

Top-Level Domain Crawling for Producing Comprehensive Monolingual Corpora from the Web

Dirk Goldhahn
Steffen Remus
Uwe Quasthoff
Chris Biemann

CMLC-2 Reykjavik



WORTSCHATZ
UNIVERSITÄT LEIPZIG



Natural Language Processing Group
Department of Computer Science
University of Leipzig

Language Technology Group
Computer Science Department
TU Darmstadt

- Extraordinary growth of information in the WWW
- Online documents increasingly become the major source for creating high quality corpora
- Amount of data + technologies that make the information conveniently available → crawling and processing becomes hard
- Researchers often rely on data provided by companies specialized in crawling the web
 - Googleology has its limitations or „is bad science“

- Creating corpora from the web
 - with little effort
 - using only freely available, open-source software
- Data processing executed in a highly distributed environment (Hadoop)
 - parallel - very good scalability
- Researchers are able to create large-scale high quality corpora for their own needs

- WaCky
- COW
- Leipzig Corpora Collection
- Common crawl (very comprehensive)

- a) Comprehensiveness for low-resourced languages - minimal effort by crawling entire TLDs
- b) Generic, easy to use, fast and **distributed** processing pipeline for automatic corpus creation and annotations
- c) Availability of the entire processing chain as open-source software components – partially provided by us

- Open-source, web-scale crawler of the Internet Archive
- Output: standard Warc file format
- Compared with other crawling software (wget, HTTrack, or Nutch), it offers several general advantages:
 - Easy to use
 - Versatile tool, many options to *configure crawling behavior*
 - GUI tools for management of crawls
 - Pause and resume crawls
 - Single crawl jobs can cover hundreds of millions of pages
 - Every page visited only once
 - Stable, fast and follows more links than other comparable tools (better handling of Java-script links)

- Initialized with a list of webpages – called seeds
- Heritrix extracts links to other webpages → they are subsequently downloaded and processed accordingly
- Crawling of TLDs
 - Starting point: Few domains
 - Minor influence on results
 - Hubs reached within first steps
 - Restrict Heritrix not to leave desired TLD
 - Disadvantage: Misses text on other TLDs such as .com
- Lists of domains
 - Sources such as www.abyznewslinks.com
 - Successful domains from previous crawls

- Using 8 CPU-cores, 20GB of RAM and a 1Gbit/s Internet connection → crawling speeds of up to 200 URLs/s per job running 3 jobs parallel
- Politeness: One query per domain per seven seconds - high crawling speed only when many servers are queued
- Basic configurations for performance:
 - To avoid link farms and spider traps we use a maximum crawl depth
 - To reduce bandwidth we exclude certain file types like images, media files...
 - URLs containing keywords (download, files, image, pics, upload, redir or search) are excluded
 - To reduce the amount of lists or computer-generated content, max. file size is 1MB

- **WebCorpus project** (TU Darmstadt)

