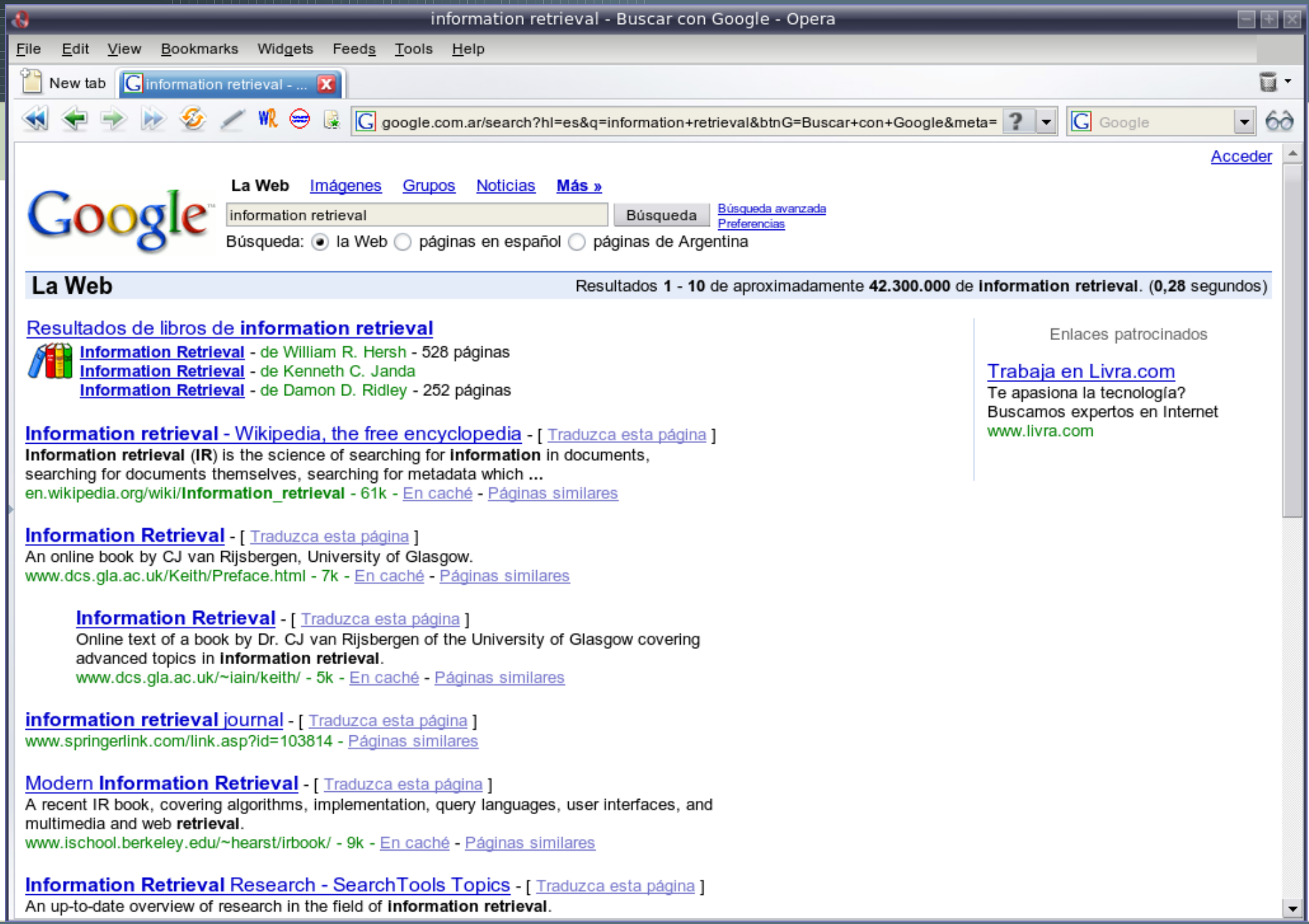

University of North Texas Libraries: UNT Library Catalog - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://iii.library.unt.edu/
Go Links
Google Search Web Search Site Page Info Up Highlight
University of North Texas
Libraries
Library Catalog
Basic Search Options:
Title Author Subject Author/Title Keyword Journal Title
Course Reserves:
Course Name Instructor
Number Search Options:
Call Numbers Other Numbers
Additional Options
Media Library Catalog
View Patron Record
New Acquisitions
Other Libraries' Catalogs
Music Special Collections Catalog
Interlibrary Loan Requests
Remote Storage Requests
Ask A Librarian
Submit Book Recommendations
This page is maintained by III Administrator
This page last updated on September 23, 2002
UNT Libraries Home Page / Resources for Library Research / Library Services / Exhibits & Collections / About the Libraries / New & Noteworthy / UNT Libraries Catalog / Electronic Resources / Library Services for UNT Off-Campus Users / Ask a Librarian / How Do I Begin / How to Find Books & Articles / Site Map / Search This Web Site / AA/EOE/ADA
University of North Texas Libraries
P.O. Box 305190
Denton, TX 76203-5190
Telephone: (940) 565-2413
Internet
start
12:39 AM

