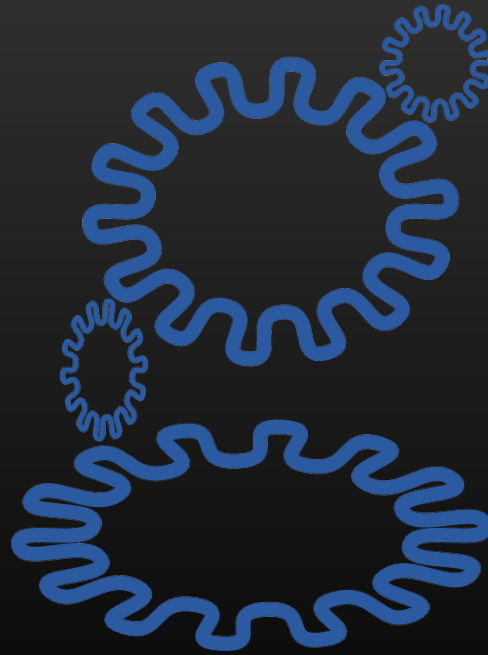
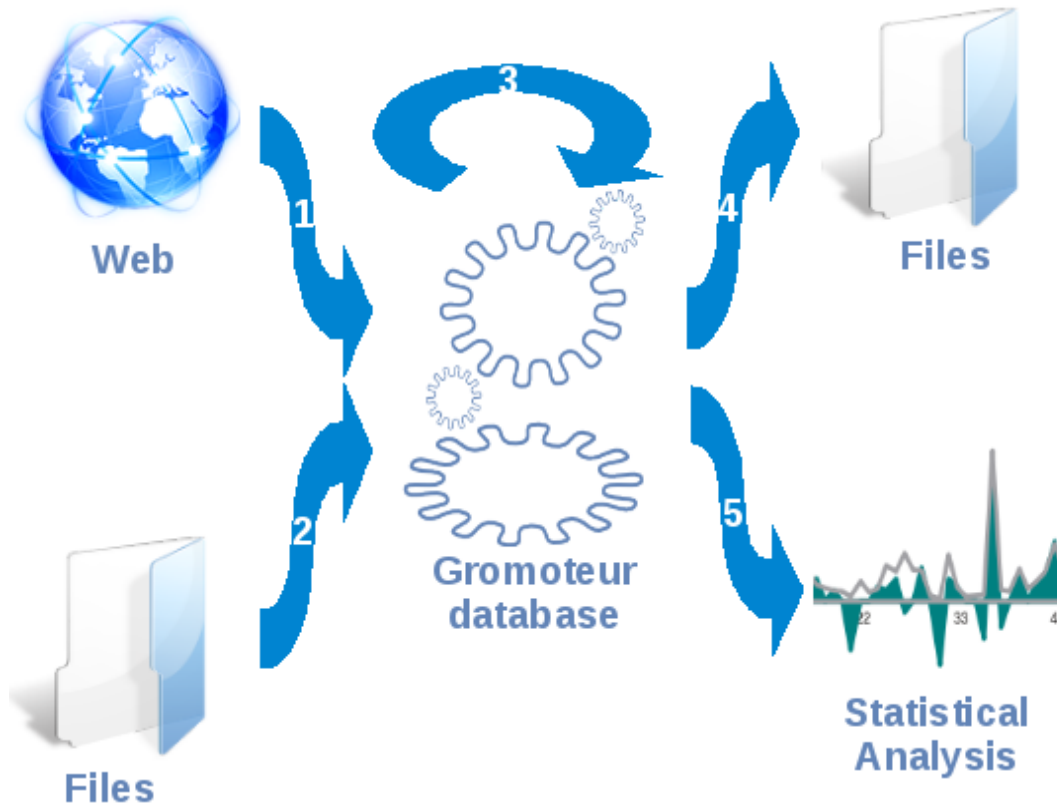


Collection et analyse de corpus  
pour des linguistes :

# Le Gromoteur



# Les chemins du Gromoteur

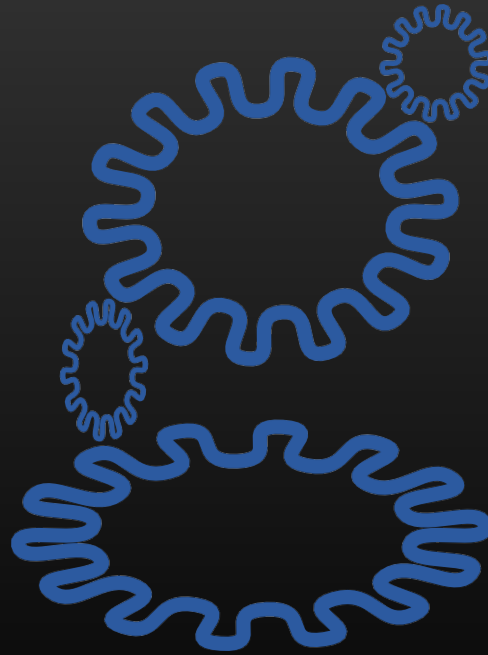


- Explications en français : fichier pdf Gro&Rapide

# Les chemins du Gromoteur

- **Téléchargement : [gromoteur.ilpga.fr](http://gromoteur.ilpga.fr)**
  - **Versions M\$-Windows et Linux**
  - **Open source (GPL) : sur launchpad**
  - **Explications en français**
    - **fichier pdf Gro&Rapide**
- **But principal : Datamasse pour un linguiste sans formation en informatique**
  - **Clickodrome**
  - **Seul difficulté :**
    - **expressions régulières**

# Fenêtre principale du Gromoteur



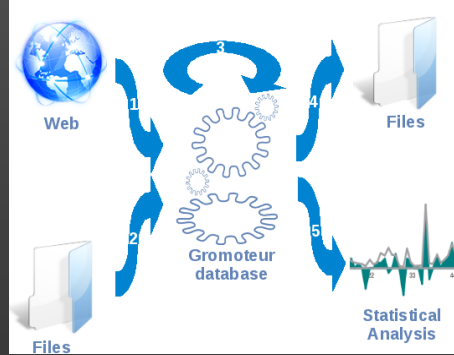
# Fenêtre principale du Gromoteur

The screenshot displays the Gromoteur application window. The title bar reads "Gromoteur". The menu bar includes "Database", "Edit", and "Help". The toolbar contains various icons for database operations. On the left, a sidebar shows a tree view of databases, with "hitch" selected. The main area is titled "Database content" and shows a table with the following columns: "url", "title", "text", and "text\_lem". The table lists 17 rows of document metadata, with the first row selected. Below the table, there are buttons for "database information", "cell content", "current filter", and "selections". The "cell content" button is active, showing a preview of the document text.

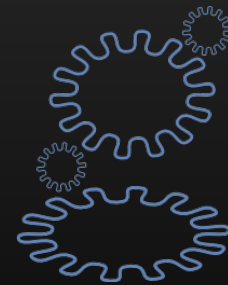
url	title	text	text_lem
/home/kim/...	hitch.000.txt	The Hitch Hi...	the hitch hik...
/home/kim/...	hitch.001.txt	1The house	1 the house ...
/home/kim/...	hitch.002.txt	2Here's what	2 here be wh...
/home/kim/...	hitch.003.txt	3On this	3 on this par...
/home/kim/...	hitch.004.txt	4Far away on	4 far away o...
/home/kim/...	hitch.005.txt	5Prostetnic	5 prostetnic ...
/home/kim/...	hitch.006.txt	6"Howl howl	6 " howl how...
/home/kim/...	hitch.007.txt	7Vogon poet...	7 vogon poe...
/home/kim/...	hitch.008.txt	8The Hitch	8 the hitch hi...
/home/kim/...	hitch.009.txt	9A computer	9 a compute...
/home/kim/...	hitch.010.txt	10The Infinite	10 the infinit...
/home/kim/...	hitch.011.txt	11The	11 the impro...
/home/kim/...	hitch.012.txt	12A loud clat...	12 a loud cla...
/home/kim/...	hitch.013.txt	13Marvin	13 marvin tr...
/home/kim/...	hitch.014.txt	14The Heart ...	14 the heart ...
/home/kim/...	hitch.015.txt	15(Excerpt fr...	15 ( excerpt ...
/home/kim/...	hitch.016.txt	16Arthur aw...	16 arthur aw...
/home/kim/...	hitch.017.txt	17After a fai...	17 after a fai...

1 the house stand on a slight rise just on the edge of the village .  
it stand on its own and look over a broad spread of west country farmland .  
not a remarkable house by any mean - it be about thirty year old , squattish , squarish , make of brick , and have four window set in  
the front of a size and proportion which more or less exactly fail to please the eye .  
the only person for whom the house be in any way special be arthur dent , and that be only because it happen to be the one he live  
in .  
he have live in it for about three year , ever since he have move out of london because it make him nervous and irritable .  
he be about thirty as well , dark haired and never quite at ease with himself .  
the thing that used to worry him most be the fact that person always used to ask him what he be look so worry about .  
he work in local radio which he always used to tell his friend be a lot more interesting than they probably think .  
it be , too - most of his friend work in advertising .  
it have n't properly register with arthur that the council want to knock down his house and build an bypass instead .  
at eight o'clock on thursday morning arthur do n't feel very good .

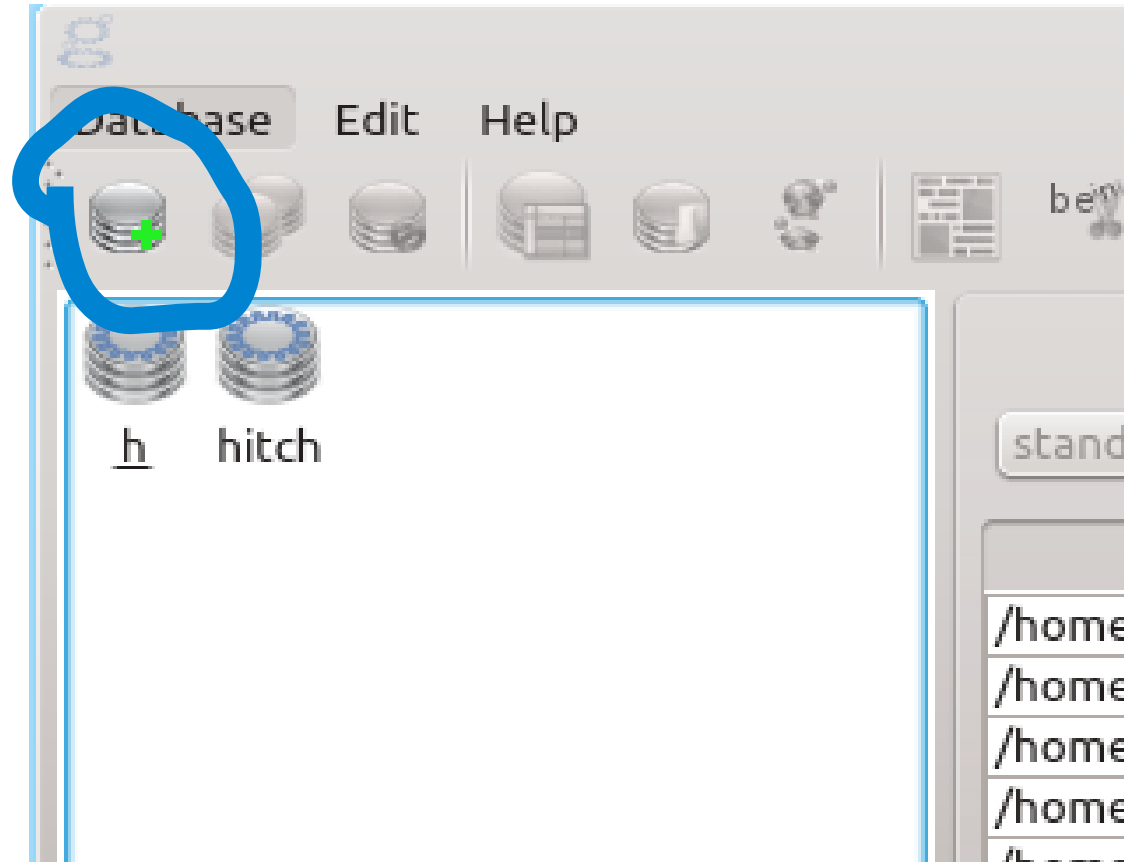
Currently showing 36 pages.



# Du web vers une base de Gromoteur



# Nouvelle base



# Le collecteur web

Database Edit

Database content

standard

url	title	text
-----	-------	------

database information cell content current filter { Groups selections

0 pages 0 links remaining 0.0 Mb no comment yet

0 sentences 0 links followed none

0 characters 6/6/14 12:50 AM

Currently showing 0 pages.



# Le collecteur web



**Control**  
Database: ars  
New spider configuration

**status**

current URL:

number of sentences captured:  current level:  encoding:

number of characters captured:  sentences/min:  characters/min:

number of pages Visited:  good pages:  pages/min:  good pages/min:

number of remaining Links:  bing visits:  running for:

**messages**

# Wizard (assistant de configuration)



# Expert?

1/5: Where to put the collected pages?  
The data base that will be used for this crawl...

selected base: ars

+ append new data    + overwrite same URL    - erase all existing content    - start from database

 Next > Cancel

1/5: Where to put the collected pages?  
The data base that will be used for this crawl...

selected base: ars

existing pages:  sentences:  words:  comment:

+ append new data    + overwrite same URL    - erase all existing content    - start from database

 Next > Cancel

# Où mettre les choses ? (1/5)



## 1/5: Where to put the collected pages?

The data base that will be used for this crawl...

selected base: ars

existing pages:  sentences:  words:  comment:

no comment yet

append new data     overwrite same URL     erase all existing content     start from database

< Back    Next >    Cancel

# Où commencer ? (2/5)



## 2/5: Where to start crawling?

Start with a URLs or use Bing to find the pages you want to get...

- A: start with one or more URLs
- B: start with a file of URLs
- C: start with a search engine

URLS:

arstechnica.com



< Back

Next >

Cancel

# Où commencer ? (2/5)



## 2/5: Where to start crawling?

Start with a URLs or use Bing to find the pages you want to get...

- A: start with one or more URLs
- B: start with a file of URLs
- C: start with a search engine

### Search Engine Setup:

Bing account id /oQLM.../rgpz/1Ma/Oce+VrWJL3mbjsJL4aA3Jo78=

keywords "statistique textuelle"

Follow only

Automatic location detection

try Bing 3190 results

get only result pages

continue surfing from results



< Back

Next >

Cancel

# Comment se balader ? (3/5)

**3/5 Which way to go?**  
Constraining the way the crawler walks...

Path:  
 Breadth first  
 Depth first

Only download page if:

URL matches

URL doesn't match

page contains

page is in

level from  to

use links only if page is OK

ignore case

Only take link if:

URL matches

URL doesn't match

take pdf files

force encoding

Only download page if:

URL matches

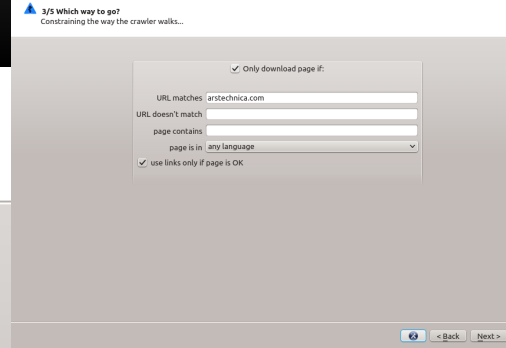
URL doesn't match

page contains

page is in

level from  to

use links only if page is OK



# Combien prendre ? (4/5)



## 4/5 How much?

Constraining the quantity of crawling results...

max pages 200

max sentences unlimited

max subdomains unlimited

max Mb of disk space unlimited

keep at least 0 Mb

of the 16670 Mb of free space in the corpus folder

Avoid spider traps

by trying to visit the same server only every

0 sec

by trying first to take from one server only

0 pages

Follow redirects

only follow if redirected URL matches download conditions

Timeout

18 sec

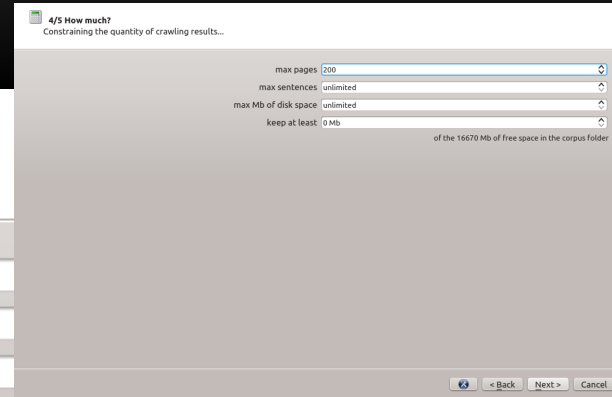
try again: 0 times



< Back

Next >

Cancel





# Vitesse ? Politesse ? Liberté ?

5/5  Go!  
Finishing up...

