PyData

How to read a million books

Clemens Neudecker
Staatsbibliothek zu Berlin
17 May 2017

Staatsbibliothek
zu Berlin
Preußischer Kulturbesitz

PyData

How to read a million books?

Clemens Neudecker
Staatsbibliothek zu Berlin
17 May 2017

Staatsbibliothek
zu Berlin
Preußischer Kulturbesitz

PyData

How to read a million ~~books~~ newspapers?

Clemens Neudecker
Staatsbibliothek zu Berlin
17 May 2017

Staatsbibliothek
zu Berlin
Preußischer Kulturbesitz

PyData

How to read* a million newspapers?

(* hint: do try this at home)

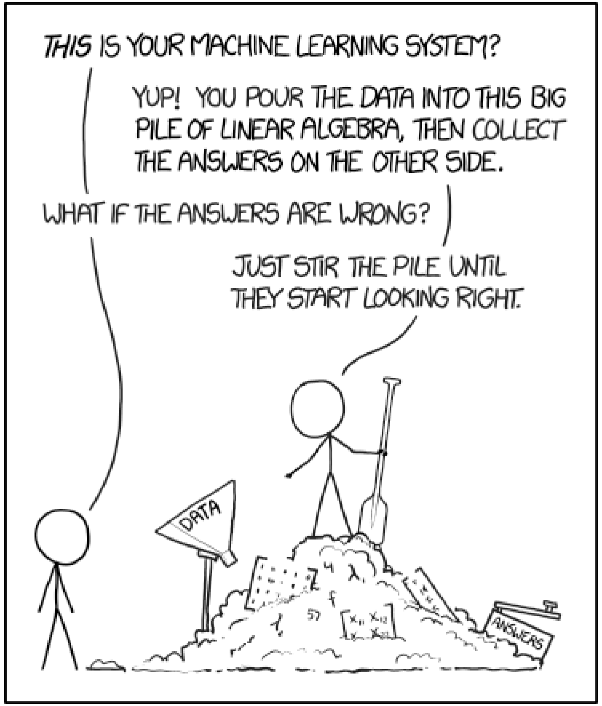
Clemens Neudecker
Staatsbibliothek zu Berlin

17 May 2017

Staatsbibliothek
zu Berlin
Preußischer Kulturbesitz

Agenda

- Digital Collections
- Data, Tools, Formats
- Europeana Newspapers
- NLP challenges
- Experiments & use cases


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

https://xkcd.com/1838

About me

- Research Coordinator @ Berlin State Library
- M.A. Philosophy, Computer Science, Political Science
- Mostly curious about
  - Optical Character Recognition, Document Analysis
  - Natural Language Processing
  - Digital Humanities
- More: @cneudecker, cneud.net

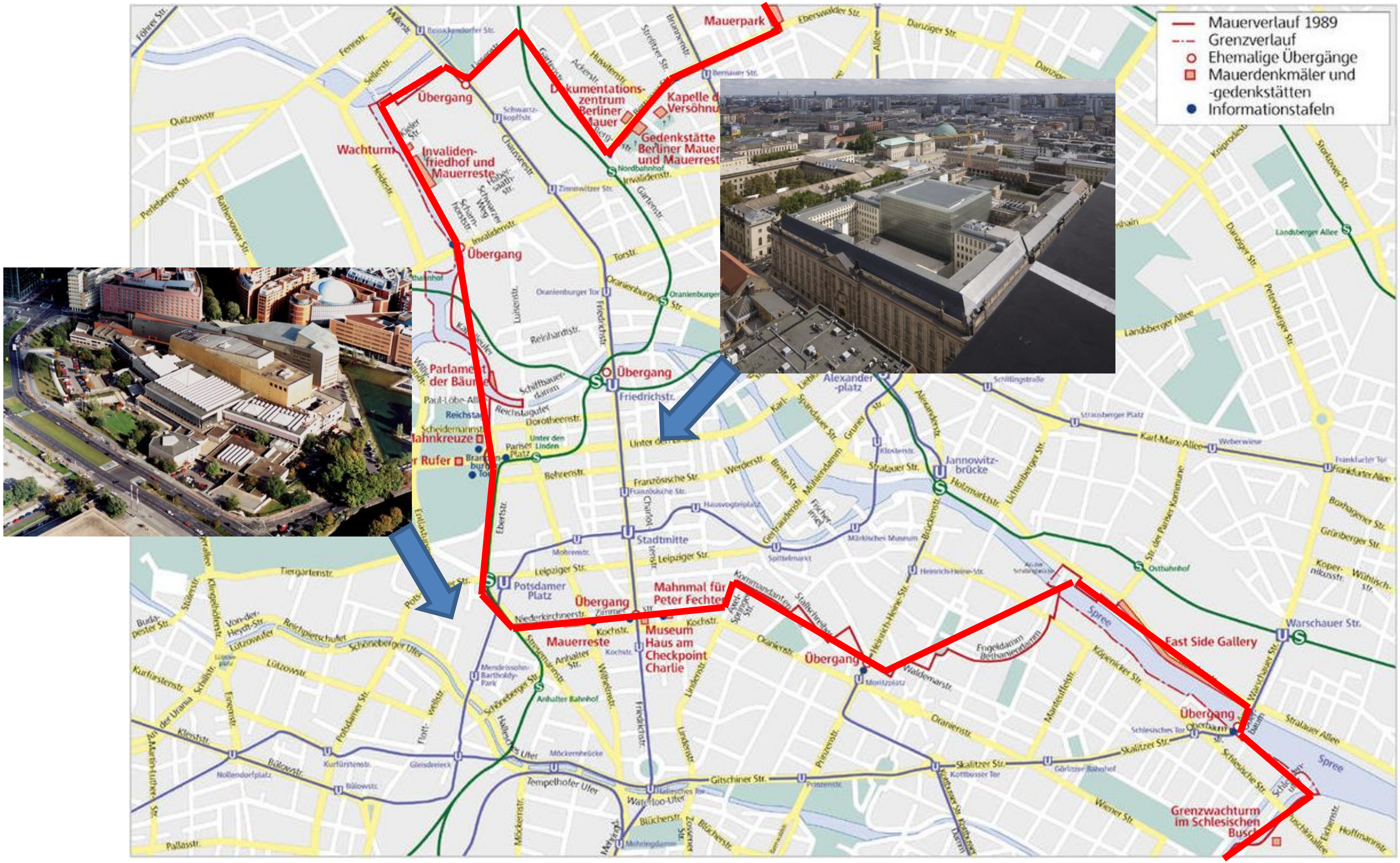
Staatsbibliothek zu Berlin

- Established 1661 as the Library of the King of Prussia

- Today largest research library in Germany,
  with approx. 11.5m volumes (23m objects)

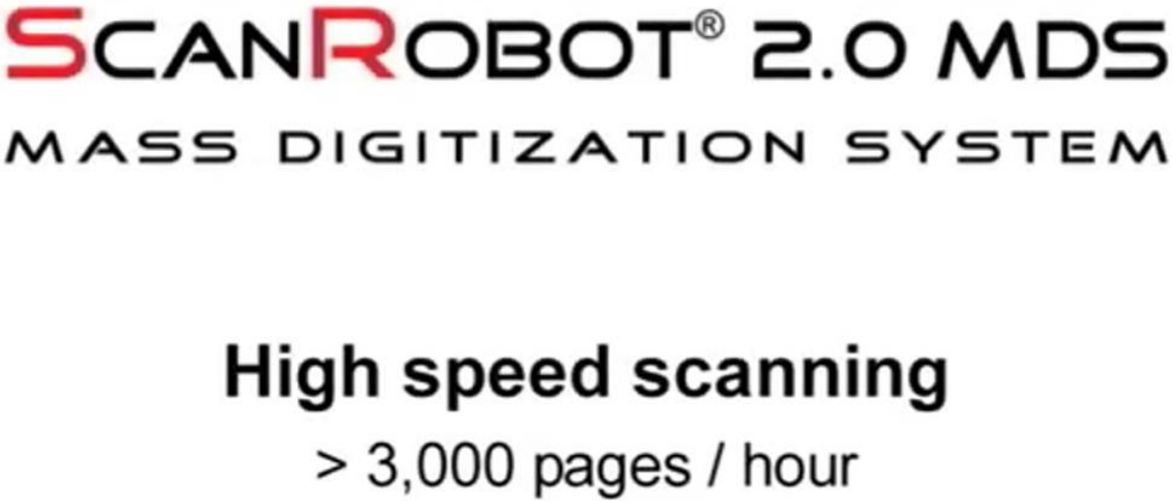
- Part of the „Stiftung Preußischer Kulturbesitz",
  a unique union of museums, archives, libraries
  and research institutes from Berlin

- http://staatsbibliothek-berlin.de/



Mauerverlauf 1989
Grenzverlauf
Ehemalige Übergänge
Mauerdenkmäler und -gedenkstätten
Informationstafeln
Mauerpark
Übergang
Dokumentationszentrum Berliner Mauer
Kapelle der Versöhnung
Gedenkstätte Berliner Mauer und Mauerreste
Wachturm
Invalidenfriedhof und Mauerreste
Nordbahnhof
Übergang
Oranienburger Tor
Oranienburger Str.
Übergang
Friedrichstr.
Alexanderplatz
Parlament der Bäume
Reichstag
Mahnkreuze
Rufer
Brandenburger Tor
Unter den Linden
Jannowitzbrücke
Stadtmitte
Potsdamer Platz
Übergang
Mahnmal für Peter Fechter
Museum Haus am Checkpoint Charlie
Mauerreste
Übergang
Moritzplatz
East Side Gallery
Ostbahnhof
Warschauer Str.
Übergang
Oberbaum
Grenzwachturm im Schlesischen Busch
Spree
Karl-Marx-Allee
Landsberger Allee
Frankfurter Allee
Leipziger Str.
Kochstr.
Anhalter Bahnhof
Tiergartenstr.
Schöneberger Ufer
Kurfürstenstr.
Bülowstr.
Gitschiner Str.
Skalitzer Str.
Oranienstr.
Köpenicker Str.
Invalidenstr.
Torstr.
Chausseestr.
Friedrichstr.
Bernauer Str.
Eberswalder Str.
Danziger Str.

Digitisation 2.0



ScanRobot® 2.0 MDS

MASS DIGITIZATION SYSTEM

High speed scanning

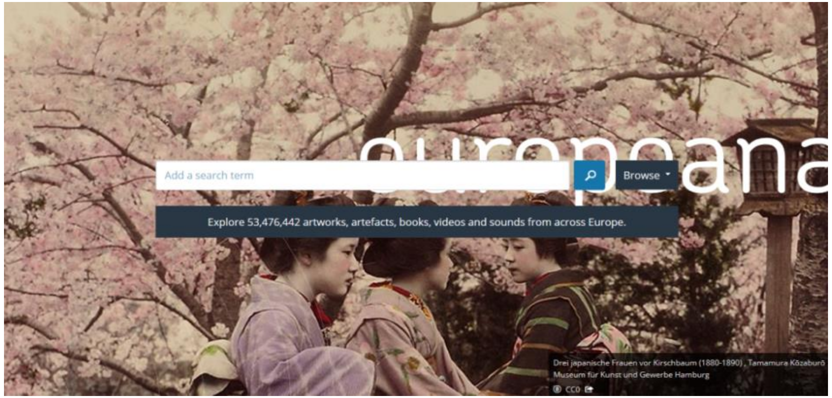
> 3,000 pages / hour

Europeana

- http://www.europeana.eu/portal/

- Europe‘s Digital Library

- > 53m objects incl.
  art, sound, fashion,...



