Índice

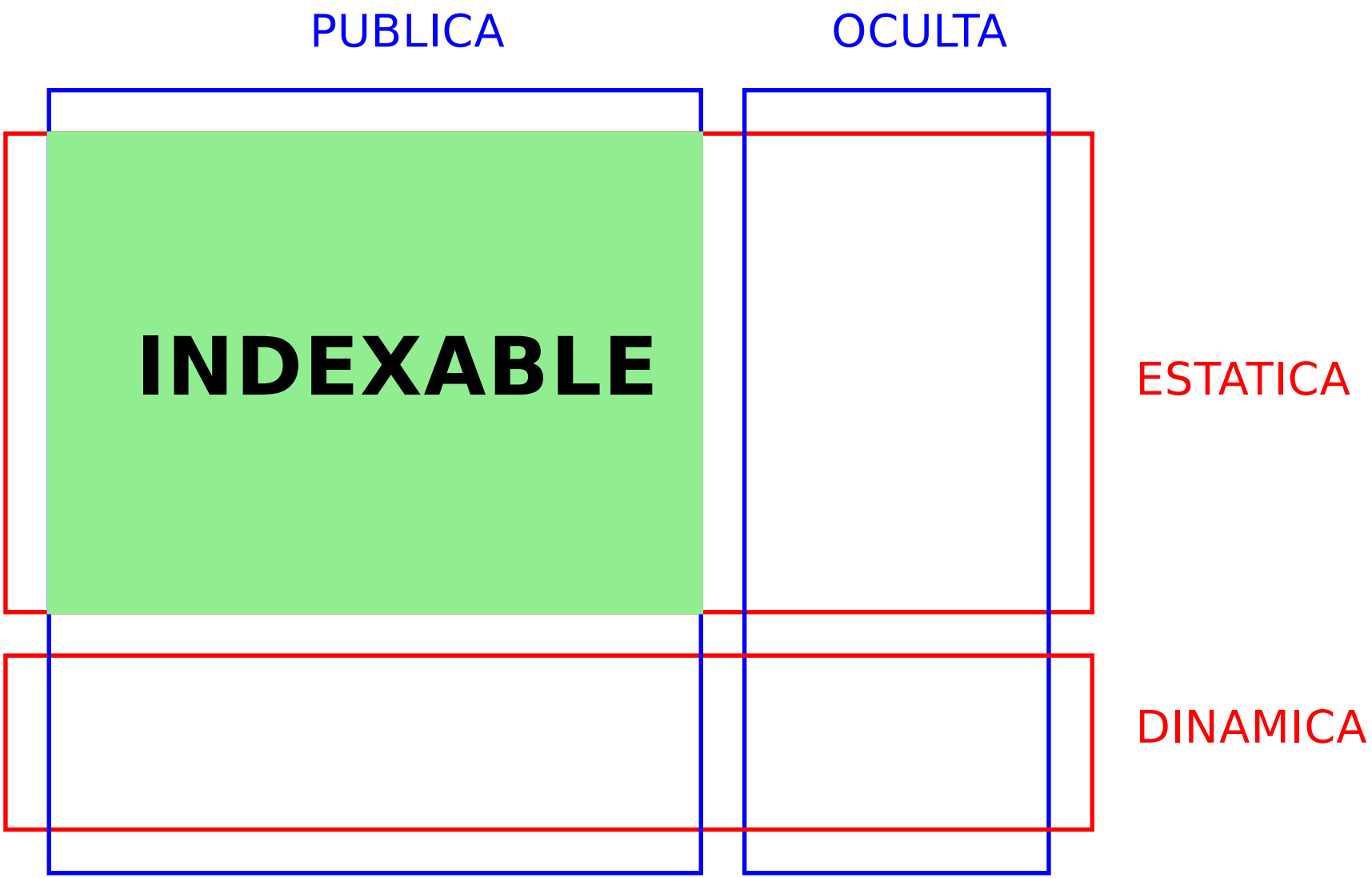
Buscadores

- Tamaño web: miles de millones de páginas (`http://www.worldwidewebsize.com/`)
- ¿Cómo encontrar algo?
- Sitios web especializados en buscar otros sitios web (**buscadores**):
  - **Directorios**: jerarquizados por temas y categorías

