

Practical Experiences and New Challenges in Web Crawling

Christopher Schröder, Martin Potthast
Leipzig University
[ASV - temir.org](http://ASV.temir.org) - webis.de

DLR Open Search Colloquium · June 8, 2021

Web Crawling for the Wortschatz

Search in more than 46 million sentences of German newspaper material

Enter a word



Welcome to the Leipzig Corpora Collection / Deutscher Wortschatz

a project of the *Natural Language Processing Group* at the Institute of Computer Science at Leipzig University.

Corpora portal

The international corpora portal offers access to more than 400 corpora of the Leipzig Corpora Collection (LCC) in more than 250 languages.



CORPORA COLLECTION
LEIPZIG UNIVERSITY

[To the corpora portal](#)

CURL portal

On this website you can contribute to corpus collection for under-resourced languages by simply entering a URL.



Crawling
Underresourced Languages

[To the CURL portal](#)

Words of the day

The words of the day based on a selection of newspaper and news services. Daily at 7 am and available as RSS!

RSS 2.0



WORDS OF
THE DAY

[To the words of the day](#)

CLARIN corpora portal

The Wortschatz's CLARIN corpora portal offers access to all corpora of the Leipzig Corpora Collection (LCC) that we already integrated into the [CLARIN infrastructure](#).

ASV Online Toolbox

The ASV Toolbox is a modular collection of tools for the exploration of written language data.

Corpus statistics

The corpus and language statistics contain analyses about various aspects of natural language based on our corpora.



Search in more than 46 million sentences of German newspaper material

Enter a word



Welcome to the Leipzig Corpora Collection / Deutscher Wortschatz

a project of the *Natural Language Processing Group* at the Institute of Computer Science at Leipzig University.

Corpora portal

The international corpora portal offers access to more than 400 corpora of the Leipzig Corpora Collection (LCC) in more than 250 languages.

CURL portal

On this website you can contribute to corpus collection for under-resourced languages by simply entering a URL.

Words of the day

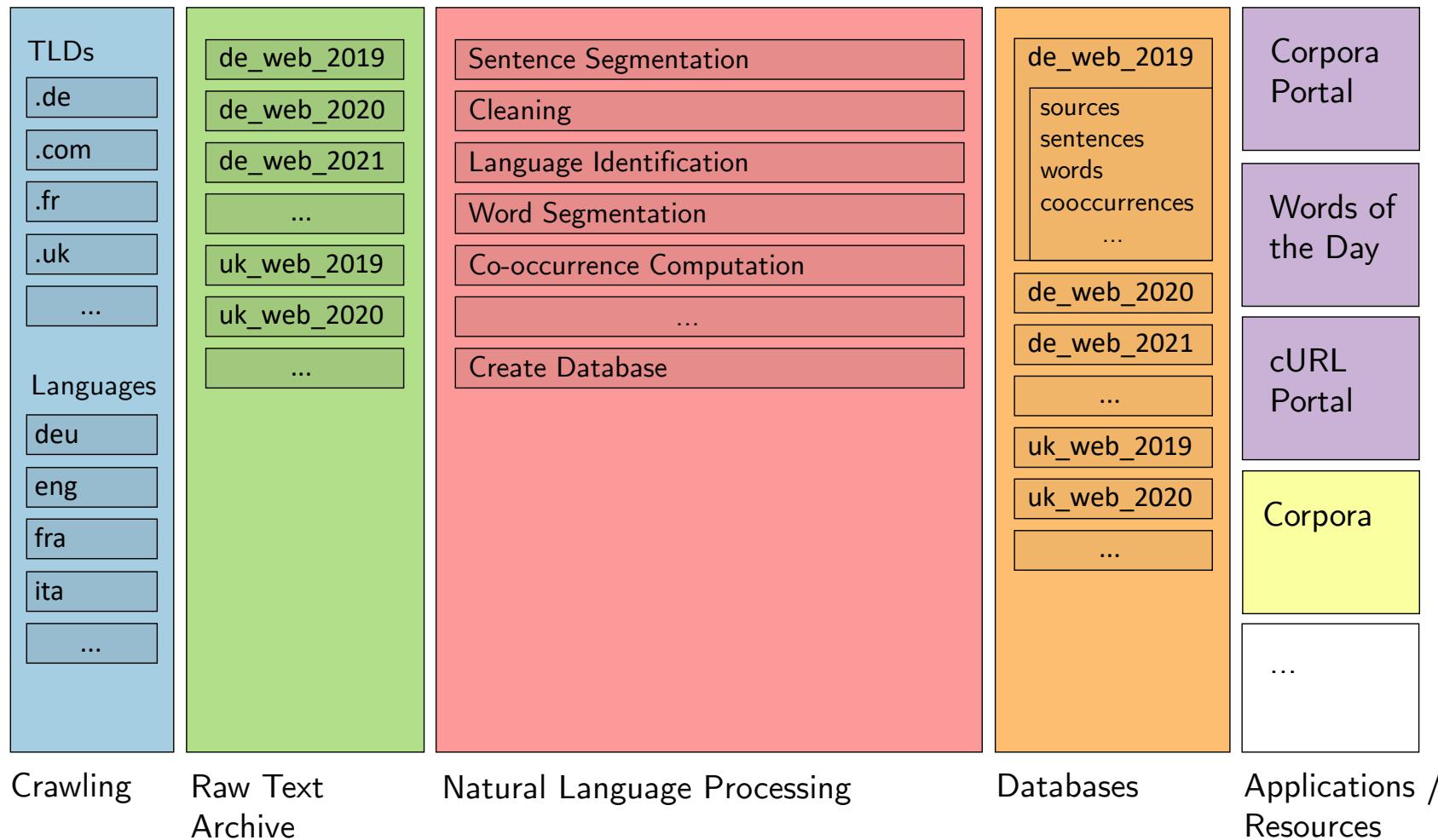
The words of the day based on a selection of newspaper and news services. Daily at 7 am and available as RSS!

RSS 2.0

- Goal: Facilitate the study of contemporary language
- Corpora in 252 languages [\[corpora.uni-leipzig.de\]](http://corpora.uni-leipzig.de)
- 8 billion unique sentences in German
- 20% share of Leipzig University's web traffic [\[wortschatz.uni-leipzig.de\]](http://wortschatz.uni-leipzig.de)

Projekt Deutscher Wortschatz

Architecture



Web Crawling

Web crawling is the process of systematically traversing the web. Usually, visited web sites, or at least parts thereof, are saved for the purpose of web indexing.

[[Wikipedia](#)]

Web Crawling

Web crawling is the process of systematically traversing the web. Usually, visited web sites, or at least parts thereof, are saved for the purpose of web indexing.

[[Wikipedia](#)]

Uses of web crawling:

- Exploring the web
- Indexing the web
- Acquiring data
- Archiving the web

Web Crawling

Selectivity

A crawler implements a **selection policy** determining which pages are crawled.

- **Crawl seeds** (Where to begin?)

Initialization of a crawler frontier with previously collected URLs, so-called seeds.

- **Crawl target** (What to crawl?)

Simple answer: Everything. More specifically: Every document for which a search engine's user might search ("Where was that document again?"). For web search engines, only few exceptions apply. In general, predicting universal non-usefulness of documents is difficult.

- **Crawl priority** (What first?)

Web pages that are *predictably* more important to the search engine's users than others. Web sites may be judged as a whole. In particular, pages comprising high-quality content.

- **Crawl filtering** (What to avoid?)

"Spider traps", and web pages from web sites whose owners harbor malicious intents toward the search engine, or its users, such as spam pages.

Web Crawling for the Wortschatz

Selection Policy

The crawler is restricted to only follow links on a specific top-level domain, or to stay within the domains of specific websites.

Web (Time Slice)

- Execution: Once a year
- Restriction: TLD (e.g., only “.de” URLs)

News (Time Slice)

- Execution: Once a year
- Restriction: Language and Seeds
(e.g., “fra” and a list of French news sites)

News (Daily)

- Execution: Daily
- Crawls a given URL list obtained via RSS/Atom

Web Crawling for the Wortschatz

Common Challenges

Stopping criteria

- After a fixed amount of time
- If less than one WARC file (1 GB compressed) is generated per day
- If crawl logs show too many low-quality / unusable links (subjective, manual)

Crawl frontier

- Typically outgrows the crawler's throughput
- For an unmonitored operation, a large storage network is required

“Low-value” sites

- Galleries, Shops, Calendars...
- Spider traps (e.g., dynamically creating many URLs pointing at the same site)
- Spam networks

Web Crawling for the Wortschatz

Maintenance

Manual orchestration

- ❑ Automation is tricky, e.g., when automating a stopping condition

Complaints

- ❑ Crawler crawls licensed material
- ❑ Crawler adds many items to the cart in a webshop
- ❑ Missing or misconfigured robots.txt

Hardware / Software issues

- ❑ Some things are just out of your control
- ❑ Recovery can be time-consuming

Web Archiving

Web Archiving

Web archiving is the process of collecting portions of the World Wide Web to ensure the information is preserved in an archive for future researchers, historians, and the public. [\[Wikipedia\]](#)