Explore 53,476,442 artworks, artefacts, books, videos and sounds from across Europe.

Drei japanische Frauen vor Kirschbaum (1880-1890) , Tamamura Kōzaburō
Museum für Kunst und Gewerbe Hamburg
CC0

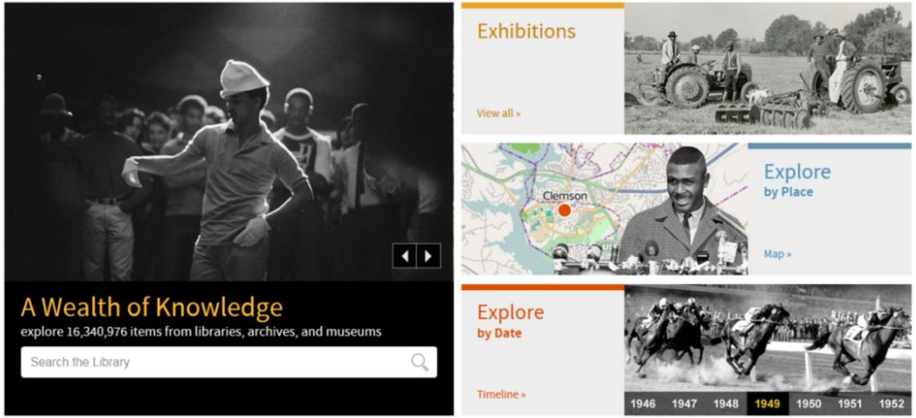
- API: http://labs.europeana.eu/api

DPLA

- https://dp.la/

- Digital Public Library of America

- > 16m objects



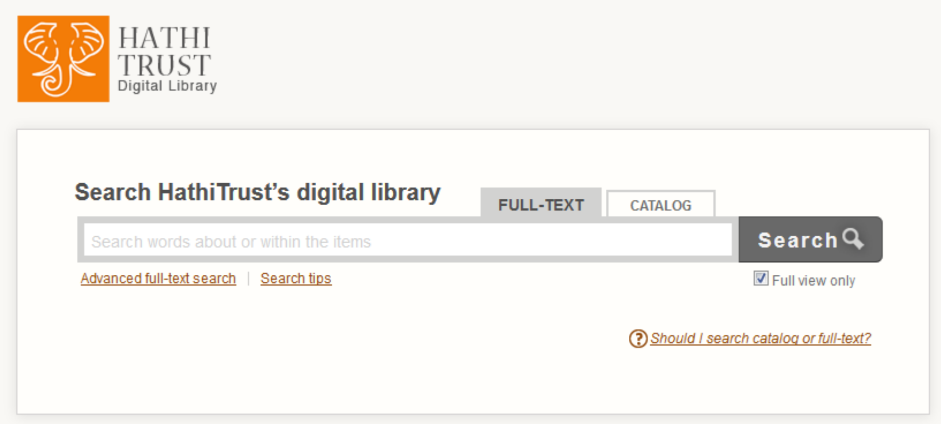
A Wealth of Knowledge

explore 16,340,976 items from libraries, archives, and museums

- API: https://dp.la/info/developers/codex/

Hathi Trust

- https://www.hathitrust.org/
- Public copy of Google Books
- > 15m volumes
- API: https://www.hathitrust.org/data



HATHI TRUST Digital Library

Search HathiTrust's digital library

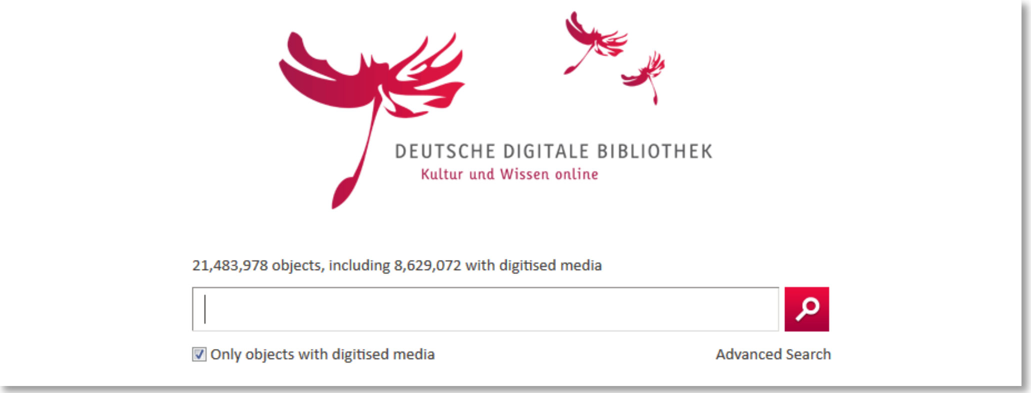
FULL-TEXT
CATALOG
Search words about or within the items
Search
Advanced full-text search
Search tips
☑ Full view only
Should I search catalog or full-text?

DDB

- http://ddb.de/

- Germany‘s federal Digital Library

- > 9m objects



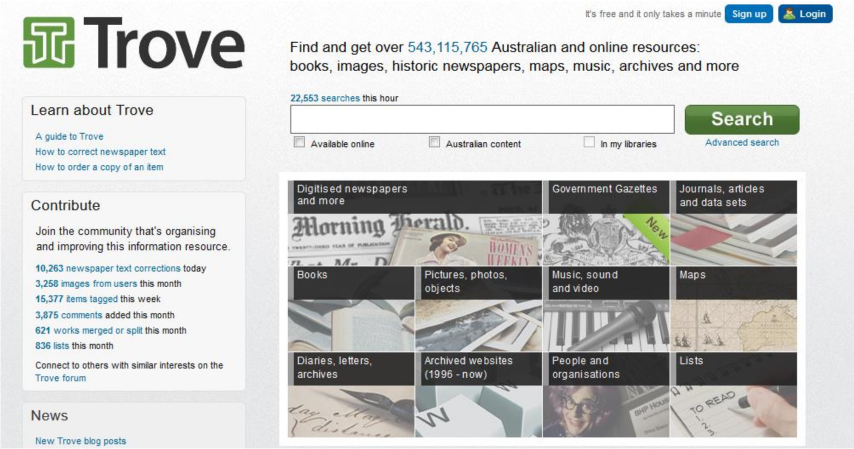
21,483,978 objects, including 8,629,072 with digitised media

☑ Only objects with digitised media

- API: https://api.deutsche-digitale-bibliothek.de/

Trove

- http://trove.nla.gov.au/



It's free and it only takes a minute

Find and get over 543,115,765 Australian and online resources: books, images, historic newspapers, maps, music, archives and more

22,553 searches this hour

Learn about Trove

A guide to Trove
How to correct newspaper text
How to order a copy of an item

Contribute

Join the community that's organising and improving this information resource.

10,263 newspaper text corrections today
3,258 images from users this month
15,377 items tagged this week
3,875 comments added this month
621 works merged or split this month
836 lists this month

Connect to others with similar interests on the Trove forum

News

New Trove blog posts

- Digital Library of Australia

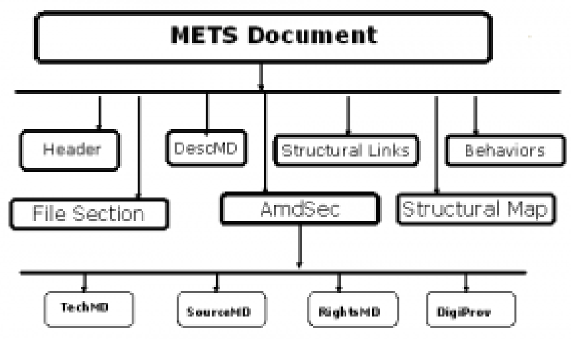
- > 540m objects

- API: http://help.nla.gov.au/trove/building-with-trove/api

Formats & Standards

- What data is available?

- Typically, a digital object is composed of:
  - Scanned Images in TIFF, JP2 or JPEG
  - Descriptive metadata in DublinCore
  - Structural metadata in METS
  - Text content in ALTO or TEI
  - Europeana in EDM
  - Linked Data in RDF or JSON-LD



Tools (1/3)

- OAI-PMH
  - https://pypi.python.org/pypi/Sickle
  - https://pypi.python.org/pypi/pyoai
  - https://pypi.python.org/pypi/oaiharvest
- METS
  - https://pypi.python.org/pypi/metsrw
  - https://pypi.python.org/pypi/pymets

Tools (2/3)

- DublinCore
  - https://pypi.python.org/pypi/pydc
  - https://pypi.python.org/pypi/dcxml
- Europeana
  - https://pypi.python.org/pypi/europeana-search
  - https://pypi.python.org/pypi/django-europeana

Tools (3/3)

- IIIF
  - https://pypi.python.org/pypi/iiif/
  - https://pypi.python.org/pypi/Flask-IIIF/
- KB NL
  - https://pypi.python.org/pypi/kb
  - https://github.com/KBNLresearch/intro-kb-apis
  - http://lab.kb.nl/

Europeana Newspapers

- EU-project to make Europe‘s historical newspapers searchable & accessible

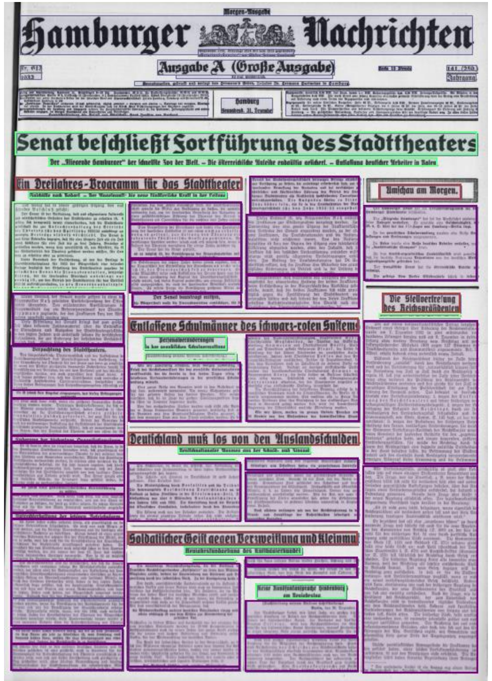
USE
Our Open-Source Tools
Designed to help libraries and others refine and improve the quality of digital historic newspapers.

