





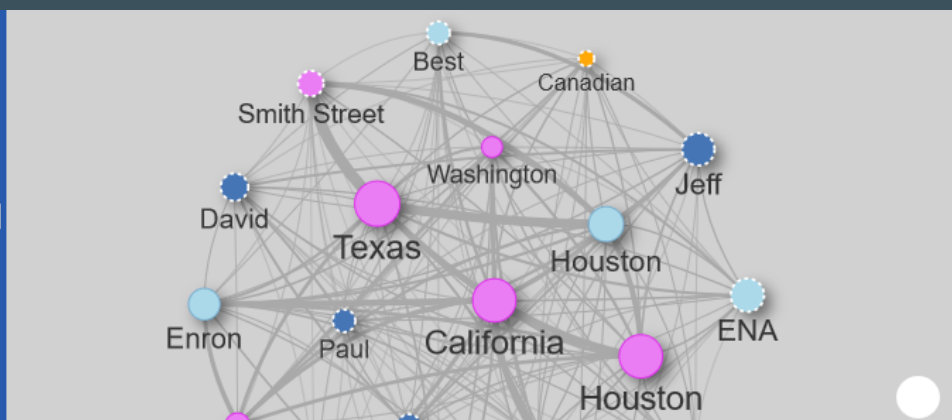


-  FW: Energy Market Report - 01/29/02 #...  
Keywords
-  RE: Spreadsheet. #322  
Keywords
-  RE: A3A1 Ad Hoc Review Team #338  
Keywords
-  RE: #1852  
Keywords
-  RE: Prior Month Adjustments going into...  
Keywords
-  FW: Important insight into our crisis #2...  
Keywords



Houston	8427	8 427
ENA	7876	7 876
Enron	7288	7 288
FERC	6187	6 187
Best	4548	4 548
LOG MESSAGES	4042	4 042
Enron	3815	3 815
From	3452	3 452
EOL	3322	3 322

Dr. Gregor Wiedemann, [gwiedemann@informatik.uni-hamburg.de](mailto:gwiedemann@informatik.uni-hamburg.de)  
26.05.2018 – #EIJC2018 – Mechelen

# new/s/leak – Information Extraction and Visualisation for Investigative Journalism

# Background



- Volkswagen Stiftung funding line „Science + Data journalism“ in 2016/17
- First phase of new/s/leak:
  - software prototype for large textual data leaks
  - exemplary use cases: Cable Gate, Enron emails, ...
- Current phase 2018:
  - consolidation of prototype into a stable release
  - real use cases

# Team



- Spiegel-Verlag, Hamburg
- **Universität Hamburg, Language Technology Group**
  - C. Biemann, G. Wiedemann, A. Panchenko, S.M. Yimam, U. Fahrer
  - Natural language processing
- Technische Universität Darmstadt  
Graphic Interactive Systems Group
  - T. von Landesberger, K. Ballweg
  - Interactive Visualization



# Motivation



- Investigative Journalism confronted not only with structured data sets provided via APIs or scrapable from the Web
- Lot's of stories buried within huge text collections (unstructured data)
  1. Disclosure of (government) official documents → reports, files, ...
  2. Answers to Freedom of Information act requests → communication
  3. Court-ordered revelation of internal communication → Enron, Tobacco
  4. Unofficial leaks of confidential information → CableGate, Paradise Papers, ...

# We need "intelligent" software!

- Natural Language Processing
  - area of computer science / artificial intelligence concerned with the interactions between computers and human (natural) languages
- Text Mining
  - (semi-)automatic extraction of semantic structures in very large amounts of texts
  - Unstructured Data → Structured Data

# Desiderata

- Current free software solutions provide fulltext search, but:
  - little/no use of natural language processing (NLP) technology for information extraction
  - limited use of visualization for data exploration
  - limited support for multi-language text collections

# Our Solution: new/s/leak

- “network of searchable leaks”: data analysis tool that combines

- natural language processing for multiple languages
- information visualization
- centered around Named Entities

- Goal:

- Enable journalists to swiftly process large collections of newly gained text documents to find interesting pieces of information.



image source: <http://www.financialdirector.co.uk/IMG/378/276378/confidential1a-370x229.jpg?1434693794>