This configuration is called: **ars**

Existing configuration will be overwritten.

parallelism

4 threads

This computer has 6 cores.

identity

user agent:

proxy

http http://127.0.0.1:8087

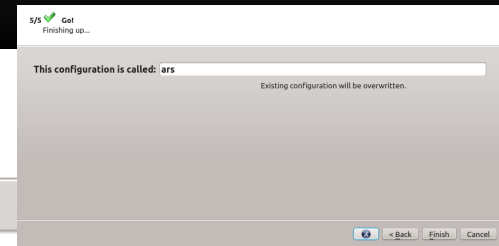
https http://127.0.0.1:8087

use the http://user:password@host/ syntax  
possible types: http:// https:// socks5://

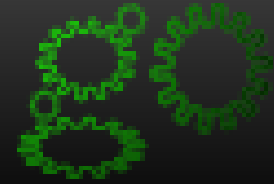
politeness

obey to robots.txt


  **Finish** 





# Gooooooooooooo !!!



**Control**  
Database: ars

ars 

current URL


number of sentences captured:  current level:  encoding

number of characters captured:  sentences/min  characters/min

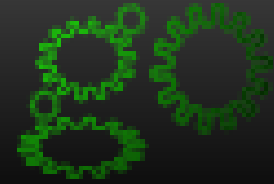
number of pages Visited:  good pages:  pages/min  good pages/min:

number of remaining Links:  bing visits:  running for

messages

 OK

# Gooooooooooooo !!!



**Control**  
Database: ars

ars

go

current URL <http://arstechnica.com/staff-directory/>

number of sentences captured:	39647	current level:	2	encoding:	UTF-8
number of characters captured:	3128011	sentences/min:	10908	characters/min:	860603
number of pages Visited:	15	good pages:	203	pages/min:	56.13
number of remaining Links:	21117	bing vis:		good pages/min:	55.85

running for 3 minutes 31 seconds

*all done!*

messages

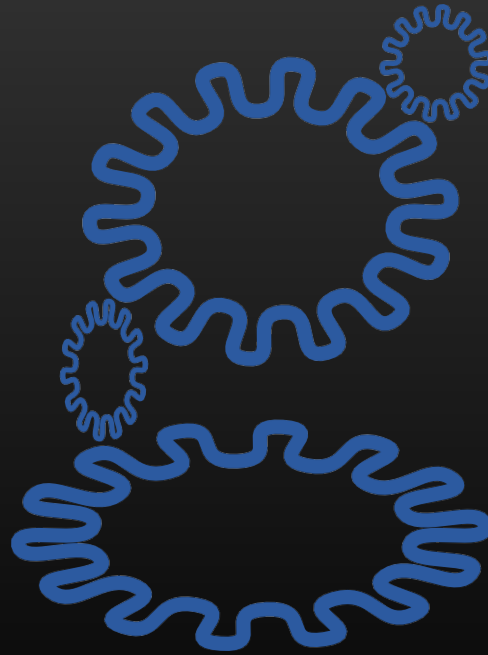
```
spider2 opens: http://arstechnica.com/?view=gri
spider1 opens: http://arstechnica.com/?view=archiv
spider0 opens: http://arstechnica.com/?theme=ligh
spider3 opens: http://arstechnica.com/?theme=dar
spider2 opens: http://arstechnica.com/reviews/
spider1 opens: http://arstechnica.com/video/
spider0 opens: http://arstechnica.com/staff/
spider1 opens: http://arstechnica.com/features/
spider2 opens: http://arstechnica.com/staff-direc
http://arstechnica.com/contact-us/ was not taken. page maximum attained
spider3 is finishing up.

Writing 21132 links to the database. Please wait!
Writing 21132 links to the database. Please wait!spider0 is done!
spider1 is done!
spider2 is done!

We finished. - All done...
```

OK

# Fenêtre principale du Gromoteur



# Fenêtre principale : Vue d'ensemble

The screenshot shows a web crawler interface with a menu bar (Database, Edit, Help), a toolbar, and a sidebar with database icons. The main area displays a table of database content with columns for url, title, and text. Below the table are tabs for database information, cell content, current filter, groups, and selections. A summary section shows statistics like 204 pages, 21117 links remaining, and 17.8 Mb. A comments section is also visible.

Database content

url	title	text
http://arstechnica.com	Ars Technica	Ars TechnicaArs Technica
http://arstechnica.com	Ars Technica	Ars TechnicaArs Technica
http://arstechnica.com/civis/ucp.php?mode=register	Register : Ars Technica OpenForum :	Register : Ars Technica OpenForum :You are curr...
http://arstechnica.com/civis/ucp.php?mode=login...	Login : Ars Technica OpenForum : 1	Login : Ars Technica OpenForum : 1Login to com...
http://arstechnica.com/information-technology	Technology Lab   Ars Technica	Technology Lab   Ars TechnicaArs Technica
http://arstechnica.com/gadgets	Gear & Gadgets   Ars Technica	Gear & Gadgets   Ars TechnicaArs Technica
http://arstechnica.com/business	Ministry of Innovation   Ars Technica	Ministry of Innovation   Ars TechnicaArs Technica
http://arstechnica.com/security	Risk Assessment   Ars Technica	Risk Assessment   Ars TechnicaArs Technica
http://arstechnica.com/tech-policy	Law & Disorder   Ars Technica	Law & Disorder   Ars TechnicaArs Technica
http://arstechnica.com/apple	Infinite Loop   Ars Technica	Infinite Loop   Ars TechnicaArs Technica
http://arstechnica.com/science	Scientific Method   Ars Technica	Scientific Method   Ars TechnicaArs Technica
http://arstechnica.com/gaming	Opposable Thumbs   Ars Technica	Opposable Thumbs   Ars TechnicaArs Technica
http://arstechnica.com/cars	Cars Technica   Ars Technica	Cars Technica   Ars TechnicaArs Technica
http://arstechnica.com/?view=gri	Ars Technica	Ars TechnicaArs Technica
http://arstechnica.com/?view=archiv	Ars Technica	Ars TechnicaArs Technica
http://arstechnica.com/reviews/	Reviews   Ars Technica	Reviews   Ars TechnicaArs Technica
http://arstechnica.com/?theme=ligh	Ars Technica	Ars TechnicaArs Technica
http://arstechnica.com/?theme=dar	Ars Technica	Ars TechnicaArs Technica
http://arstechnica.com/video/	Ars Technica Videos   Ars Technica	Ars Technica Videos   Ars TechnicaArs Technica
http://arstechnica.com/staff/	Staff   Ars Technica	Staff   Ars TechnicaArs Technica
http://arstechnica.com/features/	Features   Ars Technica	Features   Ars TechnicaArs Technica
http://arstechnica.com/staff-directory/	Staff Directory   Ars Technica	Staff Directory   Ars TechnicaArs Technica

database information | cell content | current filter | { Groups | selections

204 pages | 21117 links remaining | 17.8 Mb | no comment yet

39648 sentences | 15 links followed | ars

3128011 characters | 6/6/14 1:34 AM

terminated

# Fenêtre principale : contenu d'une page

The screenshot shows a database application window with a menu bar (Database, Edit, Help) and a toolbar. On the left is a sidebar with a tree view of databases: ars, hitch, libé, mini.test, nessuno, qsf, rrr, sarahwucorpus6, test, test.bak. The main area displays a table titled 'Database content' with columns 'url', 'title', and 'text'. The table contains 20 rows of data. Below the table are buttons for 'database information', 'cell content', 'current filter', '{ Groups', and 'selections'. A preview pane at the bottom shows the content of the selected cell, which is a page from Ars Technica.

url	title	text
<a href="http://arstechnica.com/security/2014/06/reported...">http://arstechnica.com/security/2014/06/reported...</a>	Reported Paris Hilton hacker cops to ne...	Reported Paris Hilton hacker cops to new intrus...
<a href="http://arstechnica.com/gadgets/2014/06/google-r...">http://arstechnica.com/gadgets/2014/06/google-r...</a>	Google releases Android 4.4.3 to Nexus ...	Google releases Android 4.4.3 to Nexus devices ...
<a href="http://arstechnica.com/tech-policy/2014/06/meet-...">http://arstechnica.com/tech-policy/2014/06/meet...</a>	"You could be liable for \$150k in penalti...	"You could be liable for \$150k in penalties—set...
<a href="http://arstechnica.com/tech-policy/2014/06/meet-...">http://arstechnica.com/tech-policy/2014/06/meet...</a>	"You could be liable for \$150k in penalti...	"You could be liable for \$150k in penalties—set...
<a href="http://arstechnica.com/apple/2014/06/psa-the-ide...">http://arstechnica.com/apple/2014/06/psa-the-ide...</a>	PSA: The iDevices and Macs that will sup...	PSA: The iDevices and Macs that will support iO...
<a href="http://arstechnica.com/cars/2014/06/the-past-pre...">http://arstechnica.com/cars/2014/06/the-past-pre...</a>	The past, present, and future of in-car in...	The past, present, and future of in-car infotainm...
<a href="http://arstechnica.com/security/2014/06/reported...">http://arstechnica.com/security/2014/06/reported...</a>	Reported Paris Hilton hacker cops to ne...	Reported Paris Hilton hacker cops to new intrus...
<a href="http://arstechnica.com/gadgets/2014/05/hands-o...">http://arstechnica.com/gadgets/2014/05/hands-o...</a>	Hands-on: Using Microsoft's Surface Pro ...	Hands-on: Using Microsoft's Surface Pro 3 as a L...
<a href="http://arstechnica.com/cars/2014/06/the-past-pre...">http://arstechnica.com/cars/2014/06/the-past-pre...</a>	The past, present, and future of in-car in...	The past, present, and future of in-car infotainm...
<a href="http://feeds.arstechnica.com/arstechnica/index/">http://feeds.arstechnica.com/arstechnica/index/</a>	Ars Technica	
<a href="http://arstechnica.com/rss-feeds/">http://arstechnica.com/rss-feeds/</a>	RSS Feeds   Ars Technica	RSS Feeds   Ars TechnicaArs Technica
<a href="http://arstechnica.com/gadgets/2014/05/toshiba-...">http://arstechnica.com/gadgets/2014/05/toshiba-...</a>	Toshiba P50t review: So many pixels, too...	Toshiba P50t review: So many pixels, too much ...
<a href="http://arstechnica.com/podcast/">http://arstechnica.com/podcast/</a>	Podcast   Ars Technica	Podcast   Ars TechnicaArs Technica
<a href="http://arstechnica.com/gadgets/2014/05/the-psyc...">http://arstechnica.com/gadgets/2014/05/the-psyc...</a>	The psychology of Soylent and the priso...	The psychology of Soylent and the prison of firs...
<a href="http://arstechnica.com/author/lee-hutchinson/">http://arstechnica.com/author/lee-hutchinson/</a>	Lee Hutchinson   Ars Technica	Lee Hutchinson   Ars TechnicaArs Technica
<a href="http://arstechnica.com/gadgets/2014/05/toshiba-...">http://arstechnica.com/gadgets/2014/05/toshiba-...</a>	Toshiba P50t review: So many pixels, too...	Toshiba P50t review: So many pixels, too much ...
<a href="http://arstechnica.com/page/2">http://arstechnica.com/page/2</a>	Ars Technica	Ars TechnicaArs Technica
<a href="http://arstechnica.com/about-us/">http://arstechnica.com/about-us/</a>	About Us   Ars Technica	About Us   Ars TechnicaArs Technica
<a href="http://arstechnica.com/advertise-with-us/">http://arstechnica.com/advertise-with-us/</a>	Advertise with us   Ars Technica	Advertise with us   Ars TechnicaArs Technica
<a href="http://arstechnica.com/reprints/">http://arstechnica.com/reprints/</a>	Reprints   Ars Technica	Reprints   Ars TechnicaArs Technica
<a href="http://arstechnica.com/gadgets/2014/05/the-psyc...">http://arstechnica.com/gadgets/2014/05/the-psyc...</a>	The psychology of Soylent and the priso...	The psychology of Soylent and the prison of firs...
<a href="http://arstechnica.com/newsletters/">http://arstechnica.com/newsletters/</a>	Newsletters   Ars Technica	Newsletters   Ars TechnicaArs Technica

database information cell content current filter { Groups selections

Reported Paris Hilton hacker cops to new intrusions targeting police | Ars Technica  
Ars Technica  
Register Log in  
Home Main Menu Information Technology Technology Lab Product News & Reviews Gear & Gadgets Business of Technology Ministry of Innovation Security & Hacktivism Risk Assessment Civilization & Discontents Law & Disorder The Apple Ecosystem Infinite Loop Gaming & Entertainment Opposable Thumbs Science & Exploration The Scientific Method All Things Automotive Cars Technica  
Layout:  
Grid View Article View  
Site Theme  
Dark on light Light on dark  
Explore Ars  
Reviews Video Staff Blogs Feature Archive Staff Directory Contact Us  
Featured Disciplines

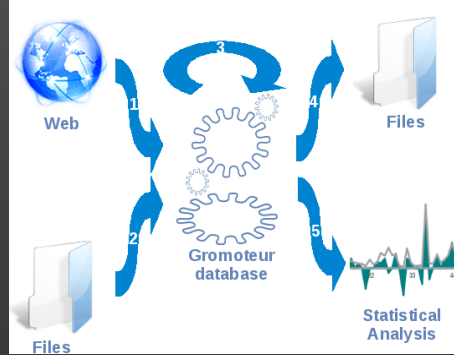
Currently showing 204 pages.