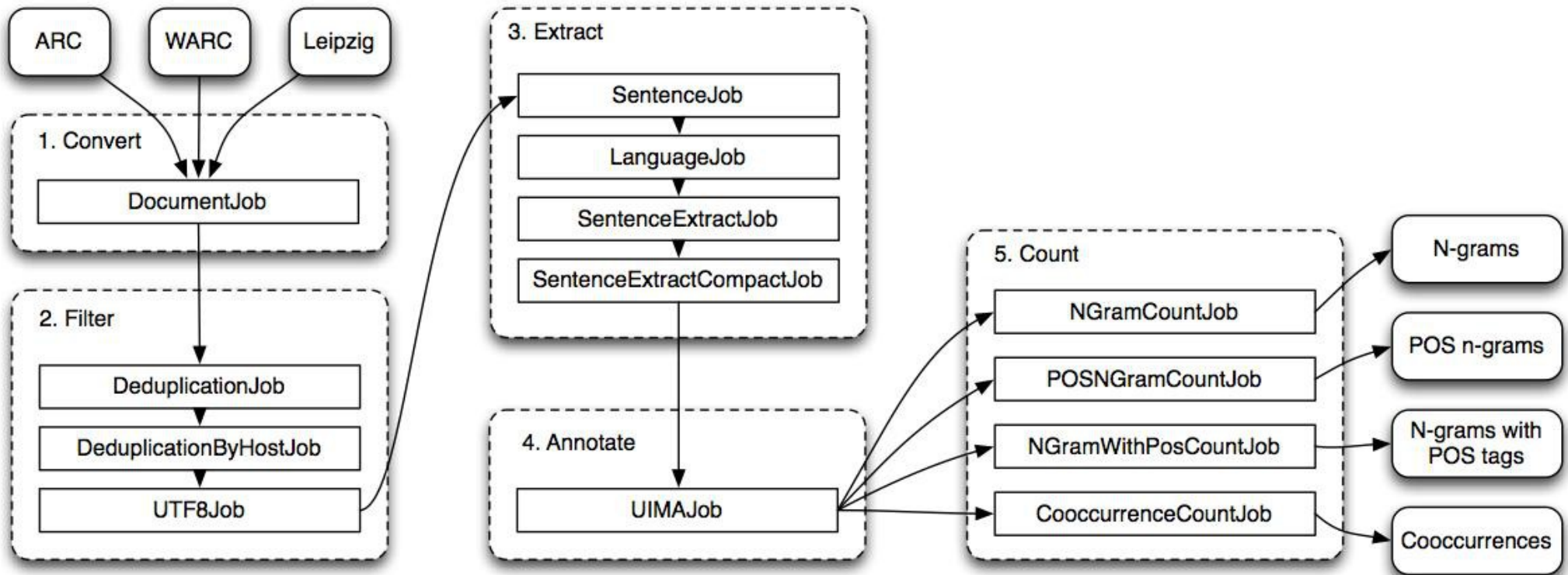
- Uses highly efficient **Hadoop** framework → Follows the MapReduce paradigm in a distributed environment
- Possible to process very large data in parallel
 - Ideally on a large number of computers
 - Or just by a single machine

- **MapReduce:**

- Idea: Split an algorithm into two phases: Map and Reduce
- Map phase: So-called key-value pairs of the input data are produced
- Reduce phase: Pairs are grouped and combined by their key to produce final result

1. Convert: Converting input data (WARC, ARC, Leipzig format) into a unified document representation
2. Filter: Removing duplicate or broken documents
3. Extract: Segmenting and filtering texts in the desired level of granularity (unique sentences, paragraphs or documents in a particular language)
4. Annotate: Using UIMA components, e.g. tokenizing, tagging and parsing
5. Count: Exploit annotations by counting n-grams, co-occurrences, subtrees of dependency parses, etc.