# Entity-Centric Approach

- **Named Entity (NE):** real-world object, such as *persons*, *locations* or *organizations* that can be denoted with a proper name
  - central aspect of every newsworthy story
  - co-occurrence of NEs in texts: trace of social interaction
- **Named Entity Recognition:** automatic process of NE identification in texts

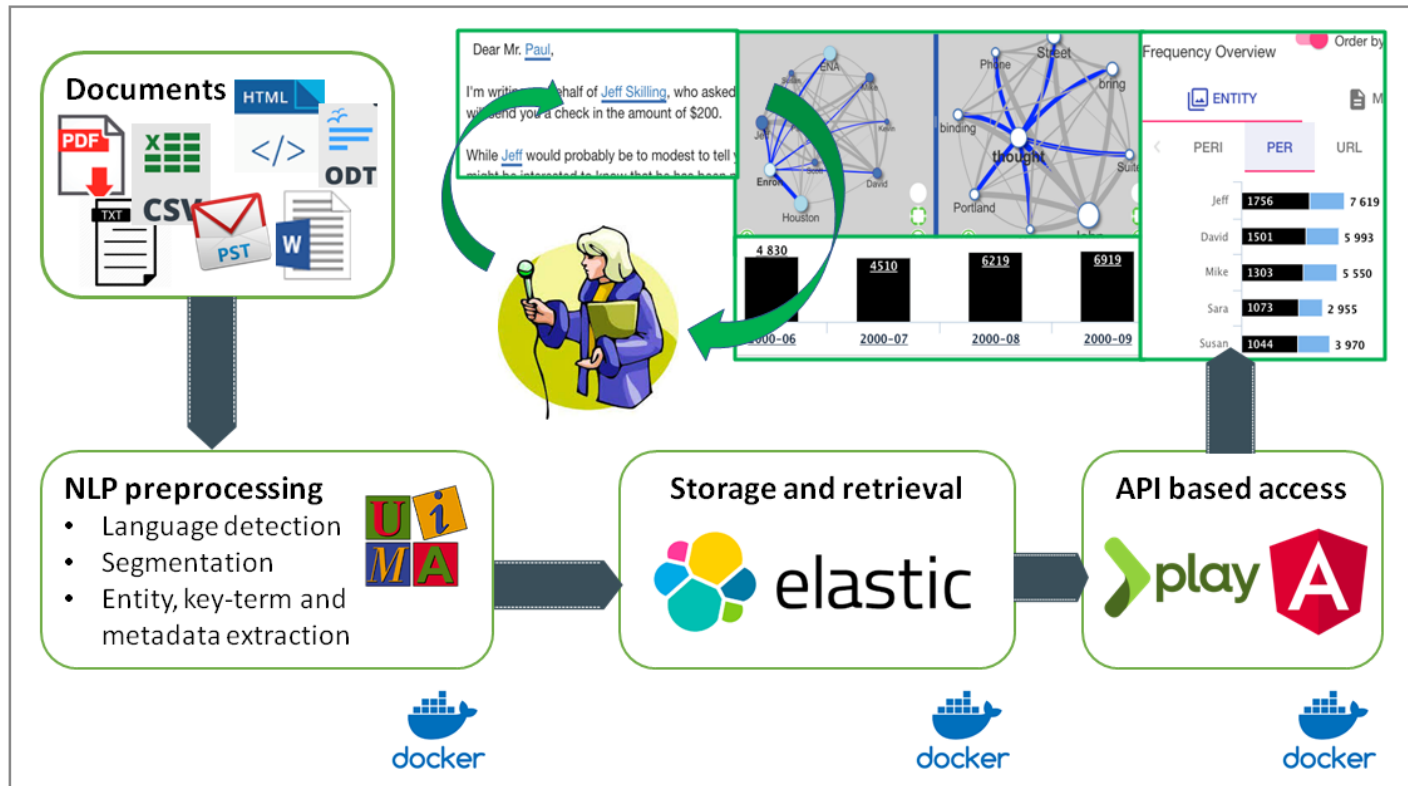


# Requirements

- Free & open-source
- Decentralized deployment
- State-of-the-art NLP
- Multi-lingual information extraction
- Multiple file formats
- Entity centric visualization
- User-defined Annotations
- Scalability
- Usability

