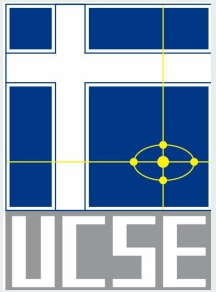
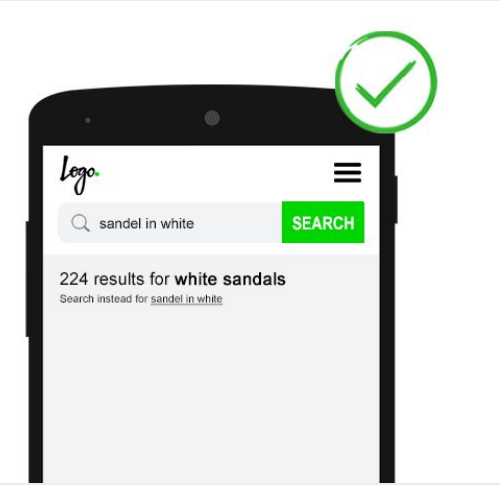
Búsqueda

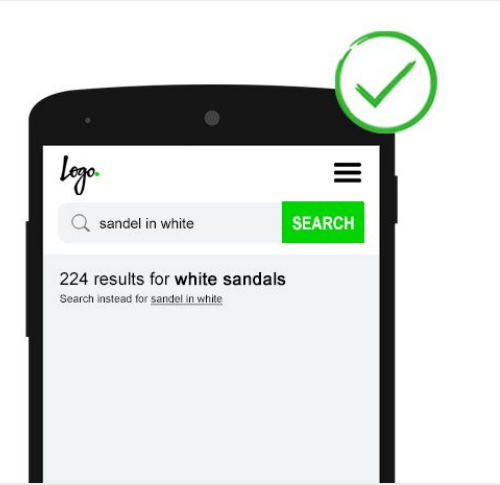
UCSE - SEIA



Búsqueda interna vs externa?



Búsqueda interna



Comunidades sustentables?

1. Magnet content authored by experts.
2. Means of collaboration.
3. **Powerful facilities for browsing and searching both magnet content and contributed content.**
4. Means of delegation of moderation.
5. Means of identifying members who are imposing an undue burden on the community and ways of changing their behavior and/or excluding them from the community without them realizing it.
6. Means of software extension by community members themselves.

Búsqueda? Por qué es tan importante?

- Queremos evitar los casos de “pregunta duplicada” o “deberías ver lo que respondieron en este hilo”, etc.
- En términos de importancia: navegación >>> búsqueda
- Sobre qué se busca?
- Qué tendría que tener una buena búsqueda?

Implementando nuestro motor de búsqueda en 5’

```
select *
from articles
where ...
```

Problemas de búsqueda con SQL tradicional

- Calidad!
- Performance!

Solución: Full-text index

Palabra	Código/s contenido/s donde aparece
Adidas	512, 71
BsAs	151, 91
Corredor	45, 76, 23
Hidratantes	19, 76, 512
...	...
Maratón	151, 91
Reloj	2, 5951, 76
Zapatillas	2, 45, 778

Solución: Full-text index

Posibles problemas:

- Alguien tiene que actualizar esa tabla (índice).
- Necesitamos un listado de ***stopwords***!
- No resolvimos cómo priorizar los resultados.

Solución: Full-text index + word-frequency histogram

Histograma de un documento/post/noticia determinado:

Palabra	Cantidad	Frecuencia
Adidas	2	2/7
BsAs	1	1/7
Corredor	3	3/7
Hidratantes	1	1/7
Maratón	1	1/7
Reloj	3	3/7
Zapatillas	2	2/7

Solución: Full-text index + word-frequency histogram

Histograma de un documento/post/noticia determinado + frecuencia general:

Palabra	Cantidad	Frecuencia	Frecuencia promedio
Adidas	2	2/7	0,12
BsAs	1	1/7	0,05
Corredor	3	**3/7**	**0,4**
Hidratantes	1	1/7	0,8
Maratón	1	1/7	0,2
Reloj	3	**3/7**	**0,03**
Zapatillas	2	2/7	0,21

Solución: Full-text index + word-frequency histogram + stemming

Stemming o lematización:

- Se trata de extraer la raíz de una palabra. Ejemplos:

  stem(running) = run

  stem(ran) = run

  stem(runners) = run

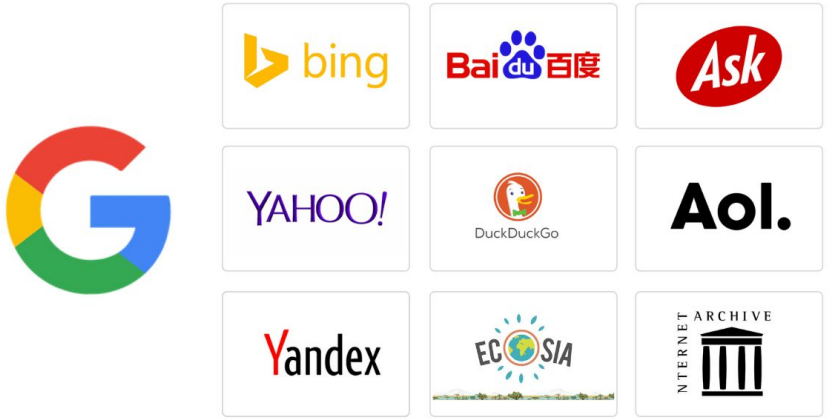
- De esta manera ganamos cobertura !
- Perdemos algo de precisión