WebCorpus – Processing Steps

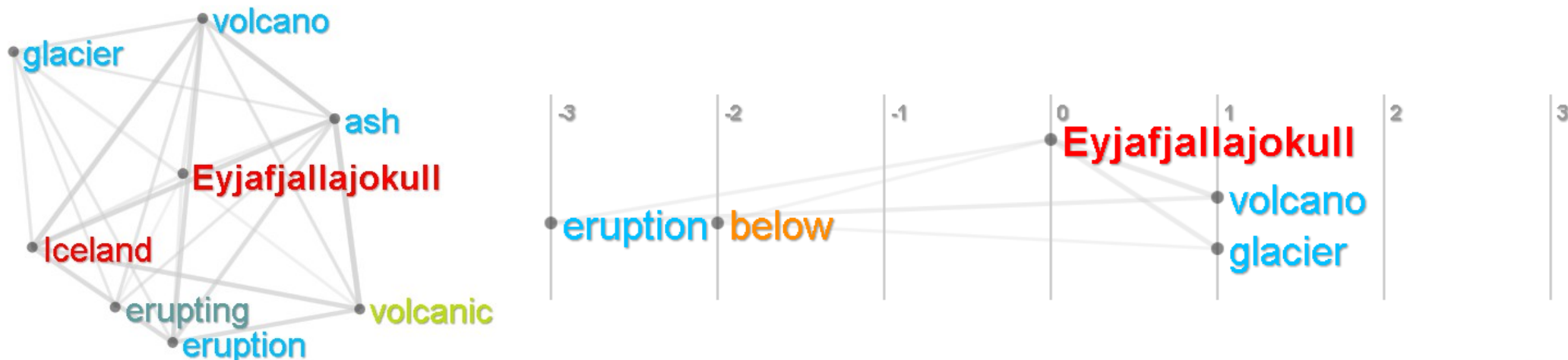


Output:

- Cleaned, language-filtered and preprocessed documents
- Various statistical outputs:
 - Statistically significant co-occurrences
 - N-grams
- CSV

- Input for a variety of applications, e.g.:
 - Distributional thesauri
 - Language models
 - Visualisation → CoocViewer

- Significant co-occurrences and concordances of words from text corpora can be explored visually
- Access to source text via full-text indexing
- CoocViewer is divided into two major components:
 1. The server-sided data management part (data stored in relational database for fast access)
 2. The web based front end running on an http server (independent of operating system etc.)



- <https://webarchive.jira.com/wiki/display/Heritrix/>

Heritrix

Erstellt von Paul Jack, zuletzt geändert von Noah Levitt am Feb 27, 2014



- [Introduction](#)
- [Browse the wiki](#)
- [Webmasters!](#)
- [Downloads](#)
- [License](#)
- [Latest Releases](#)
 - [Heritrix 3.2.0 \(Jan 2014\)](#)
 - [Heritrix 1.14.4 \(May 2010\)](#)
- [Documentation](#)
- [Mailing lists](#)
- [Development](#)

Introduction

This is the public wiki for the Heritrix archival crawler project.

Heritrix is the Internet Archive's open-source, extensible, web-scale, archival-quality web crawler project.

- <http://sourceforge.net/projects/webcorpus/>
- <http://sourceforge.net/projects/coocviewer/>
- Open source, Apache v2 License

Home / Browse / WebCorpus



WebCorpus Beta

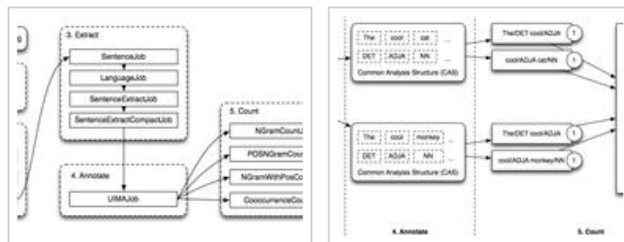
Hadoop framework for scalable processing of large web corpora

Brought to you by: [biem-tuda](#), [johannes_simon](#), [remstef](#)

[Summary](#) | [Files](#) | [Reviews](#) | [Support](#) | [Wiki](#) | [Tickets](#) | [Code](#) | [Discussion](#) | [Mailing Lists](#)

★ [Add a Review](#)
 ↓ [2 Downloads](#) (This Week)
 📅 Last Update: 2013-11-12


Download
 webcorpus-1.0.1.jar
 [Browse All Files](#)



Description

WebCorpus is a Hadoop-based framework that enables you to calculate statistics on large web corpora extracted from web crawls.

Home / Browse / Science & Engineering / Linguistics / CoocViewer



CoocViewer

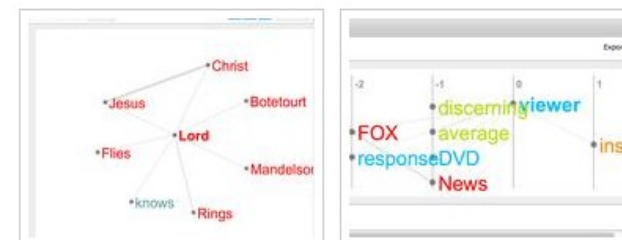
Viewer for co-occurrences and positional co-occurrences