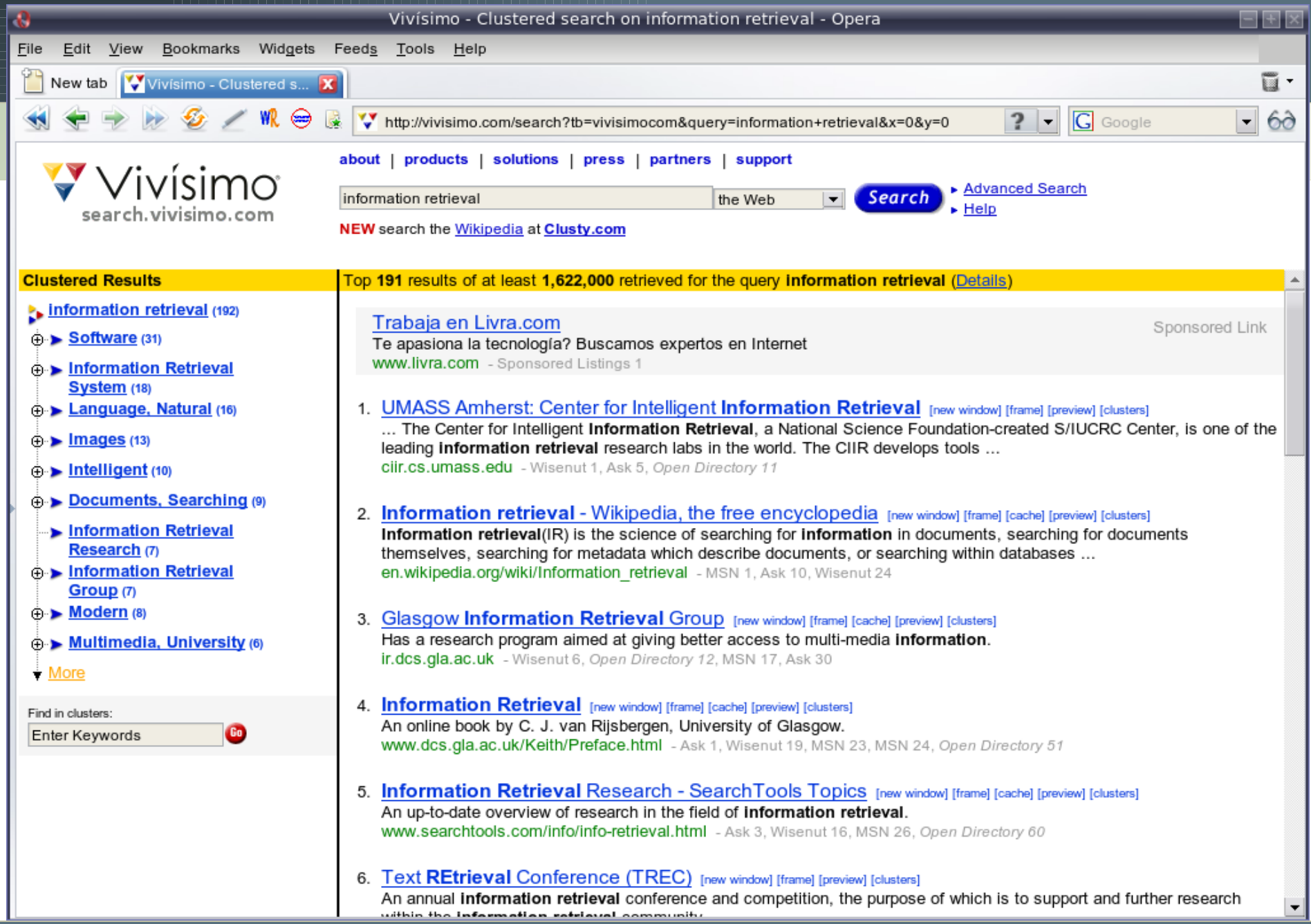
# Recuperación de Información

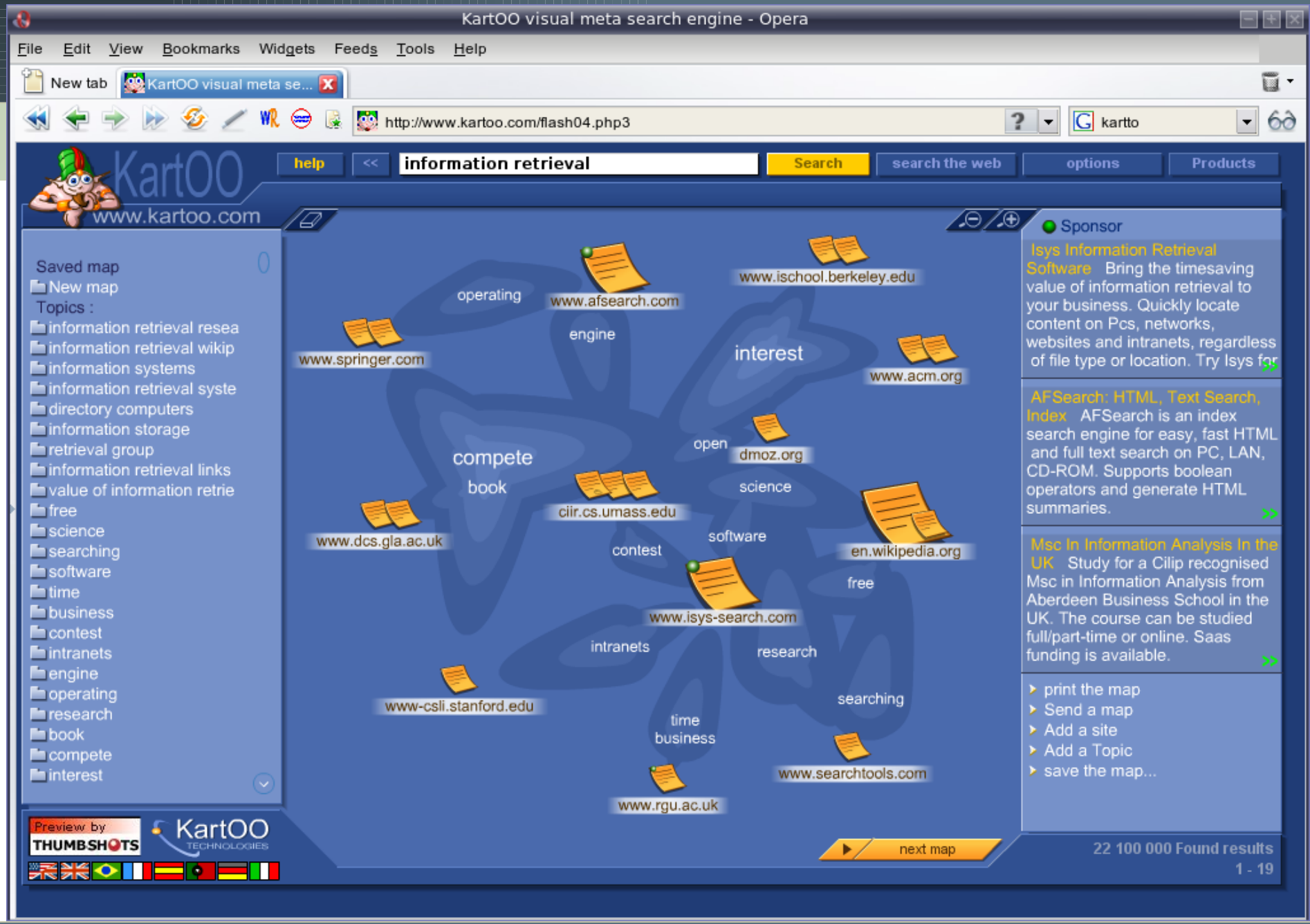
- La búsqueda de páginas en la Web es la más reciente y más ampliamente difundida aplicación de IR
  - se concentra en encontrar documentos que son relevantes a una consulta
  - trata el problema de recuperar eficientemente información dentro de una inmensa colección de documentos disponibles

# Recuperación de Información




information retrieval - Buscar con Google - Opera
File Edit View Bookmarks Widgets Feeds Tools Help
New tab
information retrieval - ...
google.com.ar/search?hl=es&q=information+retrieval&btnG=Buscar+con+Google&meta=
Google
Acceder
La Web Imágenes Grupos Noticias Más »
Google
information retrieval
Búsqueda
Búsqueda avanzada
Preferencias
Búsqueda: la Web páginas en español páginas de Argentina
La Web
Resultados 1 - 10 de aproximadamente 42.300.000 de information retrieval. (0,28 segundos)
Resultados de libros de information retrieval
Information Retrieval - de William R. Hersh - 528 páginas
Information Retrieval - de Kenneth C. Janda
Information Retrieval - de Damon D. Ridley - 252 páginas
Information retrieval - Wikipedia, the free encyclopedia - [ Traduzca esta página ]
