Ingeniería Informática

FACULTAD DE INGENIERIA
UNIVERSIDAD NACIONAL DE JUJUY

OPTATIVA

Recuperación Avanzada de Información

Dr. J. Federico Medrano
@jfedemedrano

Unidad N° 2 – Parte 2

Temas a desarrollar

- ~~*Diseño de un spider/crawler. Estrategias.*~~
- ~~*Web scraping.*~~
- ~~*Recolección y representación estructural de sitios web/dominios.*~~
- ~~*Tratamiento de grafos.*~~
- Minería de textos. Minería de datos.
- Recolección de repositorios. Protocolos de recuperación.
- Diseño de un motor de búsqueda. Diseño de un indexador de documentos.
- Optimización en motores de búsqueda web (SEO/SEM)

Minería de textos. Minería de datos

Minería de datos

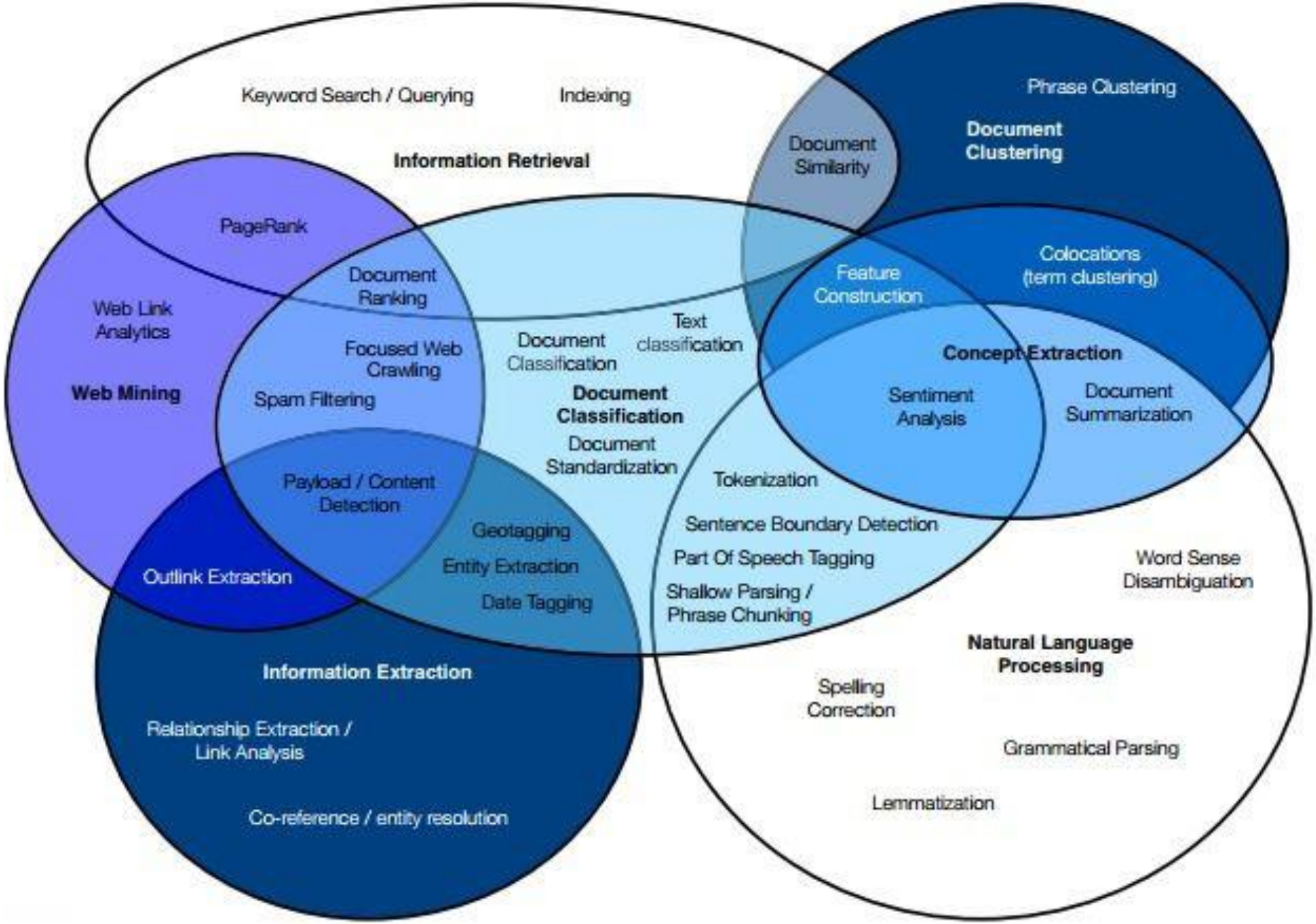
- El **Data Mining es un conjunto de técnicas y tecnologías** que permiten explorar grandes bases de datos, de manera *automática o semiautomática*, con el objetivo de encontrar patrones repetitivos que expliquen el comportamiento de estos datos.
- La minería de datos surgió con la **intención o el objetivo de ayudar a comprender una enorme cantidad de datos**, y que estos, pudieran ser utilizados para extraer conclusiones para contribuir en la mejora y crecimiento de las empresas, sobre todo, por lo que hace a las ventas o fidelización de clientes.
- Su intención es la de aportar información valiosa a las empresas para así, ayudarlas en la toma de decisiones futuras.

Minería de datos

- Es un mecanismo de explotación, consistente en la **búsqueda de información valiosa en grandes volúmenes de datos**. Está muy relacionada con las bodegas de datos que proporcionan la información histórica con la cual los algoritmos de minería tienen la información necesaria para la toma de decisiones.
- Según *Fallad* y sus coautores (1996): "La minería de datos es un proceso no trivial de **identificación** válida, novedosa, potencialmente útil y entendible **de patrones** comprensibles que se encuentran **ocultos en los datos**".
- Según *Molina* y sus colaboradores (2001): "Es la integración de un conjunto de áreas que tienen como propósito la **identificación** de un **conocimiento** obtenido a partir de las bases de datos que aporten un sesgo hacia la toma de decisión".