Brought to you by: [biem-tuda](#), [lsw2](#), [remstef](#), [riedlma](#)

[Summary](#) | [Files](#) | [Reviews](#) | [Support](#) | [Wiki](#) | [Tickets](#) | [Discussion](#) | [Code](#)

★ [Add a Review](#)
 ↓ [3 Downloads](#) (This Week)
 📅 Last Update: 2013-11-08


Download
 coocviewer20130528.zip
   [Browse All Files](#)



Description

A Demo is available at:
<http://coocviewer.sourceforge.net/coocviewer/index.php>

•Leipzig Corpora Collection

- Web interface – Corpora for >200 languages
- Download of corpora of standard sizes
- All textual data underly creative commons attribution license (cc by)
- Sentences are scrambled → original structure of documents cannot be reconstructed (German copyright legislation)



W O R T S C H A T Z
UNIVERSITÄT LEIPZIG

Search in 230 Corpus-Based Monolingual Dictionaries

Newest Dictionaries

Word: Find! ?

Active dictionary: Icelandic case sensitive search

Random words:

velkomnir listanum strákar hafna skrítið

Change Dictionary:

Abkhaz	Acholi	Afrikaans	Akan	Albanian	Amharic
Arabic	Arabic, Egyptian	Aragonese	Armenian	Aromanian	Assamese
Assyrian Neo-Aramaic	Asturian	Avar	Azerbaijani	Balkar	Bamanankan
Banjar	Bashkir	Basque	Bavarian	Belarusan	Bengali
Bicolano	Bishnupriya	Bosnian	Breton	Bulgarian	Buriat
Catalan	Cebuano	Chavacano	Chechen	Cherokee	Chinese (simplified)
Chinese, Min Dong	Chinese, Min Nan	Chuvash	Cornish	Corsican	Crimean Tatar
Croatian	Czech	Danish	Dimli	Dutch	Emiliano-Romagnolo
English	English (AU)	English (CA)	English (NZ)	English (UK)	Esperanto
Estonian	Faroese	Fijian	Finnish	French	Frisian, Northern
Frisian, Western	Friulian	Fulah	Gaelic, Irish	Gaelic, Scottish	Gagauz
Galician	Ganda	Georgian	German	German (CH)	German, Swiss
Gilaki	Goan Konkani	Greek	Greenlandic	Guarani	Gujarati
Haitian	Hausa	Hebrew	Hindi	Hindi, Fiji	Hungarian
Icelandic	Ido	Ilocano	Indonesian	Interlingua	Interlingue
Italian	Japanese	Javanese	Kölsch	Kabardian	Kabyle
Kalmyk-Oirat	Kannada	Karakalpak	Kashubian	Kazakh	Khmer, Central
Kiswahili	Klingon	Komi	Konkani	Korean	Kurdish
Kyrgyz	Ladino	Lak	Lao	Latgalian	Latin
Latvian	Ligurian	Limburgish	Lingala	Lithuanian	Lushai
Luxemburgian	Macedonian	Malagasy	Malay	Malayalam	Maldivian
Maltese	Manx	Maori	Marathi	Mari, Meadow	Mari, Western
Mingrelian	Mirandese	Moksha	Mongolian (Cyrillic)	Mongolian (traditional)	Nahuatl
Navajo	Nepali	Newari	Norse, Old	Norwegian (Bokmål)	Norwegian (Nynorsk)
Novial	Occitan	Old English	Oriya	Oromo	Oromo, West Central
Ossetian	Pampanga	Pangasinan	Panjabi	Panjabi, Western	Papiamentu
Pashto	Pennsylvanian Dutch	Persian	Picard	Piemontese	Polish
Portuguese (Brazil)	Portuguese (Macao)	Portuguese (Portugal)	Romanian	Romansch	Romany
Russian	Rusyn	Saami, North	Sami	Samoan	Samogitian
Sanskrit	Sardinian	Scots	Serbian	Shona	Sicilian

term: Eyjafjallajökull

number of occurrences: 1

class of frequency: 24 (i.e. *the* has got about 2^{24} the number of occurrences than the selected word.)

example(s):

You can also learn where **Eyjafjallajökull**, Breiðamerkurjökull, and Þórisjökull are and what they look like too.

significant cooccurrences of Eyjafjallajökull:

No words found.