    Yahoo! Directory (`http://dir.yahoo.com`)
  - **Motores de búsqueda** (o buscadores): búsqueda por palabras clave

    Google (`http://www.google.es`)
    Yahoo! (`http://www.yahoo.es`)
    Bing (`http://www.bing.es`)

Estructura de la web



El buen buscador

- Manejo sencillo e intuitivo
- Rápido
- Actualización constante de contenidos
- Resultados claros y ordenados
- *Búsquedas avanzadas*

Breve historia de los buscadores

1990	Archie	Primer buscador de Internet (FTPs)
1992 Dic	Veronica	Buscador de Gopher (menús jerarquizados)
1993 Jun	Wanderer	Primer buscador web
1993 Dic	RBSE	Primero en calcular medida relevancia
1994 Ene	Galaxy	Primer directorio
1994 Abr	Yahoo	Directorio revisado manualmente
1994 Abr	WebCrawler	Salto tecnológico: indexar texto completo
1994 Jul	Lycos	Índice masivo
1995 Feb	Infoseek	Netscape. Amigable, servicios adicionales
1995 Jun	Metacrawler	Primer metabuscador
1995 Dic	Altavista	Muy veloz. Lenguaje natural y ops. lógicos

Breve historia de los buscadores (cont.)

1996 Abr	Olé	Primer buscador hispano
1996 May	HotBot	Tecnología de búsqueda de alto rendimiento
1998	MSN Search	Buscador de Microsoft
1998 Sep	Google	Nuevo salto tecnológico: algoritmo *pagerank*
1999	Baidu	Buscador chino
2005 Nov	Live Search	Plataforma *Windows Live* de Microsoft
2006	Quaero	"Google europeo"
2009 May	Wolfram Alpha	Motor de respuestas factuales
2009 Jun	Bing	Buscador actual de Microsoft

Directorio

- Sitio web que contiene un índice o lista de páginas web estructuradas jerárquicamente en base a categorías y subcategorías temáticas

  Yahoo! Directory (`http://dir.yahoo.com`)

- Estructura navegable
- Generalmente creado/revisado a mano
  - Categorización automática
- Han ido perdiendo importancia frente a los motores de búsqueda
- Actualmente son un "complemento" a éstos
- Para búsquedas muy generales

Ejercicios

- Busca en Yahoo! Directory (`http://dir.yahoo.com`) documentos sobre *revistas de aikido*
- Ponte de acuerdo con un compañero e intentad buscar información sobre *la salud de las mascotas*, pero intentad empezar por categorías diferentes hasta llegar a los mismos contenidos. Como podréis ver una misma información puede alcanzarse en ocasiones de más de una forma.
- Intenta ahora realizar esas búsquedas empleando un uscador.

Motor de búsqueda

- Sitio web que contiene una base de datos (índice) donde las páginas web han sido indexadas en base a palabras clave y sobre la cual podemos realizar búsquedas (consultas o *queries*)

  Google (`http://www.google.es`)
  Yahoo! (`http://www.yahoo.es`)
  Bing (`http://www.bing.es`)