Minería de textos

- Debido a que el 80% de la información de una compañía está almacenada en forma de documentos, las técnicas como la categorización de texto, el procesamiento de lenguaje natural, la extracción y recuperación de la información o el aprendizaje automático, entre otras, apoyan al Text Mining (minería de texto).
- En ocasiones se confunde el Text Mining con la recuperación de la información (Information Retrieval o IR) (Hearst, 1999). Esta última consiste en la recuperación automática de documentos relevantes mediante indizaciones de textos, clasificación, categorización, etcétera. Generalmente, se utilizan palabras clave para encontrar una página relevante.
- En cambio, el Text Mining se refiere al examen de una colección de documentos y el **descubrimiento de información no contenida en ningún documento** individual de la colección; en otras palabras, trata de obtener información sin haber partido de algo. (Nasukawa y otros, 2001).


Keyword Search / Querying
Indexing
Information Retrieval
Phrase Clustering
Document Clustering
Document Similarity
PageRank
Web Link Analytics
Web Mining
Document Ranking
Focused Web Crawling
Spam Filtering
Document Classification
Text classification
Document Classification
Document Standardization
Feature Construction
Colocations (term clustering)
Concept Extraction
Sentiment Analysis
Document Summarization
Payload / Content Detection
Geotagging
Entity Extraction
Date Tagging
Outlink Extraction
Tokenization
Sentence Boundary Detection
Part Of Speech Tagging
Shallow Parsing / Phrase Chunking
Word Sense Disambiguation
Natural Language Processing
Spelling Correction
Grammatical Parsing
Lemmatization
Information Extraction
Relationship Extraction / Link Analysis
Co-reference / entity resolution

Recolección de repositorios.

Protocolos de recuperación

Interoperabilidad de los sistemas de información

- *Interoperabilidad.* Esfuerzo requerido para acoplar un sistema con otro (Pressman, 2010).
- *Interoperabilidad* es la posibilidad de que distintos tipos de ordenadores, redes, sistemas operativos, y aplicaciones trabajen juntos de forma eficaz, sin comunicación previa, de tal forma que puedan intercambiar información de manera útil y con sentido (Gomez, 2007)

Niveles de interoperabilidad

- **Sintáctica (De Giusti)**
- Hace referencia a todo lo necesario para que dos sistemas sean capaces de establecer una comunicación e intercambiar información.
- Esto incluye:
  - protocolos de comunicación y transferencia
  - codificación de caracteres
  - formatos de datos

Niveles de interoperabilidad

Sintáctica

- Elementos que corresponden a la interoperabilidad sintáctica pueden ser, por ejemplo:
  - protocolo TCP/IP
  - protocolo HTTP
  - **protocolo OAI-PMH**
  - formato XML y esquemas XML (XSD)
  - Directrices de interoperabilidad

Niveles de interoperabilidad

Semántica (De Giusti)

- Hace referencia a todo lo necesario para que el sistema receptor haga una correcta interpretación de la información recibida, de forma automática.
- Se busca que el sistema receptor "**entienda**" los datos tal como los "**entiende**" el emisor.
- ***Para contar con interoperabilidad semántica, primero debe asegurarse la interoperabilidad sintáctica***

Niveles de interoperabilidad

Semántica

Entran en juego:

- Formatos de metadatos
- Vocabularios controlados:
  - ❑Tesauros
  - ❑Sistemas de clasificación
- Ontologías
- Directrices de interoperabilidad

¿Cómo lograr que dos sistemas que no se conocen puedan comunicarse/interactuar?

La respuesta es ***hablar el mismo idioma***, en este sentido, se debe implementar un protocolo de comunicación en común (intercambio de información).

Protocolos de intercambio de información

Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)

- Establece un conjunto de reglas a partir de las cuales puede realizarse el intercambio de recursos de forma exitosa.
- Se centra en la **transferencia** de metadatos de un extremo a otro, sin establecer restricciones en cuanto a los datos que se transfieren.

OAI-PMH

Define dos perfiles de trabajo

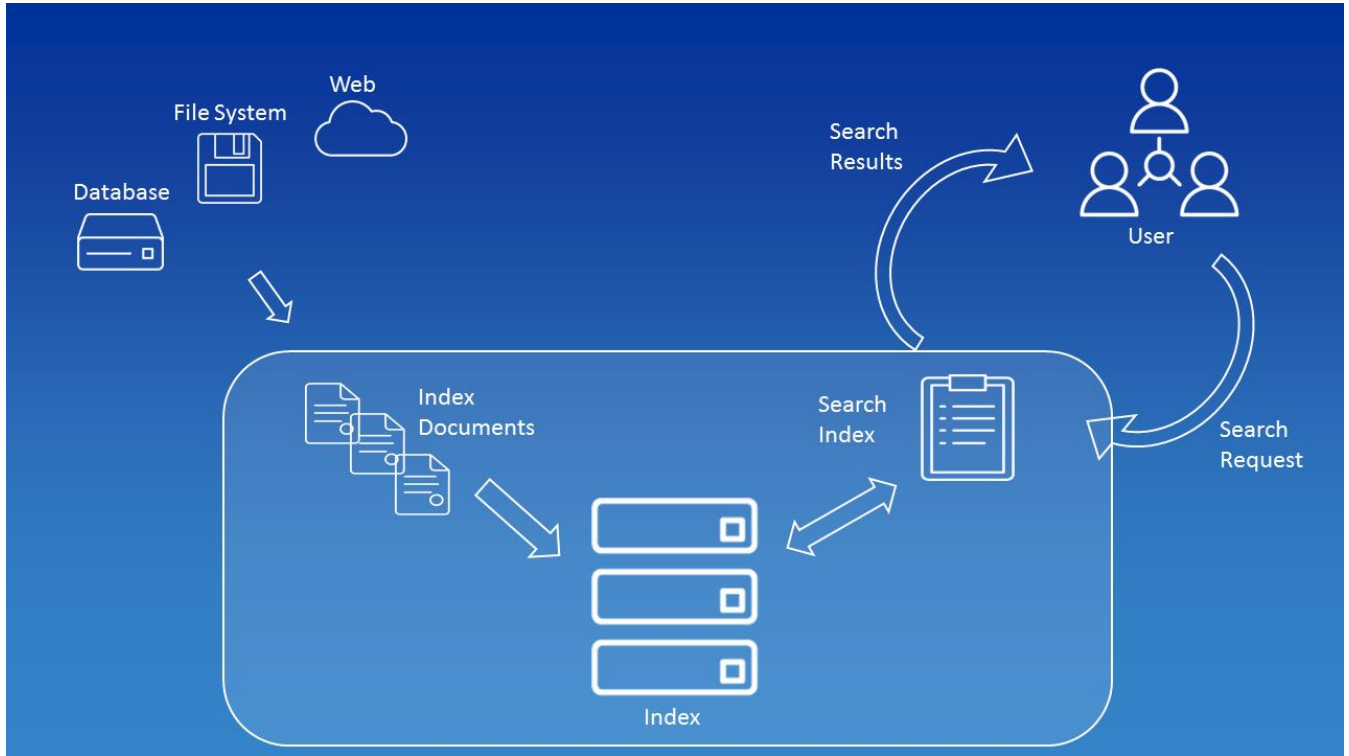
- **Data Provider:** es aquél repositorio que ofrece sus recursos bajo el protocolo OAI-PMH, para que otros los recolecten mediante cosechas.
- **Service Provider:** es aquél que recolecta recursos desde distintos Data Providers y brinda un servicio a una comunidad de usuarios en base a los recursos recolectados y el valor agregado aportado sobre los mismos (deduplicación, normalización, ordenamiento, búsquedas, etc).