Information retrieval (IR) is the science of searching for information in documents, searching for documents themselves, searching for metadata which ...
en.wikipedia.org/wiki/Information_retrieval - 61k - En caché - Páginas similares
Information Retrieval - [ Traduzca esta página ]
An online book by CJ van Rijsbergen, University of Glasgow.
www.dcs.gla.ac.uk/Keith/Preface.html - 7k - En caché - Páginas similares
Information Retrieval - [ Traduzca esta página ]
Online text of a book by Dr. CJ van Rijsbergen of the University of Glasgow covering advanced topics in information retrieval.
www.dcs.gla.ac.uk/~iain/keith/ - 5k - En caché - Páginas similares
information retrieval journal - [ Traduzca esta página ]
www.springerlink.com/link.asp?id=103814 - Páginas similares
Modern Information Retrieval - [ Traduzca esta página ]
A recent IR book, covering algorithms, implementation, query languages, user interfaces, and multimedia and web retrieval.
www.ischool.berkeley.edu/~hearst/irbook/ - 9k - En caché - Páginas similares
Information Retrieval Research - SearchTools Topics - [ Traduzca esta página ]
An up-to-date overview of research in the field of information retrieval.
Enlaces patrocinados
Trabaja en Livra.com
Te apasiona la tecnología?
Buscamos expertos en Internet
www.livra.com