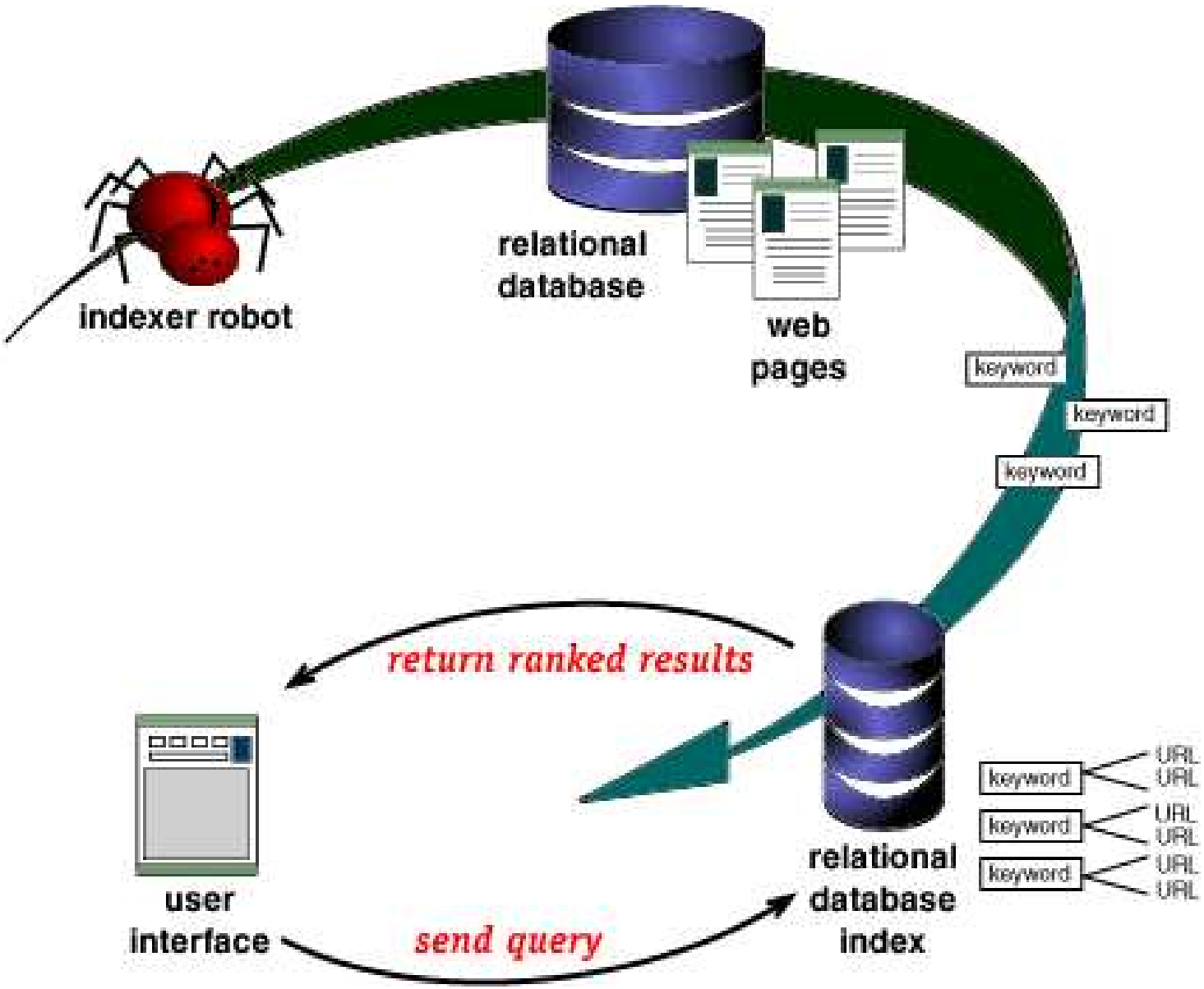
- Ante una consulta:
  1. Busca correspondencias en la base de datos
  2. Presenta las páginas web encontradas, por orden de relevancia

- Para búsquedas concretas

Ejercicios

- Lanzar en Yahoo!, en Bing y en Google algunas de las siguientes consultas:
  - `adelanto elecciones noviembre`
  - `nuevo tablet amazon`
  - *alguna noticia importante del día*