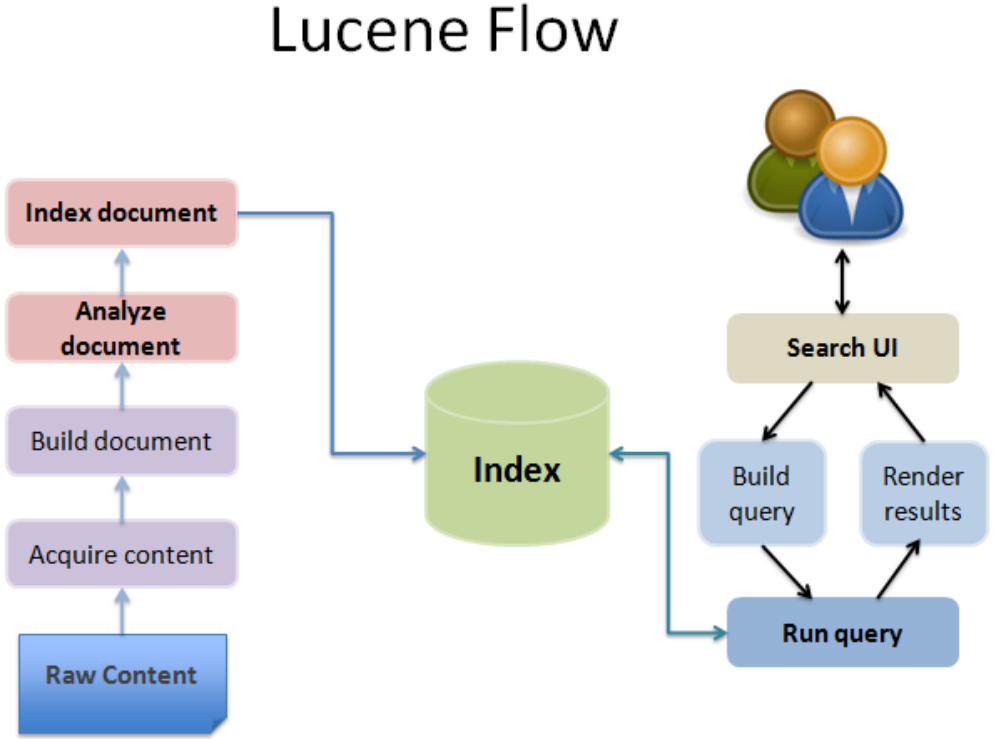
Diseño de un motor de búsqueda. Diseño de un indexador de documentos.

Lucene

- https://lucenenet.apache.org/ (C#)
- https://lucene.apache.org/ (Java)
- https://lucene.apache.org/pylucene/ (Python)

Apache Lucene es una API de código abierto para recuperación de información. Es útil para cualquier aplicación que requiera indexado y búsqueda a texto completo.





EJEMPLOS

- https://www.youtube.com/watch?v=Lv2wJRvSddw
- https://www.youtube.com/watch?v=TnYDeJkULKc
- https://howtodoinjava.com/lucene/lucene-index-search-examples/
- https://www.baeldung.com/lucene
- https://pythonhosted.org/lupyne/examples.html

Optimización en motores de búsqueda web (SEO/SEM)

SEM (*Search Engine Marketing*)

- El **concepto de SEM** (*Search Engine Marketing*) se refiere a la promoción de un sitio web en los buscadores mediante el uso de anuncios de pago a través de plataformas como **Google Ads** o **Bing Ads**. Y en ocasiones, se amplía este concepto para referirnos también a otro tipo de publicidad mediante estas y otras plataformas de display y medios sociales, donde se suele pagar generalmente en base a los clics que nos generan los anuncios.
- Mediante esta estrategia el objetivo es dar visibilidad inmediata a nuestro sitio Web, ya que desde que configuramos las campañas y pujamos por salir, nuestros anuncios tienen la posibilidad de aparecer.

SEO (*Search Engine Optimization*)

- El concepto de **SEO** (*Search Engine Optimization*) se refiere al trabajo de optimización y de aumento de la popularidad de un sitio web, con el objetivo de que dicho sitio sea **rastreable** por los motores de búsqueda, **indexado correctamente** y suficientemente relevante para que algunas o muchas de las páginas sean mostradas en las **primeras posiciones** de los buscadores para determinadas consultas de búsqueda de los usuarios.
- Por lo tanto, se trata de conseguir aparecer en los primeros resultados (lo ideal es en la primera pagina, y a ser posible, en las cinco primeras posiciones) de un buscador para un conjunto de búsquedas que nos interesan, pero sin tener que pagar un coste directo publicitario por cada visita, gracias a que somos muy relevantes y/o populares.