Vivísimo - Clustered search on information retrieval - Opera
File Edit View Bookmarks Widgets Feeds Tools Help
New tab
Vivísimo - Clustered s...
http://vivisimo.com/search?tb=vivisimocom&query=information+retrieval&x=0&y=0
Google
Vivísimo
search.vivisimo.com
about | products | solutions | press | partners | support
information retrieval
the Web
Search
Advanced Search
Help
NEW search the Wikipedia at Clusty.com
Clustered Results
Information retrieval (192)
Software (31)
Information Retrieval System (18)
Language, Natural (16)
Images (13)
Intelligent (10)
Documents, Searching (9)
Information Retrieval Research (7)
Information Retrieval Group (7)
Modern (8)
Multimedia, University (6)
More
Find in clusters:
Enter Keywords
Go
Top 191 results of at least 1,622,000 retrieved for the query information retrieval (Details)
Trabaja en Livra.com
Sponsored Link
Te apasiona la tecnología? Buscamos expertos en Internet
www.livra.com - Sponsored Listings 1
1. UMASS Amherst: Center for Intelligent Information Retrieval [new window] [frame] [preview] [clusters]
... The Center for Intelligent Information Retrieval, a National Science Foundation-created S/IUCRC Center, is one of the leading information retrieval research labs in the world. The CIIR develops tools ...
ciir.cs.umass.edu - Wisenut 1, Ask 5, Open Directory 11
2. Information retrieval - Wikipedia, the free encyclopedia [new window] [frame] [cache] [preview] [clusters]
Information retrieval(IR) is the science of searching for information in documents, searching for documents themselves, searching for metadata which describe documents, or searching within databases ...
en.wikipedia.org/wiki/Information_retrieval - MSN 1, Ask 10, Wisenut 24
3. Glasgow Information Retrieval Group [new window] [frame] [cache] [preview] [clusters]
Has a research program aimed at giving better access to multi-media information.
ir.dcs.gla.ac.uk - Wisenut 6, Open Directory 12, MSN 17, Ask 30
4. Information Retrieval [new window] [frame] [cache] [preview] [clusters]
An online book by C. J. van Rijsbergen, University of Glasgow.
www.dcs.gla.ac.uk/Keith/Preface.html - Ask 1, Wisenut 19, MSN 23, MSN 24, Open Directory 51
5. Information Retrieval Research - SearchTools Topics [new window] [frame] [cache] [preview] [clusters]
An up-to-date overview of research in the field of information retrieval.
www.searchtools.com/info/info-retrieval.html - Ask 3, Wisenut 16, MSN 26, Open Directory 60
6. Text REtrieval Conference (TREC) [new window] [frame] [preview] [clusters]
An annual information retrieval conference and competition, the purpose of which is to support and further research


KartOO visual meta search engine - Opera
File Edit View Bookmarks Widgets Feeds Tools Help
New tab
KartOO visual meta se...
http://www.kartoo.com/flash04.php3
kartto
KartOO
www.kartoo.com
help
<<
information retrieval
Search
search the web
options
Products
Saved map
New map
Topics :
information retrieval resea
information retrieval wikip
information systems
information retrieval syste
directory computers
information storage
retrieval group
information retrieval links
value of information retrie
free
science
searching
software
time
business
contest
intranets
engine
operating
research
book
compete
interest
operating
www.afsearch.com
www.ischool.berkeley.edu
engine
www.springer.com
interest
www.acm.org
open
dmoz.org
compete
book
science
ciir.cs.umass.edu
www.dcs.gla.ac.uk
software
contest
en.wikipedia.org
free
www.isys-search.com
intranets
research
www-csli.stanford.edu
searching
time
business
www.searchtools.com
www.rgu.ac.uk
Sponsor
Isys Information Retrieval Software Bring the timesaving value of information retrieval to your business. Quickly locate content on Pcs, networks, websites and intranets, regardless of file type or location. Try Isys for
AFSearch: HTML, Text Search, Index AFSearch is an index search engine for easy, fast HTML and full text search on PC, LAN, CD-ROM. Supports boolean operators and generate HTML summaries.
Msc In Information Analysis In the UK Study for a Cilip recognised Msc in Information Analysis from Aberdeen Business School in the UK. The course can be studied full/part-time or online. Saas funding is available.
print the map
Send a map
Add a site
Add a Topic
save the map...
Preview by THUMBSHOTS
KartOO TECHNOLOGIES
next map
22 100 000 Found results
1 - 19

# Recuperación de Información

- Relevancia es un juicio subjetivo del usuario que puede incluir:
  - ser del tema correcto
  - ser apropiada en tiempo (reciente)
  - ser de una fuente confiable (autoritativa)
  - satisfacer los objetivos del usuario y el uso que se pretende de ella
- La noción más simple de relevancia es que el *string* completo de la consulta aparezca en un documento
- Una noción menos estricta es que las palabras de la consulta aparezcan frecuentemente en el documento, en cualquier orden