# Fenêtre principale : filtrer

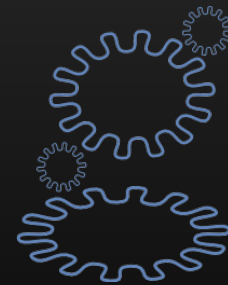


The screenshot shows a database application window with a menu bar (Database, Edit, Help) and a toolbar. On the left, there is a sidebar with a tree view of databases: ars, hitch, libé, mini.test, nessuno, qsdf, rrr, sarahwucorpus6, test, test.bak. The main area displays the 'Database content' for the 'standard' database, showing a table with columns 'url', 'title', and 'text'. A blue hand-drawn circle highlights the 'url' column in the table. Below the table, there is a 'current filter' section with a table of filter rules:

column	condition	value	remove
1 url	contains	ard gadgets	<input checked="" type="checkbox"/>
2 rowid	=		<input checked="" type="checkbox"/>

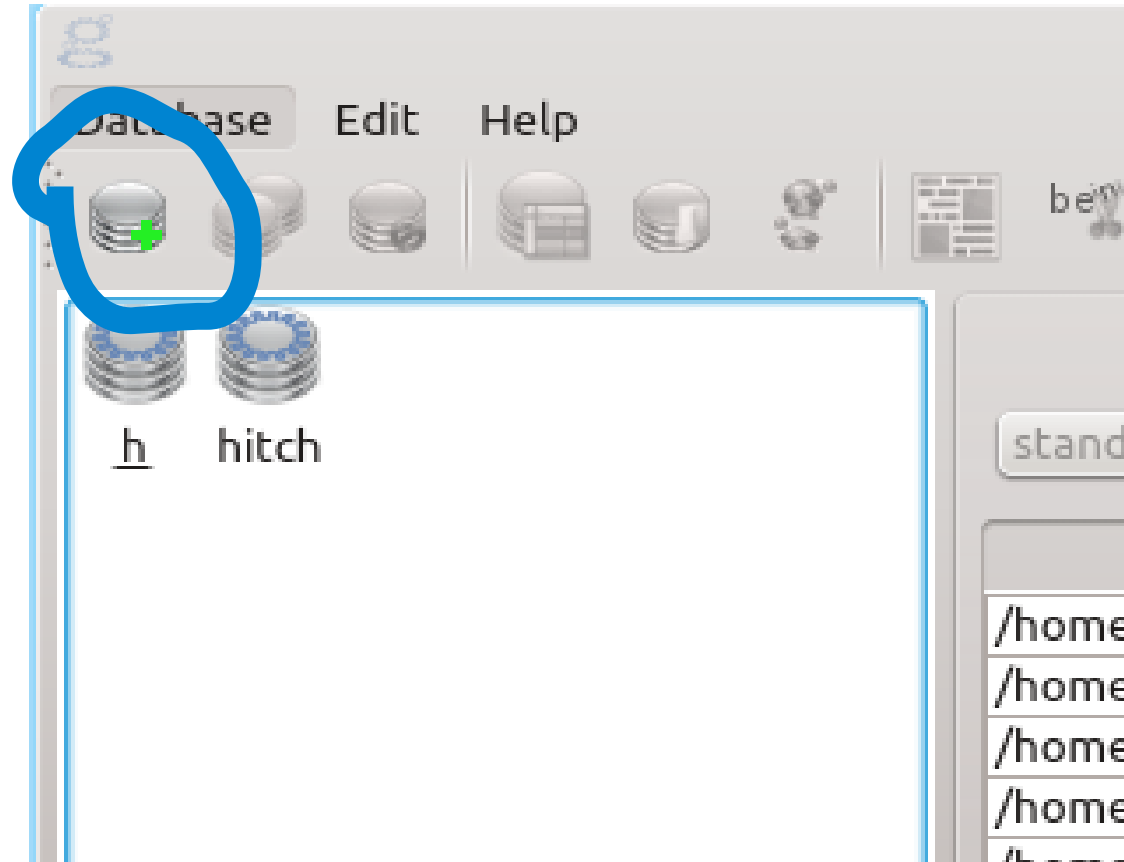


# Des fichiers sur le disque vers une base de Gromoteur





# Nouvelle base



# Le collecteur de fichiers

Database

ars hitch libé mini.test

nessuno qsdf rrr

sarahwucorpus6 test test.bak

Database content

standard

url	title	text
-----	-------	------

database information cell content current filter { Groups selections

0 pages 0 links remaining 0.0 Mb no comment yet

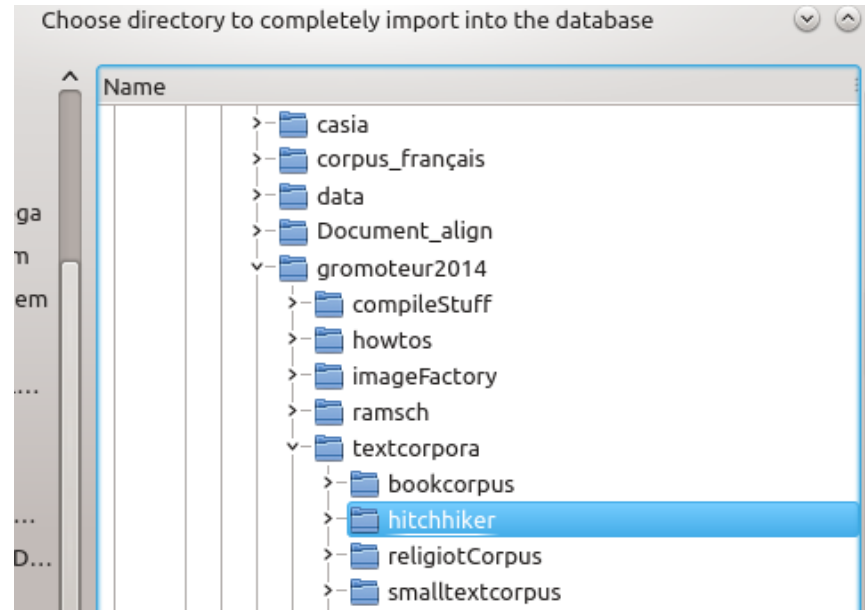
0 sentences 0 links followed none

0 characters 6/6/14 12:50 AM

Currently showing 0 pages.

# Un dossier

→ tous les fichiers txt ou pdf



# Fenêtre principale comme avant



The screenshot shows a database application window with a menu bar (Database, Edit, Help) and a toolbar. On the left is a sidebar with a tree view of databases: ars, hitch, libé, mini.test, nessuno, qsf, rrr, sarahwucorpus6, test, test.bak. The main area displays a table titled 'Database content' with columns 'url', 'title', and 'text'. The table contains 20 rows of data. The 4th row is selected, and its content is shown in a detailed view at the bottom of the window.

url	title	text
/home/kim/...	hitch.000.txt	The Hitch Hikers Guide to GalaxyFar out in the uncharted backwaters of the unfashionable end of the
/home/kim/...	hitch.001.txt	1The house stood on a slight rise just on the edge of the village. It stood on its own and looked o
/home/kim/...	hitch.002.txt	2Here's what the Encyclopedia Galactica has to say about alcohol. It says that alcohol is a colour!
/home/kim/...	hitch.003.txt	3On this particular Thursday, something was moving quietly through the ionosphere many miles above
/home/kim/...	hitch.004.txt	4Far away on the opposite spiral arm of the Galaxy, five hundred thousand light years from the star
/home/kim/...	hitch.005.txt	5Prostetnic Vogon Jeltz was not a pleasant sight, even for other Vogons. His highly domed nose rose
/home/kim/...	hitch.006.txt	6Howl howl gargle howl gargle howl howl howl gargle howl gargle howl gargle howl gargl
/home/kim/...	hitch.007.txt	7Vogon poetry is of course the third worst in the Universe.
/home/kim/...	hitch.008.txt	8The Hitch Hiker's Guide to the Galaxy is a wholly remarkable book. It has been compiled and recomp
/home/kim/...	hitch.009.txt	9A computer chatted to itself in alarm as it noticed an airlock open and close itself for no appare
/home/kim/...	hitch.010.txt	10The Infinite Improbability Drive is a wonderful new method of crossing vast interstellar distance
/home/kim/...	hitch.011.txt	11The Improbability-proof control cabin of the Heart of Gold looked like a perfectly conventional s
/home/kim/...	hitch.012.txt	12A loud clatter of gunk music flooded through the Heart of Gold cabin as Zaphod searched the sub-e
/home/kim/...	hitch.013.txt	13Marvin trudged on down the corridor, still moaning.
/home/kim/...	hitch.014.txt	14The Heart of Gold fled on silently through the night of space, now on conventional photon drive.
/home/kim/...	hitch.015.txt	15(Excerpt from The Hitch Hiker's Guide to the Galaxy, Page 634784, Section 5a, Entry: Magrathea)
/home/kim/...	hitch.016.txt	16Arthur awoke to the sound of argument and went to the bridge. Ford was waving his arms about.
/home/kim/...	hitch.017.txt	17After a fairly shaky start to the day, Arthur's mind was beginning to reassemble itself from the
/home/kim/...	hitch.018.txt	18And the next th ing that happened after that was that the Heart of Gold continued on its way perf
/home/kim/...	hitch.019.txt	19"Are we taking this robot with us?" said Ford, looking with distaste at Marvin who was standing i
/home/kim/...	hitch.020.txt	20Five figures wandered slowly over the blighted land. Bits of it were dullish grey, bits of it dul

database information | cell content | current filter | { Groups | selections

4  
Far away on the opposite spiral arm of the Galaxy, five hundred thousand light years from the star Sol, Zaphod Beeblebrox, President of the Imperial Galactic Government, sped across the seas of Damogran, his ion drive delta boat winking and flashing in the Damogran sun. Damogran the hot; Damogran the remote; Damogran the almost totally unheard of. Damogran, secret home of the Heart of Gold.  
The boat sped on across the water. It would be some time before it reached its destination because Damogran is such an inconveniently arranged planet. It consists of nothing but middling to large desert islands separated by very pretty but annoyingly wide stretches of ocean.  
The boat sped on.  
Because of this topological awkwardness Damogran has always remained a deserted planet. This is why the Imperial Galactic Government chose Damogran for the Heart of Gold project, because it was so deserted and the Heart of Gold was so secret.  
The boat zipped and skipped across the sea, the sea that lay between the main islands of the only archipelago of any useful size on the whole planet. Zaphod Beeblebrox was on his way from the tiny spaceport on Easter Island (the name was an entirely meaningless

Currently showing 36 pages.

# Explorer le tableau

The screenshot shows the Grosmoteur application window. The title bar reads "Grosmoteur". The menu bar includes "Database" and "Help". Below the menu bar is a toolbar with various icons. The main area is divided into a left sidebar and a right pane. The sidebar contains several database icons, with "hitch" selected. The right pane displays a table with two columns: "text\_nbSentences" and "text". A blue circle highlights the table content. Below the table, there are buttons for "database information", "cell content", "current filter", and "selections". At the bottom, there is a filter configuration area with a table for "column", "=", "condition", and "remove".

text_nbSentences	text
82	22He was standing with his back to Arthur watching the very last
6	23It is an important and popular fact that things do not always v
98	24Silently the aircar coasted through the cold darkness, a single
102	25There are of course many problems connected with life, of wh
7	26"Yes, very salutary," said Arthur, after Slartibartfast had relate
71	27Slartibartfast's study was a total mess, like the result of an ex
41	28It was a long time before anyone spoke.
111	29"Zaphod! Wake up!"
39	30"So there you have it," said Slartibartfast, making a feeble and
19	31It is of course well known that careless talk costs lives, but the
8	32"Emergency! Emergency!" blared the klaxons throughout Mag
45	33But the end never came, at least not then.
36	34The aircar rocketed them at speed in the direction of R17 through t
10	35That night, as the Heart of Gold was busy putting a few light y

column	=	condition	remove
1 rowid	=		

# Sélectionner des colonnes

The screenshot shows the Grosmoteur database application. On the left, a sidebar lists databases: b, bbb, h, hitch (highlighted), jiaotong, and popularscience. The main window displays the 'standard' database content as a table with two columns: 'text\_nbSentences' and 'text'. A blue circle highlights the 'text' column header. Below the table, there are buttons for 'database information', 'cell content', 'current filter', and 'selections'. At the bottom, a filter configuration area shows 'rowid' selected in the 'column' dropdown, with an equals sign in the 'condition' dropdown and a red 'X' icon in the 'remove' field.

text_nbSentences	text
82	22He was standing with his back to Arthur watching the very last
6	23It is an important and popular fact that things are not always v
98	24Silently the aircar coasted through the cold darkness, a single
102	25There are of course many problems connected with life, of wh
7	26"Yes, very salutary," said Arthur, after Slartibartfast had relate
71	27Slartibartfast's study was a total mess, the results of an ex
41	28It was a long time before anyone spoke.
111	29"Zaphod! Wake up!"
39	30"So there you have it," said Slartibartfast, making a feeble and
119	31It is of course well known that careless talk costs lives, but the
82	32"Emergency! Emergency!" blared the klaxon throughout Mag
45	33But the end never came, at least not then.
36	34The aircar rocketed them at speeds in excess of R17 through t
10	35That night, as the Heart of Gold was busy putting a few light y

# Sélectionner des colonnes

The screenshot shows the Grosmoteur database viewer interface. The window title is "Grosmoteur". The menu bar includes "Database" and "Help". The toolbar contains various icons for database operations. The left sidebar shows a tree view of databases: "b", "bbb", "h", "hitch", "jiaotong", and "popularscience". The "hitch" database is selected. The main area displays a table with the following data:

text_nbSentences	text
82	22He was standing with his back to Arthur watching the very last
6	23It is an important and popular fact that things are not always v
98	24Silently the aircar coasted through the cold darkness, a sm
10	25There are of course many problems connected with life, of w
	26"Yes, very salutary," said Arthur, after Slartibartfast had relate
71	27Slartibartfast's study was a total mess, like the results of an ex
41	28It was a long time before anyone spoke.
111	29"Zaphod! Wake up!"
39	30"So there you have it," said Slartibartfast, making a feeble and
119	31It is of course well known that careless talk costs lives, but the
82	32"Emergency! Emergency!" blared the klaxons throughout Mag
45	33But the end never came, at least not then.
36	34The aircar rocketed them at speeds in excess of R17 throu
10	35That night, as the Heart of Gold was busy putting a f

At the bottom of the window, there are tabs for "database information", "cell content", "current filter", and "selections". The "selections" tab is active, showing a table with the following structure:

column	=	condition	remove
1 rowid	=		

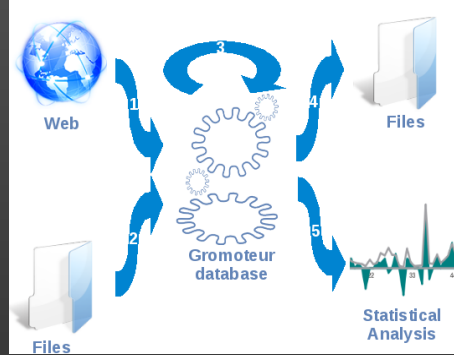
# Sélectionner des lignes

Recherche  
de mots

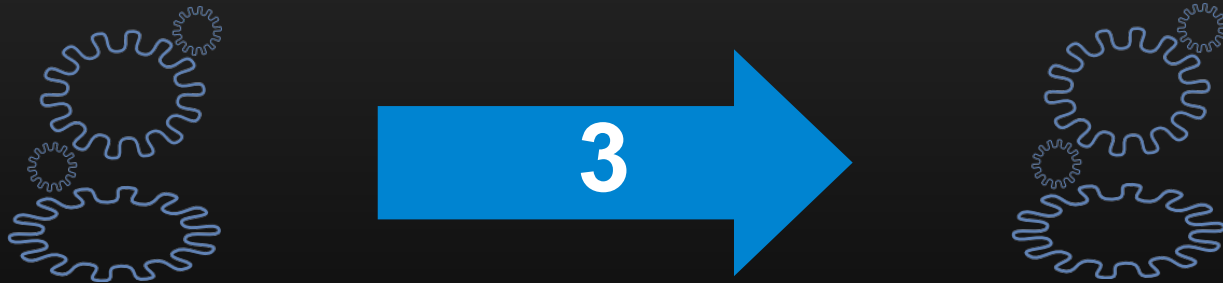
The screenshot shows the Grosmoteur application window. On the left, a sidebar displays a tree view of databases, with 'hitch' selected. The main area shows the 'Database content' for the 'standard' database, displaying a table with columns 'text\_nbSentences' and 'text'. A blue selection filter is applied to the 'rowid' column, with the value '1' entered in the filter field. The table content is as follows:

text_nbSentences	text
82	22He was standing with his back to Arthur watching the very last
6	23It is an important and popular fact that things are not always v
98	24Silently the aircar coasted through the cold darkness, a single
102	25There are of course many problems connected with life, of wh
7	26"Yes, very salutary," said Arthur, after Slartibartfast had relate
71	27Slartibartfast's study was a total mess, like the results of an ex
41	28It was a long time before anyone spoke.
111	29"Zaphod! Wake up!"
39	30"So there you have it," said Slartibartfast, making a feeble and
119	31He of course well known that careless talk costs lives, but the
	32"Emergency, emergency!" blared the klaxons throughout Mag
45	33But the end never came, at least not then.
36	34The aircar rocketed them at speed in excess of R17 through t
10	35That night, as the Heart of Gold was busy putting a few light y

