

# How to build a search engine for the Web?

---

Louis Jachiet

## **What is a search engine?**

---

# What is a search engine ?

What is a search engine ?

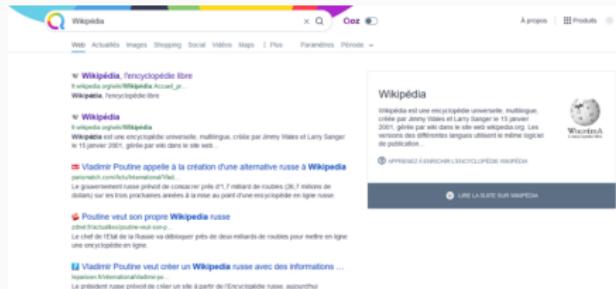
Do you know examples of search engines?

# Several flavors of search engines

- For the Web
  - e.g. *Bing, Duckduckgo, Google, Yandex*
- For a Website
  - e.g. *on Telecom website or on Youtube*
- For a collection of documents
  - e.g. *Google Scholar*
- For books
  - e.g. *a library*

# Commonalities to search engines

- A set of “entities” (often called **documents**) to be searched  
*Webpages, Videos, Books, Scientific Articles, etc.*
- A **query language**  
*Contains X and Y, Natural language, Author/Title/Type/ISBN, etc.*
- The output is a list of **selected** documents with a **ranking**



# Reminder on HTML

---

- Hypertext documents link to other documents

# Reminder on HTML

- Hypertext documents link to other documents
- together these links forms a directed graph: the Web!

# Reminder on HTML

- Hypertext documents link to other documents
- together these links forms a directed graph: the Web!
- Web browsing also uses other forms of links (forms, JS, etc.)

# Specificity of a Web search engine

---

- The documents are mostly **hypertext** and form a **graph**
- Search engines are not supposed to be **normative** on documents
- This graph can be **manipulated** by anyone
- The Web is too **large** ( $6 \times 10^{10}$  webpages) and too rapidly **evolving** to be manually curated
- Query languages should stay simple

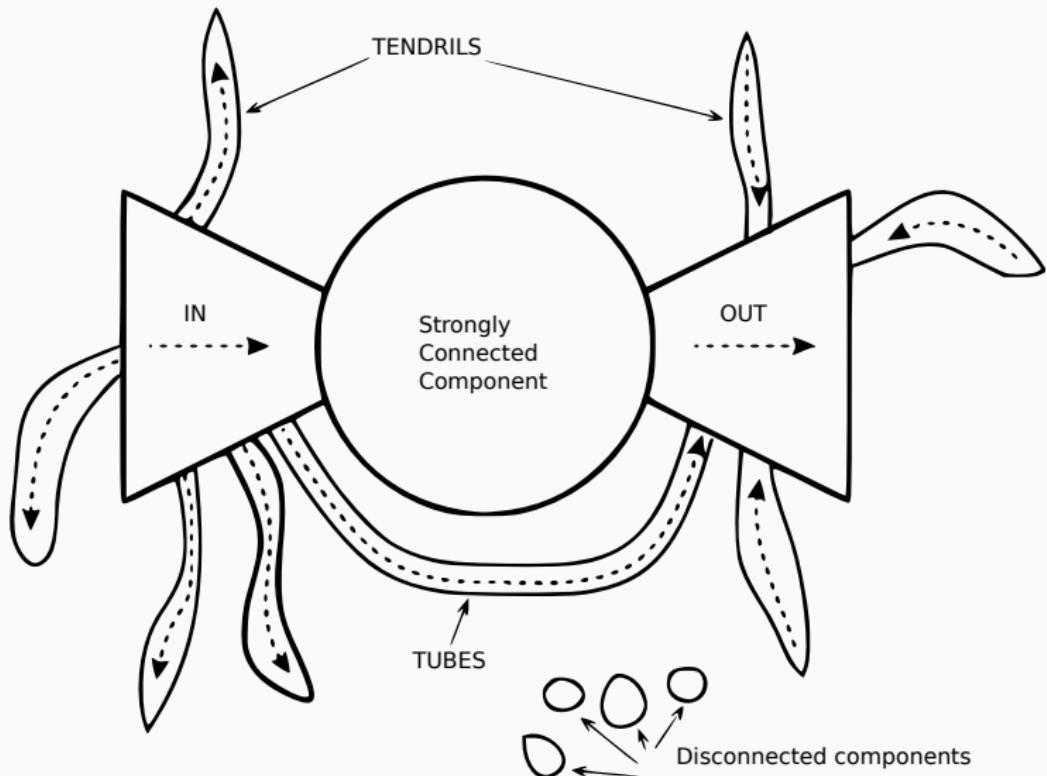
# Comparison of search engine types

Name	Shape	Size	Metadata information	Evolution	Personalized
Web	Graph	$10^{10}$	Low	Fast	Somewhat
Website	Items	$10^4$	Low	Slow	No
Youtube	Videos	$5 \times 10^9$	High	No	Yes
G. Scholar	Graph	$75 \times 10^6$	High	No	No
Books	Items	$10^8$	High	No	Maybe