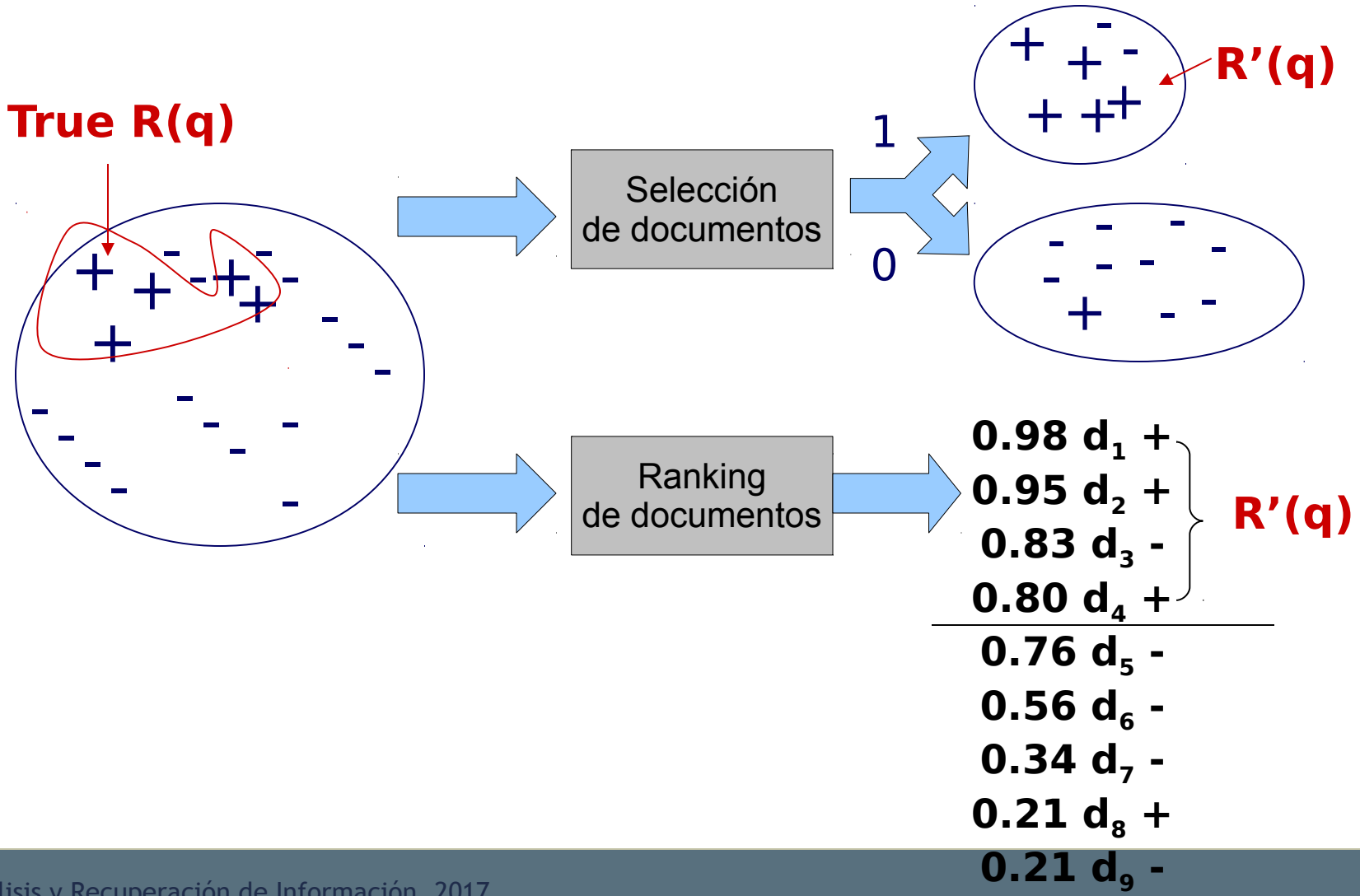
# Recuperación de Información

- La función de relevancia pocas veces es precisa
  - consulta muy restrictiva: los términos son muy específicos y no se encuentran documentos relevantes
  - consulta poco restrictiva: los términos son muy generales y se recuperan demasiados documentos
  - es difícil encontrar el equilibrio entre ambos extremos
- Aún si la función de relevancia es precisa, no todos los documentos son igualmente relevantes