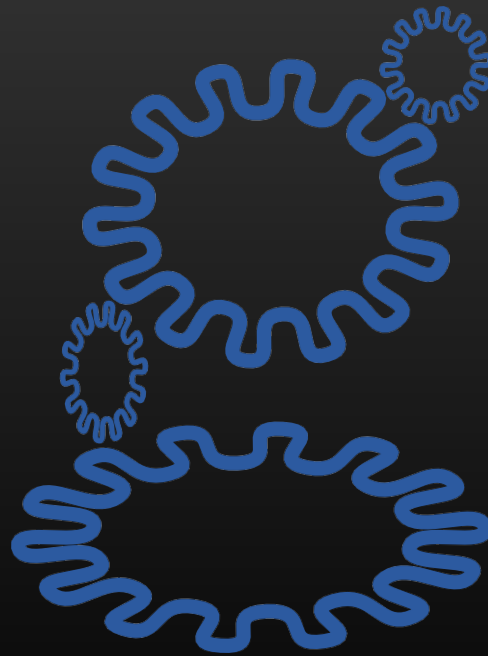




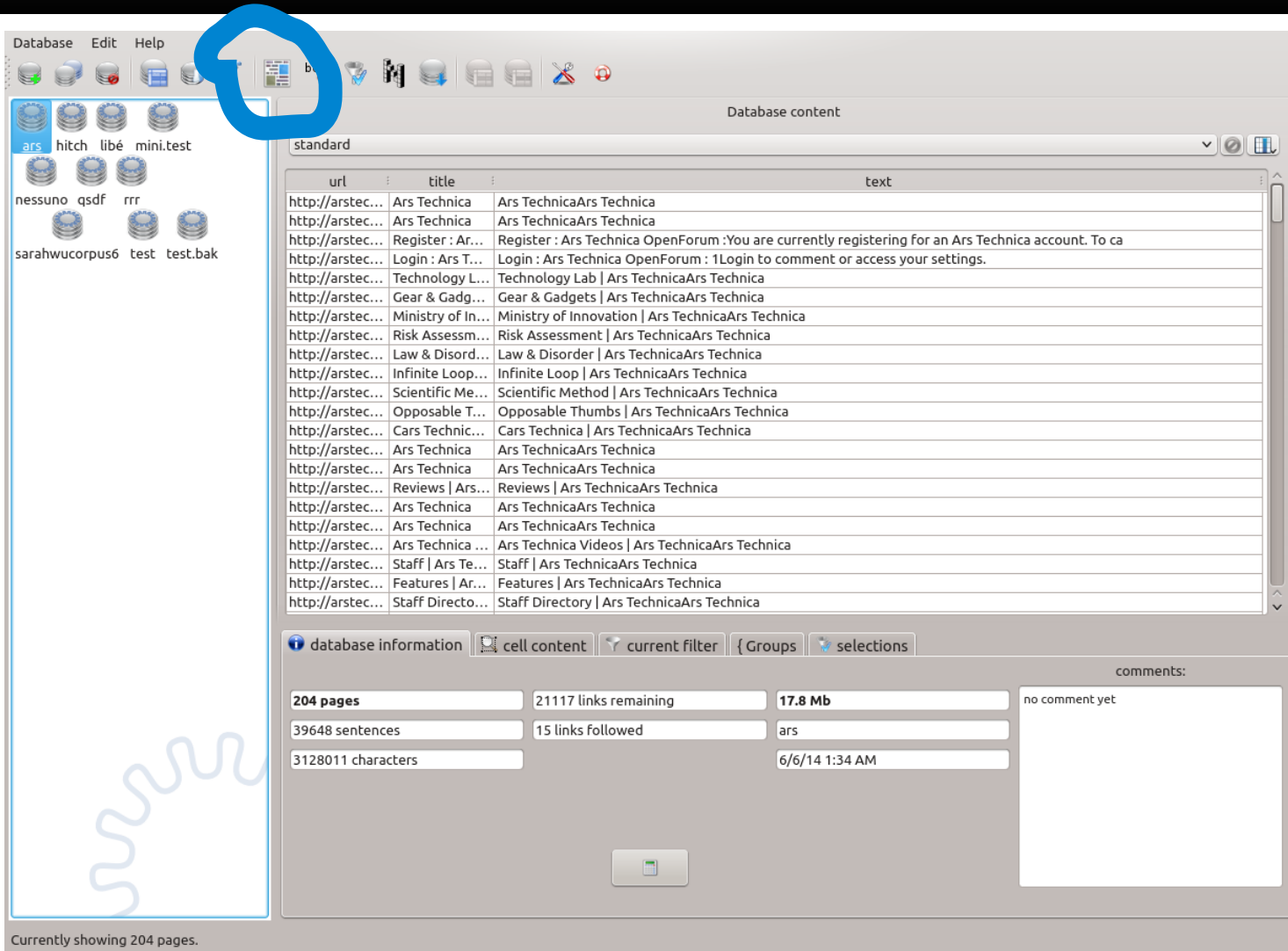
# Traitement des données



# Sélectionner des parties de contenu



# Sélectionner des parties des pages web



The screenshot shows a web crawler application interface. On the left, there is a sidebar with a tree view of databases. A blue circle highlights the 'ars' database. The main window displays a table of crawled pages. Below the table, there are statistics and a comments section.

url	title	text
http://arstec...	Ars Technica	Ars TechnicaArs Technica
http://arstec...	Ars Technica	Ars TechnicaArs Technica
http://arstec...	Register : Ar...	Register : Ars Technica OpenForum :You are currently registering for an Ars Technica account. To ca
http://arstec...	Login : Ars T...	Login : Ars Technica OpenForum : 1Login to comment or access your settings.
http://arstec...	Technology L...	Technology Lab   Ars TechnicaArs Technica
http://arstec...	Gear & Gadg...	Gear & Gadgets   Ars TechnicaArs Technica
http://arstec...	Ministry of In...	Ministry of Innovation   Ars TechnicaArs Technica
http://arstec...	Risk Assessm...	Risk Assessment   Ars TechnicaArs Technica
http://arstec...	Law & Disord...	Law & Disorder   Ars TechnicaArs Technica
http://arstec...	Infinite Loop...	Infinite Loop   Ars TechnicaArs Technica
http://arstec...	Scientific Me...	Scientific Method   Ars TechnicaArs Technica
http://arstec...	Opposable T...	Opposable Thumbs   Ars TechnicaArs Technica
http://arstec...	Cars Technic...	Cars Technica   Ars TechnicaArs Technica
http://arstec...	Ars Technica	Ars TechnicaArs Technica
http://arstec...	Ars Technica	Ars TechnicaArs Technica
http://arstec...	Reviews   Ars...	Reviews   Ars TechnicaArs Technica
http://arstec...	Ars Technica	Ars TechnicaArs Technica
http://arstec...	Ars Technica	Ars TechnicaArs Technica
http://arstec...	Ars Technica ...	Ars Technica Videos   Ars TechnicaArs Technica
http://arstec...	Staff   Ars Te...	Staff   Ars TechnicaArs Technica
http://arstec...	Features   Ar...	Features   Ars TechnicaArs Technica
http://arstec...	Staff Directo...	Staff Directory   Ars TechnicaArs Technica

database information | cell content | current filter | { Groups | selections

**204 pages** | 21117 links remaining | 17.8 Mb

39648 sentences | 15 links followed | ars

3128011 characters | 6/6/14 1:34 AM

comments: no comment yet

Currently showing 204 pages.

# Sélectionner des parties des pages web



http://arstechnica.com/security/2014/06/google-released-chrome-extension-allows-easy-in-browser-webmail-encryption/

End-to-End alpha could bring robust and easy-to-use OpenPGP crypto to Webmail.

by Dan Goodin - Jun 3, 2014 11:15 pm UTC

**PRIVACY**

Yskiflyer

Developers at Google have released an experimental tool—for Gmail and other Web-based services—that's designed to streamline the highly cumbersome task of sending and receiving strongly encrypted e-mail.

On Tuesday, the company unveiled highly unstable "alpha" code that in theory allows people to use the Google Chrome browser to generate encryption keys, encrypt e-mails sent to others, and decrypt received e-mails. Dubbed End-to-End, the Chrome extension also allows Chrome users to digitally sign and verify digital signatures of e-mails sent through Gmail and other services. The code implements a **fully compliant version of the OpenPGP standard**, which is widely regarded as providing virtually uncrackable encryption when carried out correctly.

As **Ars documented last year**, the problem with just about every e-mail encryption software available today is they require much more time and effort than sending plain-text mail. Microsoft's Outlook application, for instance, frequently crashes when working with the open-source GnuPG encryption suite. Some Outlook users, including this reporter, also experience problems when receiving encrypted e-mail from Mac users, since the encrypted messages are included in an attachment, rather in the body. End-to-End is intended to ease such burdens.

**FURTHER READING**

**ENCRYPTED E-MAIL: HOW MUCH ANNOYANCE WILL YOU TOLERATE TO KEEP THE NSA AWAY?**

How to encrypt e-mail, and why most don't bother.

**FEATURE STORY (3 PAGES)**

A fast look at Swift, Apple's new programming language

For better or worse, Apple's new language does things your way.

**WATCH ARS VIDEO**

Microsoft Surface Pro

Handling the Microsoft Surface 3, admission of its jib, and playing with its new pen.

**STAY IN THE KNOW WITH**

f t g+ e

**LATEST NEWS**

Computex lineup adds

id class : article-content clearfix previous next

tag : DIV id : article-body class : article-body mod

Textualization Name: [dropdown] [radio] text [radio] selected [radio] source [radio] inverse extract into separate column

Column Name: [dropdown]

# Sélectionner des parties des pages web



http://arstechnica.com/apple/2014/06/a-fast-look-at-swift-apples-new-programming-language/  
by John Timmer - Jun 5, 2014 1:08 pm UTC

**DEVELOPMENT** **MAC APPS**

Brendan A. Ryan

If anyone outside Apple saw Swift coming, they certainly weren't making any public predictions. In the middle of a keynote filled with the sorts of announcements you'd expect (even if the details were a surprise), Apple this week announced that it has created a modern replacement for the Objective-C, a programming language the company has used since shortly after Steve Jobs founded NeXT.

Swift wasn't a "sometime before the year's out"-style announcement, either. The same day, a 550-page language guide appeared in the iBooks store. Developers were also given access to Xcode 6 betas, which allow application development using the new language. Whatever changes were needed to get the entire Cocoa toolkit to play nice with Swift are apparently already done.

While we haven't yet produced any Swift code, we have read the entire language guide and looked at the code samples Apple provided. What follows is our first take on the language itself, along with some ideas about what Apple hopes to accomplish.

### Why were we using Objective-C?

When NeXT began, object-oriented programming hadn't been widely adopted, and few languages available even implemented it. At the time, then, Objective-C probably seemed like a good choice, one that could incorporate legacy C code and programming habits while adding a layer of object orientation on top.

But as it turned out, NeXT was the only major organization to adopt the language. This had some positive aspects, as the

**FURTHER READING**  
**APPLE SHOWS OFF SWIFT, ITS**

in penalties—settle in for \$20 per song”  
Growing copyright cop Rightscorp hope profitable alternative to "six strikes."

**WATCH ARS VIDEO**

### Microsoft Surface Pro

Handling the Microsoft Surface 3, amidst its jib, and playing with its new pen.

**STAY IN THE KNOW WITH**

**LATEST NEWS**

Computex lineup includes "five three-in-one laptop" and other

US Secret Service wants software to "detect sarcasm" on social media

tag : DIV  id  class : article-content clearfix all previous next

Brendan A. Ryan  
If anyone outside Apple saw Swift coming, they certainly weren't making any public predictions. In the middle of a keynote filled with the sorts of announcements you'd expect (even if the details were a surprise), Apple this week announced that it has created a modern replacement for the Objective-C, a programming language the company has used since shortly after Steve Jobs founded NeXT. Swift wasn't a "sometime before the year's out"-style announcement, either. The same day, a 550-page language guide appeared in the iBooks store. Developers were also given access to Xcode 6 betas, which allow application development using the new language. Whatever changes were needed to get the entire Cocoa toolkit to play nice with Swift are apparently already done. While we haven't yet produced any Swift code, we have read the entire language guide and looked at the code samples Apple provided. What follows is our first take on the language itself, along with some ideas about what Apple hopes to accomplish. Why were we using Objective-C? When NeXT began, object-oriented programming hadn't been widely adopted, and few languages available even implemented it. At the time, then, Objective-C probably seemed like a good choice, one that could incorporate legacy C code and programming habits while adding a layer of object orientation on top. But as it turned out, NeXT was the only major organization to adopt the language. This had some positive aspects, as the

Textualization: standard  
Column: content

text source invert extract into separate column

# Sélectionner des parties des pages web



http://arstechnica.com/apple/2014/06/a-fast-look-at-swift-apples-new-programming-language/

ars technica

SUBSCRIBE

## GEAR & GADGETS / PRODUCT NEWS & REVIEWS

### Google releases Android 4.4.3 to Nexus devices

The Nexus 4, 5, 7, and 10 can all manually update right now.

by Ron Amadeo - Jun 2, 2014 11:35 pm UTC

ANDROID

Google has just released a new version of Android: 4.4.3. After Sprint *totally* jumped the gun by announcing the update in April, Google has finally pushed 4.4.3 to the [Nexus Factory Image](#) page. T-Mobile has a changelog for the update, which is wildly descriptive:

Improvements

LATEST FEATURE STORY  
FEATURE STORY (3 PAGES)  
A fast look at Swift, Apple's new programming language  
For better or worse, Apple's new language does things your way.

WATCH AND VIDEO

tag : DIV  id  class : article-content clearfix all previous next

Google has just released a new version of Android: 4.4.3. After Sprint *totally* jumped the gun by announcing the update in April, Google has finally pushed 4.4.3 to the [Nexus Factory Image](#) page. T-Mobile has a changelog for the update, which is wildly descriptive:

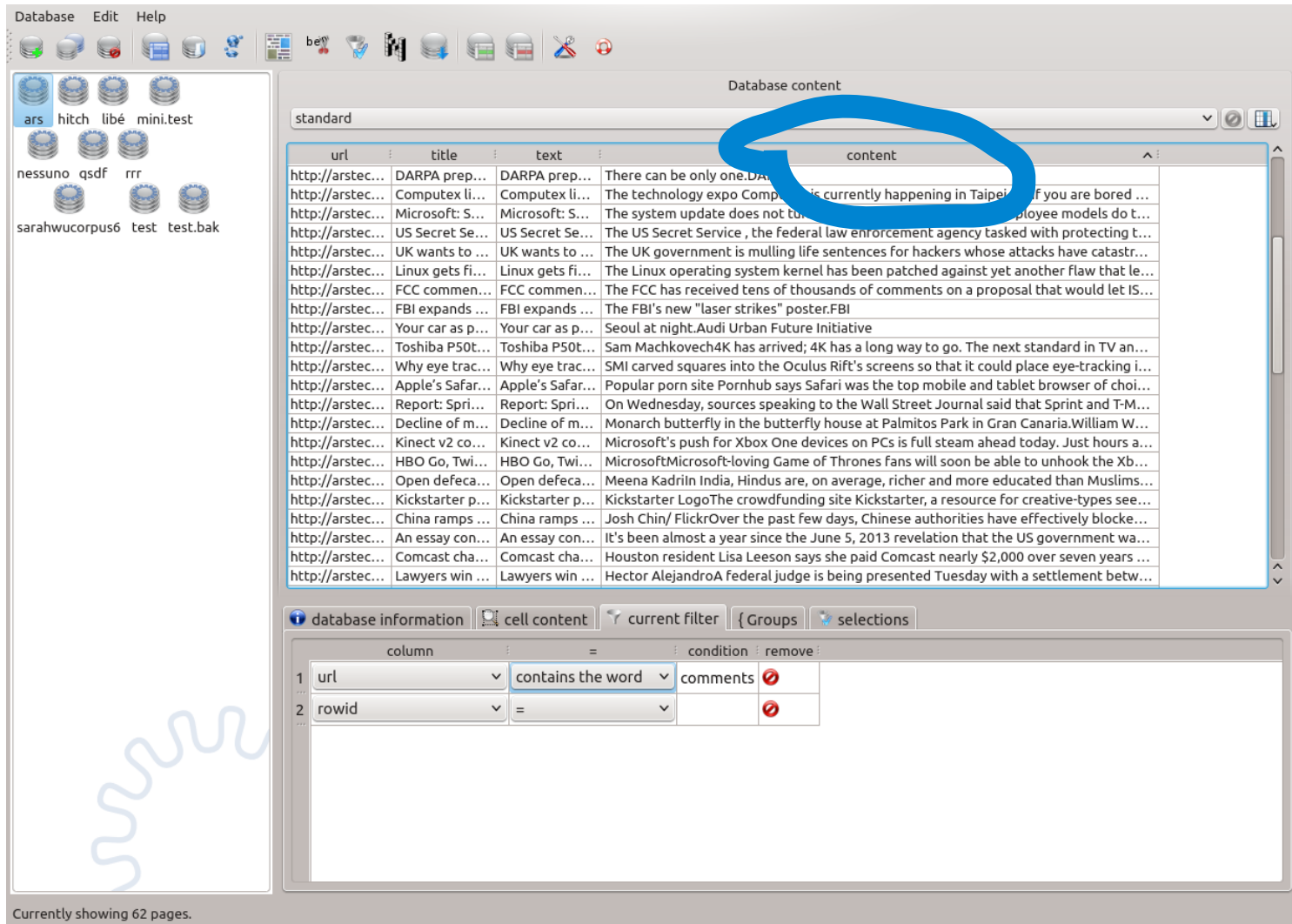
Improvements  
Security enhancements Various bug fixes Google is hosting 4.4.3 images for the Nexus 5, both Nexus 7s, Nexus 10, and Nexus 4. These are full images that must be manually applied and will erase everything on your device. Users interested in not losing their data should wait for the OTA update, which is slowly rolling out to devices now. And if you're the type of person interested in seeing the code, the AOSP code drop is going on right now, too.