Web Archiving

Web archiving is the process of collecting portions of the World Wide Web to ensure the information is preserved in an archive for future researchers, historians, and the public. [\[Wikipedia\]](#)

Providers and initiatives: [\[Wikipedia\]](#) (96 listed)

- Internet Archive: [Wayback Machine](#)
- National libraries (e.g., DNB)
- Commercial services (e.g., Google)
- Open source and activism (“archivist”)
- Societies: ICA [\[ica.org\]](#), IIPC [\[netpreserve.org\]](#)
- People: Brewster Kahle, Vint Cerf

Web Archiving

Web archiving is the process of collecting portions of the World Wide Web to ensure the information is preserved in an archive **for future researchers, historians, and the public.** [\[Wikipedia\]](#)

Providers and initiatives: [\[Wikipedia\]](#) (96 listed)

- Internet Archive: [Wayback Machine](#)
- National libraries (e.g., DNB)
- Commercial services (e.g., Google)
- Open source and activism (“archivist”)
- Societies: ICA [\[ica.org\]](#), IIPC [\[netpreserve.org\]](#)
- People: Brewster Kahle, Vint Cerf

Customers:

- Archives and Libraries
- Social science
- Humanities
- Computer science
- Business and Government

Web Archiving

Web archiving is the process of collecting portions of the World Wide Web to ensure the information is preserved in an archive for future researchers, historians, and the public. [\[Wikipedia\]](#)

Providers and initiatives: [\[Wikipedia\]](#) (96 listed)

- Internet Archive: [Wayback Machine](#)
- National libraries (e.g., DNB)
- Commercial services (e.g., Google)
- Open source and activism (“archivist”)
- Societies: ICA [\[ica.org\]](#), IIPC [\[netpreserve.org\]](#)
- People: Brewster Kahle, Vint Cerf

Customers:

- Archives and Libraries
- Social science
- Humanities
- Computer science
- Business and Government

Use cases: Artificial intelligence. Automatic HTML generation. Automatic mobile device adjustment. Change detection. Clutter reduction. Content extraction. Design optimization. Discussion analysis. Duplicate detection. Entity mining. Error page detection. Genre detection. Indexing. Language analysis. Link analysis. Person disambiguation. Phishing detection. Profiling. Ranking. Record mining. Screen reading. Snippet generation. Spam detection. Summarization. Template detection. Text reuse analysis. Trend prediction. Usage mining. Wrapper induction.

Web Archiving

Building Blocks

Four steps of web archiving:

1. Select. Decision what portion of the web shall be archived.
2. Capture. Downloading and storing live web content.
3. Preserve. Checking downloads and converting them to stable format.
4. Playback. Enable access and interaction with archived web content.

Web Archiving

Building Blocks

Four steps of web archiving:

1. Select. Decision what portion of the web shall be archived.
2. Capture. Downloading and storing live web content.
3. Preserve. Checking downloads and converting them to stable format.
4. Playback. Enable access and interaction with archived web content.

Key performance indicators:

- Representativity of selection
- Completeness of capture
- Longevity of preservation
- Accuracy and fidelity of playback

Web Archiving

Building Blocks

Four steps of web archiving:

1. Select. Decision what portion of the web shall be archived.
2. Capture. Downloading and storing live web content.
3. Preserve. Checking downloads and converting them to stable format.
4. Playback. Enable access and interaction with archived web content.

Key performance indicators:

- Representativity of selection
- Completeness of capture
- Longevity of preservation
- Accuracy and fidelity of playback

→ A web page is a complex piece of client-server software.

Web Archiving

Example: Personalization

What you see

Pinterest UTILIZZA I COOKIES PER AVVITARE LA MIGRAZIONE SPERIMENTALE POSSIBILE. OK!

Pinterest Registrati Accesso

Camicette rosa

Petali di rosa Camicetta da lavoro Camicette con ruffle Camicette Abbigliamento a fantasia floreale Stile primaverile Peplum dress Giacca impermeabile

28 idee da mettere in pratica: Come ridare NUO...
... ESKI PENYEYI ŞIK ELBİSEYE DÖNÜŞTÜREK -
https://kendin Yapışan.com/veski...
Crochet Bag, Little Bag, Little Girl Crochet Purse...
... Doma Little Little Girl Undchetto borsa
di lalainosityCrochet

Blusa stampata con colletto Rosa Camella - Camicette - Bimba : Quello

Camicette a due strati Camicetta a due strati - Rosa chiaro - DONNA | HSM IT 1

Camicette rosa 2016 Estate Donna New Fashion Casual Solid Slash collo Chiffon Camicette...
T-shirt da donna raglan stampa "3 gatti" rosa
Camicette, Top & Tuniche - T-SHIRT DA DONNA RAGLAN STAMPA "3 GATTI" ROSA...

Tunica da donna "gatti e cuori" rosa
Camicette, Top & Tuniche - TUNICA DA DONNA "GATTI E CUORI" ROSA + un...
Tunica da donna stampa "miao righe" rosa
Camicette, Top & Tuniche - TUNICA DA DONNA STAMPA "MIAO RIGHE" ROSA + un...

LilyFashion.com

Web Archiving

Example: Personalization

What you see

What someone else sees

A screenshot of a Pinterest search results page for "Pink blouses". The top navigation bar includes "Pinterest", "Sign up", "Log in", and a link to "How we use cookies". Below the navigation is a breadcrumb trail: "Women's fashion > Pink blouses". The main title "Pink blouses" is displayed in large, bold black font. Below the title is a horizontal menu with links: "Blouse designs silk", "Pink saree blouse", "Silk saree blouse designs", "Blouse designs catalogue", "Saree blouse designs 2017", and "Women's pink style". The search results are arranged in a grid format. Each result includes an image of the blouse, its name, and a "See More" link. Some results also include a "from polyvore" badge. The results are categorized into several groups: 1) Blouse designs silk: "Chloé Pink Contrast Pointed Collar Blouse" (pink sleeveless blouse with contrasting white pointed collar). 2) Pink saree blouse: "Missoni Silk Tie Neck Blouse" (pink blouse with a large white bow tie at the neck). 3) Silk saree blouse designs: "Carven's" "poplin blouse is ultra feminine in a light pink hue and" (pink blouse with a subtle texture). 4) Blouse designs catalogue: "Lena Hoschek and suspenders, lie..." (pink blouse paired with brown leather suspender belt). 5) Saree blouse designs 2017: "Elegant Navy Blue Chanderi Saree with Bright Pink Blouse Mehr" (blue and pink saree with a matching pink blouse). 6) Women's pink style: "Pierced V Neck Cold" (pink blouse with a cold-shoulder design and a belt). The bottom of the page shows a footer with social media links: "Pinterest", "Facebook", "Twitter", "Instagram", and "RSS".

Web Archiving

Example: Capturing HTML only

Original

Pinterest using cookies to help give you the best experience we can. [Get it](#)

[Sign up](#) [Log in](#)

Women's fashion > Pink blouses

Pink blouses

[Blouse designs silk](#) [Pink saree blouse](#) [SILK saree blouse designs](#) [Blouse designs catalogue](#) [Saree blouse designs 2017](#) [Women's pink style](#)

Grey Blouse · Crepe Mat · Details · See More
Choice Pink Contrast Printed Collar Blouse (322) • liked on See More
from polyvore

Off Shoulder Blouse · Off Shoulder Neck · Details · See More
Milly off-the-shoulder blouses and pink top

Tie Blouse · Tie Neck Blouse · Details · See More
M Missoni Silk Tie Neck Blouse (6,665 INR) • liked on Polyvore See More
from polyvore

Pierced V Neck Cold · Formal Blouse · Pink Blouses · Details · See More
Pierced (Cold Shoulder) V Neck Blouse, sweet lace and crochet

Pink Wrap Top · Pink Dresses · Details · See More
Carven's poplin blouse is ultra feminine in a light pink hue and See More
from polyvore

Girls in Suspender · Suspenders Outfit · Details · See More
Lena Hoschek and suspenders, tie.