Web Spam

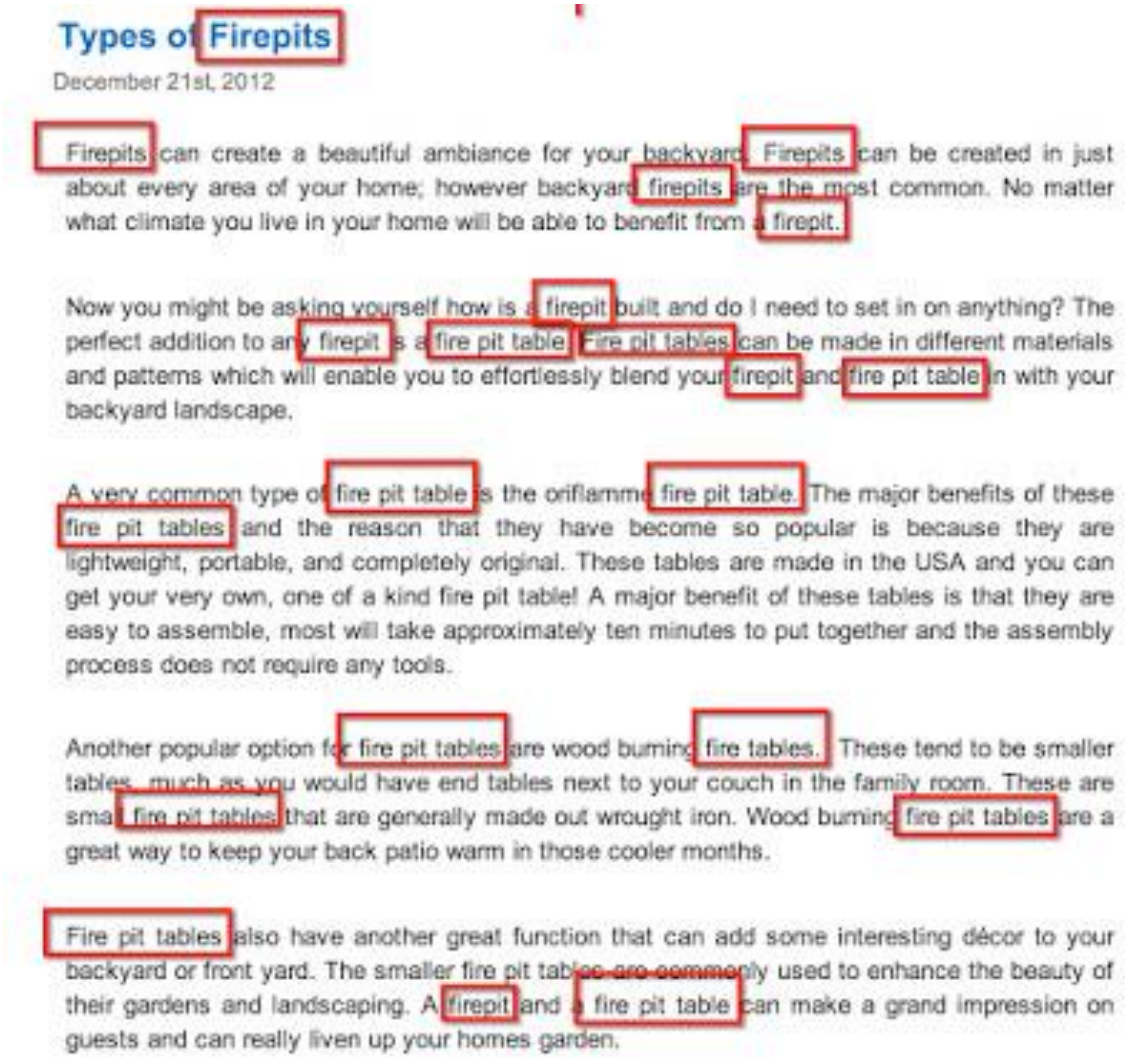
- Técnicas de SEO *Black Hat* o Sombrero Negro.
- Una práctica para conseguir una posición elevada, pero injustificada, en los resultados de los motores de búsqueda, utilizando técnicas para engañar a los algoritmos de clasificación.
- **Spamdexing** es una práctica poco ética de SEO que puede terminar con penalizaciones de los motores de búsqueda. Esto significa que los sitios web que practican spamdexing incluyen una gran cantidad de información en las páginas con el fin de indexarse y posicionarse más arriba en el resultado de la página del motor de búsqueda, pero no proporcionan información relevante o cualitativa para los usuarios.

Spamdexing

- Agregado automático de BackLinks (Enlaces entrantes que apuntan a un determinado sitio web o página que proviene de otro sitio web.)
- Keyword stuffing
- Link farms
- Spam blog
- Cloaking
- Doorway page

*Puedes **comprar zapatos rojos** en nuestra tienda online. Si necesitas **comprar zapatos rojos** para una ocasión especial, aprovéchate ahora de nuestra oferta en **zapatos rojos**. No encontrarás **zapatos rojos** de tanta calidad a este precio.*

El Keyword Stuffing es una técnica de **Black Hat SEO** que **consiste en el uso excesivo de palabras clave dentro de un texto** con el objetivo mal enfocado de darle más relevancia a esta palabra. Google penaliza con mucha frecuencia este tipo de sobre-optimización.