Textualization Name: standard  
Column Name: content

text  selected  source  inverse extract into separate column

80%

# Sélectionner des parties des pages web

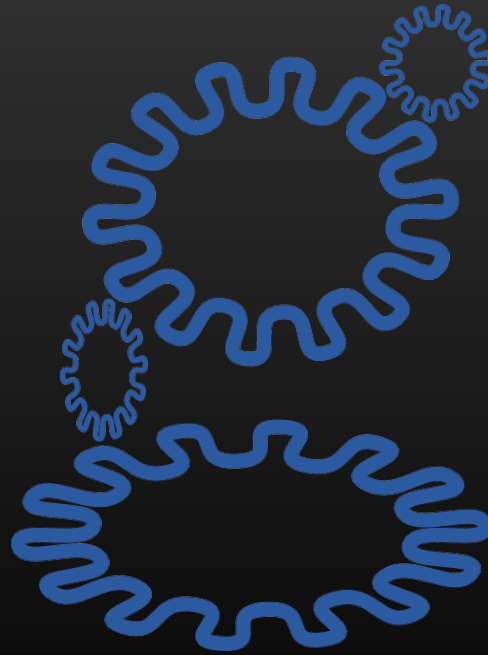


The screenshot shows a database application window with a menu bar (Database, Edit, Help) and a toolbar. On the left is a sidebar with a tree view of databases including 'ars', 'hitch', 'libé', 'mini.test', 'nessuno', 'qsdf', 'rrr', 'sarahwucorpus6', 'test', and 'test.bak'. The main area displays a table titled 'Database content' with columns 'url', 'title', 'text', and 'content'. A blue selection filter is applied to the 'content' column, indicated by a blue circle around the column header and a blue wavy line in the bottom left corner. Below the table is a filter configuration area with tabs for 'database information', 'cell content', 'current filter', and 'Groups'. The 'current filter' tab is active, showing a table with two rows of filter conditions.

column	condition	remove
1 url	contains the word	comments
2 rowid	=	

Currently showing 62 pages.

# Étiquetage & lemmatisation





# Lemmatiser

- Sélectionner seulement la colonne à lemmatiser (peu importe si les colonnes numériques sont incluses ou non)



Grosmoteur

Database Help

b bbb h hitch jiaotong

popularscience

Database content

standard

text_nbSentences	
82	22He was standing with his back to Ap...
6	23It is an important and popular fact...
98	24Silently the aircar coasted throug...
102	25There are of course many probl...
7	26"Yes, very salutary," said Arthur...
71	27Slartibartfast's study was a tota...
41	28It was a long time before anyone...
111	29"Zaphod! Wake up!"
39	30"So there you have it," said Slartib...
119	31It is of course well known that care...
82	32"Emergency! Emergency!" blared the k...
45	33But the end never came, at least not...
36	34The aircar rocketed them at speeds in...
10	35That night, as the Heart of Gold was...

- rowid
- url
- time
- linkOrigin\_rowid
- title\_nbCharacters
- title\_nbSentences
- title
- text\_nbCharacters
- text\_nbSentences
- text

database information cell content current filter selections

column	=	condition	remove
1 rowid	=		

# Lemmatiser

- Sélectionner seulement la colonne à lemmatiser (peu importe si les colonnes numériques sont incluses ou non)

standard
contenu
C'est la première visite d'Etat en Turquie d'un président français depuis celle qu'effectua il y a v ...
Sur le même sujet Tchat aujourd'hui à 15h30 C'est quoi, le pacte de responsabilité ?Début des n...
Sur le même sujet Le Dr Dukan définitivement radié de l'ordre des médecinsDukan, option latin ...
L'air de famille ne saute pas aux yeux. Pourtant Gérard Deguise raconte que parfois, même les h...
Une relique religieuse, une fiole contenant du sang de l'ancien pape polonais Jean Paul II, a été ...
Les partisans d'une candidature française aux JO-2024 ont une occasion rêvée de prendre le po...
Arcole Industries «menace de retirer son offre» de reprise du transporteur Mory Ducros si quelq...
Une enquête a été ouverte à Paris sur le financement d'un des meetings de Nicolas Sarkozy épín...
Portrait zappé Le député des Yvelines ne décolère pas contre les «clans», les «arrangements» et...
Sur le même sujet Réactions «Jour de colère» : le Crif condamne les slogans racistes et antisémi...
Le président de l'UDI, Jean-Louis Borloo, a été hospitalisé dimanche après-midi à Paris pour une ...
Une enseignante du collège Rosa-Bonheur du Châtelet-en-Brie (Seine-et-Marne) s'est vu prescri...
Sur le même sujet info libé L'étrange recrutement de Thomas Le Drian, fils de ministrePar Tonin...
Vincent Peillon a annoncé lundi que les professeurs seraient à l'avenir mieux formés et dotés de...
Au moins un des deux adolescents toulousains qui avaient frappé les esprits en quittant l'école ...
Les touristes étrangers ont dépensé 59 milliards d'euros en Espagne en 2013, une somme recor...
Une enquête a été ouverte à Paris sur le financement d'un meeting de Nicolas Sarkozy, à Toulon...
Les négociations de paix à Genève sur le conflit syrien sont bloquées après l'incapacité des émis...
Les dirigeants tunisiens ont paraphé lundi la nouvelle Constitution, un texte historique devant p...
Harlem Désir, Premier secrétaire du PS, a pris la défense du président François Hollande contre ...
Nathalie Kosciusko-Morizet, candidate UMP à la mairie de Paris, a vu lundi dans l'annonce par Fr...
La victoire de Stanislas Wawrinka dimanche en finale de l'Open d'Australie, aussi paradoxale soit...



# Lemmatiser

The screenshot shows the Grosmoteur application window. The left sidebar contains a tree view of databases: 'b', 'bbb', 'h', 'hitch', 'jiaotong', and 'popularscience'. The 'hitch' database is selected. The main area displays the 'Database content' for the 'standard' table, showing a table with two columns: 'text\_nbSentences' and 'text'. The table contains 35 rows of data, with the first row being: 82 | 22He was standing with his back to Arthur watching the very last...

text_nbSentences	text
82	22He was standing with his back to Arthur watching the very last
6	23It is an important and popular fact that things are not always v
98	24Silently the aircar coasted through the cold darkness, a single
102	25There are of course many problems connected with life, of wh
7	26"Yes, very salutary," said Arthur, after Slartibartfast had relate
71	27Slartibartfast's study was a total mess, like the results of an ex
41	28It was a long time before anyone spoke.
111	29"Zaphod! Wake up!"
39	30"So there you have it," said Slartibartfast, making a feeble and
119	31It is of course well known that careless talk costs lives, but the
82	32"Emergency! Emergency!" blared the klaxons throughout Mag
45	33But the end never came, at least not then.
36	34The aircar rocketed them at speeds in excess of R17 through t
10	35That night, as the Heart of Gold was busy putting a few light y

At the bottom of the window, there are tabs for 'database information', 'cell content', 'current filter', and 'selections'. The 'current filter' tab is active, showing a table with one row: 1 | rowid | = | [red circle with slash].



# Lemmatiser

The image shows a screenshot of the Grosmoteur software interface. The main window is titled "Grosmoteur" and has a menu bar with "Database" and "Help". Below the menu bar is a toolbar with various icons. On the left side, there is a "Database content" panel showing a list of databases: "b", "bbb", "h", "hitch", "jiaotong", and "popularscience". The "hitch" database is selected. In the center, a "Gros Tools" dialog box is open, showing the "Lemmatizer" tab. The dialog box has a title bar with "Gros Tools" and standard window controls. Below the title bar, it says "source columns: title, text". There are three tabs: "Lemmatizer", "Word Segmenter", and "Replacer". The "Lemmatizer" tab is active. It has a "Language" dropdown menu set to "English". Below that, there is a text field "added to column name:" with the value "\_lem". At the bottom of the dialog, there are four radio buttons: "lemma" (selected), "tag", "lemma-tag", and "complete". A large blue circle is drawn around the "Lemmatizer" dialog box. At the bottom of the dialog, there is a "go" button with a green gear icon. The background of the main window shows a text document with the following text: "spend so much of the intervening time wearing digital watches? Many many millions of years ago a race of hyperintelligent pan-dimensional beings (whose physical manifestation in their own pan-dimensional universe is not dissimilar to our own) got so fed up with the constant tinkering about the meaning of life which used to interrupt their favourite pastime of Brockian D... (which involved suddenly hitting people for no readily apparent reason and then running away) that they decided to sit down and solve their problems once and for all."

# Lemmatiser

The image shows a screenshot of the Grosmoteur software interface. The main window is titled "Grosmoteur" and has a menu bar with "Database" and "Help". Below the menu bar is a toolbar with various icons, including a "be" logo. The main area is divided into two panes. The left pane shows a "Database content" view with a list of databases: "b", "bbb", "h", "hitch", "jiaotong", and "popularscience". The "hitch" database is selected. The right pane shows a table with one row containing the number "9".

Overlaid on the main window is a "Gros Tools" dialog box. The dialog box has a title bar with "Gros Tools" and standard window controls. It contains the following elements:

- source columns: title, text
- Three tabs: "Lemmatizer" (selected), "Word Segmenter", and "Replacer".
- A "Language" dropdown menu set to "English".
- An "added to column name:" text box containing "\_lem".
- Four radio buttons: "lemma" (selected), "tag", "lemma-tag", and "complete".
- A "GO" button with a green gear icon, which is circled in blue.
- Text at the bottom: "standard is the input view".

Below the dialog box, a text area contains the following text:

some of the they want to spend so much of the intervening time wearing digital watches? Many many millions of years ago a race of hyperintelligent pandimensional beings (whose physical manifestation in their own pan-dimensional universe is not dissimilar to our own) got so fed up with the constant bickering about the meaning of life which used to interrupt their favourite pastime of Brockian Ultra Cricket (a curious game which involved suddenly hitting people for no readily apparent reason and then running away) that they decided to sit down and solve their problems once and for all.

# Lemmatiser

went → go

The screenshot shows the Grosmoteur application window. The main window has a menu bar with 'Database' and 'Help', and a toolbar with various icons. Below the toolbar is a 'Database content' section with a dropdown menu set to 'standard' and a table with one row containing the number '9'. A 'Gros Tools' dialog box is open in the foreground, featuring a close button circled in blue. The dialog box has three tabs: 'Lemmatizer', 'Word Segmenter', and 'Replacer'. The 'Lemmatizer' tab is active, showing a 'Language' dropdown set to 'English', an 'added to column name:' field with the value '\_lem', and four radio button options: 'lemma' (selected), 'tag', 'lemma-tag', and 'complete'. At the bottom of the dialog is a 'go' button with a green gear icon. The background of the application shows a list of database tables including 'b', 'bbb', 'h', 'hitch', 'jiaotong', and 'popularscience'. A text area at the bottom of the window contains the following text:

some of the they want to spend so much of the intervening time wearing digital watches? Many many millions of years ago a race of hyperintelligent pandimensional beings (whose physical manifestation in their own pan-dimensional universe is not dissimilar to our own) got so fed up with the constant bickering about the meaning of life which used to interrupt their favourite pastime of Brockian Ultra Cricket (a curious game which involved suddenly hitting people for no readily apparent reason and then running away) that they decided to sit down and solve their problems once and for all.

# Un texte lemmatisé

**le négociation de paix à genève sur le conflit syrien être bloquer** après le incapacité des émissaire de bachar al-assad et de le opposition à discuter lundi de la délicate question de transfert de pouvoir .

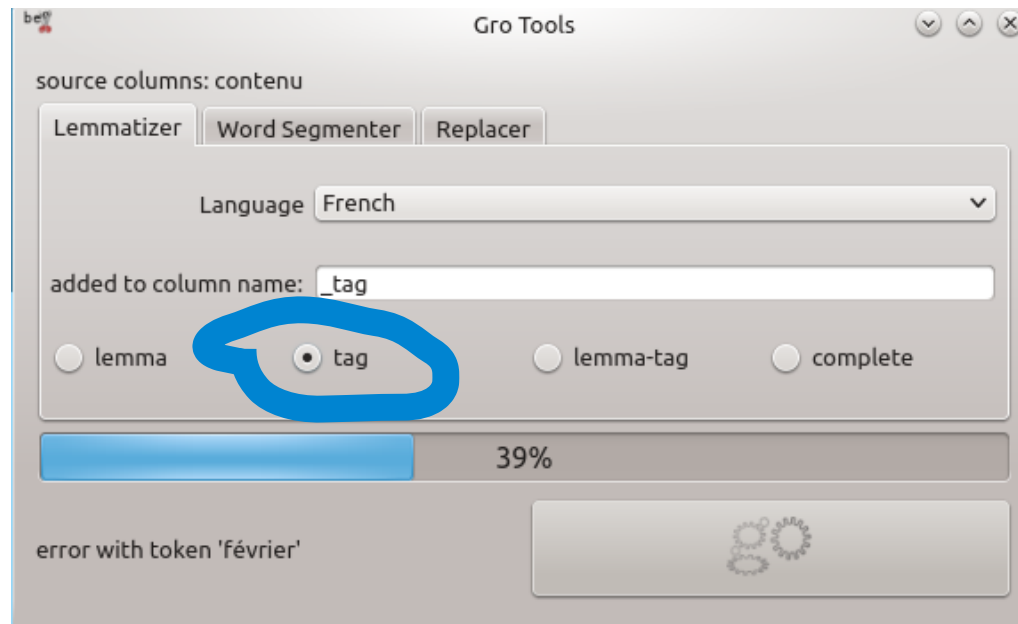
réunir pour la troisième journée consécutif au siège de le onu à genève , le deux délégation devoir aborder le question politique après avoir discuter pendant le week-end de question humanitaire , comme la situation à homs et le problème des millier de prisonnier et de disparus de conflit .

mais la réunion de lundi matin avoir tourner court .

«les discussion ne avoir pas être constructif aujourd' hui en raison de le attitude de régime qui avoir vouloir dévier des discussion qui devoir porter sur le application de genève i» , le texte rédiger en juin 2012 par le grand puissance , avoir déclarer rima fleyhane , membre de la délégation de le opposition .

# Étiquetage (taggeur)

- Went → V





# Les étiquettes d'un texte

NN ; VB DT CD NN IN NNP IN NNP IN DT NN JJ IN PRP PRP VB PRP PRP VB CD NNS NNP  
NNP , CC DT NN PRP NN JJ .

NN ; NN IN NN IN NNS VB RB DT NN IN VB DT NN , IN DT NN IN DT NNP NNP JJ IN DT NN IN  
NN IN DT NNP JJ ( DT NNS VB VBN IN CD ) .