Full-text index: Ejemplo completo

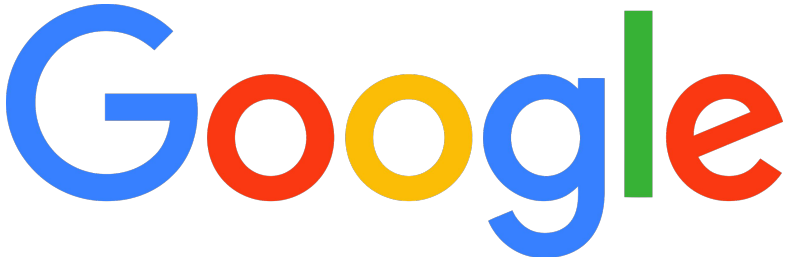
“Zapatillas para correr”

1) Tokenización (separar en unigramas)
2) Eliminar “para”
3) Realizar stemming: “zapatill” - “corre”
4) Query index
5) Priorizar resultados utilizando TF-IDF o alguna métrica similar

Búsqueda externa

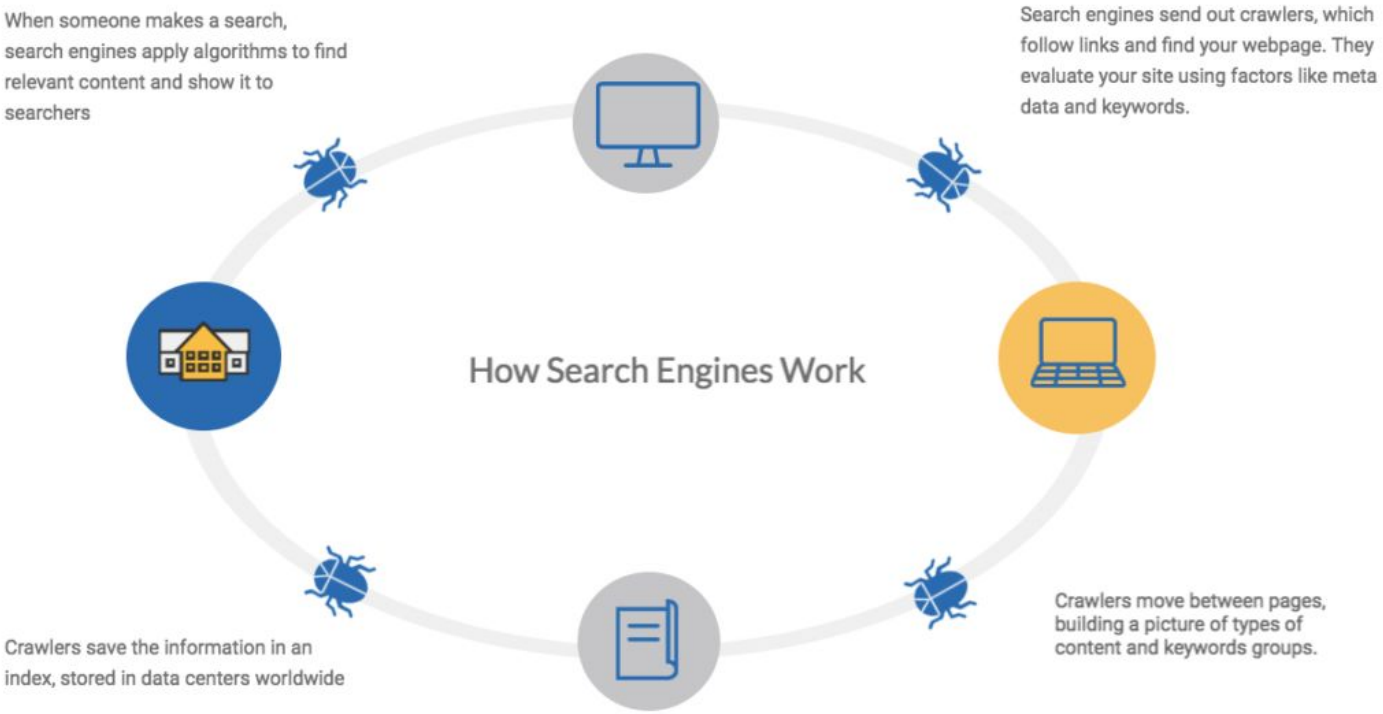


Motores de búsqueda



DuckDuckGo

Motores de búsqueda



https://blog.alexa.com/what-is-seo/

Motores de búsqueda

Cómo priorizar los resultados de una búsqueda?:

Sergey Brin, Lawrence Page. **The Anatomy of a Large-Scale Hypertextual Web Search Engine**. 1998, Stanford.

Algoritmo PageRank.

Motores de búsqueda

Qué cosas necesitamos?:

1) Que sepan que existimos ! (links externos o dándonos de alta)
2) Que puedan leer el contenido en nuestro sitio (no imagenes, no JS, no Flash, etc.)
3) Navegar por todas las páginas de nuestro sitio
4) Definir meta-tags en el HEAD de cada página (“keywords”, “description”, “title”)

Motores de búsqueda: robots.txt

- Es un archivo de texto donde le damos información extra a las arañas para indicar cómo navegar nuestro sitio.
- Ejemplo:

```
User-agent: *
# let's keep the robots away from our half-baked stuff
Disallow: /staging
```

Motores de búsqueda

- Cómo mejorar el orden en el que aparecemos en los resultados de búsqueda?: SEO
- Cosas a tener en cuenta: estructura del HTML, URLs, manejo de status codes, y un gran etc.
- Tips: Google Search Engine Optimization Starter Guide
- White hat .vs black hat SEO
- Son buenas prácticas, no hay garantías!

Búsqueda

UCSE - SEIA