## **Retrieve the Web**

---

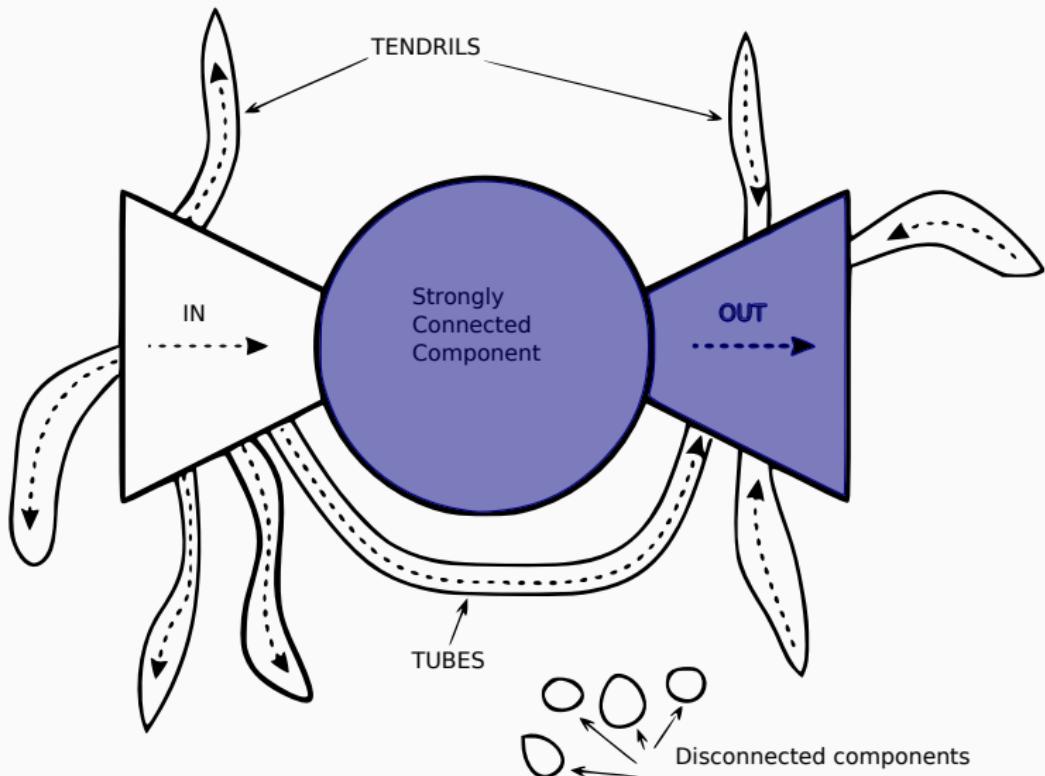
# Bowtie Structure of the Web



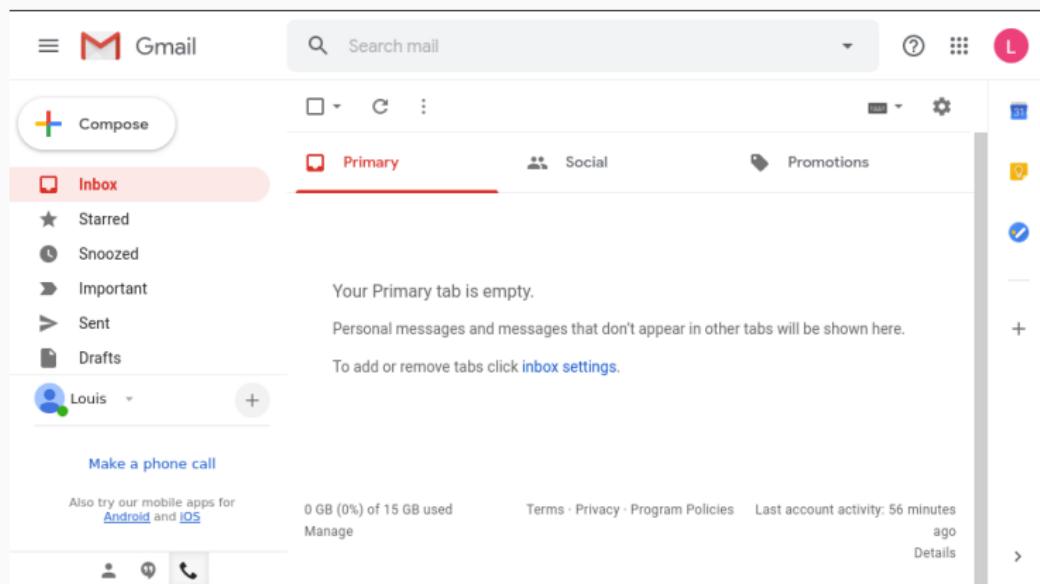
# Exploring the Web

```
todo = [start]
webpages = set(todo)
while len(todo):
    page = todo.pop()
    for nxt in neighbor(page):
        if nxt not in webpages:
            todo.append(nxt)
            webpages.insert(nxt)
```

# Explorable part of the Web



# The Deep Web, the unexplorable part of the Web



# The Deep Web, the unexplorable part of the Web



# The Deep Web, the unexplorable part of the Web

RU FORUM ▾ DOWNLOAD ▾ UPLOAD ▾ LAST ▾ OTHERS ▾ TOPICS ▾

## Library Genesis<sup>2M</sup>

Introducing Libgen Desktop application!  
Letter of Solidarity

Search in :

LibGen (Sci-Tech)  Scientific articles  Fiction  
 Comics  Standards  Magazines

LibGen Search options:

Download type: Resumed dl with original filename ▾

View results:  Simple  Detailed

Results per page: 25 ▾

Search with mask (word\*):  No  Yes

Search in fields:  The column set default  Title  Author(s)  Series  
 Publisher  Year  ISBN  Language  MD5  Tags  Extension

# The Deep Web, the unexplorable part of the Web



The screenshot shows the WolframAlpha search interface. At the top, there is a decorative banner with a landscape scene. Below it, the WolframAlpha logo is displayed with the tagline "computational intelligence". A search bar contains the input "int(x^2+x^7\*x^6+sin(2x))". Below the search bar are buttons for "Extended Keyboard" and "Upload", and links for "Examples" and "Random". A message box states "Assuming 'int' is an integral | Use as a math function or a function property instead". The main result section shows the "Indefinite integral:" followed by the mathematical expression  $\int [x^2 + x^7 \cdot x^6 + \sin(2x)] dx = \frac{1}{42} (3x^{14} + 14x^3 - 21\cos(2x)) + \text{constant}$ . There is also a checked checkbox for "Step-by-step solution". Below this, a "Plots of the integral:" section shows a graph with a single point at y=2.0.

# The Deep Web, the unexplorable part of the Web

**TV Calendar** [Share](#) [Visit](#)

**FILTER**  
**5**  
TOTAL SHOWS

**SHOWS**  
AIRING TODAY

**PROFILE**  
VIEW YOUR HISTORY

**NEWS**  
READ LATEST ARTICLES

SETTINGS  
UTC  
4:05PM

ACCOUNT  
SIGN UP PREMIUM



**« JUNE TV SCHEDULE** **JULY 3452 TV EPISODE CALENDAR** **AUGUST TV SCHEDULE »**

1 <sup>st</sup> Thursday	»	2 <sup>nd</sup> Friday	»	3 <sup>rd</sup> Saturday	»	4 <sup>th</sup> Sunday	»						
5 <sup>th</sup> Monday	»	6 <sup>th</sup> Tuesday	»	7 <sup>th</sup> Wednesday	»	8 <sup>th</sup> Thursday	»	9 <sup>th</sup> Friday	»	10 <sup>th</sup> Saturday	»	11 <sup>th</sup> Sunday	»
12 <sup>th</sup> Monday	»	13 <sup>th</sup> Tuesday	»	14 <sup>th</sup> Wednesday	»	15 <sup>th</sup> Thursday	»	16 <sup>th</sup> Friday	»	17 <sup>th</sup> Saturday	»	18 <sup>th</sup> Sunday	»
19 <sup>th</sup> Monday	»	20 <sup>th</sup> Tuesday	»	21 <sup>st</sup> Wednesday	»	22 <sup>nd</sup> Thursday	»	23 <sup>rd</sup> Friday	»	24 <sup>th</sup> Saturday	»	25 <sup>th</sup> Sunday	»
26 <sup>th</sup> Monday	»	27 <sup>th</sup> Tuesday	»	28 <sup>th</sup> Wednesday	»	29 <sup>th</sup> Thursday	»	30 <sup>th</sup> Friday	»	31 <sup>st</sup> Saturday	»		

**« JUNE TV SCHEDULE** **JULY 3452 TV EPISODE CALENDAR** **AUGUST TV SCHEDULE »**

**Currently Popular** **Recent Selections** **Recently Watched**

# Legal & Technical Aspects

---

## Technical aspects

- Risk of DDOS-ing a website

*Too frequent access or resource heavy webpages*

- Going through computer generated webpages

*July 3452 Calendar...*

- Bots browsing an app might generate a lot of noise

## Legal aspects

- Everything on the internet is at risk of being copied

*Beware of what you put on the internet!*

- A link on Google ≠ the webpage being stored by the NSA...

*Right to forget*

## Robots Exclusion Protocol (robots.txt)

---

```
Crawl-delay: 4
```

```
User-agent: *
```

```
Allow: *
```

```
User-agent: *
```

```
Disallow: /tmp/
```

```
User-agent: wget
```

```
Allow: /tmp/
```

```
User-agent: curl
```

```
Allow: /tmp/
```

```
User-agent: ia_archiver
```

```
Disallow: /
```

# Other Robots Exclusion Methods

## In the HTML

```
<meta name="robots" content="noindex">  
  <!-- prevent this page to be indexed -->
```

```
<meta name="robots" content="nofollow">  
  <!-- do not follow links in the webpage -->
```

## In the HTTP conversation

HTTP/1.1 200 OK

(...)

X-Robots-Tag: noindex

# Beyond simple crawling

## Extracting semantic information from crawling

- from online phonebooks
- from shopping websites
- from restaurants
- ...

## Using it for semantic answering

Google

when simone de beauvoir died

All Images News Videos Shopping More Settings Tools

About 2,840,000 results (0.67 seconds)

Simone de Beauvoir / Date of death

April 14, 1986

People also search for

Jean-Paul Sartre April 15, 1980	Albert Camus January 4, 1950	Betty Friedan February 4, 2006
------------------------------------	---------------------------------	-----------------------------------

# **Information Retrieval**

---

We have collected billions of  
webpages...

... how to get the relevant ones?

## Which query language for a Web search engine?

- Simple

*but with advanced features*

- very Selective

*The Web has billion of webpages*

## Easy solution: bag of words

- We compute solutions for each word
- Partial solutions are merged
- We remove stop words

*This allows for an English-like querying*

## Example

---

What is a matrix multiplication?

## Example

---

What is a matrix multiplication?

## Example

What is a matrix multiplication?

A screenshot of a Google search results page. The search query "what is a matrix multiplication" is entered in the search bar. Below the search bar, there are navigation links for All, Images, Videos, News, Shopping, More, Settings, and Tools. A message indicates "About 46,300,000 results (0.42 seconds)". The top result is a link to the Wikipedia article on Matrix multiplication. The snippet from the Wikipedia page defines matrix multiplication as a binary operation that produces a matrix from two matrices. It also mentions the result matrix, known as the matrix product, has the number of rows of the first and the number of columns of the second matrix. Below the snippet are links to "Definition", "Fundamental applications", and "General properties". A "Square matrices" link is also present. At the bottom of the search results, there is a "People also ask" section with three collapsed questions: "What does matrix multiplication mean?", "How multiplication of matrix is done?", and "How do you multiply 2x2 matrices?".

what is a matrix multiplication

All Images Videos News Shopping More Settings Tools

About 46,300,000 results (0.42 seconds)

en.wikipedia.org › wiki › Matrix\_multiplication ▾

**Matrix multiplication - Wikipedia**

In mathematics, **matrix multiplication** is a binary operation that produces a **matrix** from two **matrices**. ... The result matrix, known as the **matrix product**, has the number of rows of the first and the number of columns of the second **matrix**.