NN ; RB IN DT NNP PRP VB IN NN JJ NN RB JJ IN PRP\$ NN JJ IN IN DT NN IN PRP\$ NN .  
NNP NN .

CC DT NN IN CD NNS WP VB NN NN VB RB IN RB IN NNS CC IN NNS NNS PRP VB NN NN .  
IN DT NN IN NNS , IN NN , DT NN IN CD NN NN , NNP NNP NNP , CC IN DT NN NN VBG NN ,  
NN CC NN JJ VB RB VBN IN RB .

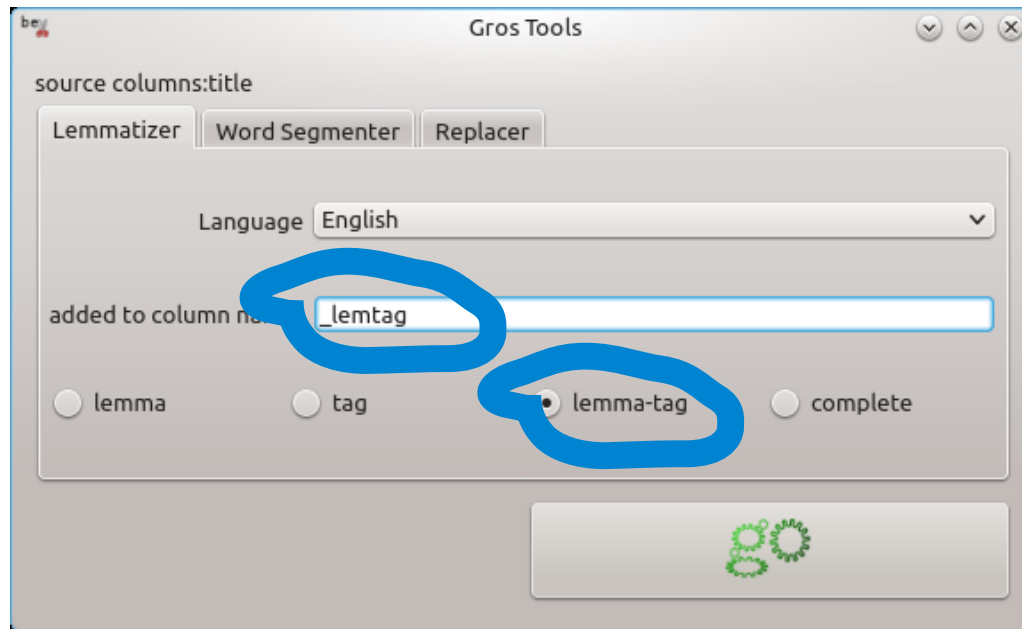
IN NN , DT NN IN NNP CC IN NN IN DT JJ NN IN DT NNP RB PRP VB VBN IN NNP .  
DT NN VB IN PRP NN .

IN RB VB VBN , NNP , VBN IN NN NNS CC IN JJ NNS IN NN , VB IN DT NN VBN IN DT JJ NN JJ  
.

NNP NN VB IN DT NN NN IN PRP RB VB RB RB VB IN NN JJ CC RB RB IN NNS NN , VB DT JJ  
NN IN NNP IN DT NN CC IN NN ( NN ) , RB NN IN DT NN IN NN JJ IN NN .

# Étiquetage (taggeur)

- Went → go-v



# Lemma-catégorie

le-DT négociation-NNS de-IN paix-NN à-IN genève-NNP sur-IN le-DT conflit-NN syrien-JJ être-VB bloquer-VBN après-IN le-DT incapacité-NN des-IN émissaire-NNS de-IN bachar-NNP al-assad-NN et-CC de-IN le-DT opposition-NN à-IN discuter-VB lundi-NN de-IN la-DT délicate-JJ question-NN de-IN transfert-NN de-IN pouvoir-NN .-.

réunir-VBN pour-IN la-DT troisième-CD journée-NN consécutif-JJ au-IN siège-NN de-IN le-DT onu-NN à-IN genève-NNP ,-, le-DT deux-CD délégation-NNS devoir-VB aborder-VB le-DT question-NNS politique-JJ après-IN avoir-VB discuter-VBN pendant-IN le-DT week-end-NN de-IN question-NNS humanitaire-NNS ,-, comme-IN la-DT situation-NN à-IN homs-NNP et-CC le-DT problème-NN des-IN millier-NNS de-IN prisonnier-NNS et-CC de-IN disparus-NN de-IN conflit-NN .-.

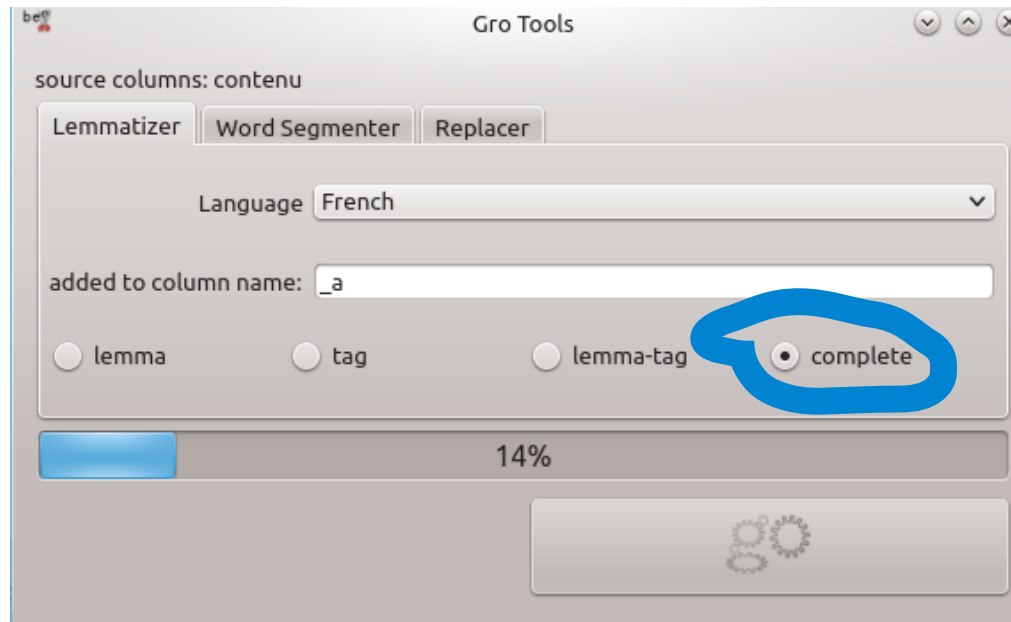
mais-CC la-DT réunion-NN de-IN lundi-NN matin-RB avoir-VB tourner-VBN court-RB .-.

«les-NNP discussion-NNS ne-RB avoir-VB pas-RB être-VBN constructif-JJ aujourd'-NN hui-NN en-IN raison-NN de-IN le-DT attitude-NN de-IN régime-NN qui-WP avoir-VB vouloir-VBN dévier-VB des-IN discussion-NNS qui-WP devoir-VB porter-VB sur-IN le-DT application-NN de-IN genève-NNP i»-NNP ,-, le-DT texte-NN rédiger-VBN en-IN juin-NN 2012-CD par-IN le-DT grand-JJ puissance-NNS ,-, avoir-VB déclarer-VBN rima-NNP fleyhane-NNP ,-, membre-NN de-IN la-DT délégation-NN de-IN le-DT opposition-NN .-.

我们/PN 所/MSP 要/VV 介绍/VV 的/DEC 是/VC 祥子/NN , /PU 不/AD 是/VC 骆驼/NN , /PU 因为/P "/PU 骆驼/NN "/PU 只/AD 是/VC 个/M 外号/NN

# Analyse complète

- Went → go-v

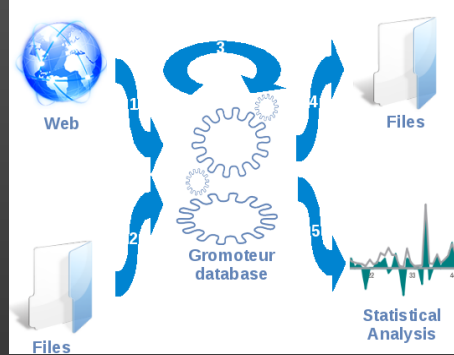


# Un texte analysé avec segmentation en chunks

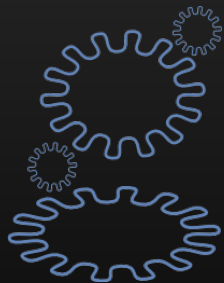
Les/DT/B-NP/O/O/le négociations/NNS/I-NP/O/O/négociation de/IN/B-PP/B-PNP/O/de paix/NN/B-NP/I-PNP/O/paix à/IN/B-PP/B-PNP/O/à Genève/NNP/B-NP/I-PNP/O/genève sur/IN/B-PP/B-PNP/O/sur le/DT/B-NP/I-PNP/NP-SBJ-1/le conflit/NN/I-NP/I-PNP/NP-SBJ-1/conflit syrien/JJ/I-NP/I-PNP/NP-SBJ-1/syrien sont/VB/B-VP/O/VP-1/être bloquées/VBN/I-VP/O/VP-1/bloquer après/IN/B-PP/B-PNP/O/après l'/DT/B-NP/I-PNP/O/le incapacité/NN/I-NP/I-PNP/O/incapacité des/IN/B-PP/B-PNP/O/des émissaires/NNS/B-NP/I-PNP/O/émissaire de/IN/B-PP/B-PNP/O/de Bachar/NNP/B-NP/I-PNP/O/bachar al-Assad/NN/I-NP/I-PNP/O/al-assad et/CC/O/O/O/et de/IN/B-PP/B-PNP/O/de l'/DT/B-NP/I-PNP/O/le opposition/NN/I-NP/I-PNP/O/opposition à/IN/B-PP/O/O/à discuter/VB/B-VP/O/VP-2/discuter lundi/NN/B-NP/O/NP-OBJ-2/lundi de/IN/B-PP/B-PNP/O/de la/DT/B-NP/I-PNP/O/la délicate/JJ/I-NP/I-PNP/O/délicate question/NN/I-NP/I-PNP/O/question du/IN/B-PP/B-PNP/O/de transfert/NN/B-NP/I-PNP/O/transfert de/IN/B-PP/B-PNP/O/de pouvoir/NN/B-NP/I-PNP/O/pouvoir ././O/O/O/.

# Outils pour différentes langues

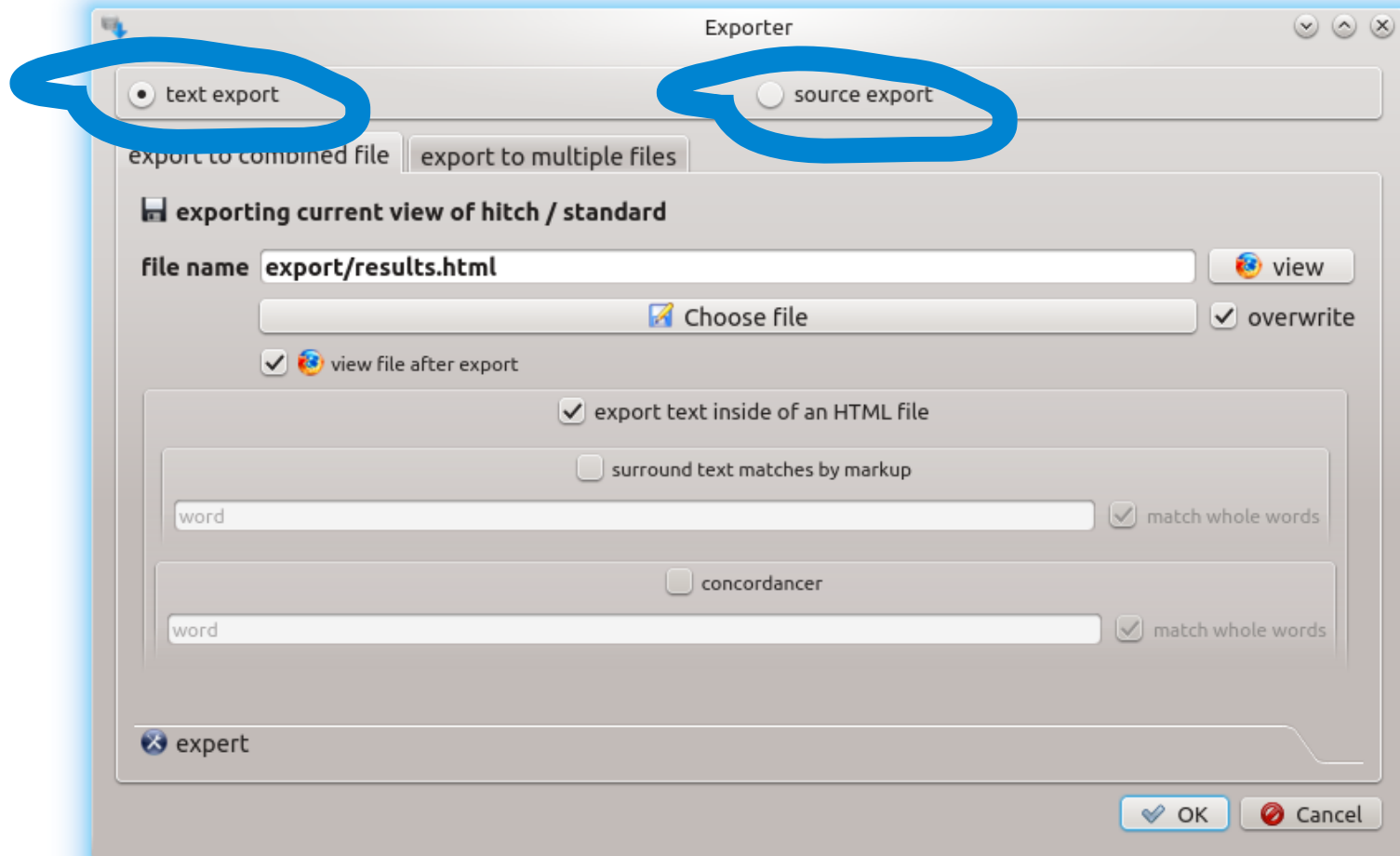
- Lemmatisation + étiquetage :
  - Français, anglais, allemand, espagnol, italien, néerlandais, chinois
- Segmentation pour le chinois
- Corrections globales
  - Par expressions régulières



# Exportation



# Export





# Export



20.	To Send People On A One-Way Trip To Mars 73418 359 7 Why We Can't Stop Eating Frosting From The	Can	73348 359 8 Wing And A Scare 72205 359 9 Dear Congress: Why Are You So Anti-Science? 73263 359 10
<a href="http://www.popsci.com/science/article/2013-05/how-avoid-meeting-neanderthals-fate">http://www.popsci.com/science/article/2013-05/how-avoid-meeting-neanderthals-fate</a>			
21.	What Modern Humans	Can	Learn From The Neanderthals' Extinction   Popular Science
22.	What Modern Humans	Can	Learn From The Neanderthals' Extinction   Popular Science Login/Register Newsletter Subscribe
23.	science 73784 What Modern Humans	Can	Learn From The Neanderthals' Extinction It's a fact of the archaeological record: Modern humans
24.	bones, tools, and pieces of art—along with some DNA that modern humans inherited from them. How	can	we avoid meeting the Neanderthals' fate? That depends on what you think wiped out these early
25.	Extermination and Assimilation The complicated debate over what happened to Neanderthals	can	be boiled down to two dominant theories: Either H. sapiens destroyed the other humans, or joined
26.	ancestry back to a single H. sapiens woman from Africa, nicknamed Mitochondrial Eve. If all of us	can	trace our roots back to one African woman, then how could we be the products of crossbreeding? We
27.	More evidence for Hawks's claims comes from Neanderthal DNA. Samples of their genetic material	can	reveal just what happened after all that Pleistocene hanky-panky. A group of geneticists at the
28.	Several of those regions contain genes connected to the neurological connections that humans	can	form in their brains. In other words, it's possible that H. sapiens ' greater capacity for
29.	many times over. And it spawned deadly famines, too. Humanity's old community-building habits	can	become pathological on a mass scale. Thousands of years after the merging of Neanderthals and H.
30.	Group (Canada), a division of Pearson Canada, Inc. Previous Article: Electrical Brain Stimulation	Can	Help You Learn Math Next Article: FYI: Which Emotion Is The Hardest To Fake? 16 Comments Link to
31.	and battles. They were hunter gatherers, not the civilization builders of the later B.Cs. Why	can	't PopSci writers actually research what they write about? Your dedicated readers are people who
32.	that. Link to this comment mike13323 05/16/13 at 6:58 pm If you want me to go into details John I	can	provide you with evidence supporting all of my criticisms. First, recent carbon dating has moved
33.	40,000 years ago. Only with an open mind and a willingness to look at all angles of the equation,	can	we hope to GUESS at what our ancestors thought. Ideas and beliefs are such an ethereal thing that
34.	to this comment GodLikesComedy 05/17/13 at 12:13 pm Sorry I didn't read everything, but how we	can	survive? 1 STOP WARS 2 STOP FIAT CURRENCIES 3 STOP BORDERS 4 STOP CREATING WEAPONS 5 S
35.	and theorize about the past is quite the debate in archaeology and paleoanthropology. The best we	can	do at the moment is try to not corrupt the past with modern notions like we've mentioned above.
36.	Of Darker Skin Makes Them Less Racist Space Tourism's Black Carbon Problem What Modern Humans	Can	Learn From The Neanderthals' Extinction Untouched For The Last Billion Years, Water In Canadian
37.	Climate Modeling Method Wanna Know How You're Going To Survive The Apocalypse? This Bacterium	Can	Do Division. Compute Logarithms And Take Square Roots Most Viewed Science The Weak In Numbers:

# Export

eccentrica gallumbits , the triple-breasted whore of eroticon 6 .  
" arthur follow ford 's finger , and see where it be point .  
for a moment it still do n't register , then his mind nearly blow up .  
" what ?