Chicwish Kiss Me Bow Top In Candy Pink · Details · See More
from polyvore

Pink Scarf · Navy Blue Lehenga · Details · See More
Elegant Navy Blue Chanderi Saree with Bright Pink Blouse Mehr See More

Front · Details · See More
from polyvore

Front · Details · See More
from polyvore

Web Archiving

Example: Capturing HTML only

Original

Pinterest is using cookies to help give you the best experience we can. [Get it!](#)

[Sign up](#) [Log in](#)

Women's fashion > Pink blouses

Pink blouses

[Blouse designs silk](#) [Pink sari blouse](#) [Silk saree blouse designs](#) [Blouse designs catalogue](#) [Saree blouse designs 2017](#) [Women's pink style](#)

Grey Blouse **Crossed Neck** **Gold** >
Choice Pink Contrast Painted Collar Blouse (\$32) • Liked on See More [from polyvore](#)

Off Shoulder Blouse **Off Shoulder Neck** >
Milly off-the-shoulder blouses and pink top See More

Teal Blouse **Teal Neck Blouse** **White** >
M Missoni Silk Tie Neck Blouse (\$665 HNL) • Liked on Polyvore See More [from polyvore](#)

Pierced V Neck Cold
Festive Blouses **Pink Blouses** **White** >
Pierced Cold Shoulder V Neck Blouse, sweet lace and crochet See More

Pink Party Tops **Pink Dresses** **Gold** >
Carven's poplin blouse is ultra feminine in a light pink hue and See More [from polyvore](#)

Girls in Suspender **Suspender Outfit** >
Lena Hoschek and suspender, tie, See More [from polyvore](#)

Chicwish Kiss Me Bow Top In Candy Pink
See More [from polyvore](#)

Pink Saree Silk **Heavy Blouse Length** >
Elegant Navy Blue Chanderi Saree with Bright Pink Blouse Molt See More

Off Shoulder Blouse **Off Shoulder Neck** >
Milly off-the-shoulder blouses and pink top

Pinterest is using cookies to help give you the best experience we can. [Get it!](#)

[Sign up](#) [Log in](#)

Women's fashion > Pink blouses

Pink blouses

[Blouse designs silk](#) [Pink sari blouse](#) [Silk saree blouse designs](#) [Blouse designs catalogue](#) [Saree blouse designs 2017](#) [Women's pink style](#)

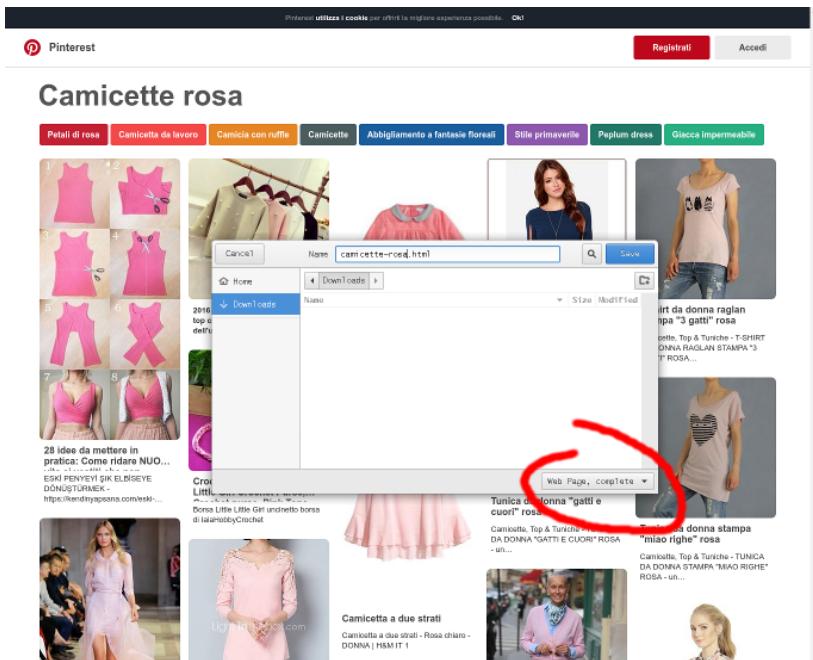
Grey Blouse **Crossed Neck** **Gold** >
Choice Pink Contrast Painted Collar Blouse (\$32) • Liked on Polyvore featuring tops, blouses, pink, pink top and pink blouse See More [from polyvore](#)

Off Shoulder Blouse **Off Shoulder Neck** >
Milly off-the-shoulder blouses and pink top

Web Archiving

Example: Capturing via a Browser's Save Page Feature

Original



Web Archiving

Example: Capturing via a Browser's Save Page Feature

Original

Pinterest utilizza i cookie per offrirti la migliore esperienza possibile. Ok!

Pinterest Registrati Accedi

Camicette rosa

Petali di rosa Camicetta da lavoro Camiccia con ruffle Camicette Abbigliamento a fantasia floreali Stile primaverile Peplum dress Giacca impermeabile

2016 t-shirt per le donne moda rosa top camicette della signora dell'ufficio Casual lavoro top stili... Blusa stampata con colletto Rosa Camilla - Camicette - Bimba : Oshello

28 idee da mettere in pratica: Come ridare NUO... ESKI PENYEŞİK ELBİSEYE DÖNÜŞTÜREK: https://kendiyapason.com/vesti... Crochet Bag, Little Bag, Little Girl Crochet Purse... Boncuk Little Girl uncinetto borsa di lalainosityCrochet

Camicetta a due strati Camicetta a due strati - Rosa chiaro - DONNA | H&M IT 1

Lightinthebox.com

Archive

Pinterest utilizza i cookie per offrirti la migliore esperienza possibile. Ok!

Pinterest Registrati Accedi

Camicette rosa

Petali di rosa Camicetta da lavoro Camiccia con ruffle Camicette Abbigliamento a fantasia floreali Stile primaverile Peplum dress Giacca impermeabile

2016 t-shirt per le donne moda rosa top camicette della signora dell'ufficio Casual lavoro top stili... Blusa stampata con colletto Rosa Camilla - Camicette - Bimba : Oshello

28 idee da mettere in pratica: Come ridare NUO... ESKI PENYEŞİK ELBİSEYE DÖNÜŞTÜREK: https://kendiyapason.com/vesti... Crochet Bag, Little Bag, Little Girl Crochet Purse... Boncuk Little Girl uncinetto borsa di lalainosityCrochet

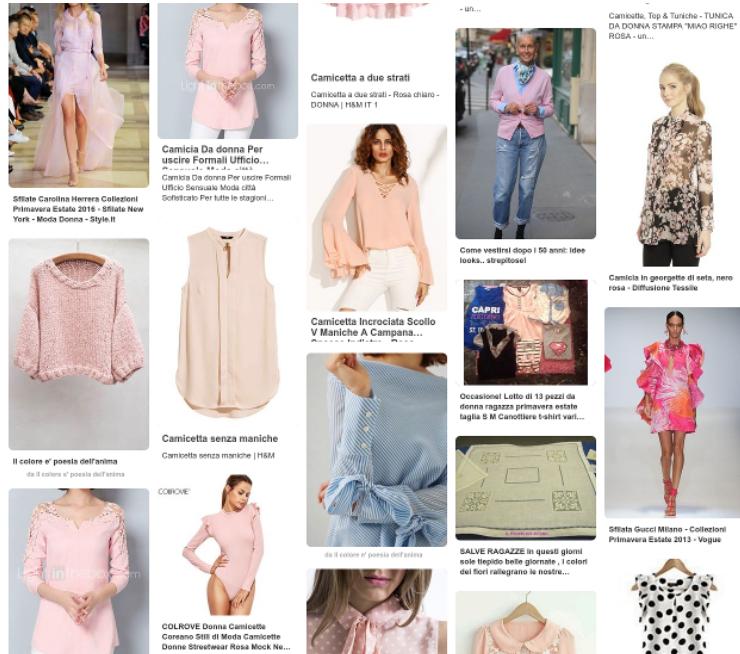
Camicetta a due strati Camicetta a due strati - Rosa chiaro - DONNA | H&M IT 1