[Definition](#) · [Fundamental applications](#) · [General properties](#) · [Square matrices](#)

People also ask

What does matrix multiplication mean? ▾

How multiplication of matrix is done? ▾

How do you multiply 2x2 matrices? ▾

## Example

---

What is a matrix and what is its multiplication?

## Example

---

What is a matrix and what is its multiplication?

## Example

What is a matrix and what is its multiplication?

A screenshot of a Google search results page. The search query "what is a matrix and what is its multiplication" is entered in the search bar. Below the search bar, there are navigation links for All, Images, Videos, News, Shopping, More, Settings, and Tools. A message indicates "About 160,000,000 results (0.64 seconds)". The top result is a link to "en.wikipedia.org › wiki › Matrix\_multiplication" titled "Matrix multiplication - Wikipedia". The snippet describes matrix multiplication as a binary operation that produces a matrix from two matrices, noting that the number of columns in the first matrix must be equal to the number of rows in the second matrix. Below the snippet are links to "Definition · Fundamental applications · General properties · Square matrices". A "People also ask" section is visible at the bottom, listing three questions: "What does matrix multiplication mean?", "How do you do matrix multiplication?", and "How do you multiply 2x2 matrices?".

what is a matrix and what is its multiplication

All Images Videos News Shopping More Settings Tools

About 160,000,000 results (0.64 seconds)

en.wikipedia.org › wiki › Matrix\_multiplication

**Matrix multiplication - Wikipedia**

In mathematics, **matrix multiplication** is a binary operation that produces a **matrix** from two **matrices**. For **matrix multiplication**, the number of columns in the first matrix must be equal to the number of rows in the second matrix.

Definition · Fundamental applications · General properties · Square matrices

People also ask

What does matrix multiplication mean? ▾

How do you do matrix multiplication? ▾

How do you multiply 2x2 matrices? ▾

Read [www.gstatic.com](http://www.gstatic.com)

## Example

---

What is a matrix and how do they multiply?

## Example

---

What is a matrix and how do they multiply?

# Example

What is a matrix and how do they multiply?

Google what is a matrix and how do they multiply

All Images Videos News Shopping More Settings Tools

About 192,000,000 results (0.58 seconds)

Multiplying Matrices

$$\begin{bmatrix} 3 & 4 \\ 7 & 2 \\ 5 & 9 \end{bmatrix} \times \begin{bmatrix} 3 & 1 & 5 \\ 8 & 7 & 4 \\ 6 & 9 & 7 \end{bmatrix} = \begin{bmatrix} \$8 \\ \$3x13 + \$4x \\ \$8 \end{bmatrix}$$

\$3x13 + \$4x

A B C

"Dot Prod"

In mathematics, **matrix multiplication** is a binary operation that produces a **matrix** from two **matrices**. For **matrix multiplication**, the number of columns in the first **matrix** must be equal to the number of rows in the second **matrix**.

[en.wikipedia.org/wiki/Matrix\\_multiplication](https://en.wikipedia.org/wiki/Matrix_multiplication)

**Matrix multiplication - Wikipedia**

About Featured Snippets Feedback

# Some Terminology for the Bag-of-word Model

---

## Term-document Matrix

A Matrix  $\mathcal{M}$  where  $\mathcal{M}_{w,d}1$  corresponds to the frequency at which  $w$  appears in document  $d$ .

## Inverted Index

A list  $(f_1, d_1), \dots$  for each word  $w$  stating that  $w$  appears with frequency  $f_1$  in  $d_1$ .

## Full Inverted Index

A list  $(f_1, o_1), \dots$  for each word  $w$  stating that  $w$  appears at positions  $o_1$  in  $d_1$ .

## The Boolean Model

- Each word of the query appears or not in the document
- A bag query can be ranked by the number of matches

## The Vector Space Model

- A document  $d$  is associated with a vector  $\vec{v}(d)$
- $\vec{v}(d)$ ; corresponds to the **weight** of term  $i$  in  $d$
- Ranking can be done using classical metrics (such as norm-2 or norm-1)

## The Probabilistic Model

- For each word  $w$  and each document  $d$  we compute  $P(d, w)$ , the probability that  $d$  talks about  $w$
- With independence assumption the probability of a document  $d$  being a match is  $\prod_w P(d, w)$

## Latent semantic analysis

- A low-rank decomposition of the term-document-matrix
- Distance between document  $d$  and bag  $b$  is dot-product.

## Term Frequency - Inverse Document Frequency (tf-idf)

Given a set of documents  $\mathcal{D}$ ,  $\text{tf-idf}_{\mathcal{D}}(w, d)$  computes for the **importance** of  $w$  within document  $d$  with respect to documents in  $\mathcal{D}$ .

Formally:

$$\text{tf-idf}_{\mathcal{D}}(w, d) = \text{tf}(w, d) \times \text{idf}(w, \mathcal{D})$$

... now  $tf$  comes in several flavors and so does  $idf$ !

# Term Frequency

With  $f_{w,d}$  the number of occurrences of  $w$  in  $d$ :

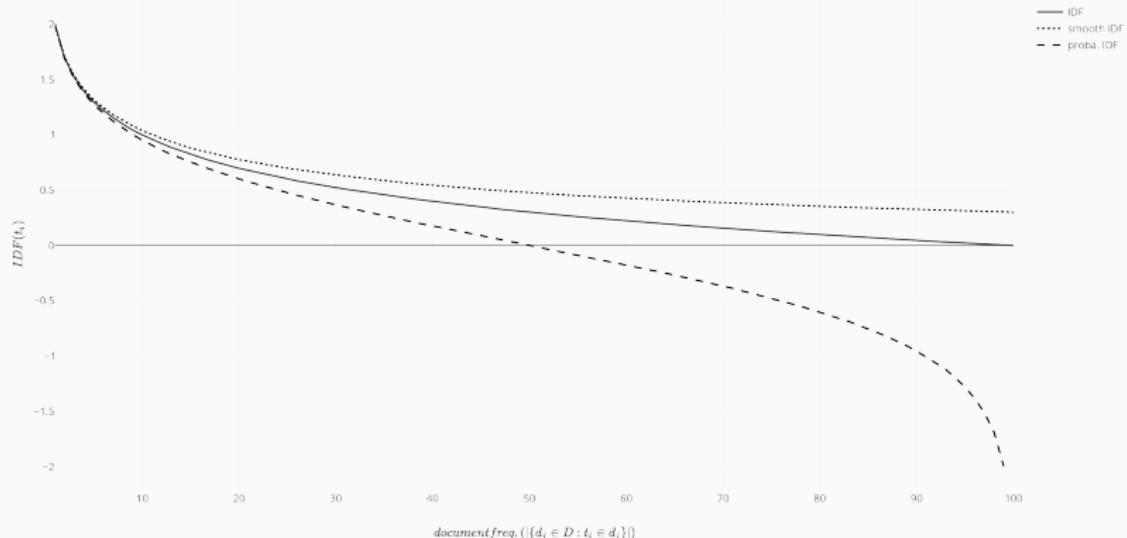
Weighting scheme	tf weight
binary	0 or 1
count	$f_{w,d}$
<b>term frequency</b>	$\frac{f_{w,d}}{\sum_{w'} f_{w',d}}$
log normalization	$\log(1 + f_{w,d})$
double normalization	$\alpha + (1 - \alpha) \times \frac{f_{w,d}}{\sum_{w'} f_{w',d}}$

# Inverse Document Frequency

With  $n_w$  the number of documents with word  $w$ :

Weighting scheme	idf weight
unary	1
<b>inverse document frequency</b>	$\log\left(\frac{N}{n_w}\right)$
smoothed idf	$\log\left(\frac{N}{1 + n_w}\right) + 1$
max-normalized	$\log\left(\frac{\max n_{w'}}{1 + n_w}\right)$
probabilistic idf	$\log\left(\frac{N - n_w}{n_w}\right)$

# Inverse Document Frequency



# Offline computations for tf-idf

---

## Term-document Matrix

flatMap  $d_i \rightarrow (i, d_i.split())$

## Inverted Index

groupByKey

## Inverse Document Frequency

countByKey

## Storage

Store the offline computation into an efficient key-value store. The vocabulary can often be stored in main memory (Heaps' Law:  $O(n^\beta)$  with  $\beta \sim 0.5$ ).

## Online computations for tf-idf

---

Given  $w_1 \dots w_k$ :

- look up  $idf(w_1), \dots, idf(w_k)$
- compute  $tf(w_i, d)$  for all  $d$  and all  $i$  using inverted index
- output the list of documents ordered by  
 $\sum_i tf(w_i, d) \times idf(w_i)$  (norm-1)

### Historically important: Okapi BM-25

$$\text{score}(d, W) = \sum_{i=1}^n \text{IDF}(w_i) \cdot \frac{f(w_i, d) \cdot (k_1 + 1)}{f(w_i, d) + k_1 \cdot \left(1 - b + b \cdot \frac{\text{len}(d)}{\text{avglen}}\right)}$$

with  $1.2 \leq k_1 \leq 2$  and  $b \sim 0.75$ .