Funcionamiento de un motor de búsqueda



Arquitectura de un motor de búsqueda

- **Robots:** Programas que recorren la red buscando documentos:
  - Analizan su contenido (total o parcial)
  - Devuelven las palabras clave o descriptores que lo describen (a indexar)
- **Base de datos:** índice de palabras clave o descriptores asociados a cada documento
  - Actualización periódica (robots)
- **Interfaz de consulta:** parte que ve el usuario
  - Introducir consulta
  - Presentar resultados

Palabras clave

- Representación del tema buscado
- Buena búsqueda = palabras clave adecuadas

- Evitar palabras demasiado generales, emplear términos específicos

  `sillas del siglo XIX` vs. `muebles antiguos`

- Evitar palabras ambiguas:

$$\texttt{reparaciones de pisos} \rightarrow \begin{cases} \texttt{viviendas ?} \\ \texttt{suelos ?} \end{cases}$$

- Evitar palabras "demasiado" comunes o "vacías" (*stopwords*): preposiciones, artículos, etc.
  - Ignoradas o eliminadas automáticamente
  - Evitamos correspondencias casuales con otros idiomas

Palabras clave (cont.)

- Comprobar ortografía: `coches`≠`cches`
  - Autocorrección o sugerencias
- En minúsculas (salvo propios) y sin tildes
  - Algunos buscadores diferencian
- El orden puede influir
  - Primero las palabras más específicas y/o relevantes
- **Casi nunca se acierta a la primera**
  - Refinar la consulta sucesivamente
    - Google Instant (`http://www.google.es/instant/`)

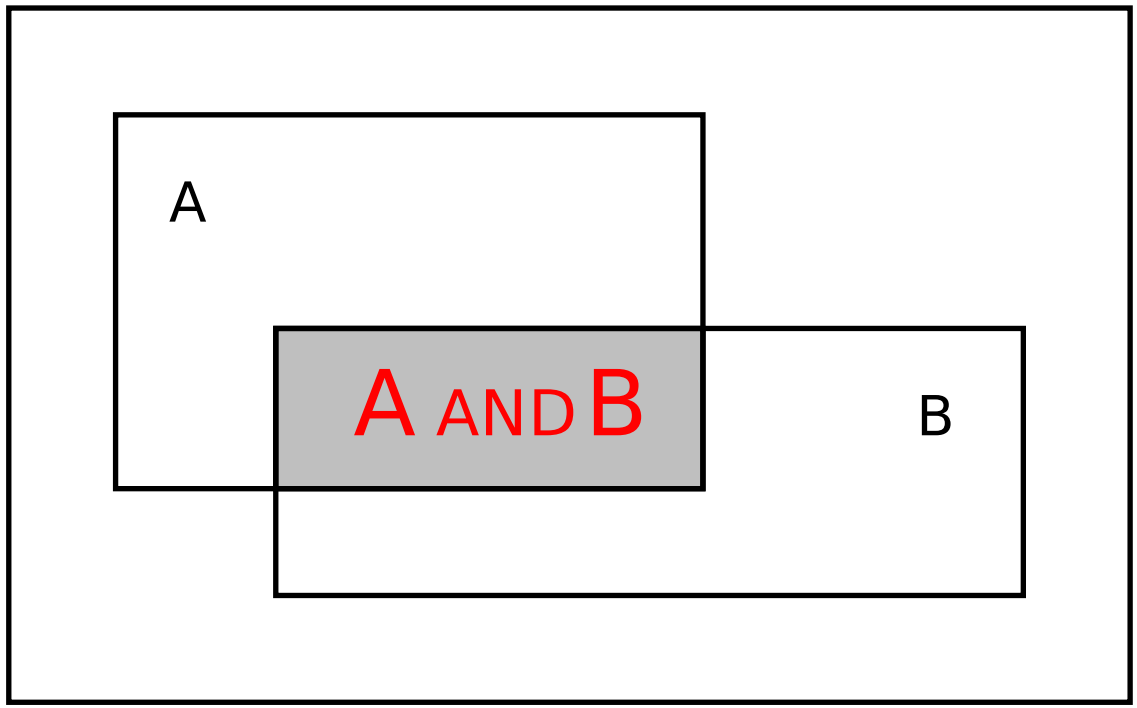
Operadores básicos

- Permiten ampliar, acotar y dirigir la búsqueda
- Disponible en `Búsquedas avanzadas` o con *sintaxis* particular