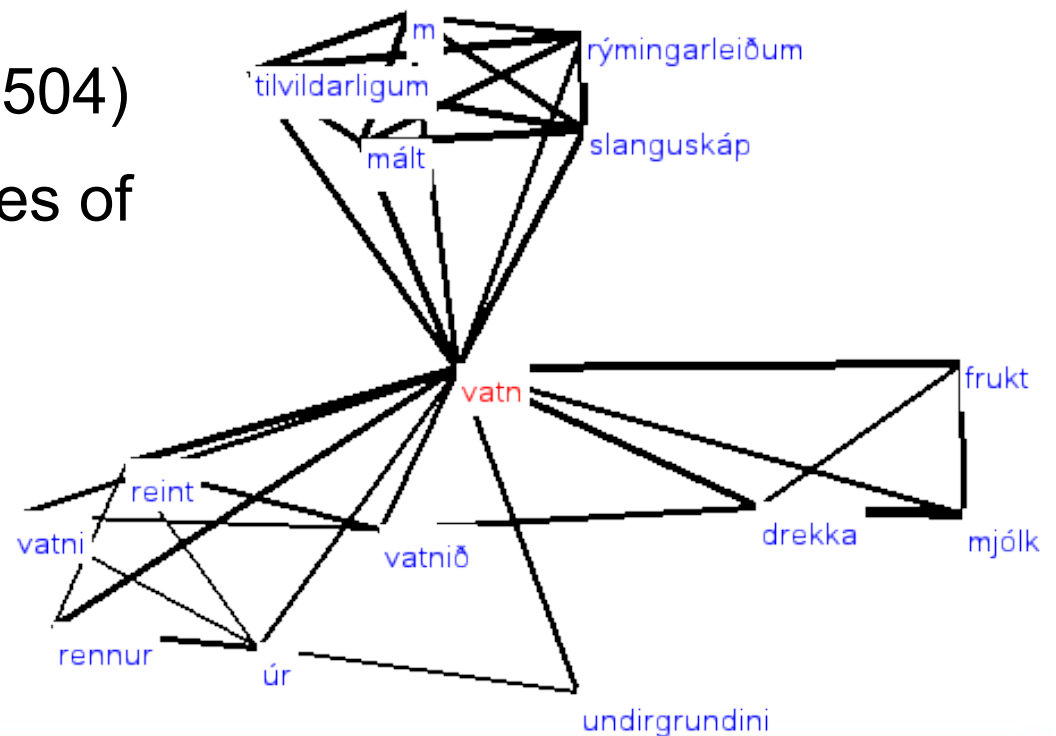
- .fo domain (Faroe Islands)
- 1.2 million websites
- “Politeness” - crawling took three days
- Language separation:

Language	Percentage
Faroese	60.63
English	14.69
Icelandic	11.24
Danish	10.29
French	0.64

- Results of processing:
 - 888,255 unique sentences



- “vatn” (engl. “water”, frequency 1,504)
- Co-Occurrences and concordances of Webcorpus + Coocviewer (top) vs. LCC (bottom)



- .ke domain (Kenya)
- 3.1 million websites
- “Politeness” - crawling took five days
- Language separation:

Language	Percentage
English	98.20
Kiswahili	0.84
Russian	0.21
Latin	0.12
Spanish	0.08

- Results of processing:
 - 7,873 unique sentences

- German:
- >7 TB of raw HTML-data
- ~140 Million Websites (type text/html)
- <2 weeks
 - Sudden stop
- Problem: > 2TB of (necessary) temporary files

- We presented a workflow for producing Web corpora
 - Using free, open-source software
 - Corpora targeted to own needs
 - Comprehensive (TLD crawling)
 - Scalable und fast (Hadoop)

Thank you!