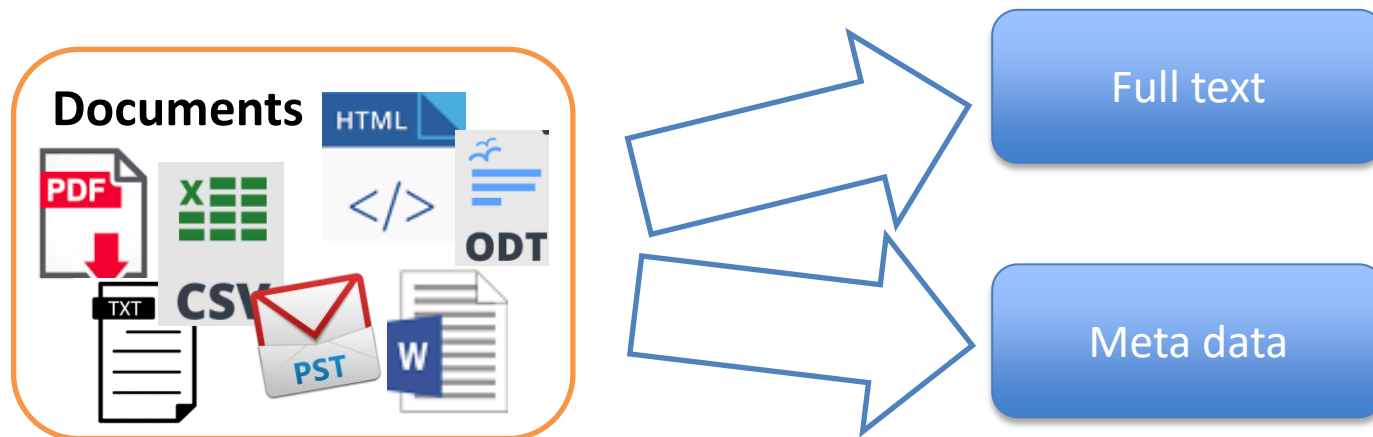
# Architecture

Hoover



# Hoover

- BY: EIC investigative network: <https://hoover.github.io>
  - data extraction
  - Fulltext search





# Hoover

\*



Refine your search using [this handy guide](#).

[Batch search](#)

Results per page 10

Sort by Relevance

- ☒ Testdata
- ☒ Enron
- ☒ CableGate



517479 email  
243488 pdf  
3558 folder  
162 archive  
23 text  
21 image  
13 doc  
3 html  
1 email-archive

« 764630 hits (**testdata** 210, **enron** 520947, **cablegate** 243473) (page 2/76463) »

11. [08DAMASCUS233.pdf](#)

[08DAMASCUS233.pdf](#)

1072 words

12. [05PARIS4918.pdf](#)

[05PARIS4918.pdf](#)

441 words

13. [08DAMASCUS321.pdf](#)

[08DAMASCUS321.pdf](#)

1446 words

14. [09JAKARTA1831.pdf](#)

[09JAKARTA1831.pdf](#)

434 words

15. [08TASHKENT701.pdf](#)

[08TASHKENT701.pdf](#)

1303 words

16. [07DHAKA284.pdf](#)

[07DHAKA284.pdf](#)

#468422: 08DAMASCUS233.pdf

[Open in new tab](#) [Original file](#)

## META

Path	<a href="#">08DAMASCUS233.pdf</a>
Filename	<a href="#">08DAMASCUS233.pdf</a>
Type	<a href="#">pdf</a>
MD5	<a href="#">be52d1e228a54f4cb03ee828e79ba478</a>
SHA1	<a href="#">0892a6b26abe306fc8be6f9398622b691d6146c8</a>
Created	2015-11-11T02:29:28+00:00
Modified	2015-11-11T04:29:28

## TEXT

bÿ

VZCZCXYZ0026  
OO RUEHWEB

DE RUEHDM #0233/01 0991420  
ZNY CCCCC ZZH  
O 081420Z APR 08  
FM AMEMBASSY DAMASCUS  
TO RUEHC/SECSTATE WASHDC IMMEDIATE 4815  
INFO RUEHAK/AMEMBASSY ANKARA PRIORITY 5516  
RUEHGB/AMEMBASSY BAGHDAD PRIORITY 0824  
RUEHLB/AMEMBASSY BEIRUT PRIORITY 4890

# Information Extraction

- Temporal Expressions
- Named Entities (Person, Location, Organization)
- Keyterms
- User-defined Dictionaries