Types of Firepits

December 21st, 2012

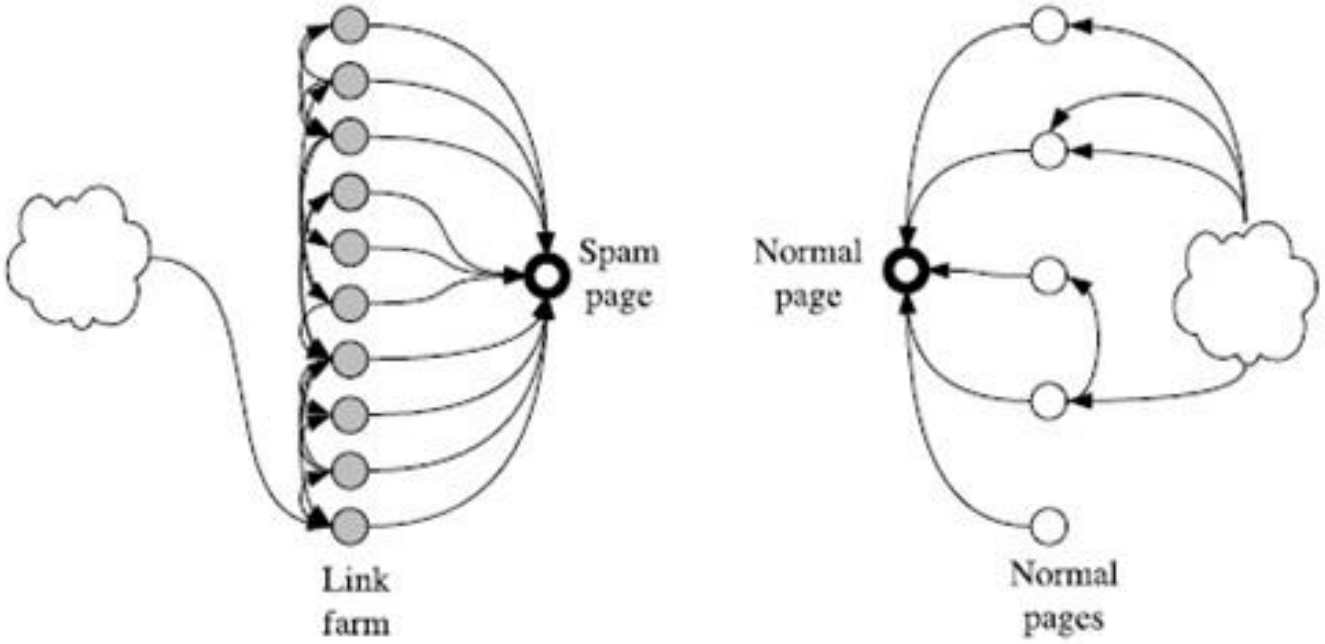
Firepits can create a beautiful ambiance for your backyard. Firepits can be created in just about every area of your home; however backyard firepits are the most common. No matter what climate you live in your home will be able to benefit from a firepit.

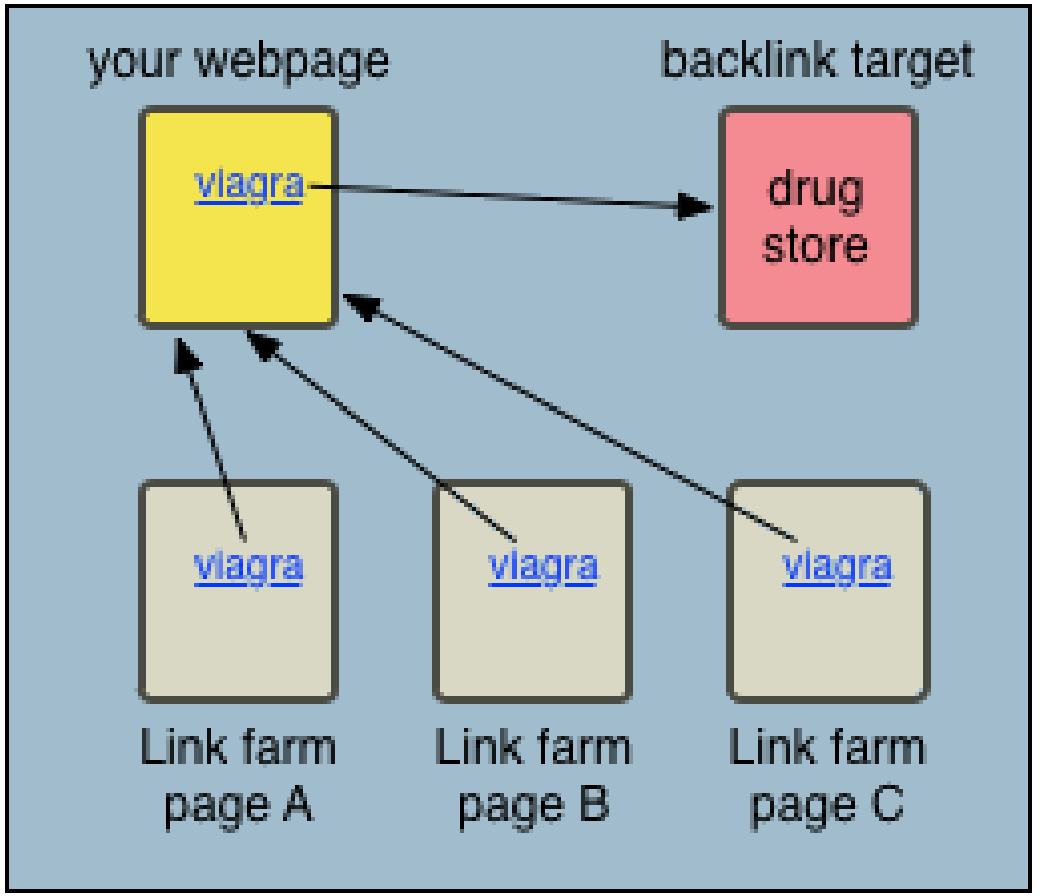
Now you might be asking yourself how is a firepit built and do I need to set in on anything? The perfect addition to any firepit is a fire pit table. Fire pit tables can be made in different materials and patterns which will enable you to effortlessly blend your firepit and fire pit table in with your backyard landscape.

A very common type of fire pit table is the oriflamme fire pit table. The major benefits of these fire pit tables and the reason that they have become so popular is because they are lightweight, portable, and completely original. These tables are made in the USA and you can get your very own, one of a kind fire pit table! A major benefit of these tables is that they are easy to assemble, most will take approximately ten minutes to put together and the assembly process does not require any tools.