NOTA: los operadores que describiremos son de uso general entre los buscadores, pero a veces las funcionalidades, su disponibilidad, o su sintaxis varía de un buscador a otro o incluso dentro del mismo buscador a lo largo del tiempo o dependiendo de su localización.
Es conveniente revisar de vez en cuando la sección de AYUDA del buscador:

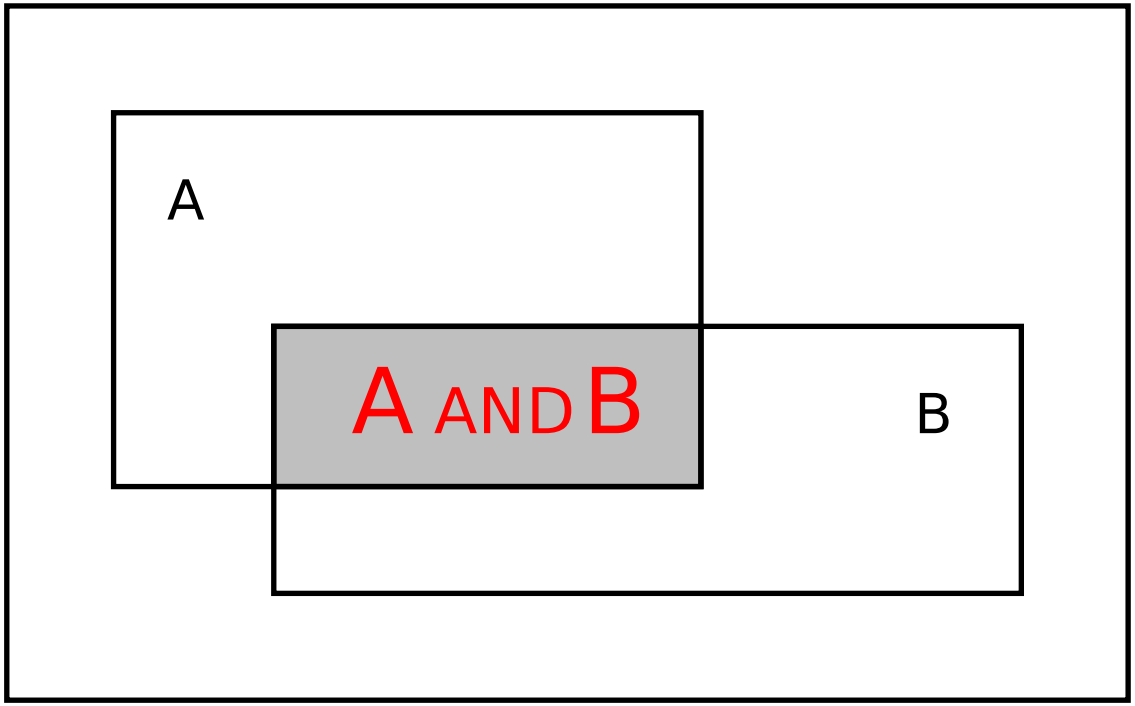
- Google → todo acerca de Google → Ayuda → Ayuda de Búsqueda web
- Yahoo! → Ayuda → Search → Información General
- Bing → Ayuda

Operadores básicos (cont.): AND (*Y lógico*)