# Temporal Expressions

- „In **January**, I bought a cat! It was born **November 2017**. **Next year**, it will be a star on the internet!“

Posted: **2018-03-27**, 11:00am

- 2018-01
- 2017-11
- 2019
- 2018-03-27

<https://github.com/HeidelTime>  
Rule-based temporal expression  
extraction for 200+ languages



# Named Entity Recognition

- Polyglot NLP
  - Python library for NER and other NLP tasks
  - <http://polyglot.readthedocs.io>
- Machine learning based Sequence Classification
  - Weakly supervised training data based on Wikipedia
    - „**Mechelen** is one of **Flanders**' prominent cities of historical art, with **Antwerp** and **Bruges**. The poet **John Heywood** moved here after having some trouble with the **Church of England**. “
  - 40+ Languages

# Keyterm extraction

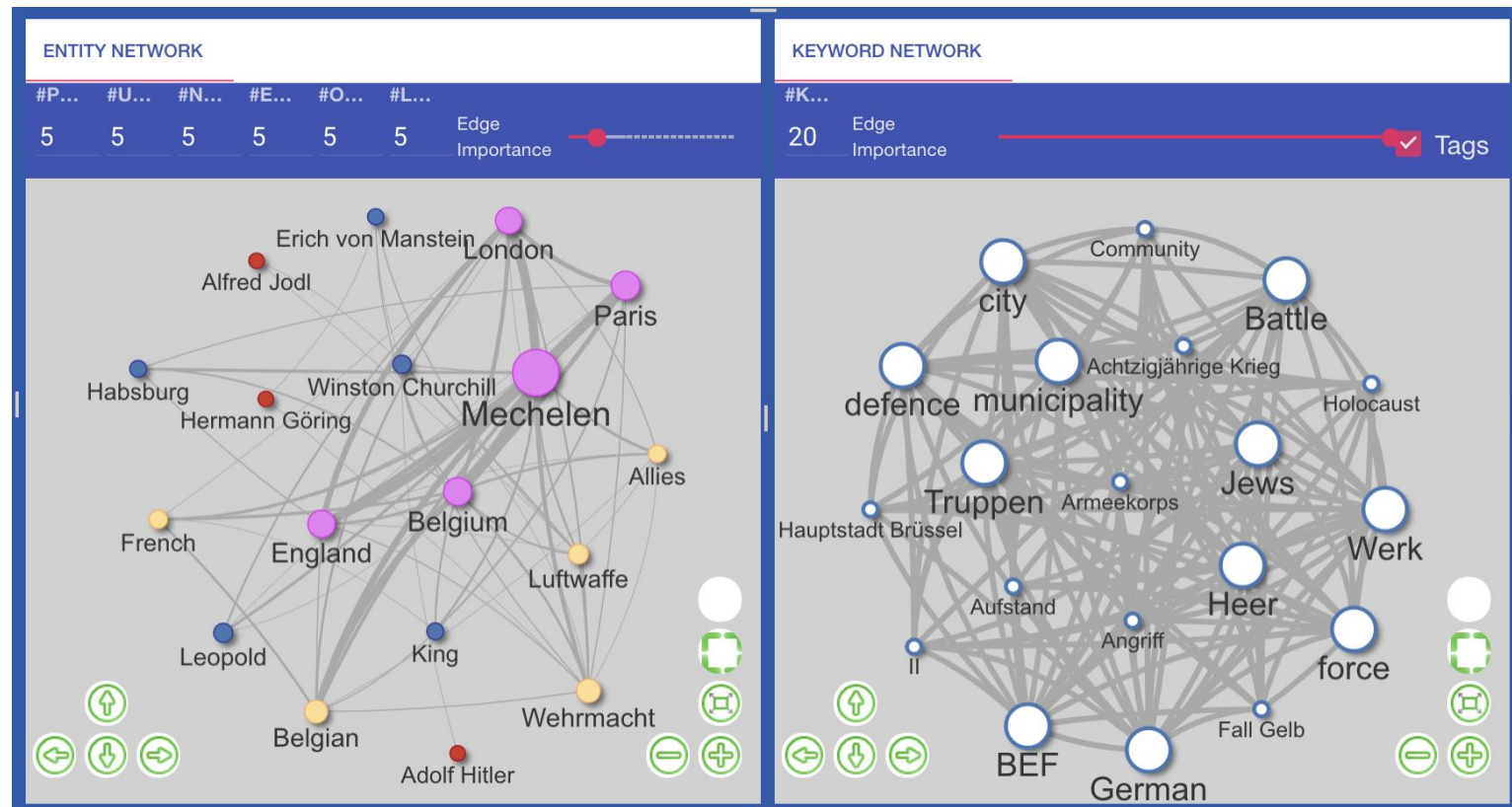
- Automatic process to identify most relevant terms in texts
  - Own development based on reference corpus comparison
    - Network visualization
    - Document summary
  - Support for 40 languages
  - MIT license: <http://github.com/uhh-lt/lt-keyterms>