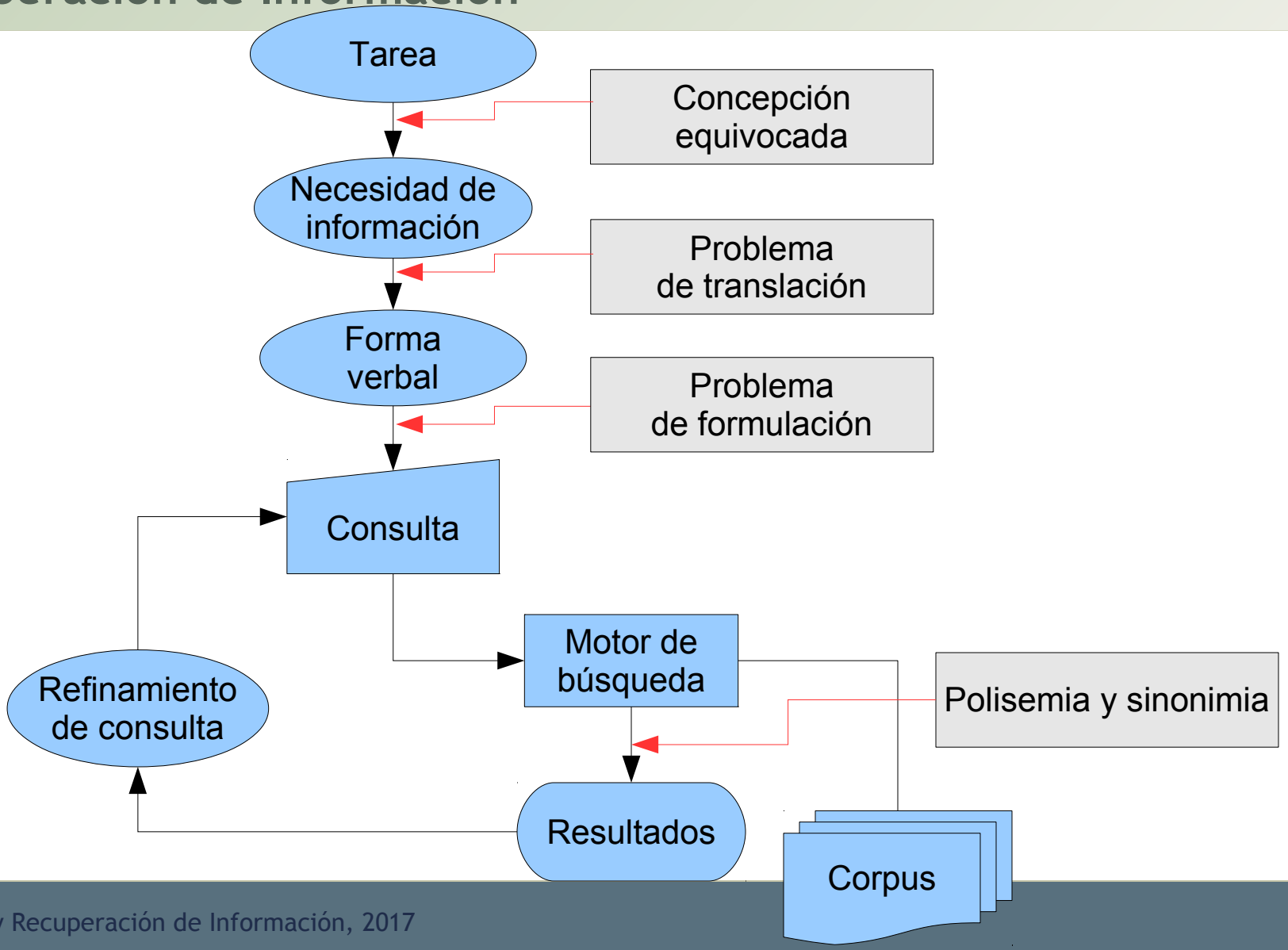
# Recuperación de Información

- Los sistemas de IR no sólo necesitan localizar documentos relevantes sino que también ordenarlos en base a este concepto de relevancia
  - Un ranking es un ordenamiento de los documentos recuperados que refleja su relevancia para el usuario (consulta)

# Recuperación de Información



# Recuperación de Información



# Recuperación de Información

- IR se enfoca al estudio de datos no estructurados
  - usualmente texto, también audio, imágenes, etc.
- Los datos se consideran no estructurados cuando:
  - la estructura es desconocida y/o la semántica de cada componente es desconocida
- Los sistemas de IR explotan regularidades estadísticas de los datos, no tratan de “entender” el significado
- Enfoques difieren:
  - DBMS trabajan sobre datos estructurados, la semántica es clara tanto en los datos con en las consultas, y se enfoca en el procesamiento eficiente de consultas bien definidas en lenguajes como SQL
  - NLP (Procesamiento de Lenguaje Natural) busca encontrar el significado (semántica) del texto no estructurado

# Recuperación de Información

- Recuperación de Información
  - Encontrar información relevante en una fuente de información desestructurada, usualmente textual

| | |
|---|---|
| **Fuente** | Desestructurada |
| **Búsqueda** | Goal-oriented |
| **Entidad atómica** | Documento |
| **Ej. Necesidad de información** | "Encontrar un restaurante Japonés en Boston que sirva comida vegetariana" |
| **Ej. Consulta** | "Restaurante japonés en Boston" or Boston->Restaurantes->Japoneses |

# Recuperación de Información

- Recuperación de Datos
  - Encontrar registros en una base de datos estructurada

| | |
|---|---|
| **Fuente** | Estructurada |
| **Búsqueda** | Goal-oriented |
| **Entidad atómica** | Registro |
| **Ej. Necesidad de información** | "Encontrar un restaurante Japonés en Boston que sirva comida vegetariana" |
| **Ej. Consulta** | SELECT * FROM restaurantes WHERE ciudad = boston AND tipo = japonés AND vegetariano = true |

# Recuperación de Información

- Minería de Datos
  - Descubrir nuevo conocimiento a partir del análisis de los datos