- http://www.europeana-newspapers.eu/

Europeana Newspapers Collection

- 12 million historic newspaper pages text (> 10.000.000.000 tokens)
- 40 languages, 4 alphabets
- 400 years (1618 – 2016)
- http://www.theeuropeanlibrary.org/tel4/newspapers

OCR / OLR



Morgen-Ausgabe

Hamburger Nachrichten

Ausgabe A (Große Ausgabe)

Hamburg Sonnabend, 31. Dezember

Senat beſchließt Fortführung des Stadttheaters

Ein Dreijahres-Programm für das Stadttheater

Umſchau am Morgen.

Verpachtung des Stadttheaters.

Die Stellvertretung des Reichspräſidenten

Entlaſſene Schulmänner des ſchwarz-roten Syſtem

Perſonalveränderungen in der preußiſchen Schulverwaltung

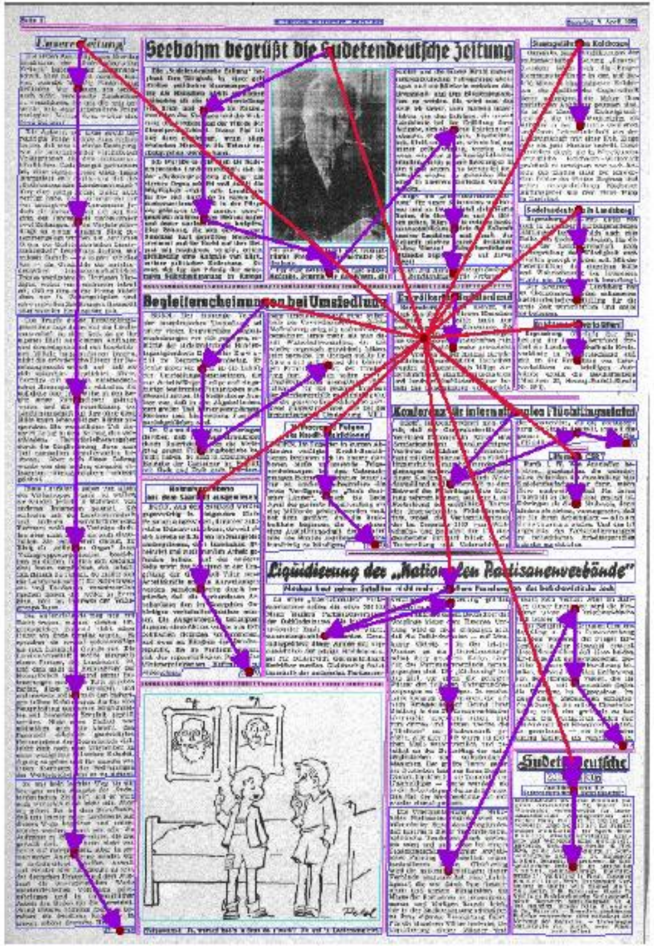
Deutſchland muß los von den Auslandsſchulden

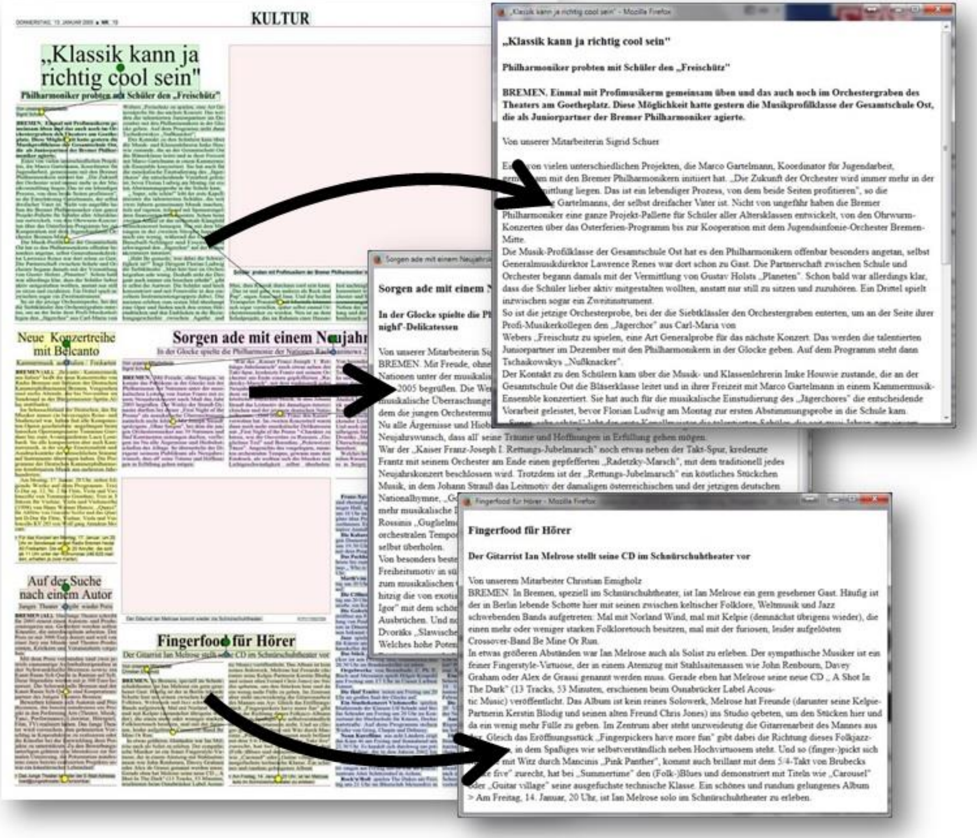
Soldatiſcher Geiſt gegen Verzweiflung und Kleinmu

(U.lag nul «chestttetrung- ■geeinoel II, Setch«it,zen I—Ig Ufr sterntpeechee g» U II. für ftrene-geingelpilche: 13 01191 nnd 13 03 11 io"gl f l««lt-beOeu; OetHn *1, blnftraße IS IZeinsptechee; H I Sanemeinummet gurfilrft 8MB); ««de«: gdn.o(tio||e III (ZemlpreAei 284.3»). Iie.gonlen nur nnier heimnnn » Erden bei der veutlchen Bonl »n« Vtdennld-Getelltchoil gttloto bumduig. Commerz- nndprinoldonlN voINchrSomI bomduig u 189 ß>, .»ontbeegee Kochelchlen- eitchelne» 12 mal wSchenNIch. täglich zweimal — morgen« nnd ndendn —, Sonntage nnr morgen». ToonInge nur abend» Zn den Kochdorerlen wird die Ndend-Nuegode noch am üben!

Dieser Entwurf ist. wie Bürgermeister Roß mitteilte, den Fraktionen zur Stellungnahme vorgelegt worden. Zum Donnerstag war eine zweite Sitzung der Fraktionsfübrer vom Vertreter des Senats angeordnet worden, zu der ober zwei Fraktionen, die Teutschnationalen und die Nationalsozialisten, nicht erschienen waren. Von den Nationalsozialisten ist kurz vor Beginn der Sitzung eine telephonische Erklärung abgegeben worden, etwa des Inhalts, daß die Fraktion sich den sachlichen Verhandlungen entziehen müsse, solange nicht gewisse Vorbedingungen erfüllt sein.

http://www.theeuropeanlibrary.org/tel4/newspapers/issue/Hamburger_Nachrichten/1932/12/31


Seebohm begrüßt die Sudetendeutsche Zeitung
Begleiterscheinungen bei Umsiedlung
Liquidierung der „Nationalen Partisanenverbände"
Sudetendeutsche


KULTUR
„Klassik kann ja richtig cool sein"
Philharmoniker probten mit Schüler den „Freischütz"
Neue Konzertreihe mit Belcanto
Sorgen ade mit einem Neujahr
Auf der Suche nach einem Autor
Fingerfood für Hörer
„Klassik kann ja richtig cool sein" - Mozilla Firefox
Sorgen ade mit einem Neujahr
Fingerfood für Hörer - Mozilla Firefox

„Klassik kann ja richtig cool sein"

Philharmoniker probten mit Schüler den „Freischütz"

BREMEN. Einmal mit Profimusikerm gemeinsam üben und das auch noch im Orchestergraben des Theaters am Goetheplatz. Diese Möglichkeit hatte gestern die Musikprofilklasse der Gesamtschule Ost, die als Juniorpartner der Bremer Philharmoniker agierte.

Von unserer Mitarbeiterin Sigrid Schuer

Ei von vielen unterschiedlichen Projekten, die Marco Gartelmann, Koordinator für Jugendarbeit,
gem am mit den Bremer Philharmonikern initiiert hat. „Die Zukunft der Orchester wird immer mehr in der
mittlung liegen. Das ist ein lebendiger Prozess, von dem beide Seiten profitieren", so die
Gartelmanns, der selbst dreifacher Vater ist. Nicht von ungefähr haben die Bremer
Philharmoniker eine ganze Projekt-Pallette für Schüler aller Altersklassen entwickelt, von den Ohrwurm-Konzerten über das Osterferien-Programm bis zur Kooperation mit dem Jugendsinfonie-Orchester Bremen-Mitte.

Die Musik-Profilklasse der Gesamtschule Ost hat es den Philharmonikern offenbar besonders angetan, selbst Generalmusikdirektor Lawrence Renes war dort schon zu Gast. Die Partnerschaft zwischen Schule und Orchester begann damals mit der Vermittlung von Gustav Holsts „Planeten". Schon bald war allerdings klar, dass die Schüler lieber aktiv mitgestalten wollten, anstatt nur still zu sitzen und zuzuhören. Ein Drittel spielt inzwischen sogar ein Zweitinstrument.

So ist die jetzige Orchesterprobe, bei der die Siebtklässler den Orchestergraben enterten, um an der Seite ihrer Profi-Musikerkollegen den „Jägerchor" aus Carl-Maria von
Webers „Freischutz zu spielen, eine Art Generalprobe für das nächste Konzert. Das werden die talentierten Juniorpartner im Dezember mit den Philharmonikern in der Glocke geben. Auf dem Programm steht dann Tschaikowskys „Nußknacker".

Der Kontakt zu den Schülern kam über die Musik- und Klassenlehrerin Imke Houwie zustande, die an der Gesamtschule Ost die Bläserklasse leitet und in ihrer Freizeit mit Marco Gartelmann in einem Kammermusik-Ensemble konzertiert. Sie hat auch für die musikalische Einstudierung des „Jägerchores" die entscheidende Vorarbeit geleistet, bevor Florian Ludwig am Montag zur ersten Abstimmungsprobe in die Schule kam.