Lightinthebox.com

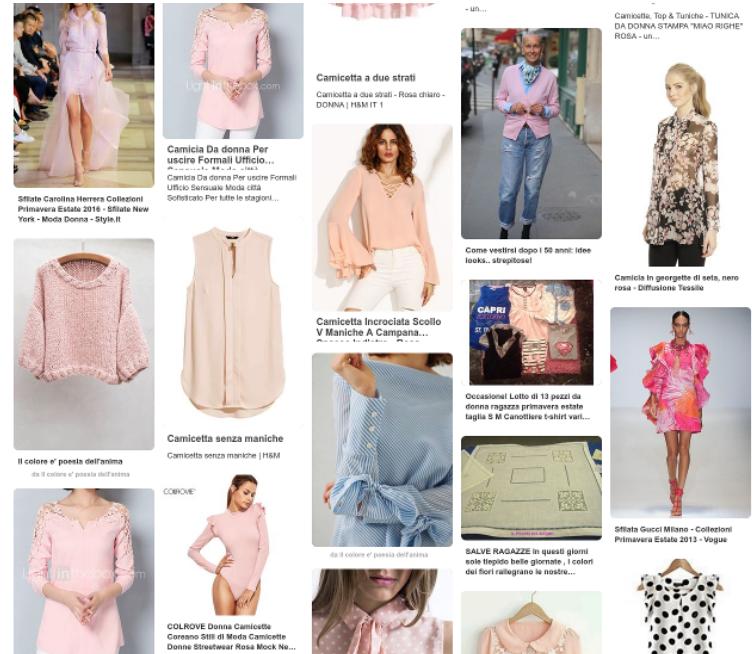
Web Archiving

Example: Capturing via a Browser's Save Page Feature

Original



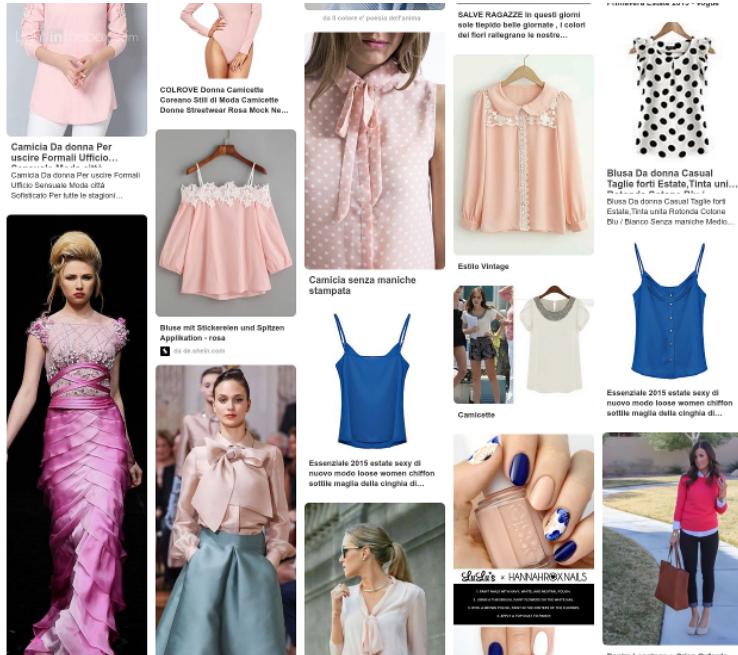
Archive



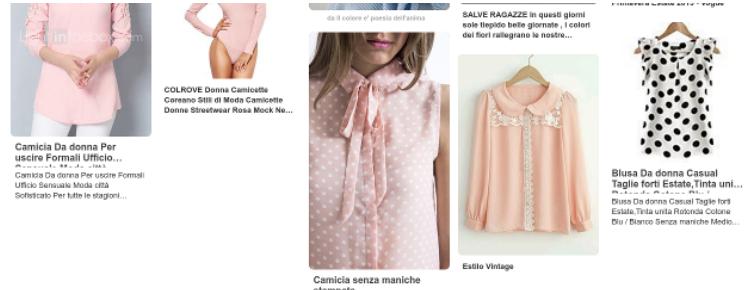
Web Archiving

Example: Capturing via a Browser's Save Page Feature

Original



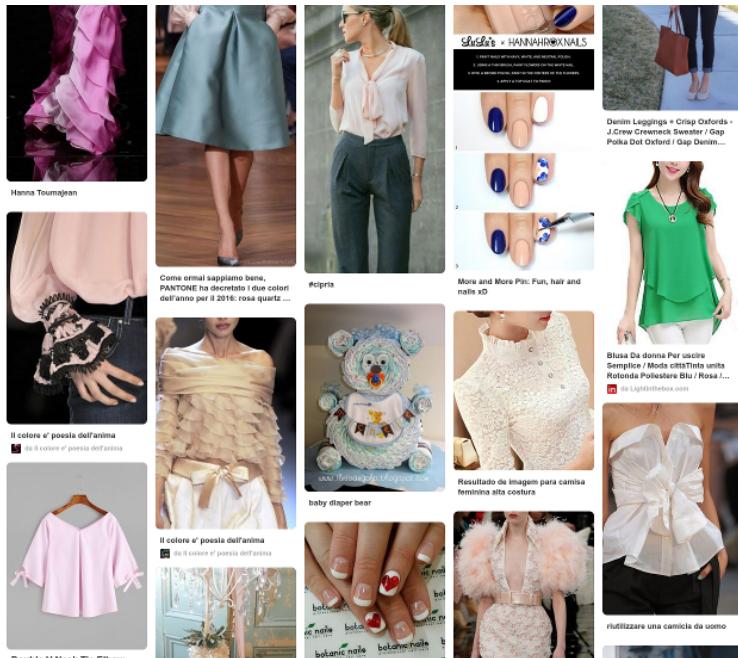
Archive



Web Archiving

Example: Capturing via a Browser's Save Page Feature

Original

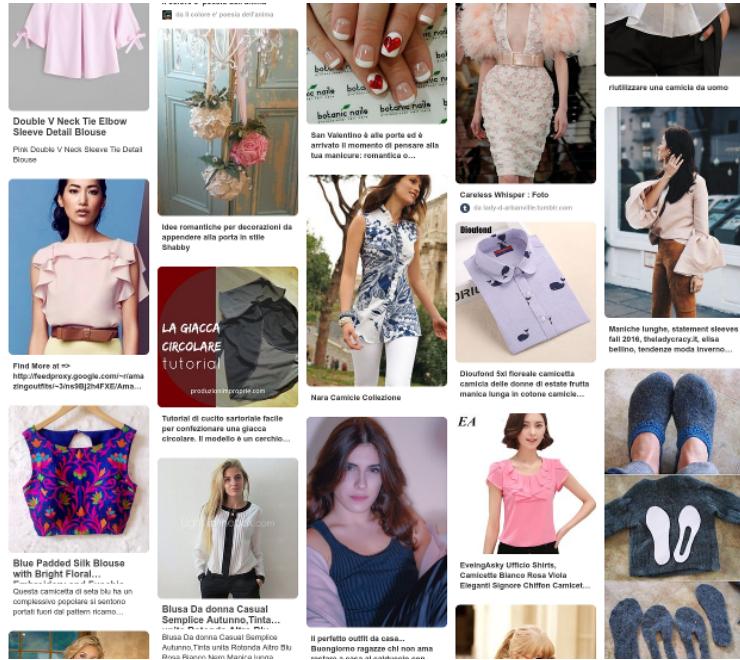


Archive

Web Archiving

Example: Capturing via a Browser's Save Page Feature

Original

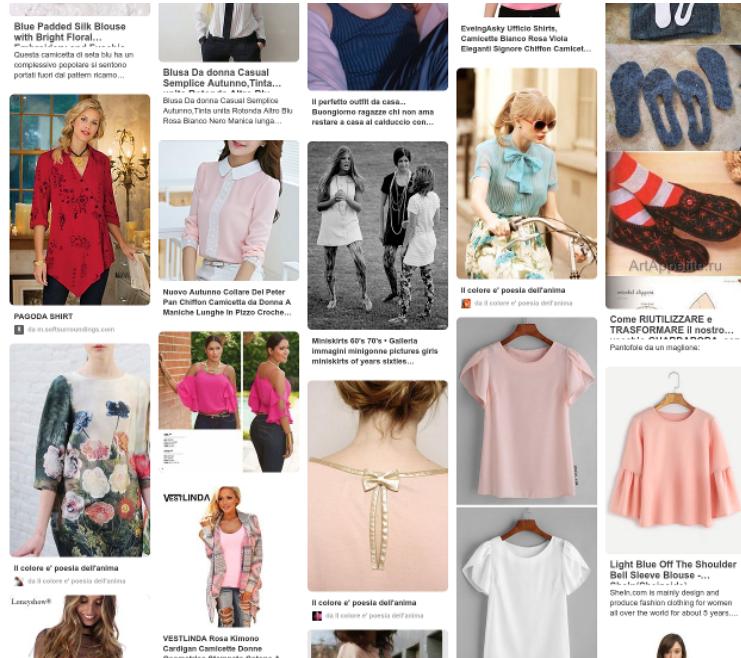


Archive

Web Archiving

Example: Capturing via a Browser's Save Page Feature

Original

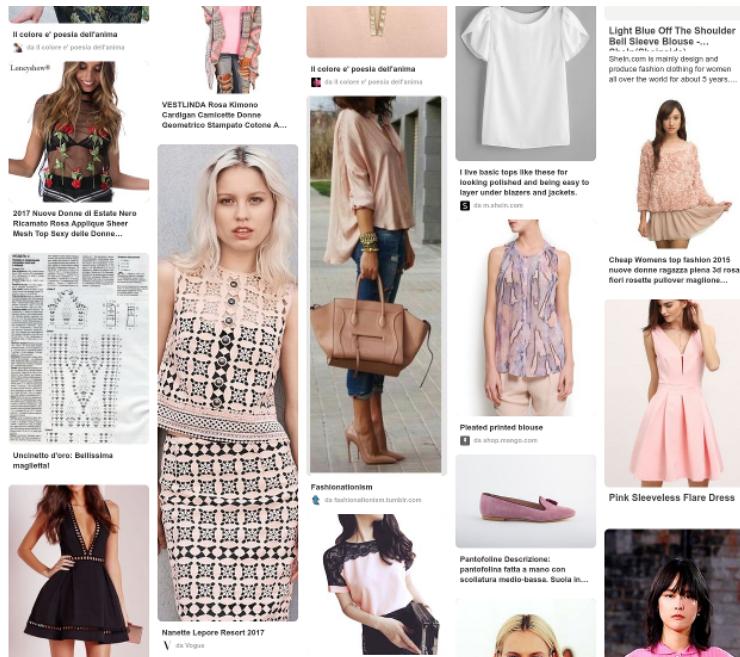


Archive

Web Archiving

Example: Capturing via a Browser's Save Page Feature

Original

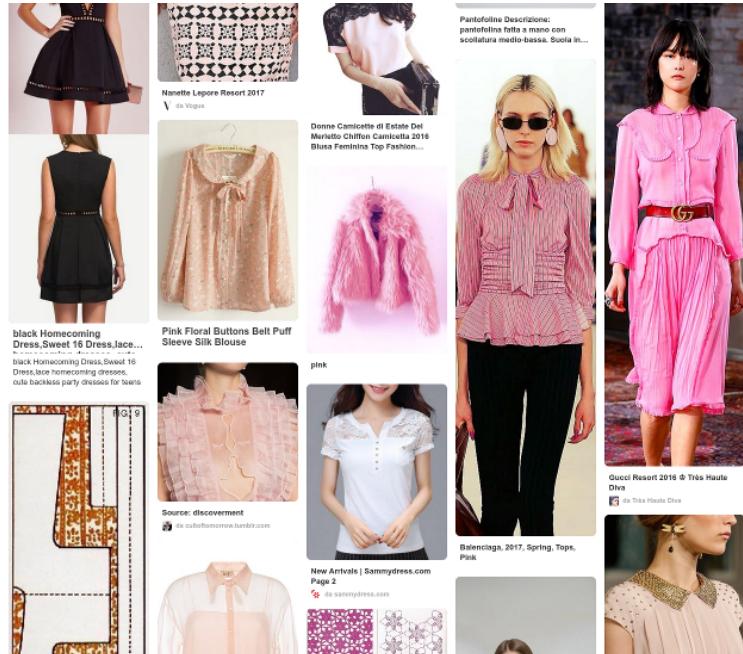


Archive

Web Archiving

Example: Capturing via a Browser's Save Page Feature

Original



Archive

Web Archiving

Example: Capturing via a Browser's Save Page Feature

Original



KIMONO TESSUTO E CONFEZIONE
da Pinterest



Light Blue Off The Shoulder Bell Sleeve Blouse ...
Shin.com is many design and produce fashion clothing for women all over the world for about 5 years....



Regata * No21
SPASHINE

2016 Womens Long Sleeve Chiffon Bianco Camicetta Della Signore Della Canotta Verde Rosa Bianca...