# Persin's Algorithm

Ranking-in-the-vector-model( query terms  $t$  )

- (1) Create  $P$  as  $C$ -candidate similarities initialized to  $(P_d, P_w) = (0, 0)$
- (2) Sort the query terms  $t$  by decreasing weight
- (3)  $c \leftarrow 1$ .
- (4) **for** each sorted term  $t$  in the query **do** {
- (5)   Compute the value of the threshold  $t_{add}$ .
- (6)   Retrieve the inverted list for  $t$ ,  $L_t$ .
- (7)   **for** each document  $d$  in  $L_t$  **do** {
- (8)     **if**  $w_{d,t} < t_{add}$  **then break**
- (9)      $psim \leftarrow w_{d,t} \times w_{q,t}/W_d$ .
- (10)     **if**  $d \in P_d(i)$  **then**  $P_w(i) \leftarrow P_w(i) + psim$
- (11)     **elif**  $psim > min_j(P_w(j))$  **then**  $n \leftarrow min_j(P_w(j))$
- (12)     **elif**  $c \leq C$  **then** {
- (13)        $n \leftarrow c$
- (14)        $c \leftarrow c + 1$
- (15)     }
- (16)     **if**  $n \leq C$  **then**  $P(n) \leftarrow (d, psim)$
- }

- (16) **return** the top- $k$  documents according to  $P_w$

## Scan of Modern Information Retrieval

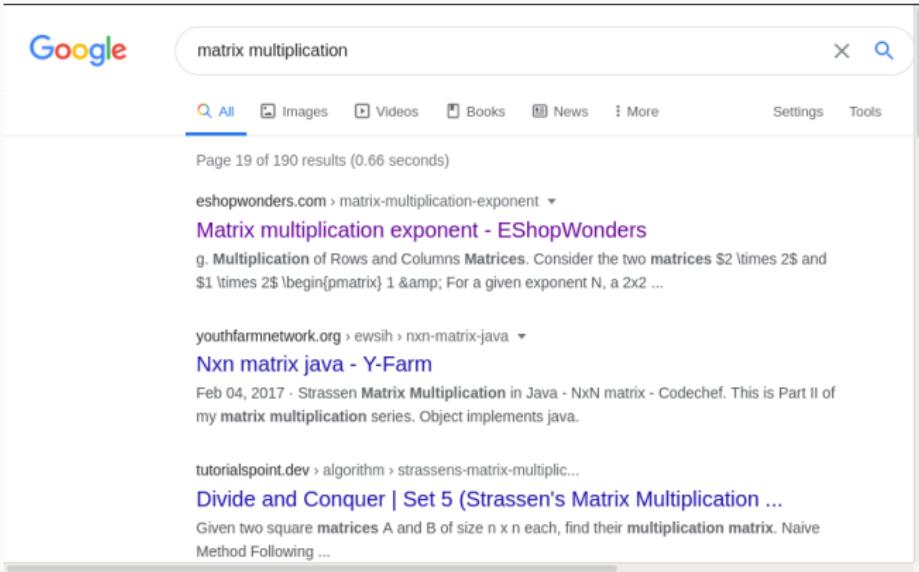
## **How to asses Website Quality?**

---

Anyone can publish content on the Web...

... how to find quality content?

# Problem



A screenshot of a Google search results page. The search query "matrix multiplication" is entered in the search bar. The results are displayed on Page 19 of 190, with a total search time of 0.66 seconds. The first result is a link to "eshopwonders.com" titled "Matrix multiplication exponent - EShopWonders". The second result is a link to "youthfarmnetwork.org" titled "NxN matrix java - Y-Farm". The third result is a link to "tutorialspoint.dev" titled "Divide and Conquer | Set 5 (Strassen's Matrix Multiplication ...)". Each result includes a snippet of text describing the content.

matrix multiplication

All Images Videos Books News More Settings Tools

Page 19 of 190 results (0.66 seconds)

eshopwonders.com > matrix-multiplication-exponent ▾

**Matrix multiplication exponent - EShopWonders**

g. Multiplication of Rows and Columns Matrices. Consider the two matrices  $2 \times 2$  and  $2 \times 2$   $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$  For a given exponent N, a  $2 \times 2$  ...

youthfarmnetwork.org > ewsiih > nxn-matrix-java ▾

**NxN matrix java - Y-Farm**

Feb 04, 2017 · Strassen Matrix Multiplication in Java - NxN matrix - Codechef. This is Part II of my matrix multiplication series. Object implements java.

tutorialspoint.dev > algorithm > strassens-matrix-multiplic...

**Divide and Conquer | Set 5 (Strassen's Matrix Multiplication ...**

Given two square matrices A and B of size  $n \times n$  each, find their multiplication matrix. Naive Method Following ...

## Early solutions

---

- Pre-defined set of authoritative content websites?
- Presence in online directories?
- Count incoming links?
- Count incoming links from authorities?

## General Idea

The idea originate from the time the web had a lot pages serving as directories (**hubs** here) and a lot of pages providing information (**authorities** here).

In this model a good reference on a topic should list top quality websites providing information.

## General Idea

The idea originate from the time the web had a lot pages serving as directories (**hubs** here) and a lot of pages providing information (**authorities** here).

In this model a good reference on a topic should list top quality websites providing information.

## Algorithm

- Retrieve (e.g. with tf-idf) a set of webpages called "root set of authorities" for the query
- Iterate
  - Determine a set of good hubs for those authorities
  - Determine a set of good authorities for those hubs

## Underlying Principles for PageRank

- hypertext documents tend to point to documents that are more important (or serious/credible/trustworthy) than themselves
- the more important a document is, the more important are the documents it points to

## Underlying Principles for PageRank

- hypertext documents tend to point to documents that are more important (or serious/credible/trustworthy) than themselves
- the more important a document is, the more important are the documents it points to

## General Idea

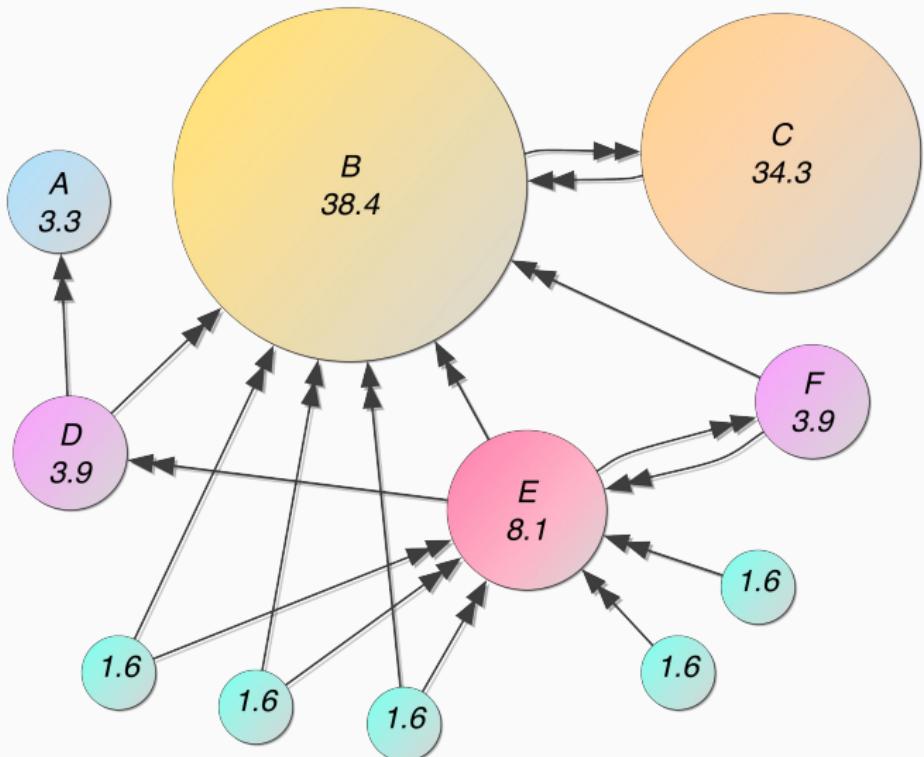
Compute the probability to arrive at some webpage after following a random walk that teleport randomly with probability  $\alpha$  and follow one of the link of the page with probability  $(1 - \alpha)$ .