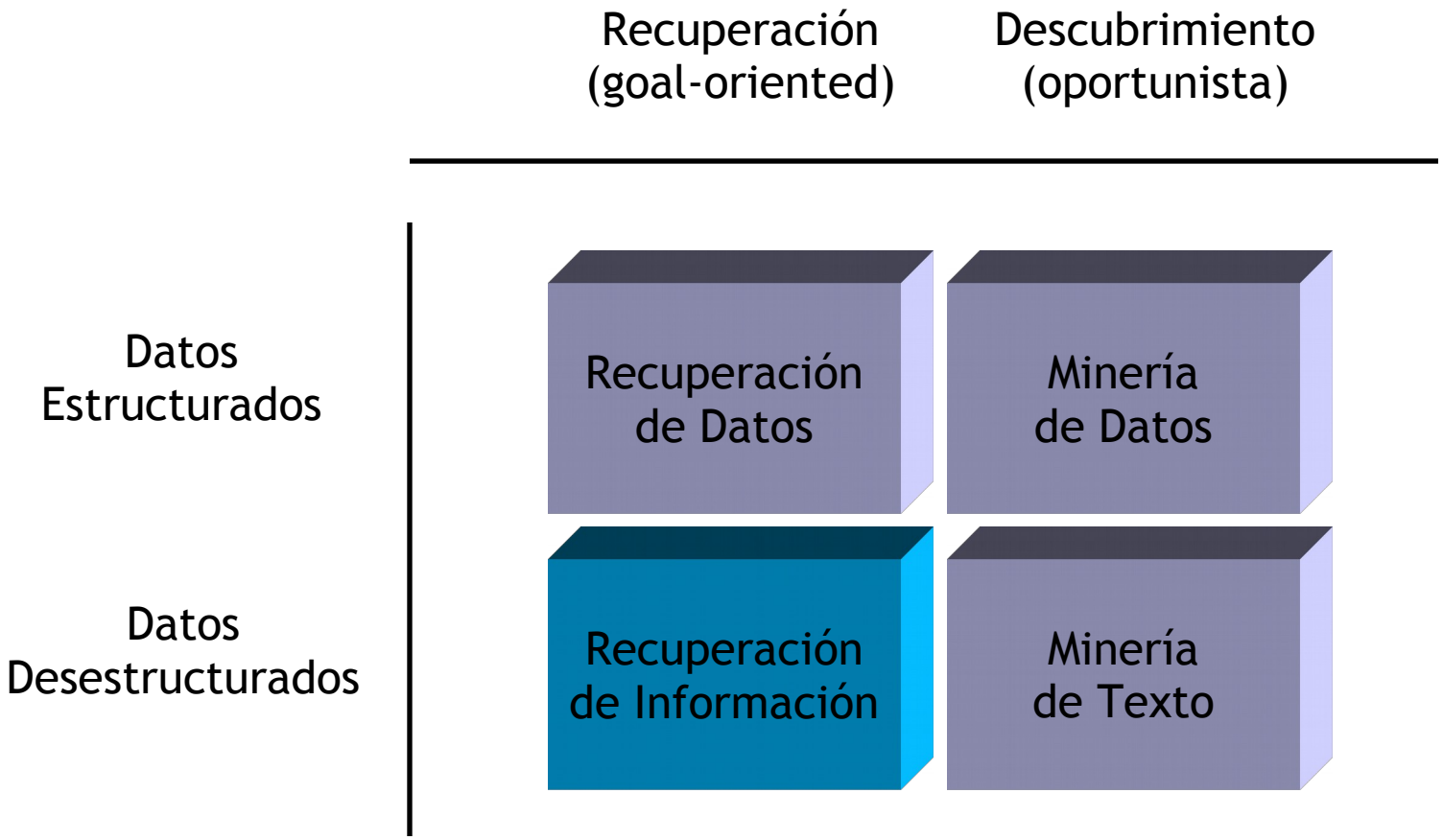
| | |
|---|---|
| **Fuente** | Estructurada |
| **Búsqueda** | Oportunista |
| **Entidad atómica** | Dimensiones |
| **Ej. Necesidad de información** | "Mostrar la tendencia del número de visitantes a restaurantes japoneses en Boston" |
| **Ej. Consulta** | SELECT SUM(visitantes) FROM restaurantes WHERE ciudad = boston AND tipo = japonés ORDER BY date |

# Recuperación de Información

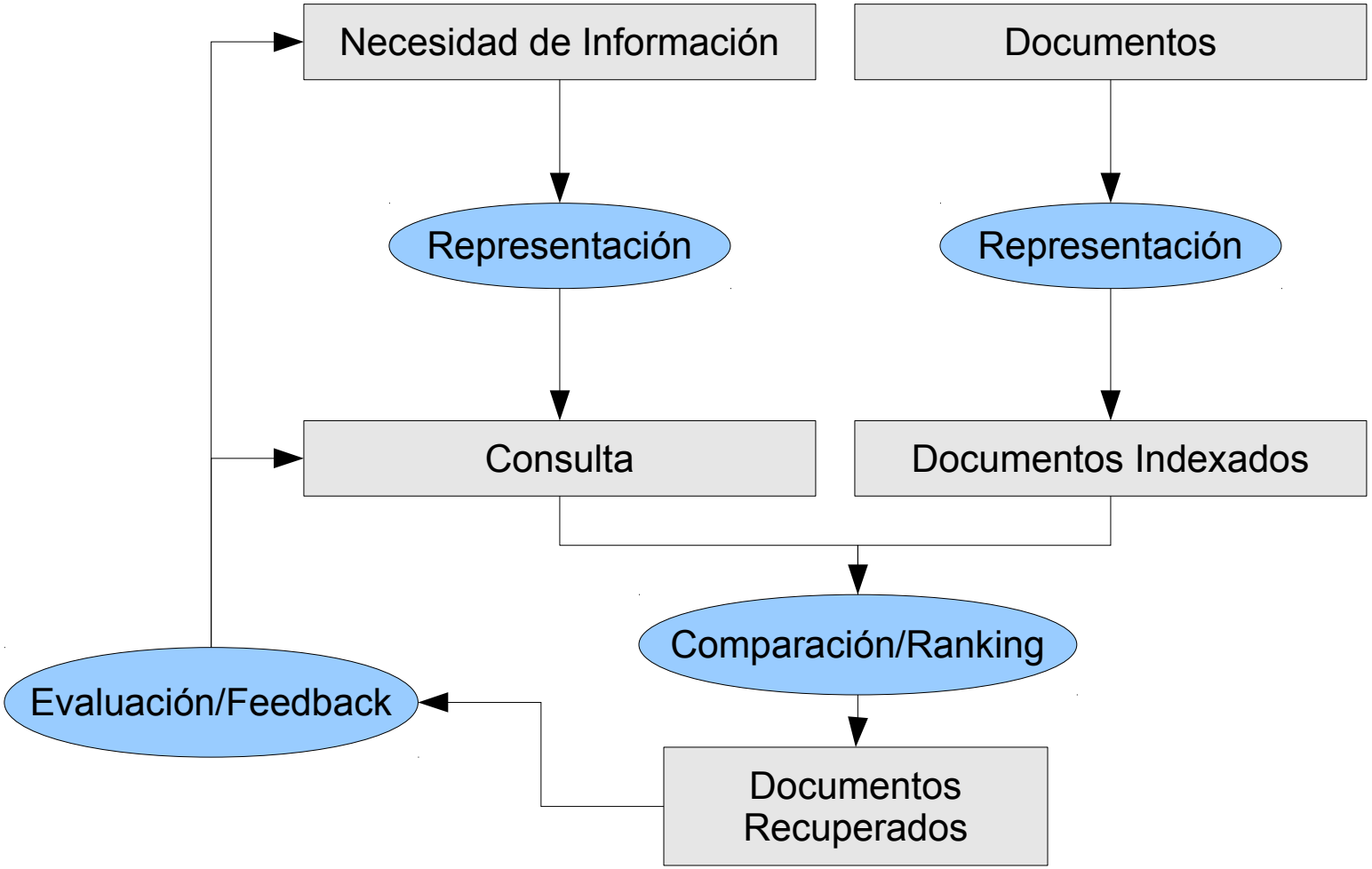
- Minería de Texto
  - Descubrir nuevo conocimiento a partir del análisis de textos