# User defined dictionaries

- Annotation of specific entities provided as lists:
  - Terms: swear words/negative sentiment words
  - Person names: company/party representatives
  - Event indicators
  - ...
- Language-dependent or cross-lingual

# Entity-Centric Visualization



# Multi-lingual Information Extraction

- Automatic language detection: Document Level ⇔  
Paragraph Level (planned)

---

Arabic	Finnish	Korean	Serbian
Bulgarian	French	Latvian	Slovak
Catalan	<b>German</b>	Lithuanian	Slovene
Chinese	Greek	Malay	<b>Spanish</b>
Croatian	Hebrew	Norwegian	Swedish
Czech	Hindi	Persian	Tagalog
Danish	Hungarian	Polish	Thai
Dutch	Indonesian	Portuguese	Turkish
<b>English</b>	<b>Italian</b>	Romanian	Ukrainian
Estonian	Japanese	Russian	Vietnamese

---

# Workflows

- Full-text search
- Faceted search
- Tagging: mark documents with user-defined tags
- History: save/load filter status to share with colleagues
- Annotate: add new entities while reading documents

# Exemplary Case Studies

- Study 1: WWII collection
  - Wikipedia crawl of articles linking (back) to World War II
  - Four languages: English, German, Spanish, Hungarian
  - ca. 27,000 articles
  - Scenarios:
    1. Uncover interesting details
    2. Check for "distant view" patterns

# Exemplary Case Studies

- Study 2: Parliamentary reports
  - NSU murder case: 7 parliamentary enquiry commissions
  - 7 reports, ca. 12,000 pages, language: German
  - Scenario:
    1. Follow hypothesis based on external knowledge → use list of former NSDAP party members as dictionary

# Installation

## 1. Docker Deployment

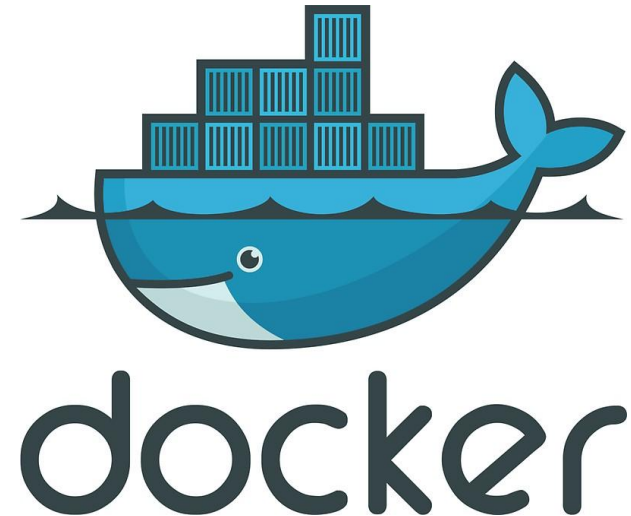
- Install docker & docker-compose

## 2. Hoover data wrangling

- Install Hoover: <https://github.com/hover/docker-setup>
- Import your collection