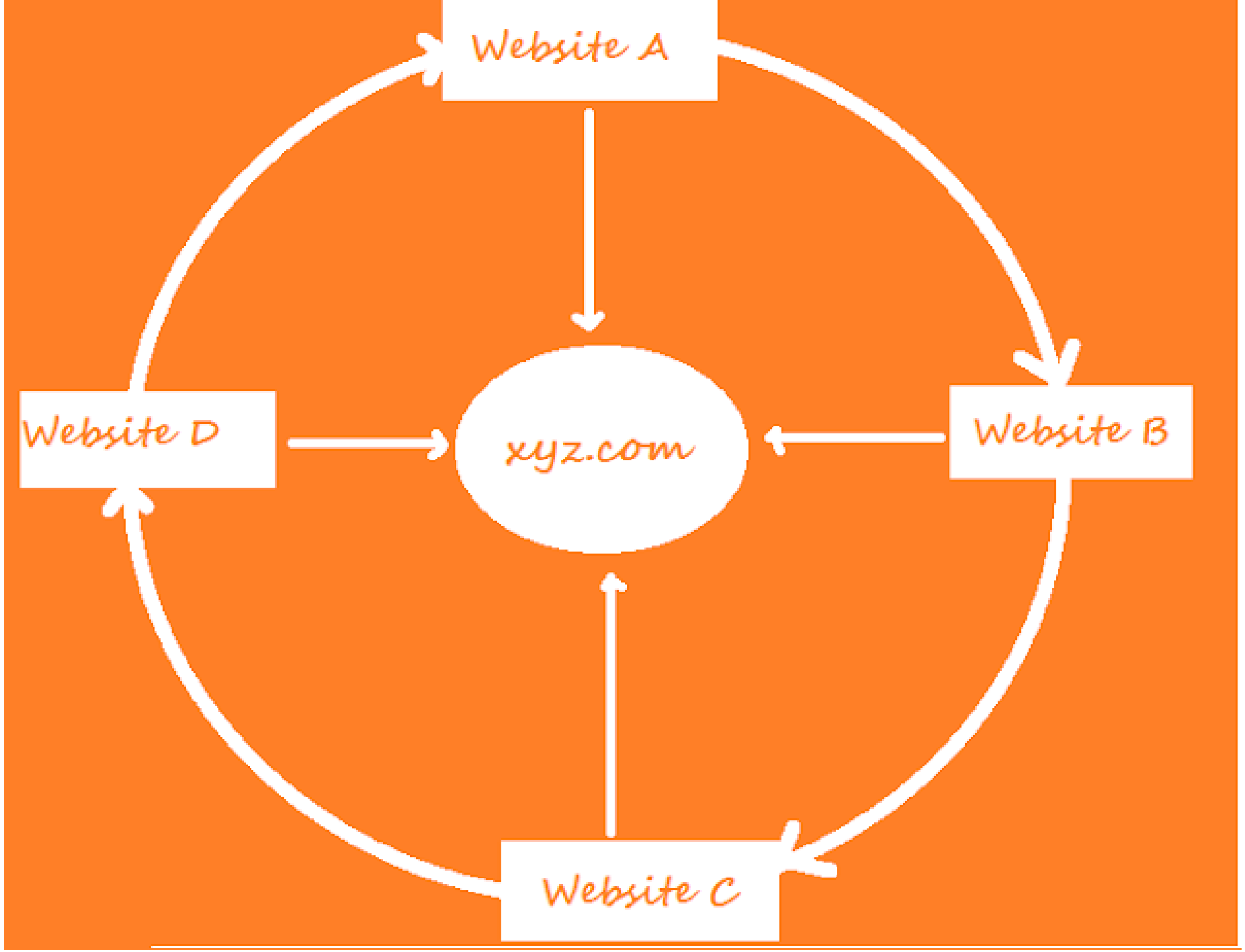
Another popular option for fire pit tables are wood burning fire tables. These tend to be smaller tables, much as you would have end tables next to your couch in the family room. These are small fire pit tables that are generally made out wrought iron. Wood burning fire pit tables are a great way to keep your back patio warm in those cooler months.

Fire pit tables also have another great function that can add some interesting décor to your backyard or front yard. The smaller fire pit tables are commonly used to enhance the beauty of their gardens and landscaping. A firepit and a fire pit table can make a grand impression on guests and can really liven up your homes garden.