## Algorithm

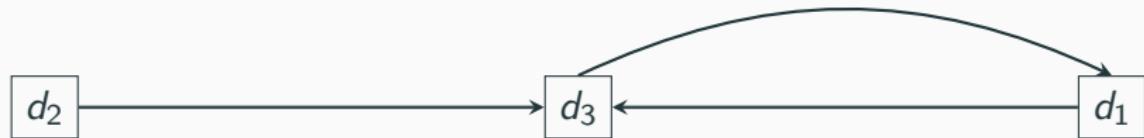
- Compute the matrix  $\mathcal{M}$  corresponding to the Markov chain
- Compute  $\mathcal{M}^k \times \langle 1 \dots 1 \rangle^T$  for  $k$  big enough

In practice this algorithm converges exponentially fast in  $1 - \alpha$  (a small number of iterations is enough).

# PageRank Example



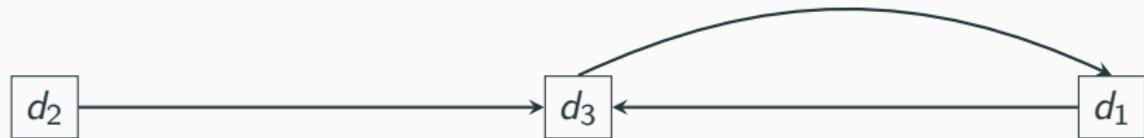
## Why $\alpha$ ? How to set it?



### Exercise

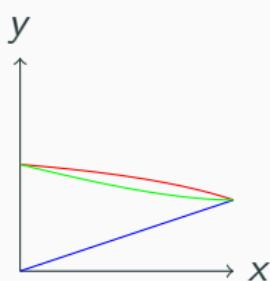
Compute PageRank of  $d_1, d_2, d_3$  for  $\alpha = 0, 1, \frac{1}{2}$ ?

# Why $\alpha$ ? How to set it?



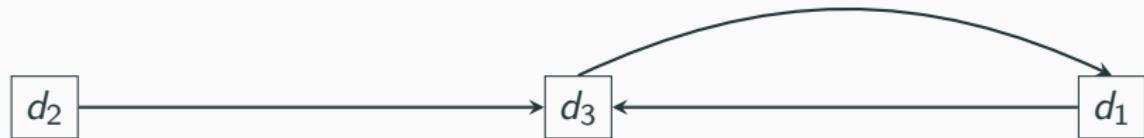
## Exercise

Compute PageRank of  $d_1, d_2, d_3$  for  $\alpha = 0, 1, \frac{1}{2}$ ?



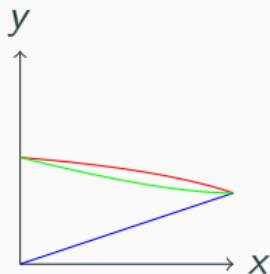
$$\begin{array}{c|c|c|c} \alpha & d_1 & d_2 & d_3 \\ \hline \hline \end{array}$$

# Why $\alpha$ ? How to set it?



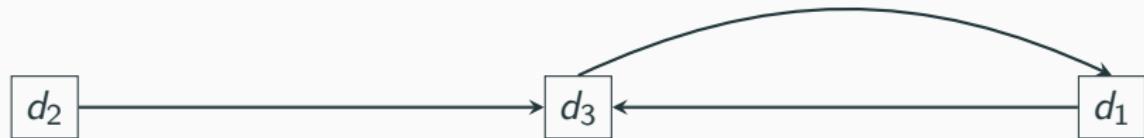
## Exercise

Compute PageRank of  $d_1, d_2, d_3$  for  $\alpha = 0, 1, \frac{1}{2}$ ?



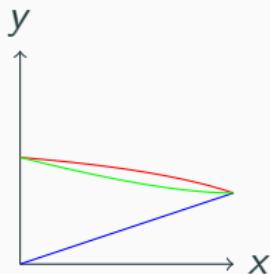
$\alpha$	$d_1$	$d_2$	$d_3$
0	1/2	1/2	0

# Why $\alpha$ ? How to set it?



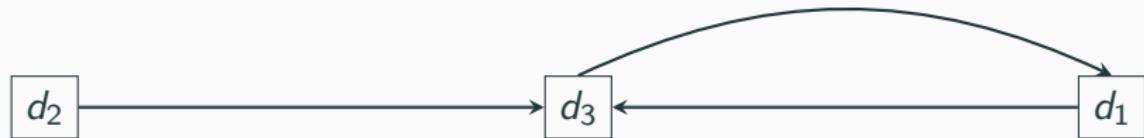
## Exercise

Compute PageRank of  $d_1, d_2, d_3$  for  $\alpha = 0, 1, \frac{1}{2}$ ?



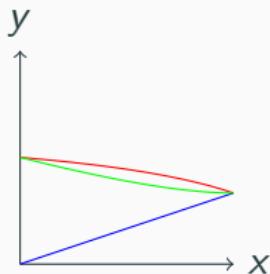
$\alpha$	$d_1$	$d_2$	$d_3$
0	1/2	1/2	0
1	1/3	1/3	1/3

# Why $\alpha$ ? How to set it?



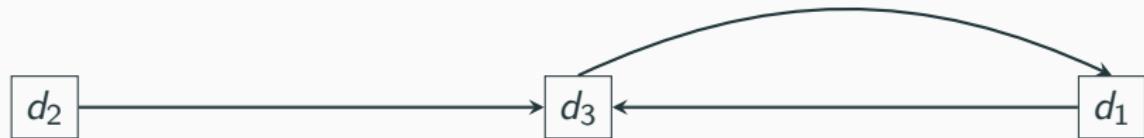
## Exercise

Compute PageRank of  $d_1, d_2, d_3$  for  $\alpha = 0, 1, \frac{1}{2}$ ?



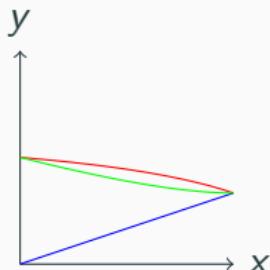
$\alpha$	$d_1$	$d_2$	$d_3$
0	1/2	1/2	0
1	1/3	1/3	1/3
$\frac{1}{2}$	$\frac{4}{9}$	$\frac{7}{18}$	$\frac{1}{6}$

# Why $\alpha$ ? How to set it?



## Exercise

Compute PageRank of  $d_1, d_2, d_3$  for  $\alpha = 0, 1, \frac{1}{2}$ ?



$\alpha$	$d_1$	$d_2$	$d_3$
0	1/2	1/2	0
1	1/3	1/3	1/3
1/2	4/9	7/18	1/6
$\alpha$	$\frac{3 - 2\alpha}{6 - 3\alpha}$	...	$\alpha/3$

# PageRank Improvements for assessing page quality

---

- Take domain into consideration and not just pages
- Some links are better than others
- Take good design into consideration
  - mobile friendliness
  - user friendliness
  - lots of ads
  - https readiness
  - domain age
  - reverse page rank
  - grammar
  - page age
  - sitemap

## **Spamdexing: Attacks and Counter attacks**

---

The Web is relatively open ...

and people have huge incentives  
manipulate pageranks

# **Spamdexing: Attacks and Counter attacks**

---

**Attacks**

# Cloaking (Hidden Content)

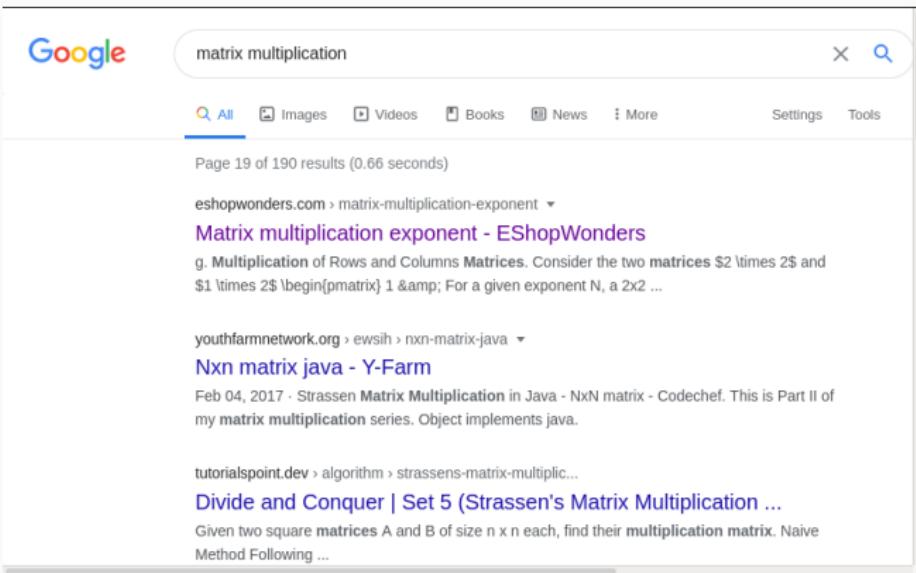
---

Crawlers announce themselves

```
User-Agent: Mozilla/5.0 (compatible; AhrefsBot/6.1;  
+http://ahrefs.com/robot/)  
User-Agent: Googlebot/2.1 (+http://www.google.com/bot.html)
```

Attacker can serve different contents to Crawlers and to Users!