Sorgen ade mit einem N

In der Glocke spielte die Ph
nigh"-Delikatessen

Von unserer Mitarbeiterin Si
BREMEN. Mit Freude, ohne
Nationen unter der musikalis
2005 begrüßen. Die We
musikalische Überraschunge
dem die jungen Orchestermu
Nu alle Ärgernisse und Hiob
Neujahrswunsch, dass all' seine Träume und Hoffnungen in Erfüllung gehen mögen.
War der „Kaiser Franz-Joseph I. Rettungs-Jubelmarsch" noch etwas neben der Takt-Spur, kredenzte
Frantz mit seinem Orchester am Ende einen gepfefferten „Radetzky-Marsch", mit dem traditionell jedes
Neujahrskonzert beschlossen wird. Trotzdem ist der „Rettungs-Jubelmarsch" ein köstliches Stückchen
Musik, in dem Johann Strauß das Leitmotiv der damaligen österreichischen und der jetzigen deutschen
Nationalhymne, „Go
mehr musikalische I
Rossinis „Guglielm
orchestralen Tempo
selbst überholen.
Von besonders beste
Freiheitsmotiv in sü
zum musikalischen
hitzig die von exoti
Igor" mit dem schö
Ausbrüchen. Und n
Dvořáks „Slawische
Welches hohe Poten

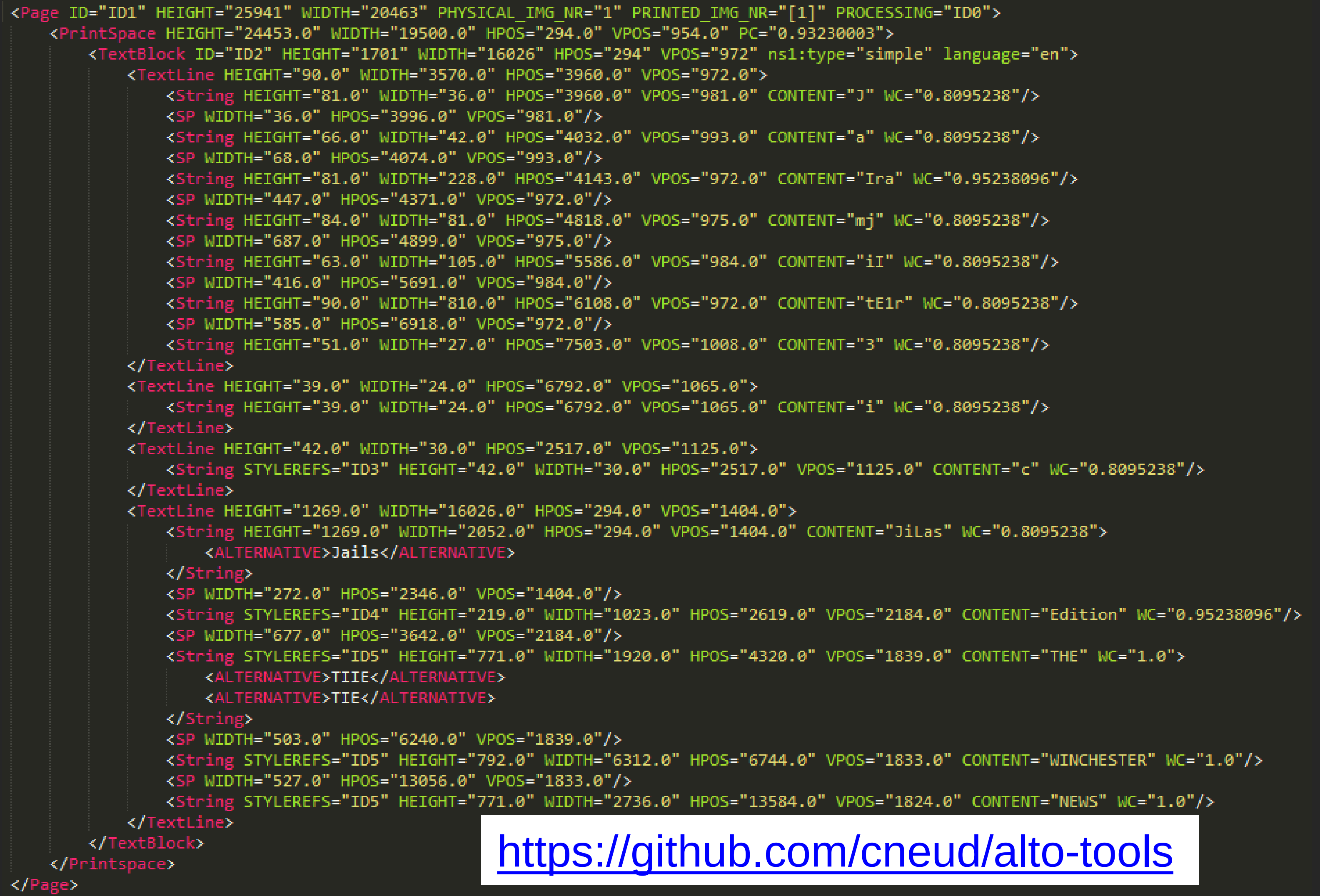
Fingerfood für Hörer

Der Gitarrist Ian Melrose stellt seine CD im Schnürschuhtheater vor

Von unserem Mitarbeiter Christian Emigholz

BREMEN. In Bremen, speziell im Schnürschuhtheater, ist Ian Melrose ein gern gesehener Gast. Häufig ist der in Berlin lebende Schotte hier mit seinen zwischen keltischer Folklore, Weltmusik und Jazz schwebenden Bands aufgetreten: Mal mit Norland Wind, mal mit Kelpie (demnächst übrigens wieder), die einen mehr oder weniger starken Folkloretouch besitzen, mal mit der furiosen, leider aufgelösten Crossover-Band Be Mine Or Run.

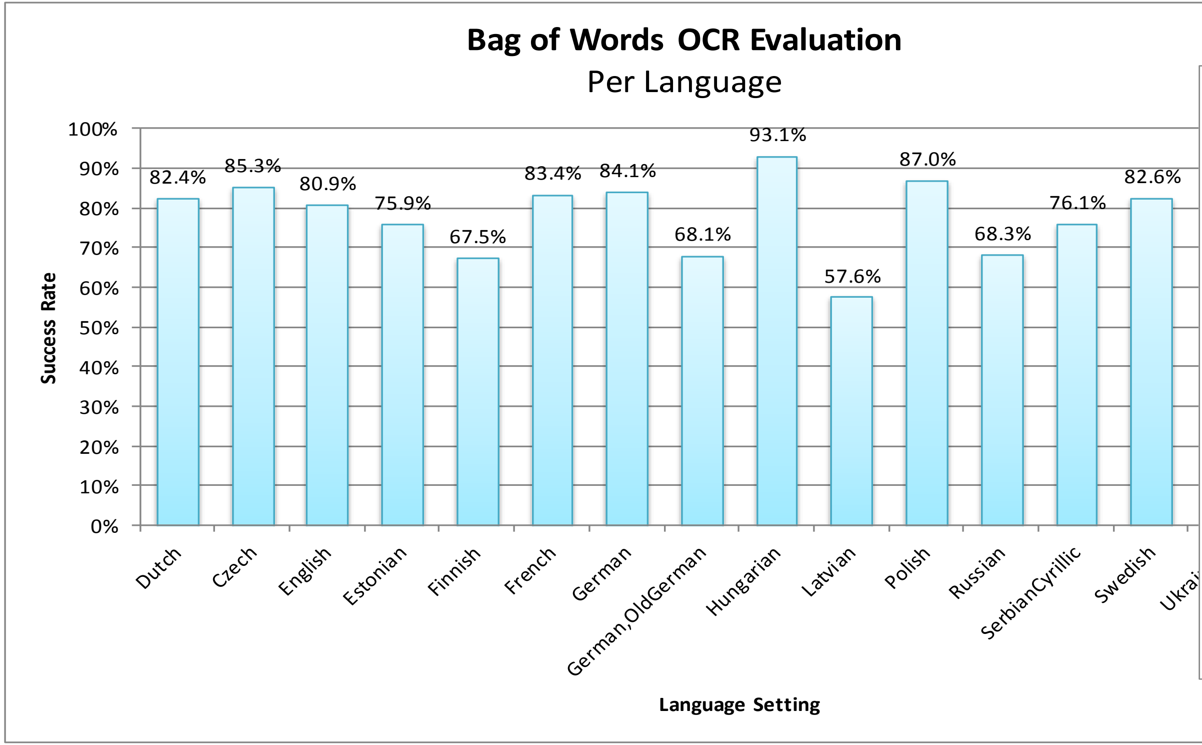
In etwas größeren Abständen war Ian Melrose auch als Solist zu erleben. Der sympathische Musiker ist ein feiner Fingerstyle-Virtuose, der in einem Atemzug mit Stahlsaitenassen wie John Renbourn, Davey Graham oder Alex de Grassi genannt werden muss. Gerade eben hat Melrose seine neue CD „ A Shot In The Dark" (13 Tracks, 53 Minuten, erschienen beim Osnabrücker Label Acous-
tic Music) veröffentlicht. Das Album ist kein reines Solowerk, Melrose hat Freunde (darunter seine Kelpie-Partnerin Kerstin Blodig und seinen alten Freund Chris Jones) ins Studio qebeten, um den Stücken hier und da ein wenig mehr Fülle zu geben. Im Zentrum aber steht unzweideutig die Gitarrenarbeit des Mannes aus
. Gleich das Eröffnungsstück „Fingerpickers have more fun" gibt dabei die Richtung dieses Folkjazz-
, in dem Spaßiges wie selbstverständlich neben Hochvirtuosem steht. Und so (finger-)pickt sich
mit Witz durch Mancinis „Pink Panther", kommt auch brillant mit dem 5/4-Takt von Brubecks
e five" zurecht, hat bei „Summertime" den (Folk-)Blues und demonstriert mit Titeln wie „Carousel"
oder „Guitar village" seine ausgefuchste technische Klasse. Ein schönes und rundum gelungenes Album
> Am Freitag, 14. Januar, 20 Uhr, ist Ian Melrose solo im Schnürschuhtheater zu erleben.



```
<Page ID="ID1" HEIGHT="25941" WIDTH="20463" PHYSICAL_IMG_NR="1" PRINTED_IMG_NR="[1]" PROCESSING="ID0">
    <PrintSpace HEIGHT="24453.0" WIDTH="19500.0" HPOS="294.0" VPOS="954.0" PC="0.93230003">
        <TextBlock ID="ID2" HEIGHT="1701" WIDTH="16026" HPOS="294" VPOS="972" ns1:type="simple" language="en">
            <TextLine HEIGHT="90.0" WIDTH="3570.0" HPOS="3960.0" VPOS="972.0">
                <String HEIGHT="81.0" WIDTH="36.0" HPOS="3960.0" VPOS="981.0" CONTENT="J" WC="0.8095238"/>
                <SP WIDTH="36.0" HPOS="3996.0" VPOS="981.0"/>
                <String HEIGHT="66.0" WIDTH="42.0" HPOS="4032.0" VPOS="993.0" CONTENT="a" WC="0.8095238"/>
                <SP WIDTH="68.0" HPOS="4074.0" VPOS="993.0"/>
                <String HEIGHT="81.0" WIDTH="228.0" HPOS="4143.0" VPOS="972.0" CONTENT="Ira" WC="0.95238096"/>
                <SP WIDTH="447.0" HPOS="4371.0" VPOS="972.0"/>
                <String HEIGHT="84.0" WIDTH="81.0" HPOS="4818.0" VPOS="975.0" CONTENT="mj" WC="0.8095238"/>
                <SP WIDTH="687.0" HPOS="4899.0" VPOS="975.0"/>
                <String HEIGHT="63.0" WIDTH="105.0" HPOS="5586.0" VPOS="984.0" CONTENT="iI" WC="0.8095238"/>
                <SP WIDTH="416.0" HPOS="5691.0" VPOS="984.0"/>
                <String HEIGHT="90.0" WIDTH="810.0" HPOS="6108.0" VPOS="972.0" CONTENT="tE1r" WC="0.8095238"/>
                <SP WIDTH="585.0" HPOS="6918.0" VPOS="972.0"/>
                <String HEIGHT="51.0" WIDTH="27.0" HPOS="7503.0" VPOS="1008.0" CONTENT="3" WC="0.8095238"/>
            </TextLine>
            <TextLine HEIGHT="39.0" WIDTH="24.0" HPOS="6792.0" VPOS="1065.0">
                <String HEIGHT="39.0" WIDTH="24.0" HPOS="6792.0" VPOS="1065.0" CONTENT="i" WC="0.8095238"/>
            </TextLine>
            <TextLine HEIGHT="42.0" WIDTH="30.0" HPOS="2517.0" VPOS="1125.0">
                <String STYLEREFS="ID3" HEIGHT="42.0" WIDTH="30.0" HPOS="2517.0" VPOS="1125.0" CONTENT="c" WC="0.8095238"/>
            </TextLine>
            <TextLine HEIGHT="1269.0" WIDTH="16026.0" HPOS="294.0" VPOS="1404.0">
                <String HEIGHT="1269.0" WIDTH="2052.0" HPOS="294.0" VPOS="1404.0" CONTENT="JiLas" WC="0.8095238">
                    <ALTERNATIVE>Jails</ALTERNATIVE>
                </String>
                <SP WIDTH="272.0" HPOS="2346.0" VPOS="1404.0"/>
                <String STYLEREFS="ID4" HEIGHT="219.0" WIDTH="1023.0" HPOS="2619.0" VPOS="2184.0" CONTENT="Edition" WC="0.95238096"/>
                <SP WIDTH="677.0" HPOS="3642.0" VPOS="2184.0"/>
                <String STYLEREFS="ID5" HEIGHT="771.0" WIDTH="1920.0" HPOS="4320.0" VPOS="1839.0" CONTENT="THE" WC="1.0">
                    <ALTERNATIVE>TIIE</ALTERNATIVE>
                    <ALTERNATIVE>TIE</ALTERNATIVE>
                </String>
                <SP WIDTH="503.0" HPOS="6240.0" VPOS="1839.0"/>
                <String STYLEREFS="ID5" HEIGHT="792.0" WIDTH="6312.0" HPOS="6744.0" VPOS="1833.0" CONTENT="WINCHESTER" WC="1.0"/>
                <SP WIDTH="527.0" HPOS="13056.0" VPOS="1833.0"/>
                <String STYLEREFS="ID5" HEIGHT="771.0" WIDTH="2736.0" HPOS="13584.0" VPOS="1824.0" CONTENT="NEWS" WC="1.0"/>
            </TextLine>
        </TextBlock>
    </Printspace>
</Page>
```