New Arrivals | Sammydress.com
Page 2
da sammydress.com



Balenciaga, 2017, Spring, Tops, Pink



Genny Resort 2017



Простой красоты - Вдохновение весело весело блог красивы



Vestidos de novia (inspiración) II
Wedding dress (inspiration) Que gusto encontrar bonitos vestidos...



Flared Cuffs Gingham Check Blouse
Flared Cuffs Gingham Check Blouse - WHITE/BLACK M

Archive

Web Archiving

Example: Capturing via a Browser's Save Page Feature

Original

DIMINUTIVE is a mature designer website
produce fashion clothing for women
at over the world for about 5 years....



Blusa A Cuadros Gingham
Manga Campana

Blusa A Cuadros Gingham Manga
Campana - Blanco y Negro L



Malón + Miní



Donna Camicette di Estate Del
Merletto Chiffon Camicette 2017
Blusa Feminina Top Fashion....



moldes de blusa campesina para
descargar gratis - Buscar con...



Product Image 1
N/A No Image



Jacquard Military Top



28 idee da mettere in
pratica: Come ridare NUO...
Customizazio...



Anne Fontaine shorts. It feels like
I should have one of them.



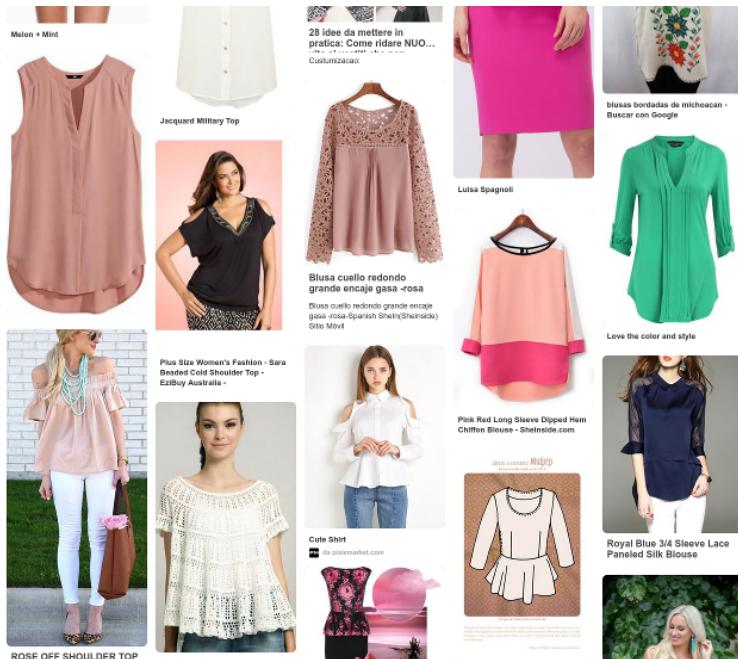
blusas bordadas de michoacan -
Buscar con Google

Archive

Web Archiving

Example: Capturing via a Browser's Save Page Feature

Original



Archive

Web Archiving

Example: Capturing via the Wayback Machine

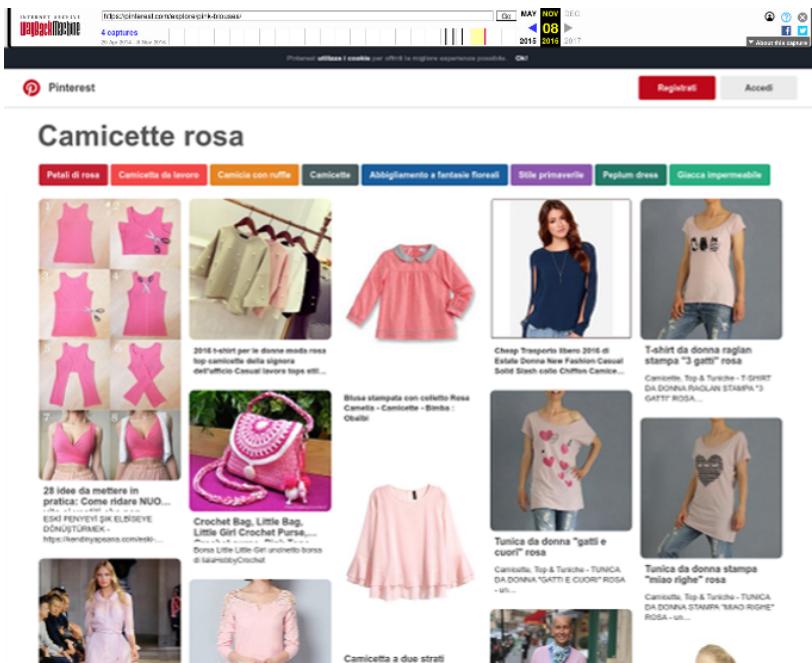
Look up page in / save page using the Internet Archive (web.archive.org)