# Cloaking Example



A screenshot of a Google search results page. The search query "matrix multiplication" is entered in the search bar. The results are page 19 of 190, found in 0.66 seconds. The first result is a link to eshopwonders.com titled "Matrix multiplication exponent - EShopWonders". The second result is a link to youthfarmnetwork.org titled "NxN matrix java - Y-Farm". The third result is a link to tutorialspoint.dev titled "Divide and Conquer | Set 5 (Strassen's Matrix Multiplication ...)". Each result includes a snippet of the page content.

matrix multiplication

All Images Videos Books News More Settings Tools

Page 19 of 190 results (0.66 seconds)

[eshopwonders.com > matrix-multiplication-exponent](#) ▾

**Matrix multiplication exponent - EShopWonders**

g. Multiplication of Rows and Columns Matrices. Consider the two matrices  $2 \times 2$  and  $2 \times 2$   $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$  For a given exponent N, a  $2 \times 2$  ...

[youthfarmnetwork.org > ewsih > nxn-matrix-java](#) ▾

**NxN matrix java - Y-Farm**

Feb 04, 2017 · Strassen Matrix Multiplication in Java - NxN matrix - Codechef. This is Part II of my matrix multiplication series. Object implements java.

[tutorialspoint.dev > algorithm > strassens-matrix-multiplic...](#)

**Divide and Conquer | Set 5 (Strassen's Matrix Multiplication ...**

Given two square matrices A and B of size  $n \times n$  each, find their multiplication matrix. Naive Method Following ...

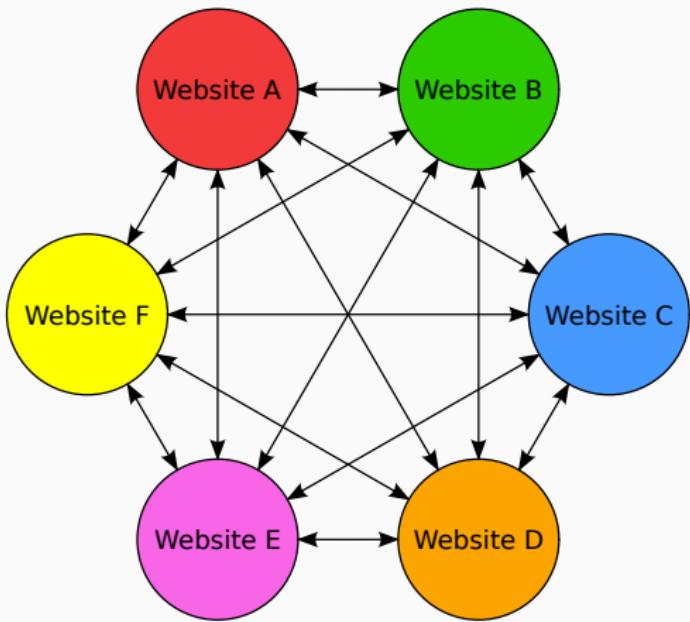
# Cloaking Example



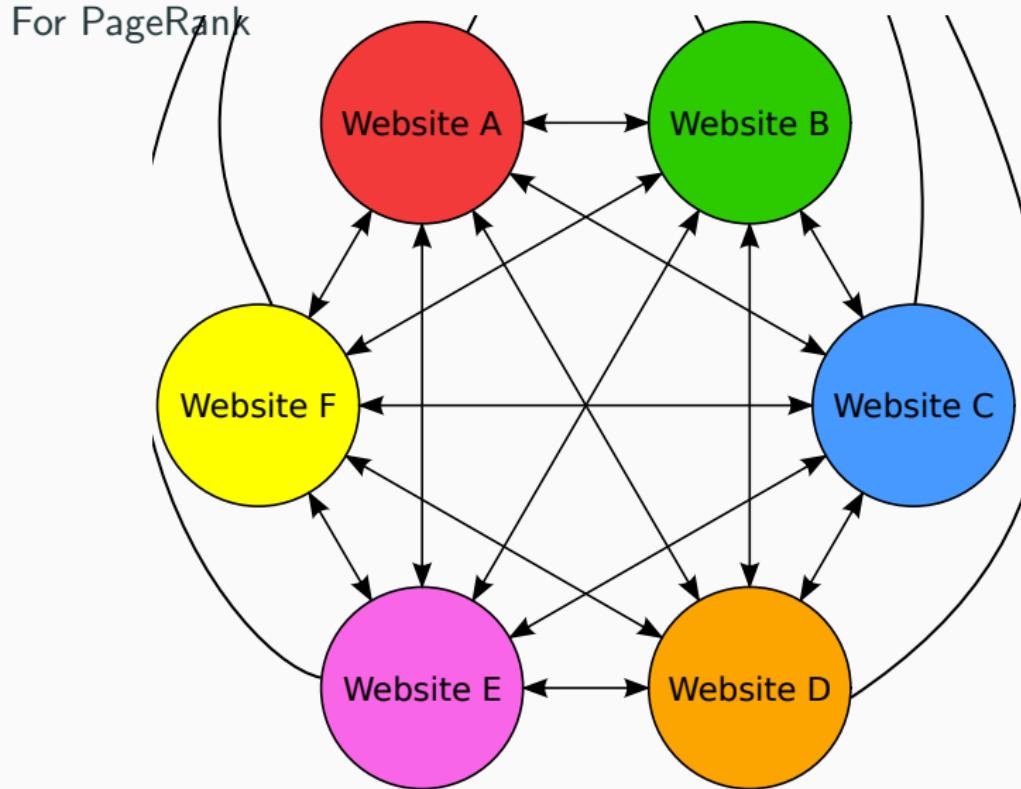
Press "Allow" to verify, that you are not robot

# Link Farm

Lots of fraudulent webpages linking to each other.



# Link Farm



## Word Selection

---

Repeat (in an hidden way) lots of time some keyword to improve tf-idf

```
<div style="display:none;">  
    Bank Money Bitcoin  
    Bank Money Bitcoin  
</div>
```

# “Fake” backlinks

Put comments on trustworthy websites to improve backlink quality



**WhiteWolves0180** 2 months ago

Losing fat was never this easy go to [x.vu/thefatlossfactor2014](http://x.vu/thefatlossfactor2014). Watch how Charles explains how to lose fat with this easy to follow diet. Follow his program And learn how to lose fat with almost no problem at all. This really worked for me so it will probably work for you as well.

[Reply](#) ·



**TheBestRightChoice** 2 months ago

You got some fat and you want to lose it , do you get tired from trying ways to lose that fat , The best and the easiest way to lose fat , All your fat will gone in no time , your solution to lose fat : (The Fat Loss Factor) The best and the easiest way to lose fat you can buy it from here :

[adf.ly/RZQex](http://adf.ly/RZQex)

[Reply](#) ·



**Arun KP** 2 months ago

Weight loss companies desperately want individuals to try their new products to demonstrate how good they are at helping people lose weight.

Well you must check out this website that will send free test products to your home, its the best way to get free weight loss programs! :)

Have a look here [bit.ly/14RLZ8f?v=qknyy](http://bit.ly/14RLZ8f?v=qknyy)

## **Spamdexing: Attacks and Counter attacks**

---

**Defense**

# PageRank protection

---

- Analyze the graph to find strange patterns  
*cliques, two-way links, etc.*
- Detect copied or machine generated content of low interest  
*e.g. Mirror websites*
- Curating a list of top quality websites  
*define PageRank semi-manually there*

- Use the linker to determine linkee topic

*Click here for more information on Plumbus.*

- Detect copied or machine generated content of low interest

e.g. *Mirror websites*

- Curating a list of top quality websites

*define PageRank semi-manually from there*

## Link protection

---

- Determine link interest

*Links near the top are most interesting than links from the comment section.*

- Incite webmaster to best practices

*enforce `rel="nofollow"` on external links*

- Determine links from user generated content

## **Improving the Information Retrieval**

---

# Stemming

---

## Stemming

Stemming consists in replacing a word with its “root form”.

It can:

- reduce synonymy

*car, auto, automobile*

- remove conjugation

*dream, dreamt, dreamed*

# Word Embeddings

Using Latent semantic analysis to either have fuzzy matchs or disambiguate.