## 3. Newsleak

- Install and import: <https://github.com/uhh-lt/newsleak-docker>



# Next steps

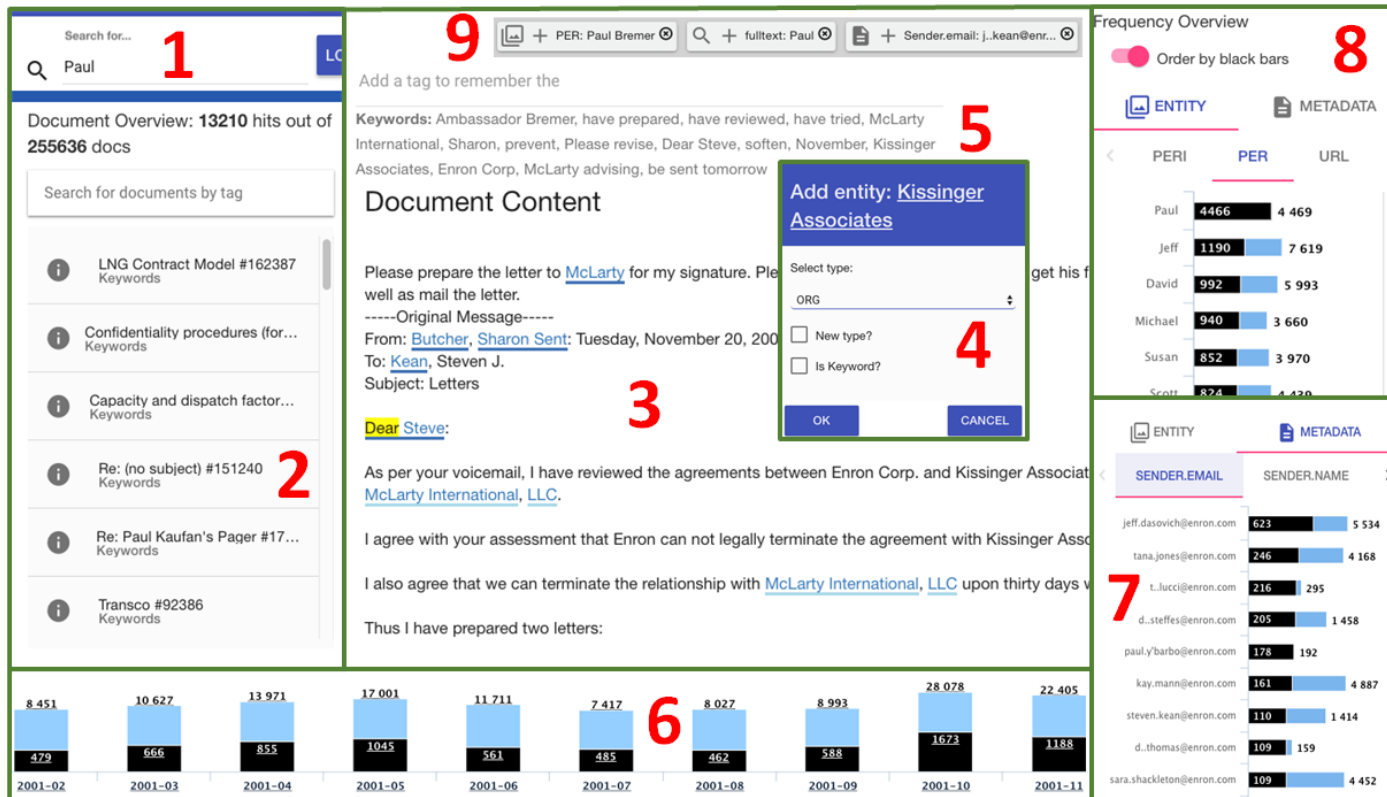
- Finalize implementation
  - Multi-Lingual document support
  - Keyword-in-context (KWIC) view
  - Multi-collection handling in UI
- Extended documentation
- Collecting user feedback!



# Try it out!

- Github:  
<https://github.com/uhh-lt/newsleak-docker>  
<https://github.com/uhh-lt/newsleak-frontend>
- Documentation: <https://uhh-lt.github.io/newsleak-frontend/>
- Project blog: <http://newsleak.io>
- Demo: <http://ltbev.informatik.uni-hamburg.de/newsleak>  
(not the latest version)
- Help/support: [gwiedemann@informatik.uni-hamburg.de](mailto:gwiedemann@informatik.uni-hamburg.de)

# Faceted Search and Full-text View



# Security

1. Install the Hoover docker setup and import the Hoover test collection.
2. Install the Newsleak docker setup (set own credentials for the DB and the newsleak app)
3. Import the testcollection extracted by Hoover.
4. **If everything works fine, disconnect form the internet.**
5. Copy your data to the Hoover collection directory.
6. Import your as a new Hoover collection.
7. Import the new Hoover collection into Newsleak
8. Analyze your content 😊