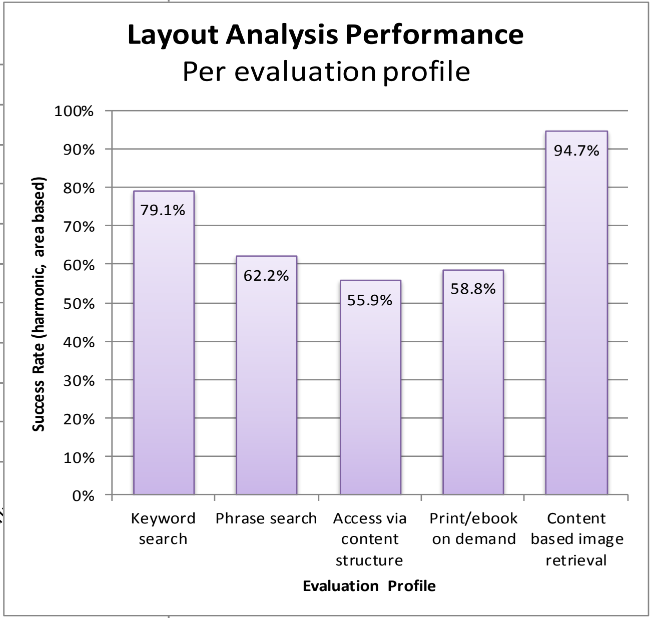
https://github.com/cneud/alto-tools

Performance



Bag of Words OCR Evaluation

Per Language



Layout Analysis Performance

Per evaluation profile

Experimental(!) Downloads

- http://data.theeuropeanlibrary.org/download/newspapers-by-country/README.html
- http://research.europeana.eu/itemtype/newspapers
- http://test-solr-mongo.eanadev.org/europeana-research-newspapers-dump/sample-2017-04-26/Staatsbibliothek_zu_Berlin_Preu%253Fischer_Kulturbesitz/titles.html

OCR

- EU: IMPACT project (2008-2012)
- US: eMOP project (2013-2015)
- DE: OCR-D project (2016-2018)

Improving Access to Text
IMPACT

EMOP
Early Modern OCR Project

OCR-D
Koordinierungsprojekt zur Weiterentwicklung von Verfahren der Optical Character Recognition (OCR)

- Google:
  - Tesseract
  - ocropy (fka OCRopus)
  - Aksara

Named Entity Recognition

- 3 Categories:
  - PERSON; LOCATION; ORGANIZATION
- 3 Languages:
  - Dutch; French; German
- Powered by Stanford CoreNLP - CRF-NER

The Stanford Natural Language Processing Group

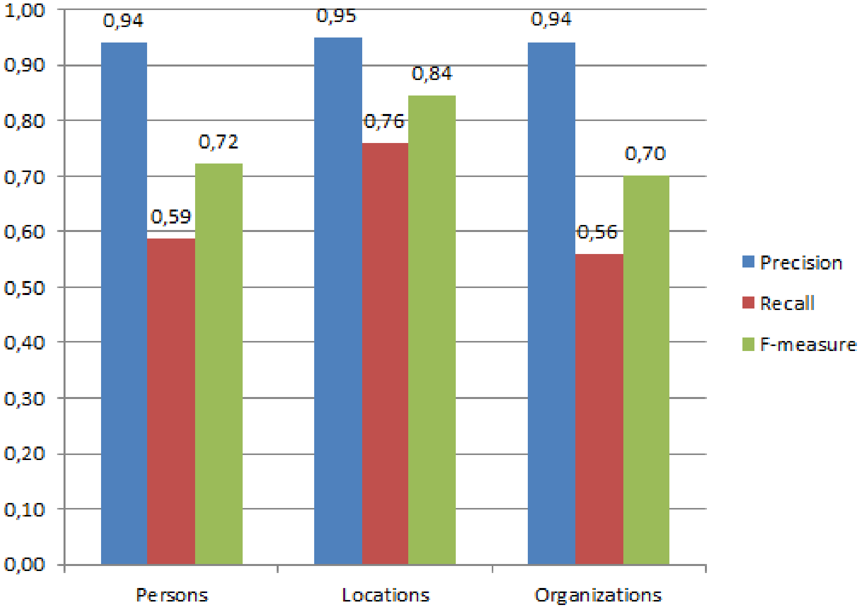
Annotations

Language	# tokens	# PER	# LOC	# ORG
French	207,000	5,672	5,614	2,574
Dutch	182,483	4,492	4,448	1,160
German	96,735	7,914	6,143	2,784

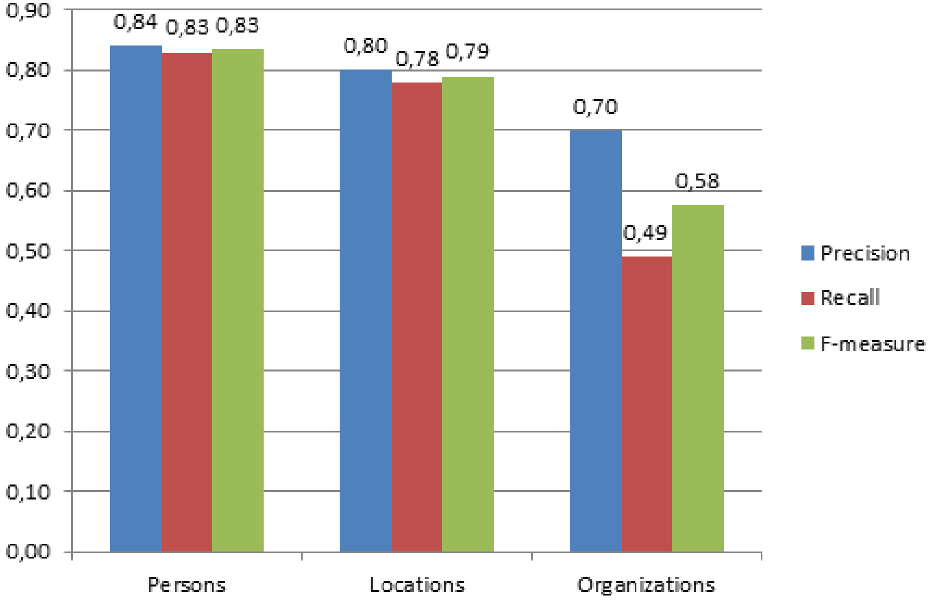
Language	# tokens	# PER	# LOC	# ORG
French	100%	2,75%	2,71%	1,24%
Dutch	100%	2,46%	2,44%	0,64%
German	100%	8,18%	6,35%	2,88%

Language	Word-Error-Rate (Bag of Words)	Reading Order Success Rate
French	16,6%	19,9%
Dutch	17,6%	23,2%
German	15,9% / 21,9%	13,6%