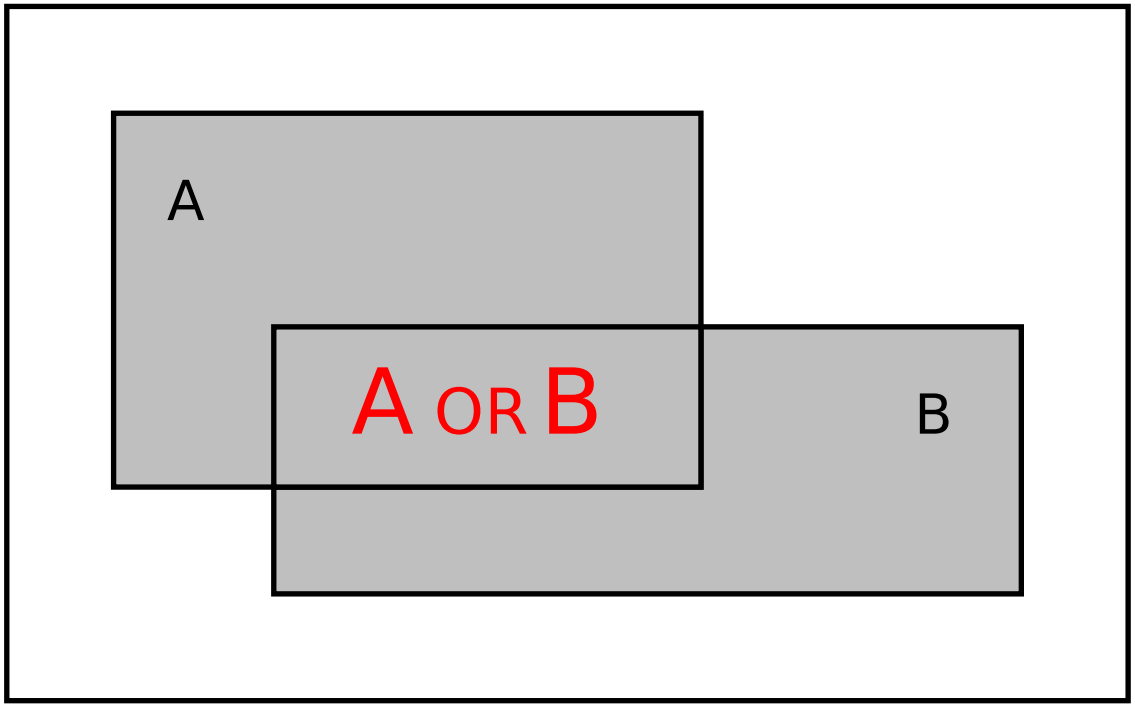
- Operador por defecto en Google, Yahoo! y Bing (otros?)
- Correspondencia simultánea con TODAS las palabras
- Restringe la búsqueda
- Sintaxis:
  - **Google/Yahoo!/Bing:** `moto Peugeot` `moto AND Peugeot`

Operadores básicos (cont.): AND (*Y lógico*) (cont.)