**harmless** ?

be that all it be get to say ?

**harmless** !

one word !

" ford shrug .

" well , there be a hundred billion star in the galaxy , and only a limited amount of space in the book 's microprocessor , " he say , " and no one know much about the earth of course .

" " well for god 's sake i hope you manage to rectify that a bit .

" " oh yes , well i manage to transmit a new entry off to the editor .

he have to trim it a bit , but it be still an improvement .

" " and what do it say now ?

" ask arthur .

" mostly **harmless** , " admit ford with a slightly embarrassed cough .

" mostly **harmless** !

" shout arthur .

" what be that noise ?

" hiss ford .

" it be me shout , " shout arthur .

" no !

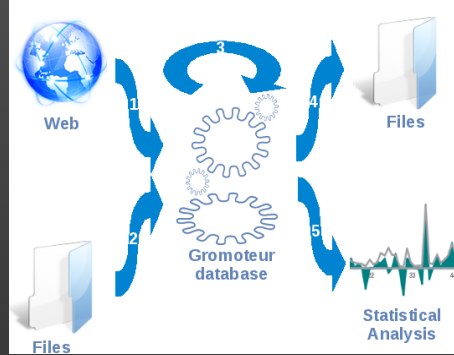
shut up !

" say ford .

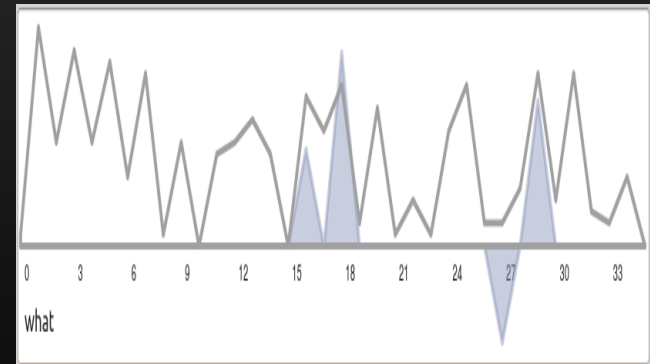
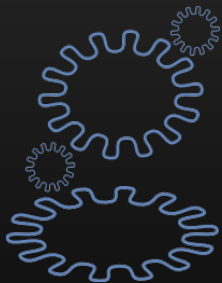
i think we be in trouble .

" " you think we be in trouble !

" outside the door be the sound of march foot .



# Analyse statistique



# Fenêtre principale : choix de la colonne à analyser

The screenshot shows a database application window with a menu bar (Database, Edit, Help) and a toolbar. On the left, a sidebar lists database files: ars, hitch, libé, mini.test, nessuno, qsd, rrr, sarahwucorpus6, test, test.bak. The main area is titled "Database content" and shows a dropdown menu with "standard" selected. Below it, a list of columns is displayed, with "text\_lem" circled in blue. The text preview for "text\_lem" shows the beginning of a document: "the hitch hikers guide to galaxy... out in the uncharted b... water of the unfashion...". Below the text preview, there are tabs for "database information", "cell content", "current filter", "Groups", and "selections". The "database information" tab is active, showing statistics: 0 pages, 0 sentences, 0 characters, 0 links remaining, 0 links followed, 0.8 Mb, none, and a timestamp of 4/6/14 12:24 AM. A "comments:" section is also visible, showing "no comment yet".

Database content

standard

text\_lem

the hitch hikers guide to galaxy... out in the uncharted b... water of the unfashion...  
1 the house stand on a slight rise... age . it stand on its own ...  
2 here be what the encyclopedia galactica have to say about alcohol . it say that alco...  
3 on this particular thursday , something be move quietly through the ionosphere m...  
4 far away on the opposite spiral arm of the galaxy , five hundred thousand light yea...  
5 prostetnic vogon jeltz be not a pleasant sight , even for other vogon . his highly do...  
6 " howl howl gargle howl gargle howl howl howl gargle howl gargle howl howl gargl...  
7 vogon poetry be of course the third worst in the universe . the second worst be tha...  
8 the hitch hiker 's guide to the galaxy be a wholly remarkable book . it have be comp...  
9 a computer chat to itself in alarm as it notice an airlock open and close itself for no...  
10 the infinite improbability drive be a wonderful new method of cross vast interstel...  
11 the improbability-proof control cabin of the heart of gold look like a perfectly con...  
12 a loud clatter of gunk music flood through the heart of gold cabin as zaphod sear...  
13 marvin trudge on down the corridor , still moan . " ...  
14 the heart of gold flee on silently through the night of space , now on conventiona...  
15 ( excerpt from the hitch hiker 's guide to the galaxy , page 634784 , section 5a , en...  
16 arthur awake to the sound of argument and go to the bridge . ford be wave his ar...  
17 after a fairly shaky start to the day , arthur 's mind be begin to reassemble itself fr...  
18 and the next th ing that happen after that be that the heart of gold continue on it...  
19 " be we take this robot with us ? " say ford , look with distaste at marvin who be st...  
20 five figure wander slowly over the blight land . bit of it be dullish grey , bit of it du...  
21 on the surface of magrathea arthur wander about moodily . ford have thoughtfull...

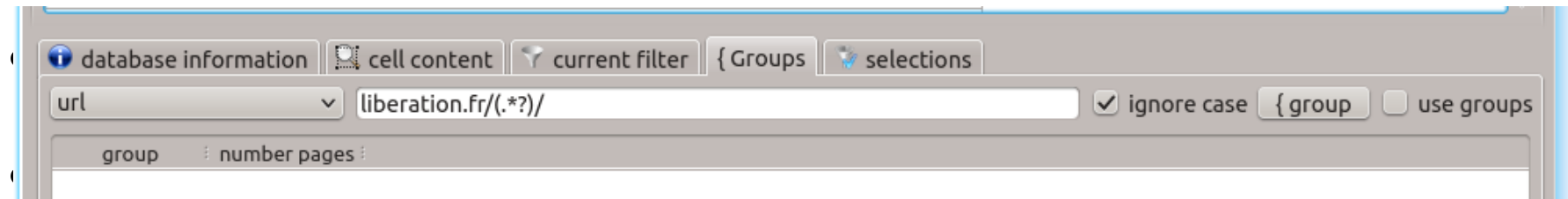
database information | cell content | current filter | Groups | selections

0 pages | 0 links remaining | 0.8 Mb | no comment yet  
0 sentences | 0 links followed | none  
0 characters | 4/6/14 12:24 AM

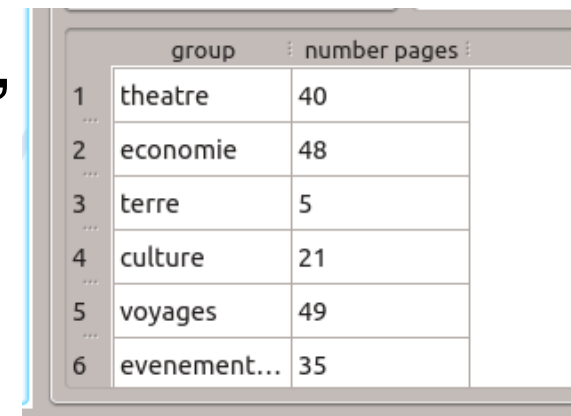
Currently showing 36 pages.

# Grouper

- Sélectionner l'onglet Groups



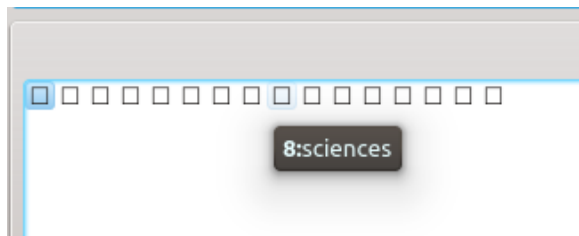
- Donner une expression régulière avec un groupe (entre parenthèses) qui correspond à la partie à grouper
- Cliquer sur le bouton “group”



	group	number pages
1	theatre	40
...		
2	economie	48
...		
3	terre	5
...		
4	culture	21
...		
5	voyages	49
...		
6	evenement...	35

# Grouper

- Si “use groups” est coché et on ouvre maintenant Nexico!, il prend comme partie pas les pages mais les groupes



- On reconnaît facilement les groupes par leurs mots spécifiques.

Specificity table for section 3.

82641 tokens

filter:

	token	totfreq	freq	spec
1	finale	96	89	51
2	joueur	62	59	51
3	ribéry	46	46	46
4	joueurs	44	44	44
5	(	1540	348	44
6	match	60	50	40
7	club	62	51	40
8	wawrinka	39	39	39
9	Federer	37	37	37
10	6	182	86	37
11	zahia	33	33	33
12	bleus	42	37	32
13	chelem	31	31	31
14	nadal	28	28	28
15	ballon	28	27	26
16	foot	32	29	26
17	mondial	54	38	26
18	sky	25	25	26
19	melbourne	24	24	25
20	buts	24	24	25
21	froome	23	23	24
22	suisse	55	37	24
23	jo	23	23	24
24	open	34	28	23
25	matchs	22	22	23
26	olympique	21	21	22
27	),	454	120	22
28	coupe	42	30	21
29	karabatic	20	20	21
30	laporte	19	19	20
31	jeux	54	33	20
32	roger	26	23	20
33	tennis	21	20	19

# Groupes & catégories

- Les sciences utilisent le plus d'adjectifs :

	token	totfreq	freq	spec
1	jj	19602	1055	15
2	nns	19582	1008	10
3	dt	41934	2016	9
4		327974	14304	4
5	.(	75	11	4
6	in	74229	3326	4
7	cc	10173	490	3
8	cd	5814	285	3
9	)	854	54	3
10	nn	74275	3278	3
11	:.:	1	1	2
12		16	1	0
13	.),	5	0	0
14	-	604	33	0
15	:+	4	0	0
16	,(	9	1	0
17	-,	56	1	0
18	.»),	2	0	0

- Les articles de la section “monde” utilisent le plus de verbes :

	token	totfreq	freq	spec
1	vbn	12481	2333	31
2	nns	19582	3301	14
3	dt	41934	6524	4
4	in	74229	11382	4
5		327974	49515	3
6	jj	19602	3045	3
7	, (...)	2	2	3
8	,	26203	4040	3
9	”;	2	2	3
10		16	1	0
11	.),	5	2	0
12	-	604	78	0
13	:+	4	2	0
14	,(	9	0	0
15	-,	56	9	0
16	.»),	2	1	0
17	:-.	1	0	0

# Nexico !


## Vue d'ensemble

Nexico!

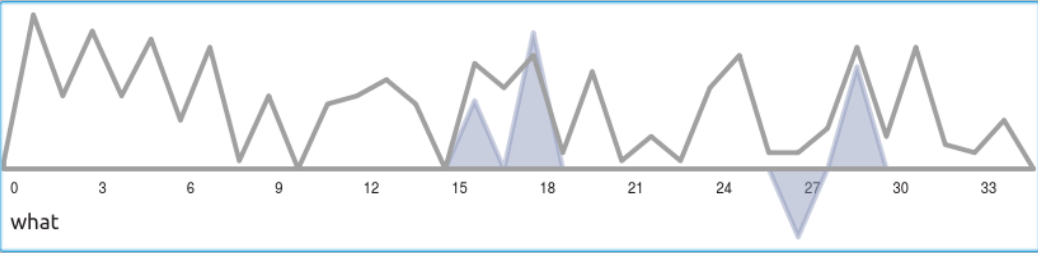
File Edit Compute Help

standard:

Text with 48398 tokens in 36 sections. Colors show occurrences (pink squares) and specificities (red +, green - background) of 'what'.



Occurrences of 'what'



Content of section 0 - hitch.000.txt - 606 tokens

the hitch hikers guide to galaxy far out in the uncharted backwater of the unfashionable end of the western spiral arm of the galaxy lie a small unregarded yellow sun . orbiting this at a distance of roughly ninety-two million mile be an utterly insignificant little blue green planet whose apedescended life form be so amazingly primitive that they still think digital watch be a pretty neat idea . this planet have - or rather have - a problem , which be this : most of the person on it be unhappy for pretty much of the time . many solution be suggest for this problem , but most of these be largely concern with the movement of small green piece of paper , which be odd because on the whole it be n't the small green piece of paper that be unhappy . and so the problem remain ; lot of the person be mean , and most of them be miserable , even the one with digital watch . many be increasingly of the opinion that they will all make a big mistake in come down from the tree in the first place . and some say that even the tree have be a bad move , and that no one should ever have left the ocean . and then , one thursday , nearly two thousand year after one man have be nail to a tree .

NEXICO!

Specificity table for section 0.

606 tokens

filter: ha

token	otfre	freq	s
that	586	8	
have	573	11	
what	273	1	
than	43	3	
happen	39	0	
hand	29	0	
thank	22	0	
half	19	0	
perhaps	15	0	
shape	14	0	
hatchway	14	0	
sharply	13	0	
shall	11	0	
shake	11	0	
whatever	11	0	
whale	10	0	
chamber	10	0	
chance	10	0	
hang	9	0	
alpha	9	0	
hate	9	0	
change	9	1	
etha	9	0	
shadow	8	0	
ghastly	8	0	
harmless	8	0	
hard	8	0	
happy	7	1	
chair	6	0	
hat	6	0	
hatch	5	0	



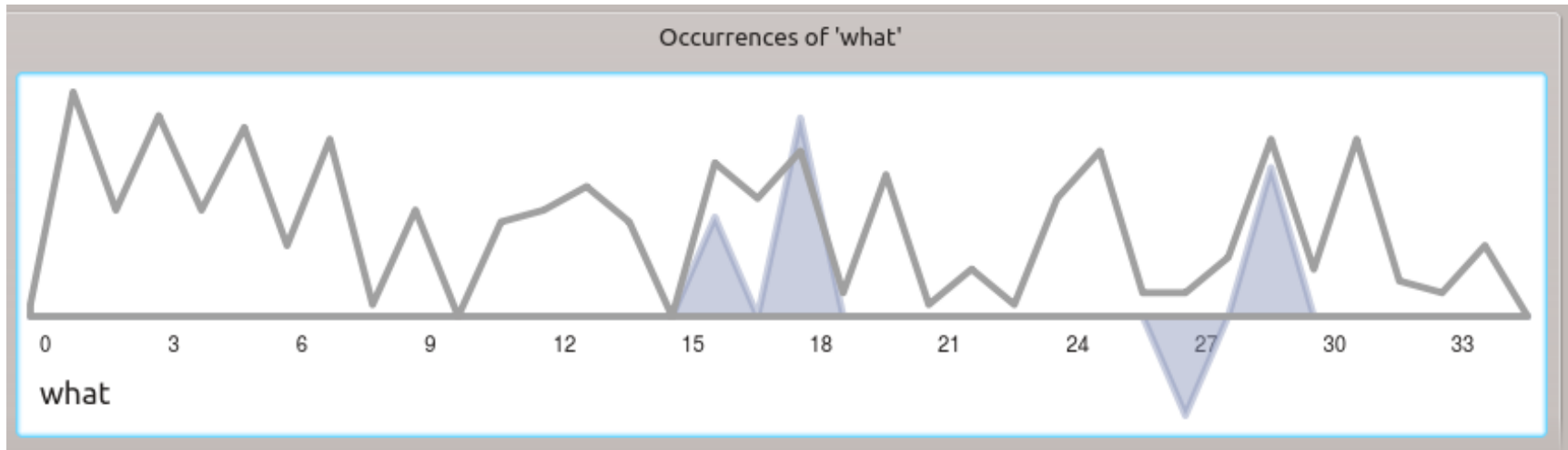
# Nexico !