Evaluation

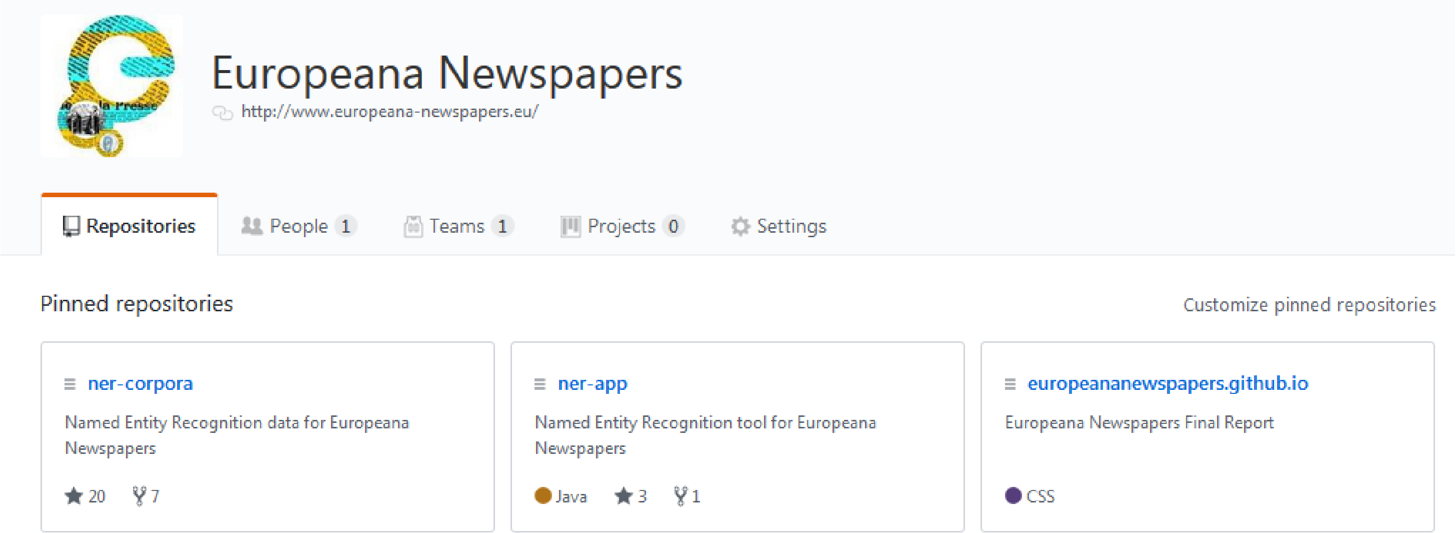


Dutch



French

Challenges



Europeana Newspapers

http://www.europeana-newspapers.eu/

Pinned repositories

ner-corpora

Named Entity Recognition data for Europeana Newspapers

★ 20 7

ner-app

Named Entity Recognition tool for Europeana Newspapers

● Java ★ 3 1

europeananewspapers.github.io

Europeana Newspapers Final Report

● CSS

- https://github.com/EuropeanaNewspapers/ner-corpora/wiki/Corpus-cleanup

Lack of metadata

Issue

There is no associated metadata for the annotated text (newspaper title, date, etc.)

Solution

Automatically match lines with newspaper pages through keyword search

OCR errors vs. historical spelling

Issue

Text contains OCR errors but also valid(!) historical spelling variants

$$teil \xrightarrow{t \to th, ei \to ey} theyl$$

Solution

Document language profiling to distinguish OCR errors and spelling variants

Sentence splits

Issue

During data pre-processing, (parts) of sentences have been erroneously cut

Solution

Reconstruct sentences through keyword search and matching procedure

Hyphenation

Issue

Text contains hyphenation to be removed but hyphens do also occur in regular text

Solution

Use a tokenizer to determine hyphens to be removed

Missing tags

Issue

Human operators forgot to tag some entities or tagged them with the wrong category

Solution

???

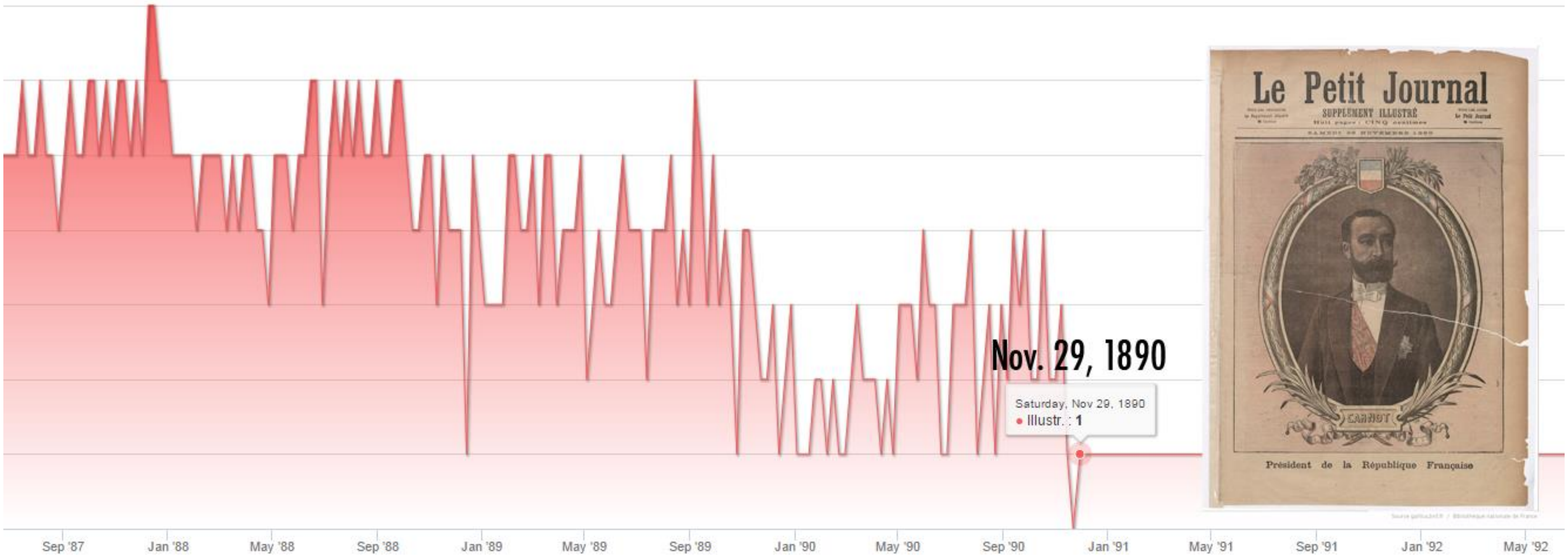
Punctuation

Issue

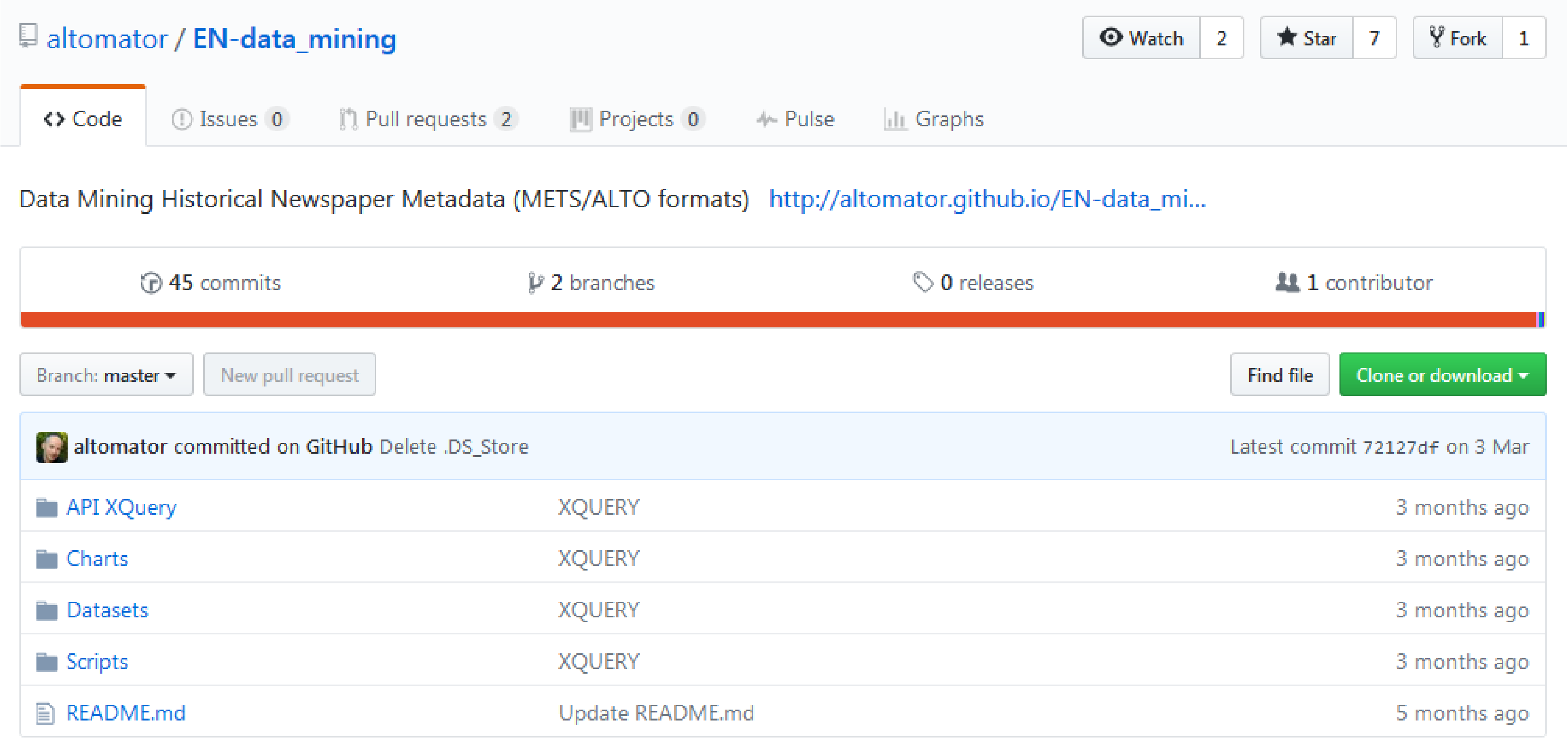
According to CONLL, punctuation should be in a separate line from the token - but abbreviations…

Solution

???



https://altomator.github.io/EN-data_mining/



altomator / EN-data_mining

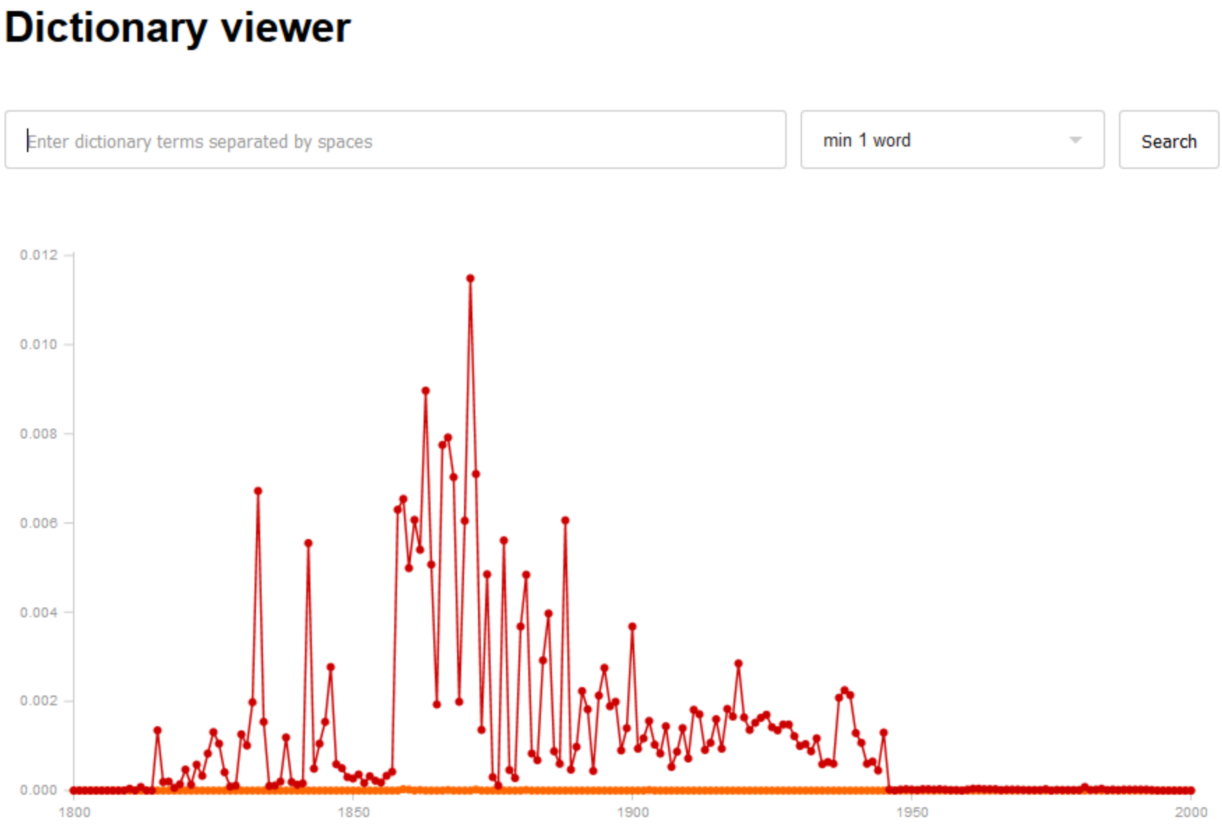
Data Mining Historical Newspaper Metadata (METS/ALTO formats) http://altomator.github.io/EN-data_mi...