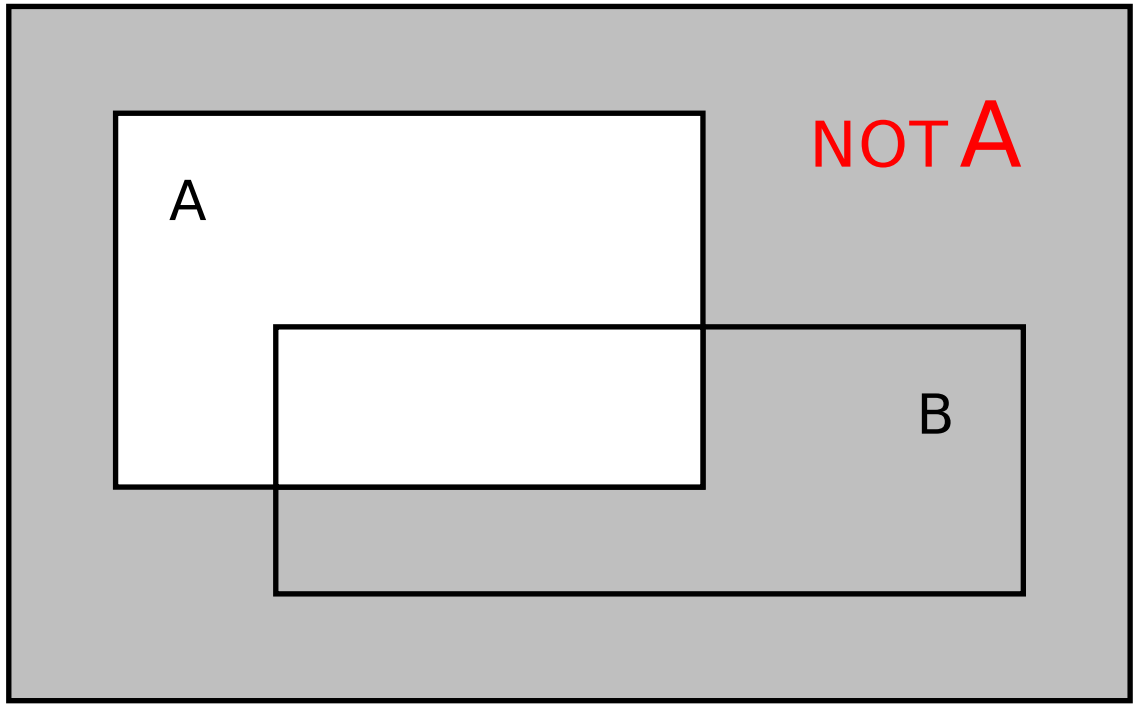
- **Variante (Google):**
- Correspondencia simultánea con TODAS las palabras aunque sean *stopwords* o *signos ortográficos*, que suelen ser ignorados en otro caso
- Busca la CORRESPONDENCIA EXACTA con el término de la consulta, sin aplicar sinónimos ni variantes
- Sintaxis:
  - `+gatas +persas` vs. `gatas persas`

Operadores básicos (cont.): OR (*O lógico*)



- Correspondencia con ALGUNA (O TODAS!) las palabras
- Amplía la búsqueda
- Sintaxis:
  - **Google/Yahoo!/Bing:** `moto OR Peugeot`

Operadores básicos (cont.): NOT (*NO lógico*)



- EXCLUYE la palabra
- Restringe la búsqueda
- Sintaxis:
  - **Google/Yahoo!/Bing:** `smartphone -iPhone`

Operadores básicos (cont.): frase

- Busca la SECUENCIA EXACTA (cita literal: mismas palabras, incluso *stopwords*, y mismo orden)
- Restringe la búsqueda
- Sintaxis:
  - **Google/Yahoo/Bing:** `"y fue entonces cuando llegó"`

- NOTA: `"profesoras"` equivale a `+profesoras` , y quiere decir que busque esa palabra obligatoriamente, aunque sea una *stopword* y tal cual está escrita, sin sinónimos ni variantes

`venta "gatas"` vs. `venta gatas`

`"profesoras"` vs. `profesoras`

Presentación de resultados

- Google como ejemplo:

conceptos basicos internet

Calculando la relevancia: indicadores

- *Peso* de la palabra
- Frecuencia de la palabra en el documento y la consulta
- Número de palabras de la consulta con correspondencia
- Longitud del documento
  - Ponderar frecuencia
- Correspondencia exacta de la palabra frente a variantes

  `gato` / `gata` / `gatos` / `gatas`
- Mismo orden

  `cumbres altas` / `altas cumbres`

Calculando la relevancia: indicadores (cont.)

- Proximidad entre sí dentro del documento
- Posición en el texto
  - Mejor en título y encabezamientos
- Presencia en las etiquetas <META> (al crear página)
- Popularidad del documento (número de enlaces que lo referencian, número de accesos)
- Valoración de los usuarios
- Enlaces patrocinados
- "Trampas" y "castigos"