Web Archiving

Example: Capturing via the Wayback Machine

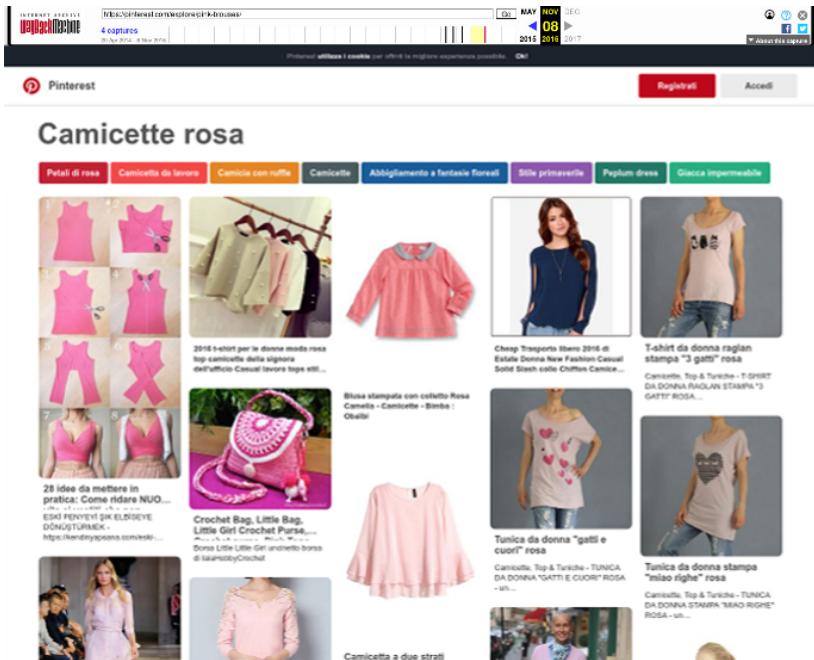
Archive



Web Archiving

Example: Capturing via the Wayback Machine

Archive



Timestamps

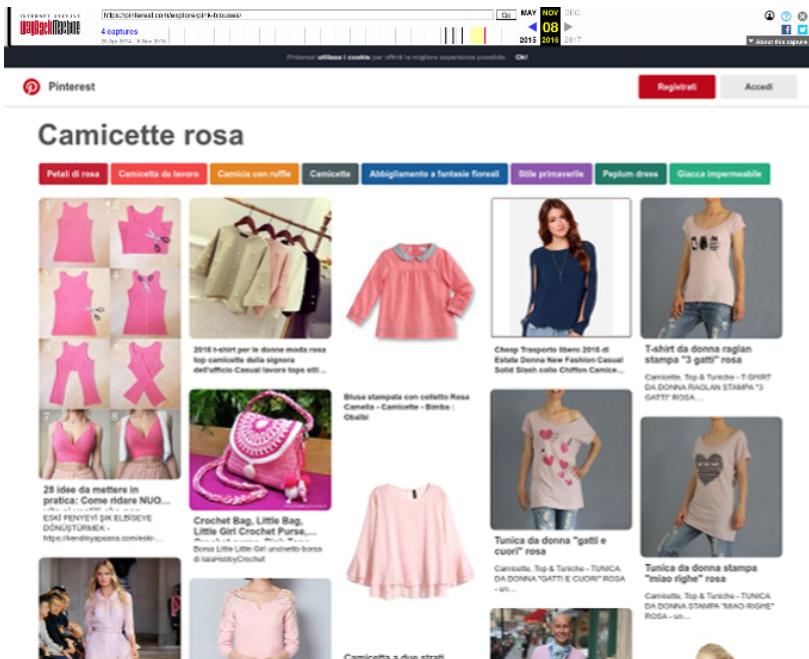
TIMESTAMPS

<https://i.pinimg.com/236x/88/db/2d/68db2d879ea9f2251f36340e29abd109.jpg> +1 year 2 months
<https://i.pinimg.com/236x/a6/e1/96/a6e1963216daac6517d8e69c55c78d67.jpg> +1 year 2 months
<https://i.pinimg.com/236x/bf/cb/f6/9682b123a08668a49d8c3e43cb.jpg> +1 year 2 months
<https://i.pinimg.com/236x/31/d2/65/31d265a067d4665c7d16ad27b10e409.jpg> +1 year 2 months
<https://i.pinimg.com/236x/12/1/7d/12117decc42286e7a43f9eba590085d.jpg> +1 year 2 months
<https://i.pinimg.com/236x/bf/c5/c6/bfc5c6adacae135696dbe43a1091753c.jpg> +1 year 2 months
<https://s-passsets-cache-ak0.pinimg.com/webapp/style/gestalt-73dea985.css> +1 hour 18 minutes
<https://s-passsets-cache-ak0.pinimg.com/webapp/s/entryChunk-core-8f1938b32.js> +1 hour 17 minutes
<https://s-passsets-cache-ak0.pinimg.com/webapp/s/external-core-1-117b51a7.js> +1 hour 17 minutes
<https://i.pinimg.com/236x/b9/62/94/b96294214709216377a1be8b4fb0f61.jpg> +1 year 2 months
<https://i.pinimg.com/236x/3c/93/69/3c938987b1409bdcd5dad300893c99d.jpg> +1 year 2 months
<https://i.pinimg.com/236x/f2/d4/c7/f2d4c71b7618d1b1c42d6cc4cdab48bf.jpg> +1 year 2 months
<https://i.pinimg.com/236x/34-c7b3eb8ecd0a5138867.js> +4 months 18 days
<https://s-media-cache-ak0.pinimg.com/236x/0/a/0/afdcabdbe0959620d5eabfd89cba43.jpg> +18 days 10 hours
<https://s-media-cache-ak0.pinimg.com/236/x/ea/3e/ea3e7a588509114633e156da5f029993.jpg> +8 days 10 hours
<https://s-passsets-cache-ak0.pinimg.com/webappjs/pjs-34-c7b3eb8ecd0a5138867.js> -7 days 13 hours
https://connect.facebook.net/en_US/sdk.js -7 hours 20 minutes
<https://s-passsets-cache-ak0.pinimg.com/webappjs/entryChunk-react-InterestFeedReactPage-557d79b3.js> -7 hours 19 minutes
<https://s-passsets-cache-ak0.pinimg.com/webappjs/pjs-70-b4eb3053cc23fc2db354.js> -2 hours 24 minutes
<https://s-passsets-cache-ak0.pinimg.com/webappjs/pjs-32-c3163c8f9c6def531648.js> -1 hour 30 minutes
<https://s-passsets-cache-ak0.pinimg.com/webapp/style/gestalt-73dea985.css> +1 hour 18 minutes

Web Archiving

Example: Capturing via the Wayback Machine

Archive



Timestamps

PS

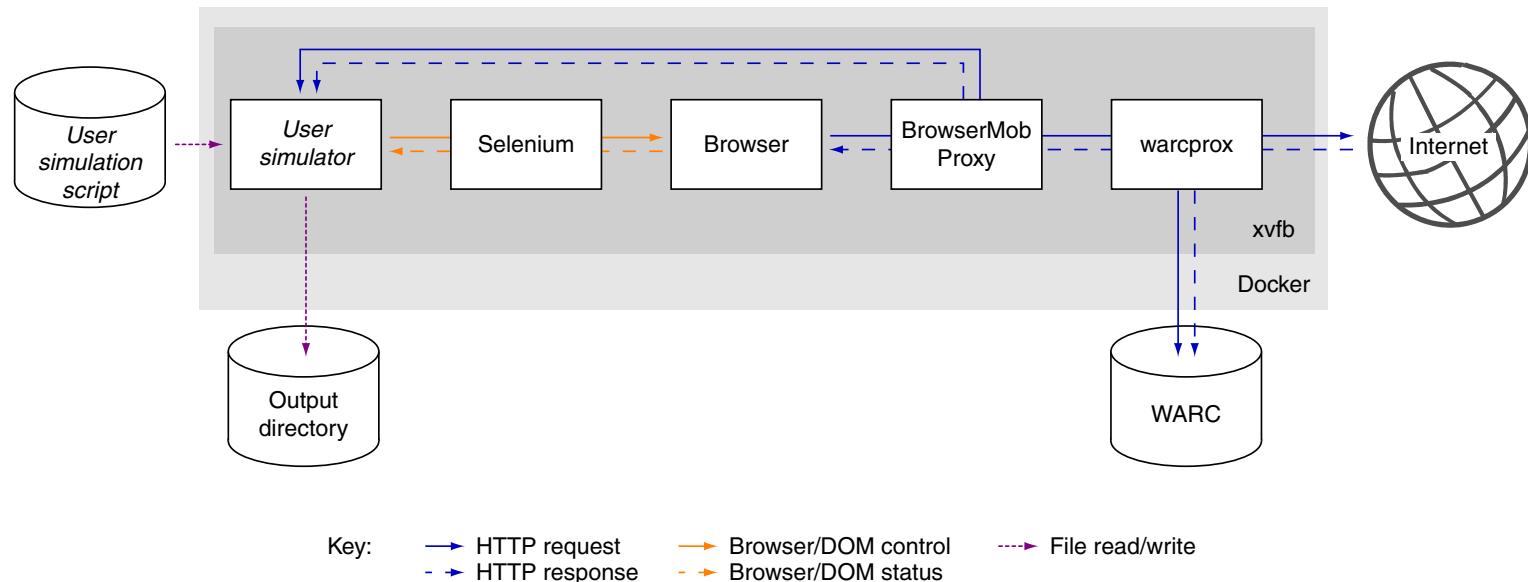
!6340e29abd109.jpg +1 year 2 months
38e69c55c78d67.jpg +1 year 2 months
!a49d8c3e43cb.jpg +1 year 2 months
!16ad27b10e9409.jpg +1 year 2 months
43f9eba590085d.jpg +1 year 2 months
!dbe43a101753c.jpg +1 year 2 months
lea985.css +1 hour 18 minutes
ore-8f1938b32.js +1 hour 17 minutes
!-1-117b51a7.js +1 hour 17 minutes
e8d88df3.js +1 hour 17 minutes
a1be8b4fb0f6f1.jpg +1 year 2 months
!dad300893c99d.jpg +1 year 2 months
!d6cc4cdab48bf.jpg +1 year 2 months
4-c7b3eb8ecd0a551388e7.js +4 months 18 days
!dcabdbe0959620d5eabfd89cba43.jpg +18 days 10 hours
!ea7588509114633e156da5f029993.jpg +8 days 10 hours
!1b962a902c7029642860b76a7a976.jpg +7 days 13 hours

Web Archiving

Webis Web Archiver [\[github.com\]](https://github.com) [\[hub.docker.com\]](https://hub.docker.com)

- Scriptable user-page interactions
- Reproducibility of web corpora, user experience, user behavior
- Compatibility with other web archiving software
- Automatic archive quality assessment

Capturing and Preservation:

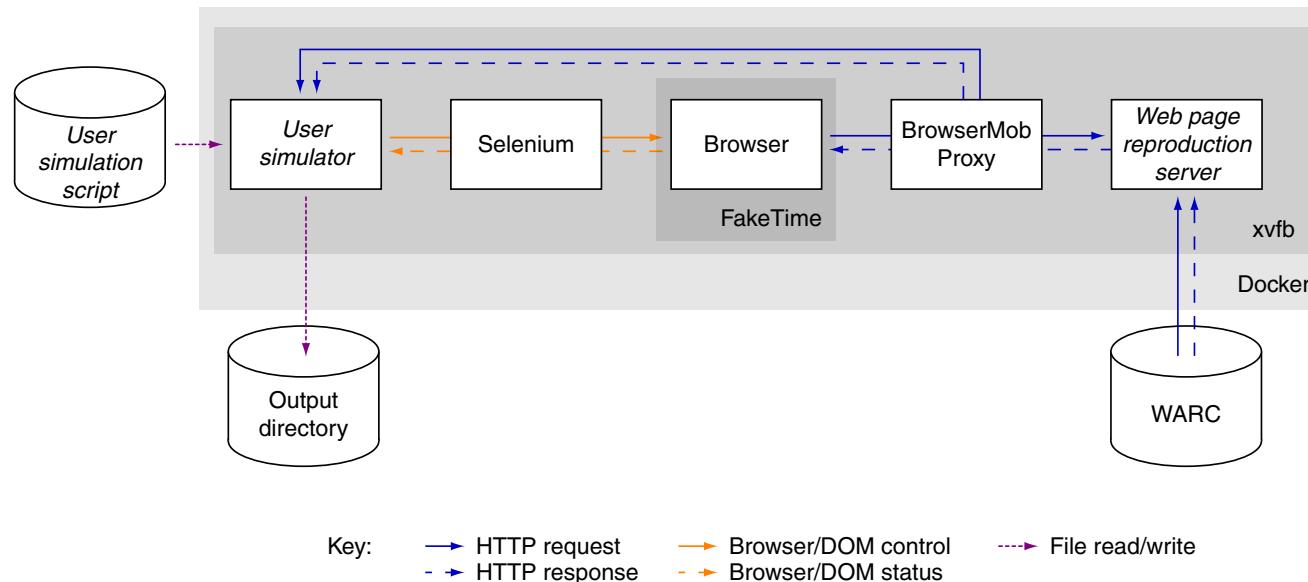


Web Archiving

Webis Web Archiver [\[github.com\]](https://github.com) [\[hub.docker.com\]](https://hub.docker.com)

- Scriptable user-page interactions
- Reproducibility of web corpora, user experience, user behavior
- Compatibility with other web archiving software
- Automatic archive quality assessment

Playback:



Remarks:

- ❑ A readily executable configuration of the archiver is encapsulated in a Docker image. Docker also ensures the reproducibility of our archiver's execution environment by fixing the versions of each software library and especially the browser. Moreover, all of 2 GB worth of fonts available to Ubuntu are installed.
- ❑ The virtual screen software xvfb is used to run the browser without requiring a physical screen, allowing for server-side execution, and with some additions even recording videos.
- ❑ The Selenium browser automation software serves as an interface between the simulation script and the browser.
- ❑ We employ a current version of Google Chromium, but others are supported as well.
- ❑ During playback, the FakeTime Preload library is used to pretend to the browser that it runs at the time of archiving, which affects all JavaScript calls that use the current date.
- ❑ The browser is set up to communicate with an instance of BrowserMob Proxy, which is used by the script to learn when network traffic ceases.
- ❑ During archiving, the BrowserMob Proxy communicates with the Internet via an instance of the warcprox proxy, which stores all requests and the corresponding responses that pass through it in a standard WARC archive file.
- ❑ During playback, a local server is started that pretends to be a proxy, but actually attempts to retrieve the previously recorded responses to requests made by a to-be-reproduced web page from its WARC files.
- ❑ Our archiver currently allows to choose from three different tools for reproducing web pages : warcprox, Python WayBack (pywb), and a custom implementation.

Remarks: (continued)

- ❑ Example of a user simulation script:

```
@Override
protected void executeInteraction(
    final Browser browser, final String url, final Path outputDirectory)
throws Throwable {
    WebDriver window = browser.openWindow(url);
    this.scrollDown(browser, window); \ 25 times
    this.saveSnapshot( \ PNG and HTML
        browser, window, outputDirectory);
    this.doCoolStuff( \ Your analysis
        browser, window, outputDirectory);
}
```

Web Archiving

Quality Assurance

- ❑ Capturing all content: embedded multimedia, dynamic content, mobile view...
- ❑ Content overlay detection and interactive removal
- ❑ Content error detection
- ❑ Fuzzy request-response matching during playback
- ❑ Playback accuracy and fidelity

Web Archiving

Quality Assurance

- Capturing all content: embedded multimedia, dynamic content, mobile view...
- Content overlay detection and interactive removal
- Content error detection
- Fuzzy request-response matching during playback
- Playback accuracy and fidelity

The screenshot shows a website for "texts from last night". The main content includes a list of texts and a "Quick! While no one's looking" modal.

texts from last night

(805):
peeing is so easy
your body to pee

f share t tweet

(360):
She looked like S
answer your ques

f share t tweet

Quick! While no one's looking

Click this little button for TFLN texts and updates

Like 3.1M

We also send dope emails

Forget to check the site? We'll send our best texts, memes and weekly shenanigans straight to your inbox.

Your email address... **SUBMIT**

already like us? well thanks! [click here.](#)

ERROR:404

Someone's chewed through something they shouldn't have...

#OutOfCheeseAgain

CONTINUE SHOPPING

Web Archiving

Quality Assurance

- Capturing all content: embedded multimedia, dynamic content, mobile view...
- Content overlay detection and interactive removal
- Content error detection
- Fuzzy request-response matching during playback
- **Playback accuracy and fidelity**

The screenshot shows a complex web interface with multiple news feeds and interactive elements. At the top, there are tabs for 'FINANCE', 'BUSINESS INSIDER', and 'LTI'. Below these are several news cards:

- London's crazy house price bubble is finally losing some air**: A headline from Business Insider. It features a photo of a woman looking at a laptop screen displaying a chart. Below the headline is a detailed paragraph about the housing market in London.
- Why Non-Financial Data is a CEO Game Changer**: A headline from Forbes. It features a photo of a person working on a laptop.
- Markets Insider**: A section from Markets Insider showing a chart of the Dow Jones Industrial Average (DJIA) with a 1.06% drop. The chart includes data for the S&P 500, Nasdaq, and FTSE 100.
- Videos You May Like**: A section featuring a video thumbnail of Floyd Mayweather.
- From the Web**: A section featuring a photo of a woman and a video thumbnail of a building.

Throughout the interface, there are numerous social media sharing buttons for platforms like Facebook, Twitter, LinkedIn, and others. The overall layout is dense and typical of a news aggregator or a specialized news portal.

Web Archiving

Quality Assurance

- Capturing all content: embedded multimedia, dynamic content, mobile view...
- Content overlay detection and interactive removal
- Content error detection
- Fuzzy request-response matching during playback
- **Playback accuracy and fidelity**

London's crazy house price bubble is finally losing some air

By Oliver Williams-Dale - 5/25/14 at 8:00 AM



Follow



Facebook



Twitter



LinkedIn



Email



Print



Mobile



Search



Help



Feedback



Report



Share



Comment



Like



Save



Delete



Edit



Archive



Anyone who's tried to buy in London knows that property prices in the capital are insane. New figures from online estate agent Rightmove released this morning show average asking prices in London are £254,074, compared to £254,151 nationally.

But prices in the capital have actually slipped so far this month, falling 2.3% compared to April. Nationally, the average asking price slipped by just 0.1% and London registered the biggest monthly fall.

Rightmove's latest dip follows a sharp drop in pre-election prices. London would have been hit hardest by Rightmove's latest analysis, which gives some buyers only short-term relief from the long-term housing shortage,

and many estate agents are reporting a resurgence in interest following the surprise election result.^[1]

NOW WATCH: Here's how Floyd Mayweather spends his millions



More: Housing Property London Rightmove



Share



Email



Print



Mobile



Search



Help



Feedback



Report



Share



Comment



Like



Save



Delete



Edit



Archive

Recommended For You



Share



Email



Print



Mobile



Search



Help



Feedback



Report



Share



Comment



Like



Save



Delete



Edit



Archive

Sponsored Financial Content

German Election: How will markets react?
Opportunities and Risks
Top 10 Stocks: Energy Materials, Tech & Financials
How to Trade Oil Prices
Top 10 Stocks from C-Suite
Stocks to Watch: Oil & Gas
Oil & Gas: What's Next?
Event Finance and Capital Moves in Latin America (Infrastructure)
How to go Long or Short with CFDs
Sprint to Profit Guide (Corporate Finance)

Featured

How Data Drives Business in a Mobile World
How to Create a Data Strategy
How to Create a Data Strategy
depending on where you stand
Move Three Percent

Finance Email & Alerts

Get the best of business news delivered to your inbox every day
Subscribe

BigData Exclusive

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

The 20 Big Data Predictions for 2017
From Gartner

Finance Deck

Web Archiving

Quality Assurance

- ❑ Capturing all content: embedded multimedia, dynamic content, mobile view...
 - ❑ Content overlay detection and interactive removal
 - ❑ Content error detection
 - ❑ Fuzzy request-response matching during playback
 - ❑ **Playback accuracy and fidelity**

Conclusion

Summary:

- Projekt Deutscher Wortschatz: Resources to study contemporary language.
- Webis Web Archiver: Tool to capture web pages with high fidelity.