| | |
|---|---|
| **Fuente** | Desestructurada |
| **Búsqueda** | Oportunista |
| **Entidad atómica** | Característica del lenguaje (palabra) o concepto |
| **Ej. Necesidad de información** | "Encontrar otros tipos de restaurantes que visita la gente a la que le gustan los restaurantes japoneses" |
| **Ej. Consulta** | Hacer un ranking de los otros tipos de restaurantes que están asociados a "restaurantes japoneses" |

# Recuperación de Información



# Recuperación de Información

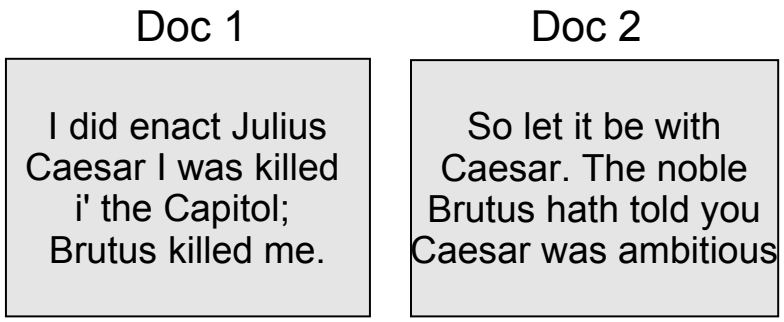


# Recuperación de Información

- Representación del texto (Indexación)
  - dado un documento textual, identificar los conceptos que describen el contenido y su importancia relativa
- Representación de la necesidad de información (Formulación de Consultas)
  - describir y refinar la necesidad de información en una consulta explícita
- Comparación de representaciones (Recuperación)
  - comparar las representaciones del texto y la consulta para determinar que documentos son potencialmente relevantes
- Evaluar los documentos recuperados (Feedback)
  - presentar los documentos al usuario y modificar la consulta en base al feedback

# Componentes de un Sistema de IR

- Operaciones sobre el texto: para obtener palabras a indexar
  - Tokenización
  - Eliminación de stop-words
  - Stemming
- Indexación: construir un índice invertido de palabras con punteros a documentos
  - mapean palabras a IDs de documentos




| Term | Doc # |
|---|---|
| I | 1 |
| did | 1 |
| enact | 1 |
| julius | 1 |
| caesar | 1 |
| I | 1 |
| was | 1 |
| killed | 1 |
| i' | 1 |
| the | 1 |
| capitol | 1 |
| brutus | 1 |
| killed | 1 |
| me | 1 |
| so | 2 |
| let | 2 |
| it | 2 |
| be | 2 |
| with | 2 |
| caesar | 2 |
| the | 2 |
| noble | 2 |
| brutus | 2 |
| hath | 2 |
| told | 2 |
| you | 2 |
| caesar | 2 |
| was | 2 |
| ambitious | 2 |

# Componentes de un Sistema de IR

Documentos a indexar

Friends, Romans, countrymen.

Tokenizador

Tokens

Friends | Romans | Countrymen

Módulos lingüisticos

Tokens modificados

friend | roman | countryman

Indexador

Indice invertido

***friend*** → 2 → 4

***roman*** → 1 → 2

***countryman*** → 13 → 16

# Componentes de un Sistema de IR

- Búsqueda: recupera los documentos que contienen un término dado de la consulta a partir del archivo invertido
- Ranking: le da un score a todos los documentos recuperados de acuerdo a una métrica de relevancia
- Interfaz de usuario: maneja la interacción con el usuario:
  - Entrada de la consulta y presentación de documentos
  - Feedback de relevancia
  - Visualización de los resultados
- Operaciones sobre la consulta: transformaciones de la consulta para mejorar la recuperación
  - expansión usando diccionarios
  - transformación usando el feedback de relevancia