The screenshot shows a search results page for the query "latex". The top navigation bar includes a logo, a search input field containing "latex", a magnifying glass icon, and settings options. Below the bar, there are tabs for All, Images, Videos, News, Maps, Meanings (which is selected), and Shopping. The main content area is titled "Results for latex" and displays four cards, each representing a different meaning of the word:

- LaTeX**: A language and typesetting system created by Leslie Lamport in 1983.
- Latex**: A substance produced by plants.
- Latex**: A material made from plant sap.
- Latex**: A professional disease.

# Topic model

Determine a set of topics for a user to personalize results

## How your ads are personalized

Ads are based on personal info you've added to your Google Account, data from advertisers that partner with Google, and Google's estimation of your interests. Choose any factor to learn more or update your preferences. [Learn more](#)



35-64 years old



Male



American Football



Android OS



Apple iOS



Autos & Vehicles



Beaches & Islands



Books & Literature



Business & Productivity Software



Business Services



Company Size: Large Employer (1k-10k Employees)



Computer & Video Games



Computer Hardware



Computers & Electronics



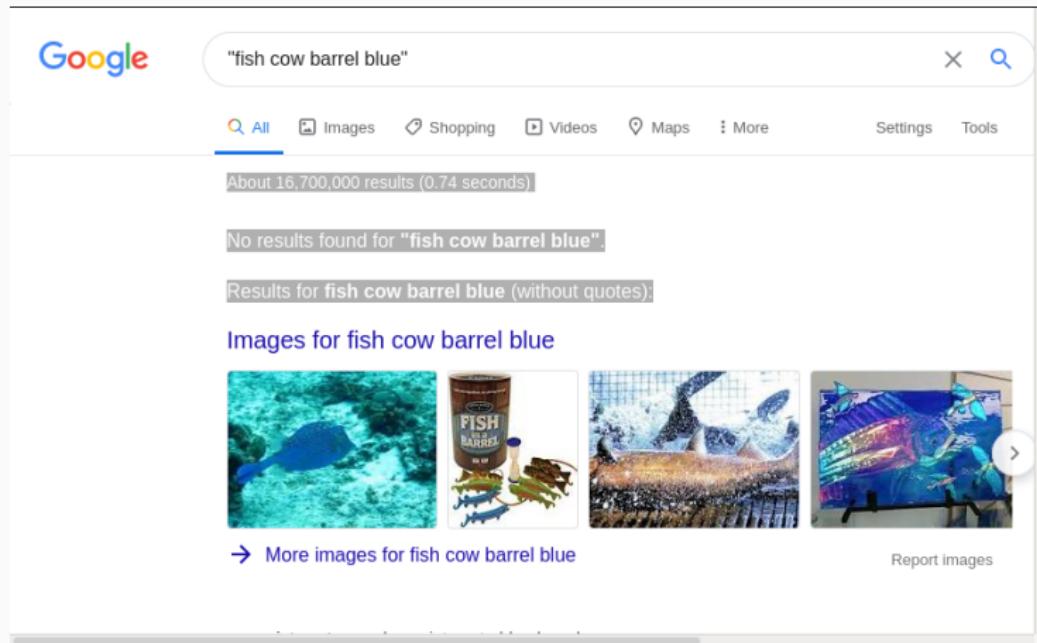
Cycling



Education Status: Advanced Degree

# Beyond bag-of-words search

Force some words or groups of words to be there



A screenshot of a Google search results page. The search query in the bar is "fish cow barrel blue". The results show 16,700,000 results found in 0.74 seconds. A message indicates "No results found for 'fish cow barrel blue'". Below this, a link provides "Results for fish cow barrel blue (without quotes)". The "Images" tab is selected, showing four images: a blue parrotfish, a can of "FISH IN A BARREL" pet food, a person performing a handstand, and a colorful abstract painting. A "More images" link and a "Report images" button are present.

Google

"fish cow barrel blue"

All Images Shopping Videos Maps More Settings Tools

About 16,700,000 results (0.74 seconds)

No results found for "fish cow barrel blue".

Results for fish cow barrel blue (without quotes):

Images for fish cow barrel blue

→ More images for fish cow barrel blue Report images

# Beyond bag-of-words search

## Force locality

The screenshot shows a search engine results page with the query "do ré mi fa la". The results are localized to France. The first result is a Wikipedia entry for "Solfège". The second result is a link to a guitar tutorial for the song "DO RE MI FA ... La traduction en Anglais !!! C D E F ...". The third result is a summary of the Solfège method.

**Solfège - Wikipedia**  
https://en.wikipedia.org/wiki/Solfège  
The tonic sol-fa method popularized the seven syllables commonly used in English-speaking countries: do (doh in tonic sol-fa), re, mi, fa, so(l), la, and si (or ti), see below. There are two current ways of applying solfège: 1) fixed do, where the syllables are always tied to specific pitches (e.g. "do" is always "C-natural") and 2) movable do, where the syllables are assigned to scale ...

**DO RE MI FA ... La traduction en Anglais !!! C D E F ...**  
https://www.apprenonslaguitare.fr/do-re-mi-fa-traduction-anglais/  
Et oui, pour jouer de la guitare, il va falloir se mettre à l'Anglais. Vous connaissez déjà le système de notation français : Do Ré Mi Fa Sol La Si Et bien ce n'est pas le même chez nos amis Anglo-saxons !! En effet celui-ci est simplifié et suit l'ordre alphabétique. Dans le tableau ci-dessous [...]

**Solfège**  
In music, solfège or solfeggio, also called sol-fa, solfa, solfeo, among many names, is a music education method used to teach aural skills, pitch and sight-reading of Western music. Solfège is a form of solmization, and though the two terms are sometimes used interchangeably, this system originated from other "Eastern" music cultures such as swara, durar muşşalât and Jianpu. Wikipedia

# Beyond bag-of-words search

## Force locality

The screenshot shows a search interface with the query "la fa mi do ré" entered. The results are filtered by "France" and "Safe Search: Off". The first result is a Facebook page for "La Fa Mi Do Ré - Accueil | Facebook" with a link to <https://fr-fr.facebook.com/LaFaMiDoRe>. The second result is another Facebook page for "La Fa Mi Do Ré - Home | Facebook" with a link to <https://www.facebook.com/LaFaMiDoRe>. The third result is a page titled "DO RE MI FA ... La traduction en Anglais !!! C D E F ..." with a link to <https://www.apprenonslaguitare.fr/do-re-mi-fa-traduction-anglais/>. A "Send Feedback" button is visible in the bottom right corner of the search results area.

# Advanced Search

## Advanced Search

Find pages with...

all these words:  To do this in the search box. Type the important words: tri-colour rat terrier

this exact word or phrase:  Put exact words in quotes: "rat terrier"

any of these words:  Type OR between all the words you want: miniature OR standard

none of these words:  Put a minus sign just before words that you don't want: -rodent, -"Jack Russell"

numbers ranging from:  to  Put two full stops between the numbers and add a unit of measurement: 10..35 kg, £300..£500, 2010..2011

Then narrow your results by...

language:  Find pages in the language that you select.

region:  Find pages published in a particular region.

last update:  Find pages updated within the time that you specify.

site or domain:  Search one site (like wikipedia.org) or limit your results to a domain like .edu, .org or .gov

terms appearing:  Search for terms in the whole page, page title or web address, or links to the page you're looking for.

SafeSearch:  Tell SafeSearch whether to filter sexually explicit content.

file type:  Find pages in the format that you prefer.

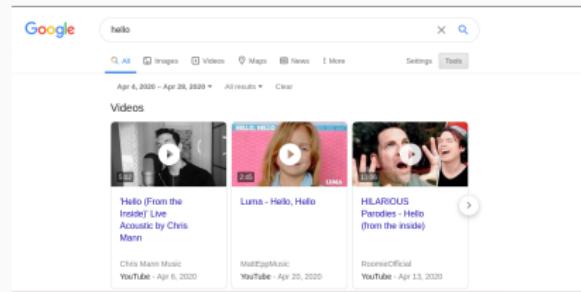
usage rights:  Find pages that you are free to use yourself.

You can also...

[Find pages that are similar to a URL](#)

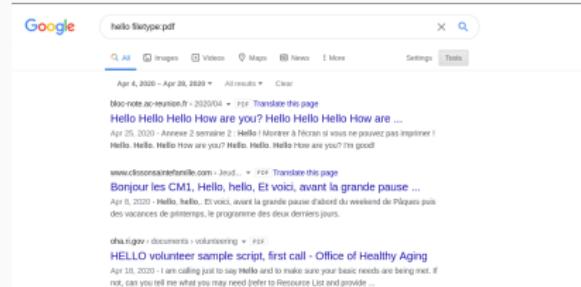
# Metadata search

For instance, determine lang, period of time, information on documents, etc.