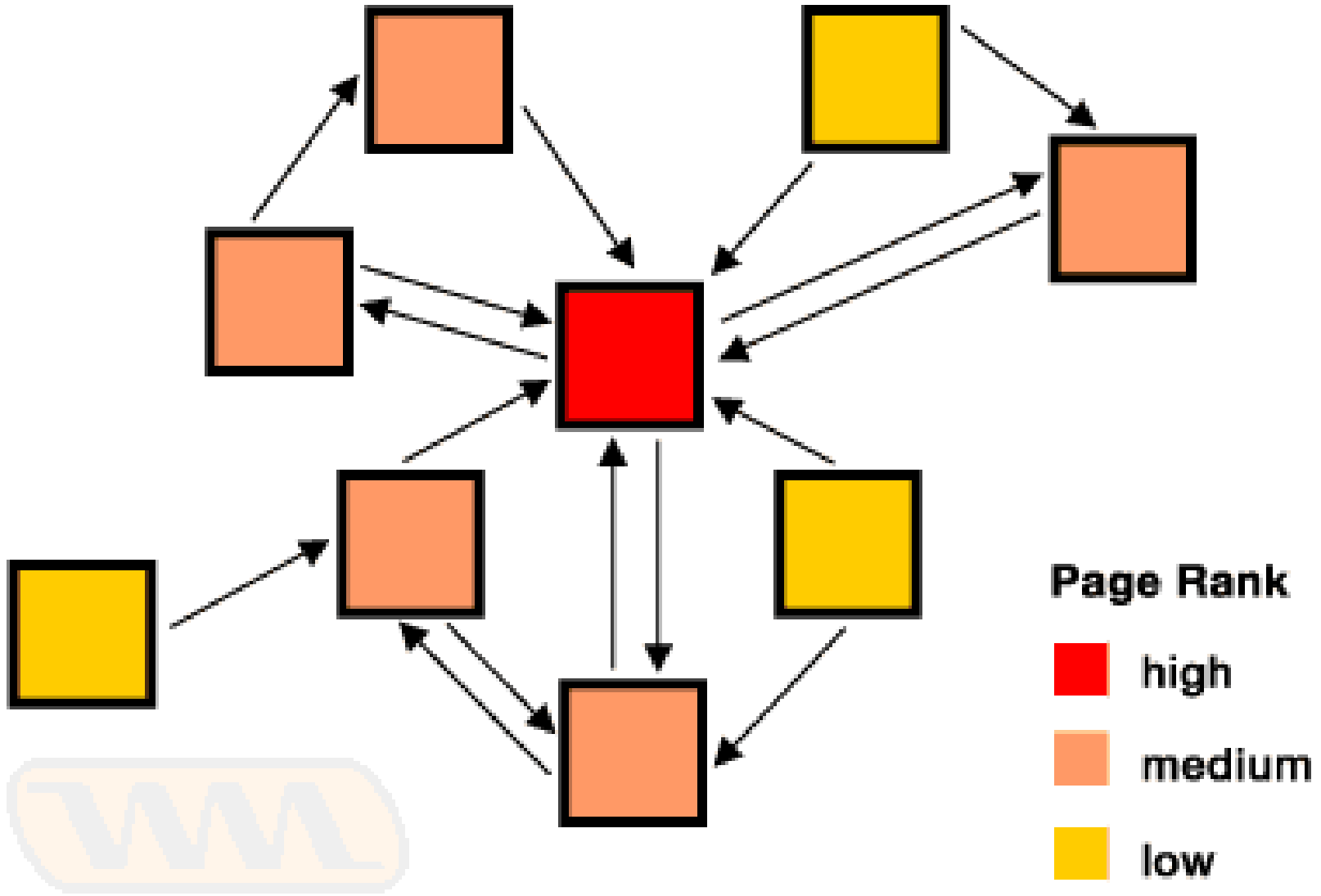




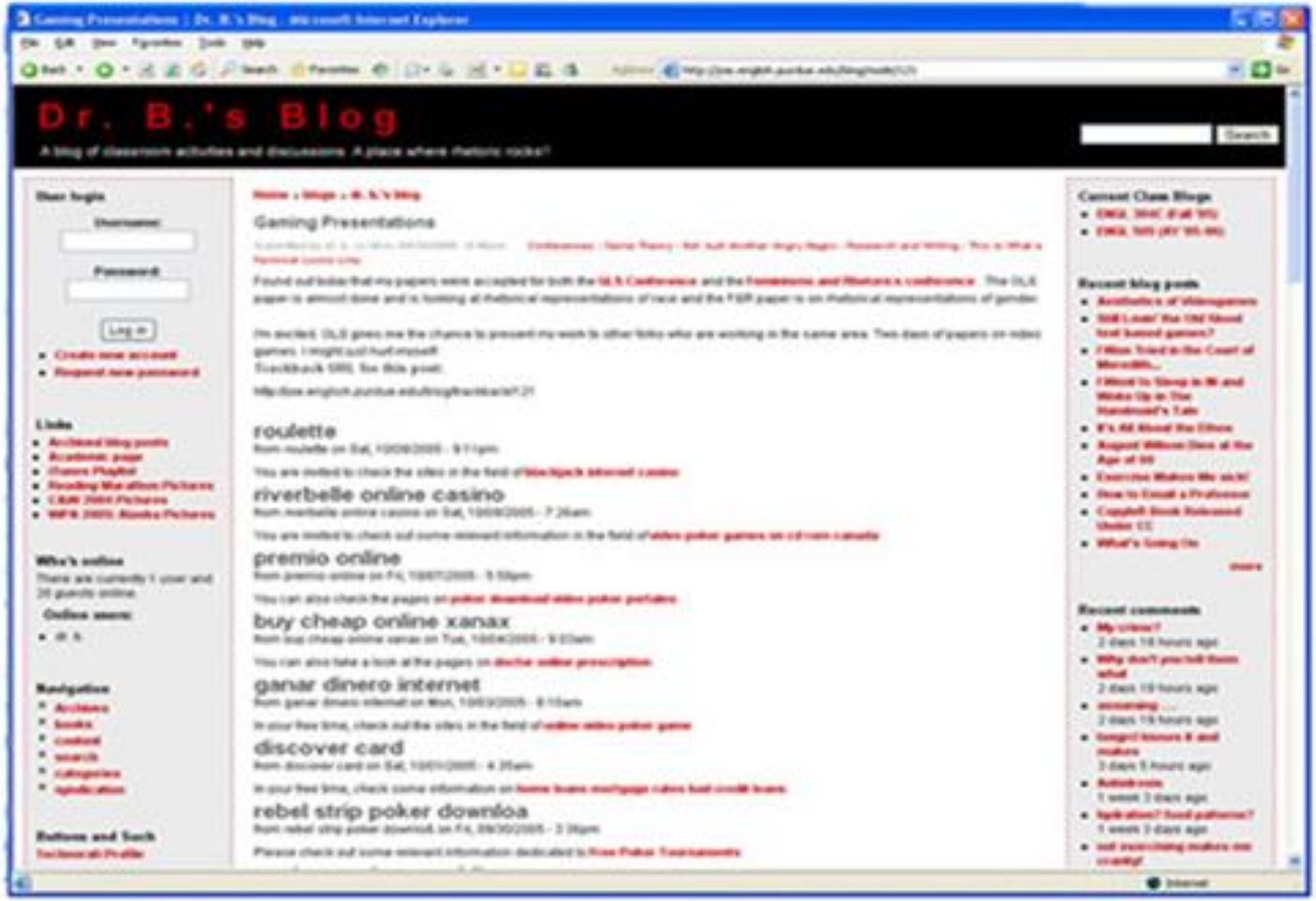
examples of Link farm backlinks


Website A
Website D
xyz.com
Website B
Website C

Links are an important factor in Page Rank



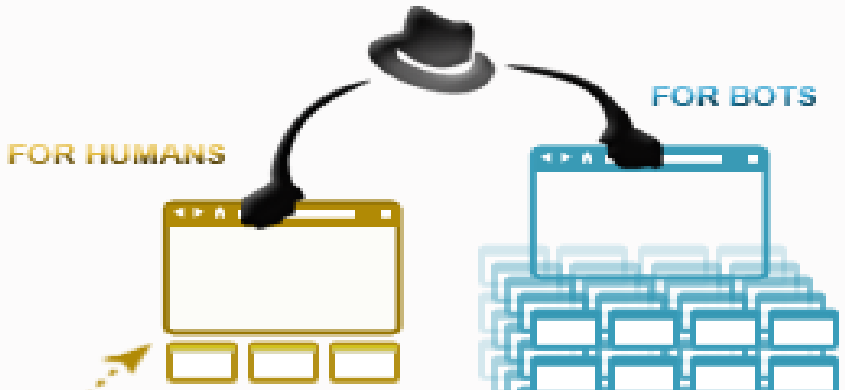
Web blog spam



Black Hat Cloaking Explained

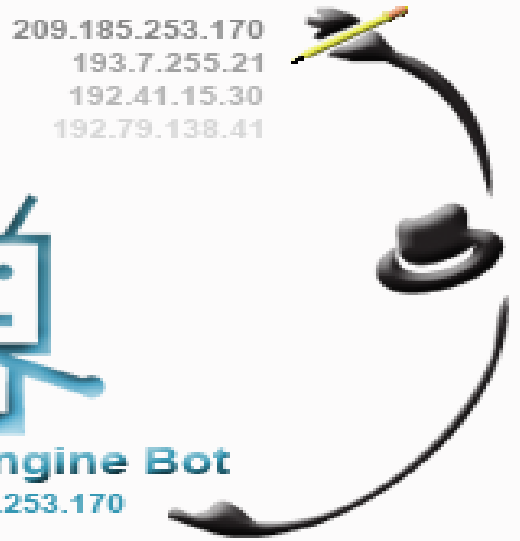
1

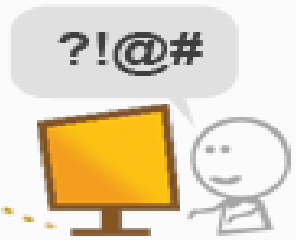
Sites engaged in black hat SEO prepare two sets of content, one targeted for bots and the other targeted for human visitors. Bots are identified by their IP address.



2