## Carte des sections



# Nexico !

## Occurrence et spécificité



# Nexico !

## Occurrence et spécificité

Specificity table for section 18.

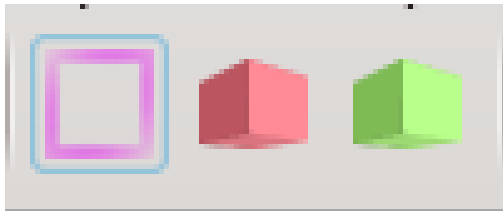
933 tokens

filter:

	token	totfreq	freq	spec
480	unharm	1	0	0
1041	character	2	0	0
1539	share	1	0	0
1563	hardest	2	0	0
1587	sharp	2	0	0
2168	hardly	5	0	0
2734	uncharted	1	0	0
2766	charge	3	0	0
2920	harmless	8	0	0
2969	harm	1	0	0
2978	hard	8	0	0
3073	characterize	1	0	0
3173	chart	2	0	0
3178	charm	1	0	0
3632	sharper	1	0	0
3707	characteristic	1	0	0
3780	charming	3	0	0
3804	sharply	13	0	0
4550	harmonic	1	0	0
4624	sharklike	1	0	0

# Nexico !

## Sélection / coccurrence



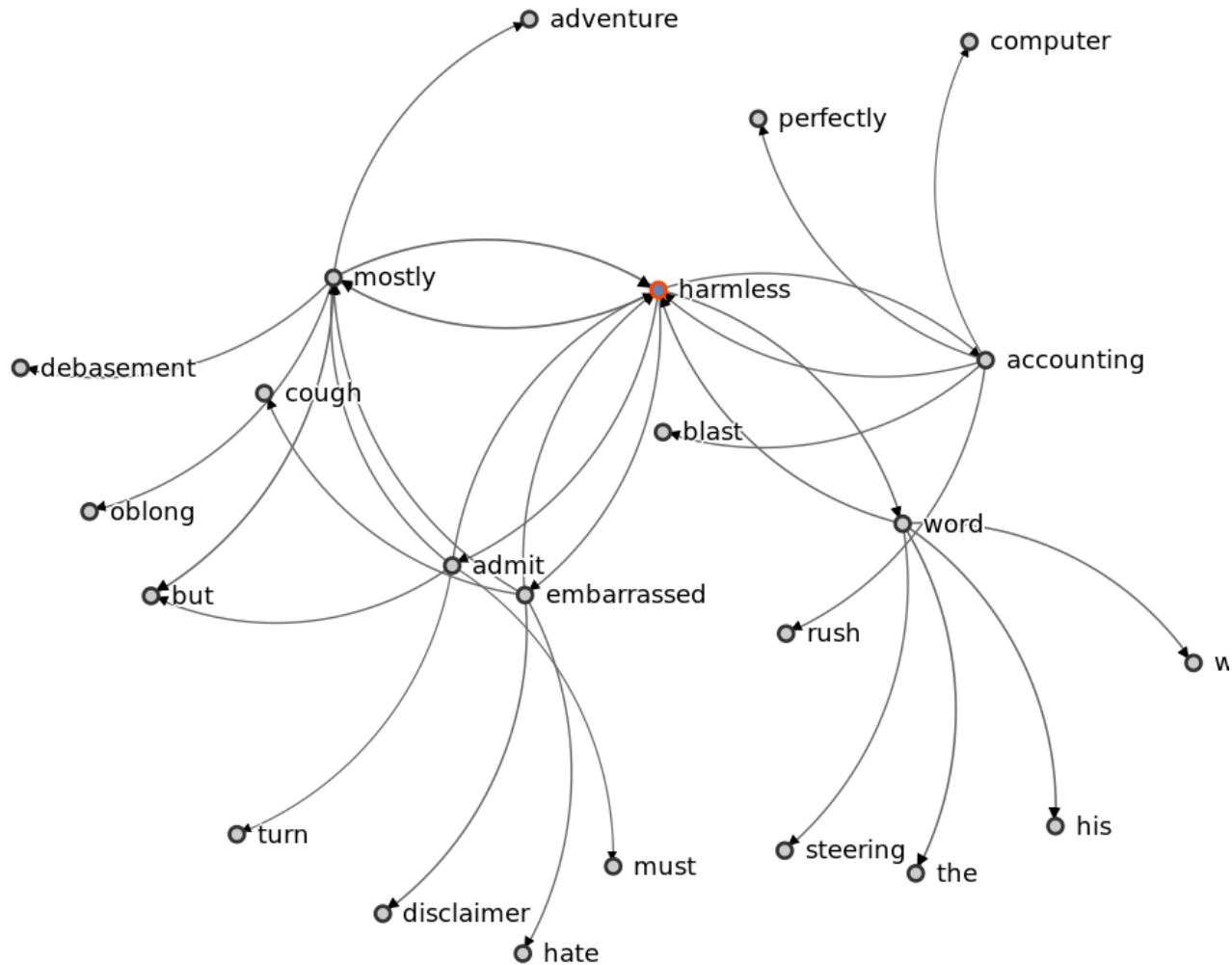
sections 4, 11, 13, 16, 17, 24, 28, 31.

13574 tokens

filter:

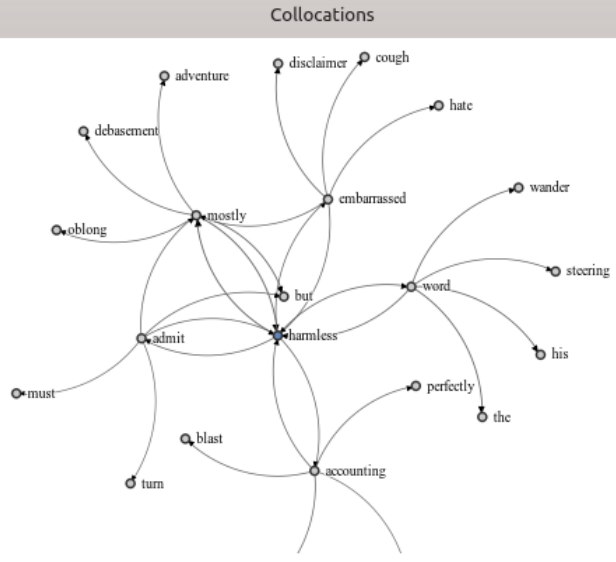
	token	totfreq	freq	spec ^
1 ...	zaphod	214	103	11
2 ...	sharply	13	13	8
3 ...	mouse	31	22	7
4 ...	marvin	49	30	7
5 ...	beebleb...	28	20	7
6 ...	president	27	19	6
7 ...	damogran	15	13	6
8 ...	trillian	79	40	6
9 ...	boat	9	9	6
10 ...	benji	16	12	5
11 ...	door	36	21	5
12 ...	arm	22	15	5
13 ...	control	21	15	5
14 ...	our	39	22	5
15 ...	infinity	10	9	5
16 ...	ultimate	11	8	4
17 ...	side	14	10	4

# Nexico ! Collocation



# Nexico ! Collocation

standard:  
Text with 48398 tokens in 36 sections. Colors show occurrences (pink squares) and specificities (red +, green - background) of 'harmless'.



Content of section 4 - hitch.004.txt - 2480 tokens

4 far away on the opposite spiral arm of the galaxy , five hundred thousand light year from the star sol , zaphod beeblesh , president of the imperial galactic government , sped across the sea of damogran , his ion drive delta boat wink and flash in the damogran sun . damogran the hot ; damogran the remote ; damogran the almost totally unheard of . damogran , secret home of the heart of gold . the boat sped on across the water . it would be some time before it reach its destination because damogran be

## NEXICO!

sections 4, 11, 13, 16, 17, 24, 28, 31.

13574 tokens

filter: harm

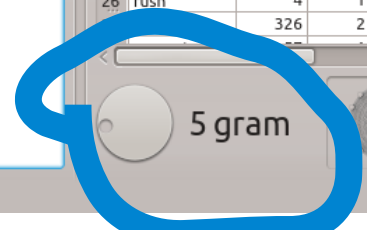
	token	totfreq	freq	spec
497	unharmd	1	1	0
2907	<b>harmless</b>	8	0	0
2959	harm	1	0	0
3166	charm	1	1	0
3756	charming	3	2	0
4506	harmonic	1	0	0



## Collocations

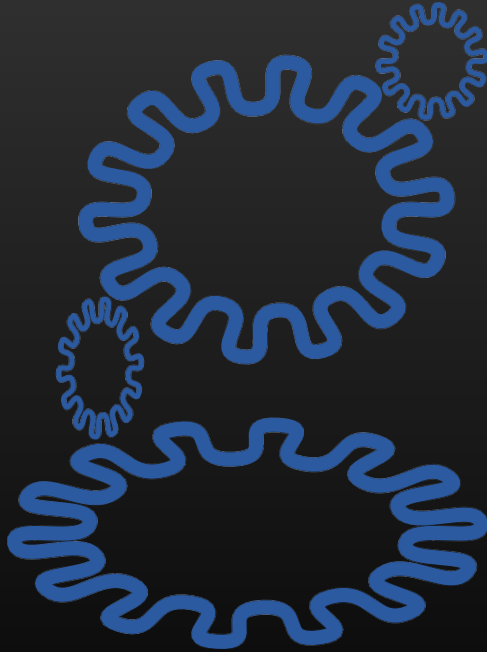
filter:

	token	totfreq	cooc	spec
1	mostly	10	4	11
2	word	26	2	5
3	embarrassed	2	1	4
4	accounting	1	1	4
5	unruly	2	1	4
6	eccentric	1	1	4
7	boozier	1	1	4
8	admit	3	1	4
9	blow	7	1	10
10	all	196	2	10
11	nearly	11	1	10
12	blast	4	1	10
13	an	155	2	10
14	gun	4	1	10
15	cough	4	1	10
16	now	117	2	10
17	be	1902	5	10
18	one	183	2	10
19	slightly	14	1	10
20	arthur	347	2	10
21	what	273	2	10
22	perfectly	22	1	10
23	left	29	1	10
24	shrug	13	1	10
25	yesterday	4	1	10
26	rush	4	1	10
		326	2	10

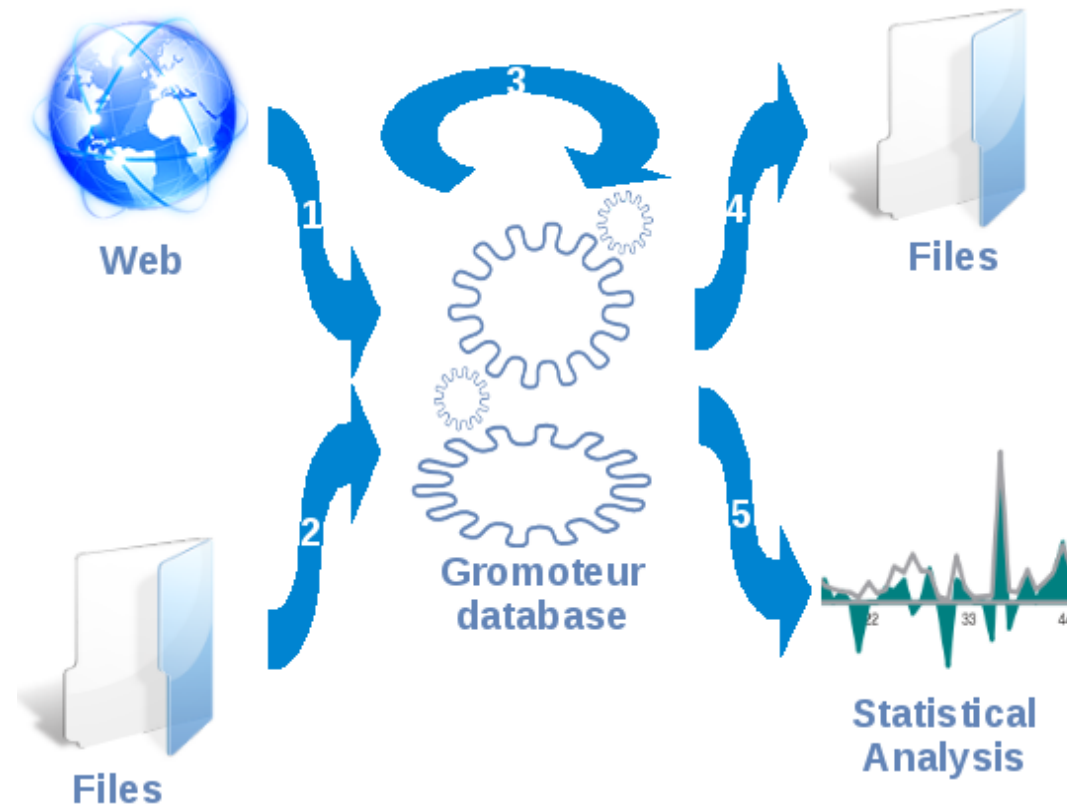


5 gram

# Résumé



# Résumé





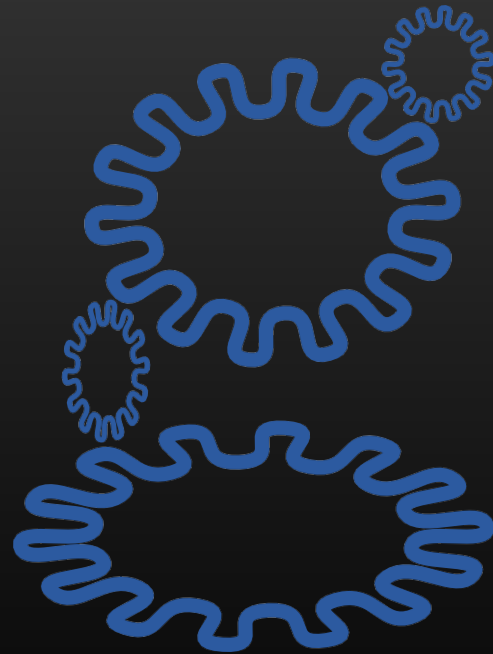
# étapes

- **Récolte / collection du corpus (web, dossiers contenant les texte)**
- **Extraction des parties de contenu**
- **Nettoyage, correction**
- **Segmentation (surtout chinois)**
- **Étiquetage :**
  - **partie du discours (catégorie syntaxique)**
  - **Lemme**
- **Analyse des corpus annotés**

# Résumé

- **Utilisable**
  - pour projet de recherche de grande taille ( >> 1 Go de texte)
  - en enseignement
- **Utilisable**
  - En tant qu'outil complet
  - En tant que étape dans un traitement global (collecteur, étiqueteur, segmenteur, ...)

Gromerci !



[gromoteur.ilpga.fr](http://gromoteur.ilpga.fr)