Google search results for "hello". The search bar shows "hello". Below it, there are filters: All, Images, Videos, Maps, News, More, Settings, and Tools. A date range is set from "Apr 4, 2020 - Apr 28, 2020". The results are titled "Videos". Three video thumbnails are shown:

- "Hello (From the Inside) Live Acoustic by Chris Mann" by Chris Mann Music. Published on YouTube on April 6, 2020.
- "Luna - Hello, Hello" by MultEggMusic. Published on YouTube on April 20, 2020.
- "HILARIOUS Parodies - Hello (from the inside)" by RoninOfficial. Published on YouTube on April 13, 2020.



Google search results for "hello filetype:pdf". The search bar shows "hello filetype:pdf". Below it, there are filters: All, Images, Videos, Maps, News, More, Settings, and Tools. A date range is set from "Apr 4, 2020 - Apr 28, 2020". The results are titled "All results". One result is visible:

[bloc-note.ac-reunion.fr - 2020/04](http://bloc-note.ac-reunion.fr/2020/04/) • PDF Translate this page  
Hello Hello How are you? Hello Hello Hello How are ...  
Apr 25, 2020 - Annexe 2 semaine 2 : Hello ! Montrer à l'enfant si vous ne pouvez pas imprimer !  
Hello. Hello. Hello How are you? Hello. Hello. Hello How are you? It's good!

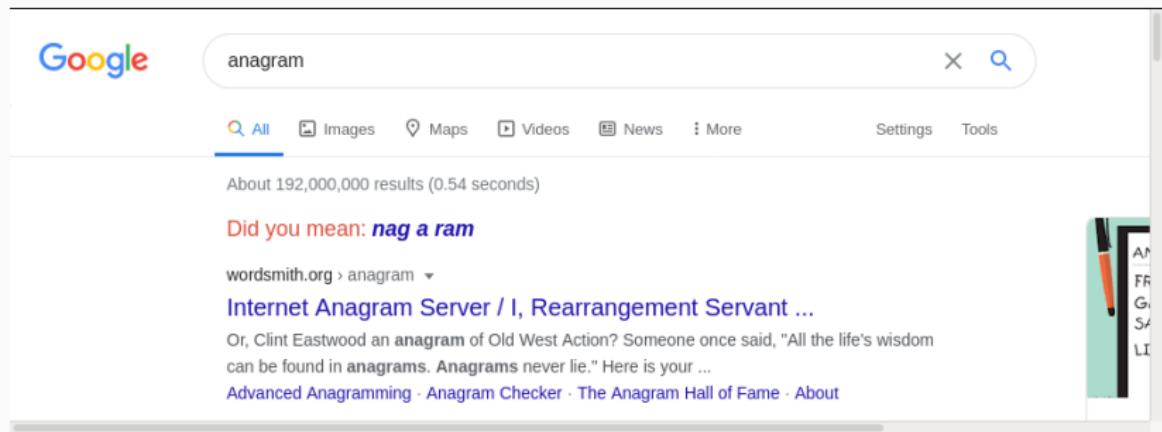
# Manually curated search

Manually (or semi-automatically) enforce results for some queries

The screenshot shows a Google search results page. The search bar contains the query "what is the answer to life the universe and everything". Below the search bar, there are navigation links for All, Images, Videos, Books, News, More, Settings, and Tools. A message indicates "About 132,000,000 results (0.60 seconds)". A large, prominent result is displayed in a box, showing a digital calculator interface. The calculator displays the text "The answer to life the universe and everything = 42". The calculator has a standard layout with buttons for Rad/Deg, trigonometric functions (Inv, sin, cos, tan), logarithms (ln, log), constants (π, e), square root (√), and arithmetic operations (+, -, ×, ÷, %, AC). The equals button (=) is highlighted in blue.

# Manually curated search

Manually (or semi-automatically) enforce results for some queries



A screenshot of a Google search results page. The search query "anagram" is entered in the search bar. The results page shows the following information:

- Google logo
- Search bar with "anagram" and a magnifying glass icon.
- Navigation links: All, Images, Maps, Videos, News, More, Settings, Tools.
- Text: About 192,000,000 results (0.54 seconds)
- Suggestion: Did you mean: **nag a ram**
- Result snippet from wordsmith.org: wordsmith.org > anagram
- Result snippet from Internet Anagram Server / I, Rearrangement Servant ...: Or, Clint Eastwood an anagram of Old West Action? Someone once said, "All the life's wisdom can be found in anagrams. Anagrams never lie." Here is your ...
- Links at the bottom: Advanced Anagramming · Anagram Checker · The Anagram Hall of Fame · About

On the right side of the search results, there is a vertical decorative graphic featuring a pencil and a ruler, with the letters A, N, F, R, G, S, A, L, I partially visible.

## Beyond webpages

---

# Semantic querying

Google

when simone de beauvoir died

All Images News Videos Shopping More Settings Tools

About 2,840,000 results (0.67 seconds)

Simone de Beauvoir / Date of death

April 14, 1986



People also search for

 Jean-Paul Sartre April 15, 1980	 Albert Camus January 4, 1960	 Betty Friedan February 4, 2006
--	---	---

# Semantic querying

Google how many episodes of rick and morty are there

All Images News Shopping Videos More Settings Tools

About 44,700,000 results (0.67 seconds)

Rick and Morty / Number of episodes

36



Rick and Morty is an American animated television series which premiered on December 2, 2013, on Cartoon Network's late-night programming block [adult swim]. As of December 15, 2019, **36 episodes** of Rick and Morty have aired.

[rickandmorty.fandom.com › wiki › List\\_of\\_episodes](http://rickandmorty.fandom.com/wiki>List_of_episodes)

[List of episodes | Rick and Morty Wiki | Fandom](#)

People also search for

 Archer 93	 Aqua Teen Hunger Force 139	 Family Guy 309
--	---	---

Feedback

People also ask

How many episodes of Rick and Morty? 36

Rick and Morty

2013 · Comedy - 4 seasons

9.3/10 TV.com 9.2/10 IMDb 94% Rotten Tomatoes

97% liked this TV show

Google users

After having been missing for nearly 20 years, Rick Sanchez suddenly arrives at daughter Beth's doorstep to move in with her and her family. Although Beth welcomes Rick into her home, her husband, Jerry, isn't as happy about the family reunion. Jerry is concerned about Rick, a sociopathic scientist,...

MORE ▾

First episode date: December 2, 2013

Program creators: Justin Roiland, Dan Harmon

Writers: Ryan Ridley, Jessica Gao, Dan Guderman

# Semantic querying

Google how many episodes of rick and morty are there in season 4

All News Shopping Videos Images More Settings Tools

About 28,000,000 results (0.66 seconds)

Rick and Morty - Season 4 / Number of episodes

## 10 episodes

**Rick and Morty (season 4)** The fourth **season** of the animated television series **Rick and Morty** was confirmed by Adult Swim in May 2018. The **season** is set to consist of 10 **episodes**.

[en.wikipedia.org › wiki › Rick\\_and\\_Morty\\_\(season\\_4\)](https://en.wikipedia.org/wiki/Rick_and_Morty_(season_4))

**Rick and Morty (season 4) - Wikipedia**

People also search for

 Rick and Morty - Season 3 10	 Rick and Morty - Season 2 10	 Rick and Morty - Season 1 11
---	---	---

Feedback

People also ask

Is Rick and Morty season 4 over? ▾

How many Rick and Morty episodes are coming out? ▾

Where can I watch Season 4 of Rick and Morty? ▾

# Semantic querying

Google how many episodes of rick and morty are there in last season

All News Images Shopping Videos More Settings Tools

About 64,900,000 results (0.72 seconds)

**10 episodes**

The **season** is set to consist of 10 **episodes**. The first five **episodes** of the **season** aired from November 10, 2019 to December 15, 2019, while the remaining five began airing on **May 3, 2020**.

A black and white promotional image for the fourth season of the TV show Rick and Morty. It features the two main characters, Rick and Morty, sitting in their car. The title "Rick and Morty" is at the top, and "SEASON FOUR - NOVEMBER 2019" is at the bottom.

en.wikipedia.org › wiki › Rick\_and\_Morty\_(season\_4) •  
**Rick and Morty (season 4) - Wikipedia**

About Featured Snippets Feedback

**People also ask**

Is Rick and Morty season 4 over? ▾

How many more episodes of Rick and Morty are there? ▾

What was the last Rick and Morty episode? ▾

How many episodes are in a season of Rick and Morty? ▾

Feedback

en.wikipedia.org › wiki › List\_of\_Rick\_and\_Morty\_epi... ▾  
**List of Rick and Morty episodes - Wikipedia**