Conclusion

Summary:

- ❑ Projekt Deutscher Wortschatz: Resources to study contemporary language.
- ❑ Webis Web Archiver: Tool to capture web pages with high fidelity.

Take-away messages:

- ❑ Web crawling setups differ considerably depending on one's goals.
- ❑ Maintaining and running crawlers comes with manual overhead.
- ❑ Web archiving presents unique challenges on top of and beyond crawling.
- ❑ For commercial search engines, crawling means archiving.

Conclusion

Summary:

- Projekt Deutscher Wortschatz: Resources to study contemporary language.
- Webis Web Archiver: Tool to capture web pages with high fidelity.

Take-away messages:

- Web crawling setups differ considerably depending on one's goals.
- Maintaining and running crawlers comes with manual overhead.
- Web archiving presents unique challenges on top of and beyond crawling.
- For commercial search engines, crawling means archiving.

Open Search Initiative:

- Can we afford not to archive, but only to crawl?

Conclusion

Summary:

- ❑ Projekt Deutscher Wortschatz: Resources to study contemporary language.
- ❑ Webis Web Archiver: Tool to capture web pages with high fidelity.

Take-away messages:

- ❑ Web crawling setups differ considerably depending on one's goals.
- ❑ Maintaining and running crawlers comes with manual overhead.
- ❑ Web archiving presents unique challenges on top of and beyond crawling.
- ❑ For commercial search engines, crawling means archiving.

Open Search Initiative:

- ❑ Can we afford not to archive, but only to crawl?

Thank you!

Appendix

Search in more than 46 million sentences of German newspaper material

Enter a word



Welcome to the Leipzig Corpora Collection / Deutscher Wortschatz

a project of the *Natural Language Processing Group* at the Institute of Computer Science at Leipzig University.

Corpora portal

The international corpora portal offers access to more than 400 corpora of the Leipzig Corpora Collection (LCC) in more than 250 languages.



[To the corpora portal](#)

CURL portal

On this website you can contribute to corpus collection for under-resourced languages by simply entering a URL.



[To the CURL portal](#)

Words of the day

The words of the day based on a selection of newspaper and news services. Daily at 7 am and available as RSS!

RSS 2.0



[To the words of the day](#)

CLARIN corpora portal

The Wortschatz's CLARIN corpora portal offers access to all corpora of the Leipzig Corpora Collection (LCC) that we already integrated into the CLARIN infrastructure.



[To the LCC's CLARIN corpora portal](#)

ASV Online Toolbox

The ASV Toolbox is a modular collection of tools for the exploration of written language data.



[To the online toolbox](#)

Corpus statistics

The corpus and language statistics contain analyses about various aspects of natural language based on our corpora.



[To the corpus statistics](#)



Search in 431 Corpus-Based Monolingual Dictionaries for 252 Languages.

Enter a word



Corpus: English (eng_news_2016)

English news corpus based on material from 2016 with 156,934,303 sentences.

Info Statistics Downloads

[German](#) [English](#) [French](#) [Arabic](#) [Russian](#) [all...](#)



Random words:

[minimum](#) [academic](#) [hour](#) [assist](#) [considered](#)

Corpus

A Abkhazian Acoli Afrikaans ▾ Akan ▾ Albanian ▾ Amharic ▾ Anaang
Arabic ▾ Aragonese Armenian ▾ Assamese ▾ Assyrian Neo-Aramaic Asturian
Avaric Aymara Azerbaijani

B Bambara Banjar ▾ Bashkir ▾ Basque Bavarian Belarusian Bengali ▾
Betawi Bikol ▾ Bishnupriya Bosnian Breton Bulgarian Buriat

C Catalan Cebuano ▾ Central Bikol ▾ Central Khmer Chavacano Chechen ▾
Cherokee Chinese ▾ Chuvaš ▾ Classical Nahuatl Cornish Corsican Crimean Tatar
Czech

D Danish ▾ Dari Dhivehi Dimli ▾ Dutch

E Eastern Mari Eastern Yiddish Egyptian Arabic Emiliano-Romagnolo English✓ ▾
Erzya Esperanto Estonian Ewe Extremaduran

F Faroese Fiji Hindi Fijian Finnish French ▾ Friulian Fulah

G Gagauz Galician ▾ Gan Chinese Ganda ▾ Georgian ▾ German ▾ Gilaki ▾
Goan Konkani ▾ Guarani ▾ Gujarati

H Haitian ▾ Halk Mongolian Hausa ▾ Hebrew Hindi Hungarian

I Icelandic ▾ Ido Igbo Iloko ▾ Indonesian Interlingua Interlingue
Iranian Persian Irish Italian

J Japanese Javanese ▾

K Kabardian ▾ Kabyle Kalaallisut Kalmyk Kannada ▾ Kara-Kalpak
Karachay-Balkar Kashubian Kazakh Kinyarwanda Kirghiz ▾ Kituba (Congo)
Komi Komi-Permyak Konkani ▾ Korean Kurdish ▾ Kölsch

L Ladino Lao Latgalian Latin ▾ Latvian Ligurian Limburga ▾

Enter a word

Q ?

deu_newscrawl-public_2018
deu_typical-mixed_2018
deu_newscrawl_2011
deu_newscrawl-hyphen_2011
deu_wikipedia_2018
deu-ch_web_2002
deu-na_newscrawl_2012
deu_mixed_2011_dan
deu-na_web_2019
deu-ac_web_2017
deu-na_web_2013

Word: Winter Number of occurrences: 278,185 Rank: 1,385 Frequency class: 9



See also: Winter, WINTER, WInter

Part of speech: Verb, Noun

Baseform of: wintered, wintering, winters, Winters, Winter

Part of: winter sports, winter wheat, winter day, winter solstice, winter squash, hard winter, winter garden, winter wind, winter crop, winter quarters, winter rye, winter annual, winter tire, winter sleep, winter barley, winter melon, winter beer, winter moth, winter flounder, depth of winter, winter golf, winter fat, winter jasmine, winter aconite, winter wren, more...

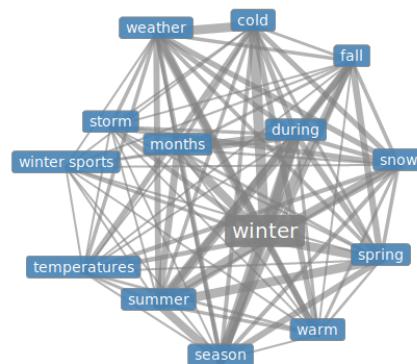
Examples:

Graph:



- The met office are predicting rain, sleet and possibly snow in winter. (www.impartialreporter.com, crawled on 25/11/2016)
- There are plans to stop there winter fuel Allowance. (www.blackelectorate.com, crawled on 01/11/2016)
- Reducing high winter bills. (www.eugeneweekly.com, crawled on 02/11/2016)
- The Salvation Army will continue to collect and distribute coats throughout the winter. (www.tyronepa.com, crawled on 07/11/2016)
- So how exactly do fashionistas bring winter style to the Golden State? (www.pccourier.com, crawled on 22/11/2016)
- Gloves, scarves and warm socks are also important layers to add to any winter ensemble. (www.kentontimes.com, crawled on 12/11/2016)
- It's the shortest east-west crossing and doesn't get closed in winter. (www.newsandstar.co.uk, crawled on 12/11/2016)
- Resident Kari Simonson said she thought the need for winter parking change was obvious. (www.snyderdailynews.com, crawled on 22/11/2016)
- Bald eagles use the area in winter. (www.idahostatesman.com, crawled on 18/11/2016)
- Paredes got good news on his application last winter. (www.modbee.com, crawled on 11/11/2016)

+10 +30 +100



Cooccurrences:



| Sentence | Left | Right |
|----------|------|-------|
|----------|------|-------|

cold (75,757), snow (73,129), summer (71,314), spring (64,363), weather (58,502), during (51,546), months (49,240), fall (40,488), warm (34,813), winter sports (31,858), season (30,453), the (23,655), temperatures (23,572), this (22,016), storm (20,887), ice (18,205), in (17,361), Winter (16,511), break (16,139), mild (15,732), winter wheat (15,165), winter day (14,451), wonderland (13,467), coats (12,273), warmer (12,111), storms (12,110), winter solstice (11,698), winter squash (10,755), conditions (10,742), for (10,404), last (10,352), harsh (9,621), snowfall (9,404), and (9,353), autumn (9,077), colder (9,062), solstice (8,966), ski (8,700), skiing (8,270), freezing (8,154), snowy (7,850), through (7,763), birds (7,739), coat (7,543), wet (7,399), long (7,218), heat (7,048), sports (6,909), dry (6,905), wheat (6,780), early (6,742), heating (6,268), days (6,015), the weather (5,999), coldest (5,882), tires (5,881), hard winter (5,759), sun (5,682), rain (5,574), Weather (5,275)