45 commits · 2 branches · 0 releases · 1 contributor

altomator committed on GitHub Delete .DS_Store — Latest commit 72127df on 3 Mar

API XQuery	XQUERY	3 months ago
Charts	XQUERY	3 months ago
Datasets	XQUERY	3 months ago
Scripts	XQUERY	3 months ago
README.md	Update README.md	5 months ago

https://github.com/altomator/EN-data_mining

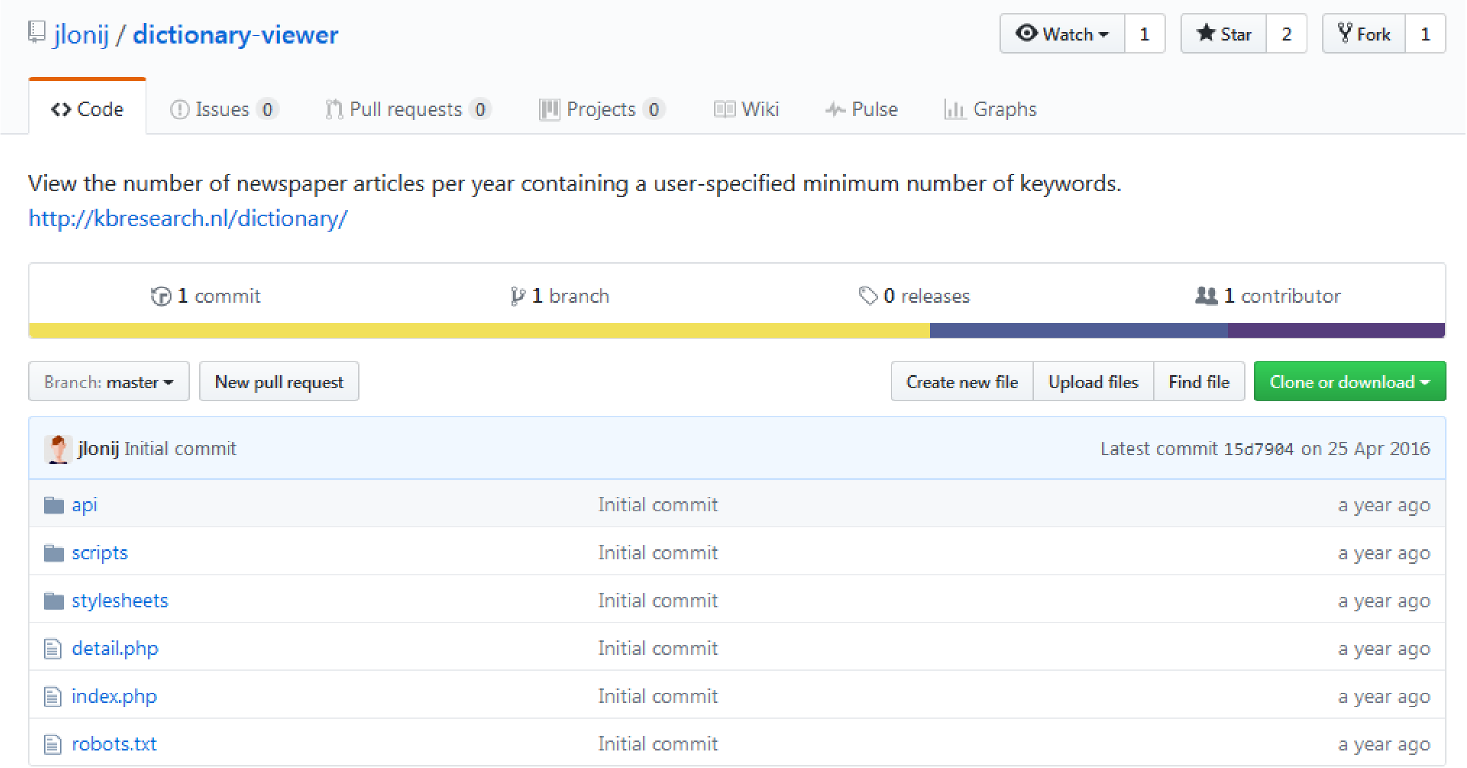


Dictionary viewer

Enter dictionary terms separated by spaces

min 1 word

http://www.kbresearch.nl/dictionary/



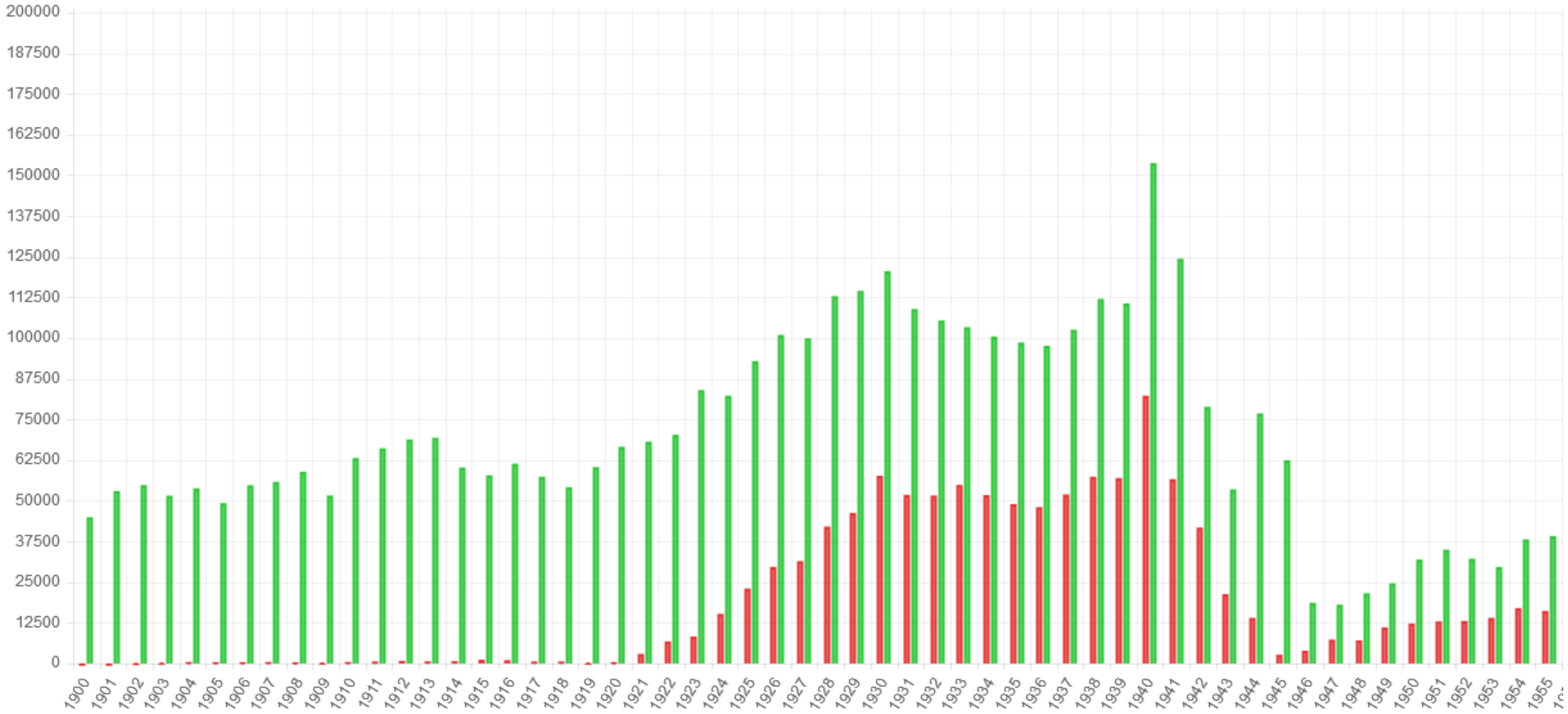
View the number of newspaper articles per year containing a user-specified minimum number of keywords.
http://kbresearch.nl/dictionary/