Since bot IP's can change, black hat informants provide a regularly updated list of bot IP addresses.



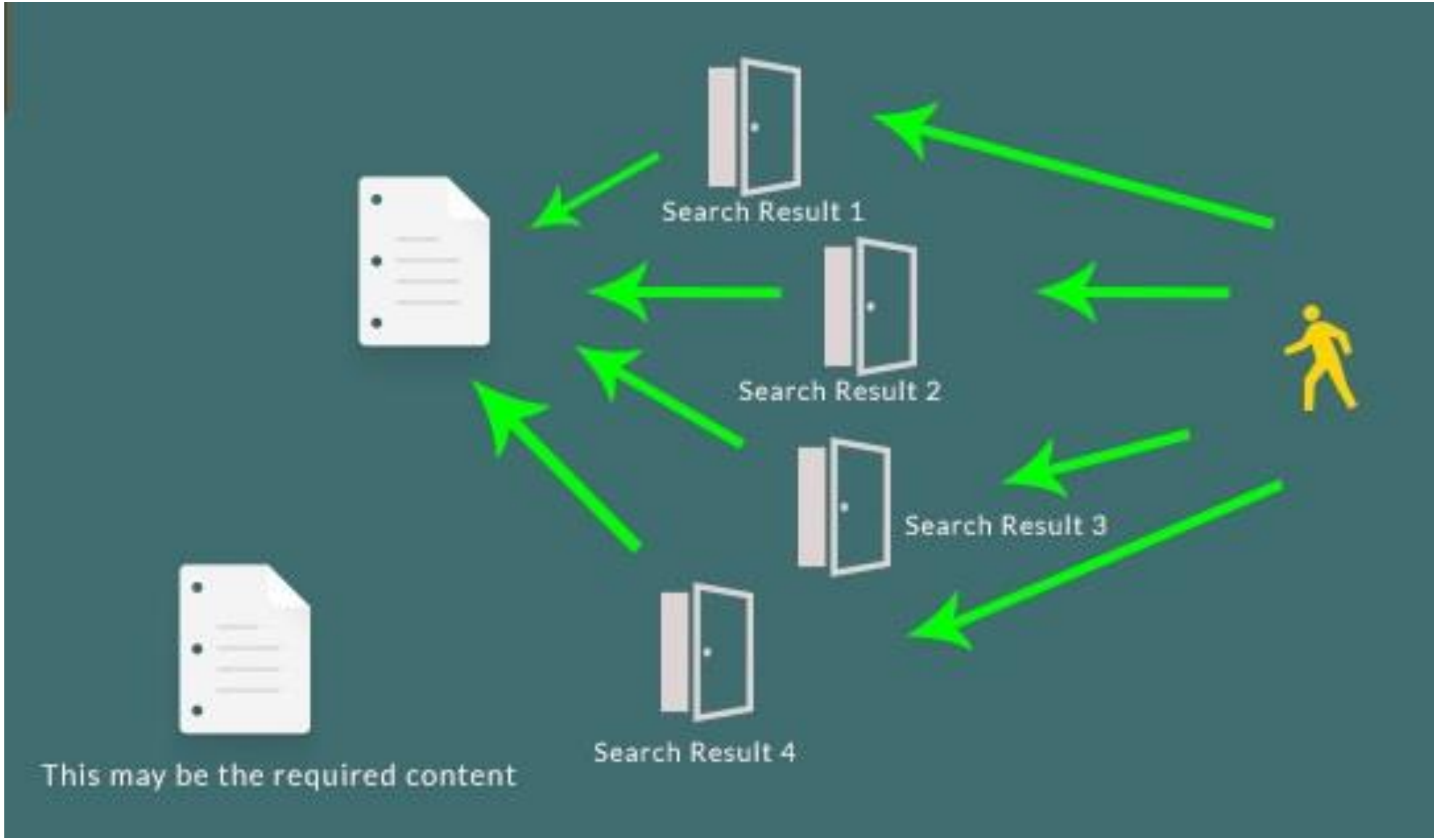


3

Bots are served abundant fabricated content packed with targeted keywords. This false information boosts rankings.

4

Human visitors often won't find the best information despite the site's high rankings.


Search Result 1
Search Result 2
Search Result 3
Search Result 4
This may be the required content

emagrecer depress
www.emagrecer-depressa.ja
www.emagrecer-rapido.ja
emagrecer sem esf
www.emagrecer-sem-esforco.ja
Fat Burner
perder peso rápido
www.perder-peso-rapido.ja
dietas milagrosas