1 commit 1 branch 0 releases 1 contributor

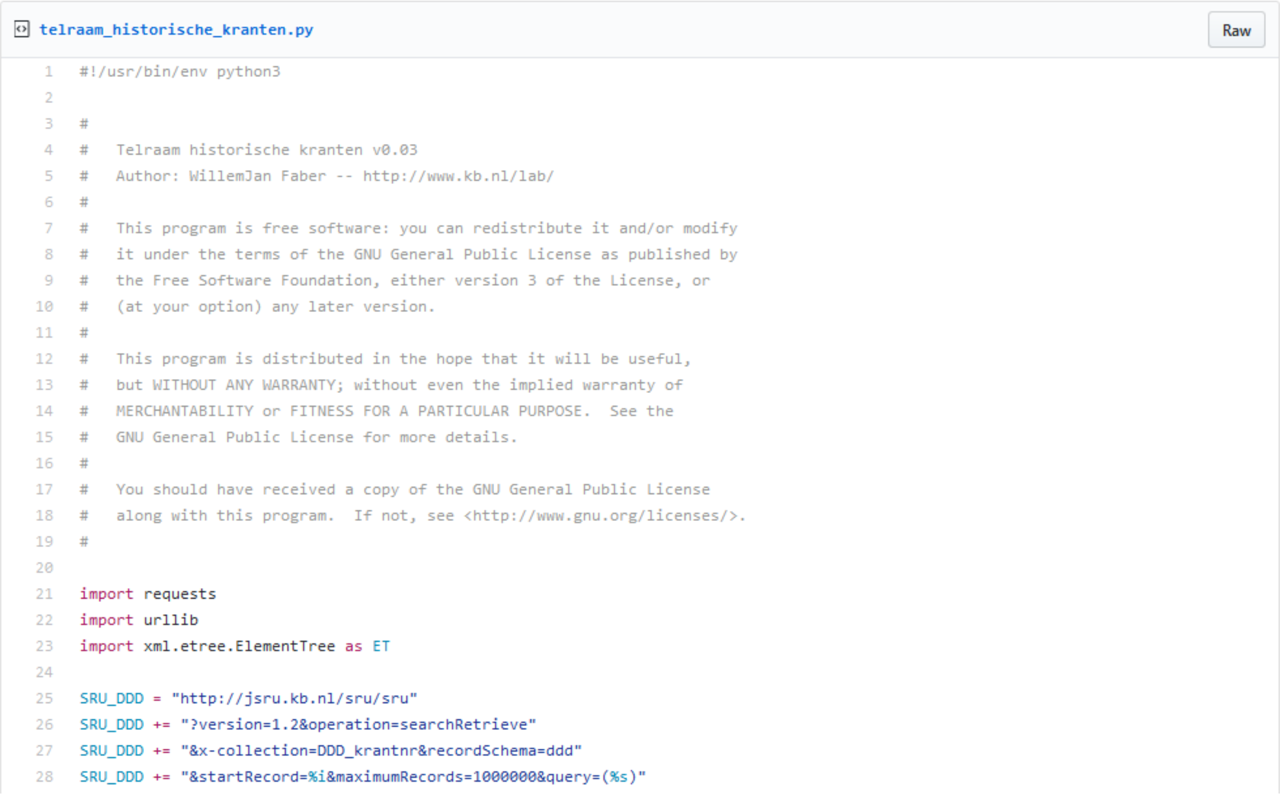
jlonij Initial commit Latest commit 15d7904 on 25 Apr 2016

api	Initial commit	a year ago
scripts	Initial commit	a year ago
stylesheets	Initial commit	a year ago
detail.php	Initial commit	a year ago
index.php	Initial commit	a year ago
robots.txt	Initial commit	a year ago

https://github.com/jlonij/dictionary-viewer



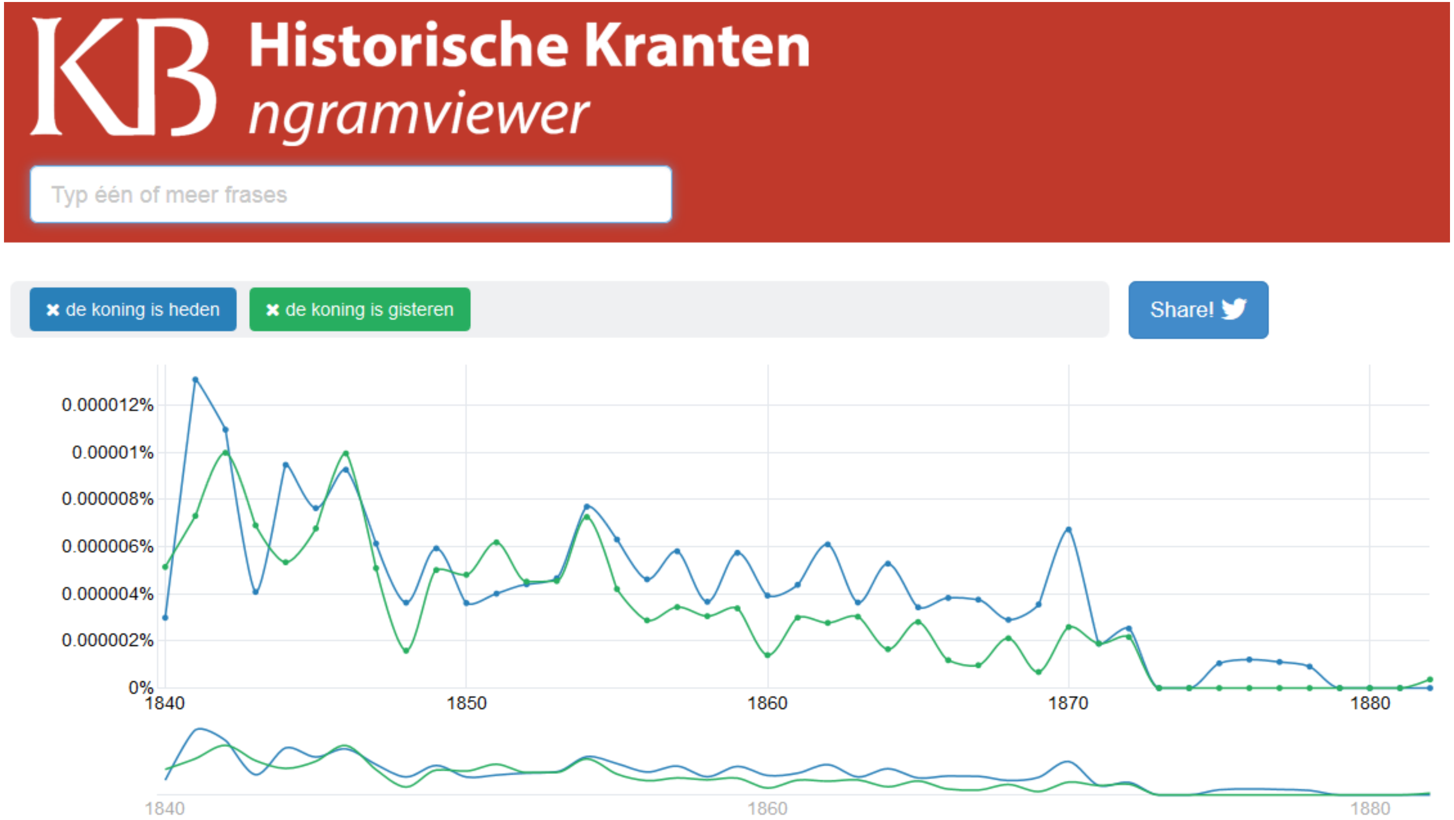
http://www.kbresearch.nl/telraam/



telraam_historische_kranten.py

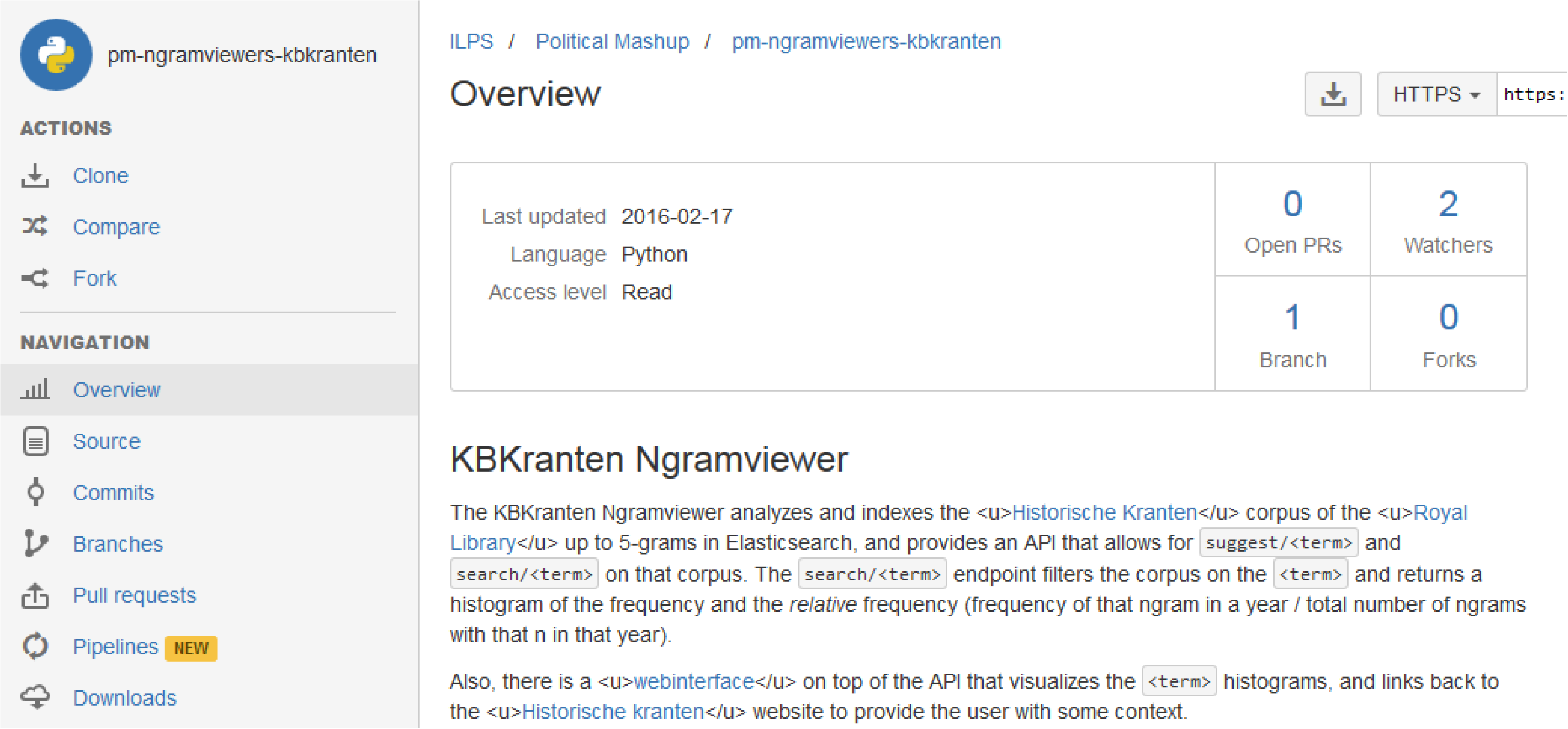
```python
#!/usr/bin/env python3

#
#   Telraam historische kranten v0.03
#   Author: WillemJan Faber -- http://www.kb.nl/lab/
#
#   This program is free software: you can redistribute it and/or modify
#   it under the terms of the GNU General Public License as published by
#   the Free Software Foundation, either version 3 of the License, or
#   (at your option) any later version.
#
#   This program is distributed in the hope that it will be useful,
#   but WITHOUT ANY WARRANTY; without even the implied warranty of
#   MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
#   GNU General Public License for more details.
#
#   You should have received a copy of the GNU General Public License
#   along with this program.  If not, see <http://www.gnu.org/licenses/>.
#

import requests
import urllib
import xml.etree.ElementTree as ET

SRU_DDD = "http://jsru.kb.nl/sru/sru"
SRU_DDD += "?version=1.2&operation=searchRetrieve"
SRU_DDD += "&x-collection=DDD_krantnr&recordSchema=ddd"
SRU_DDD += "&startRecord=%i&maximumRecords=1000000&query=(%s)"
```

https://gist.github.com/WillemJan/6ab02c48af576ba47b68



KB Historische Kranten ngramviewer

http://ngramviewer.kbresearch.nl/



Overview

Last updated 2016-02-17
Language Python
Access level Read

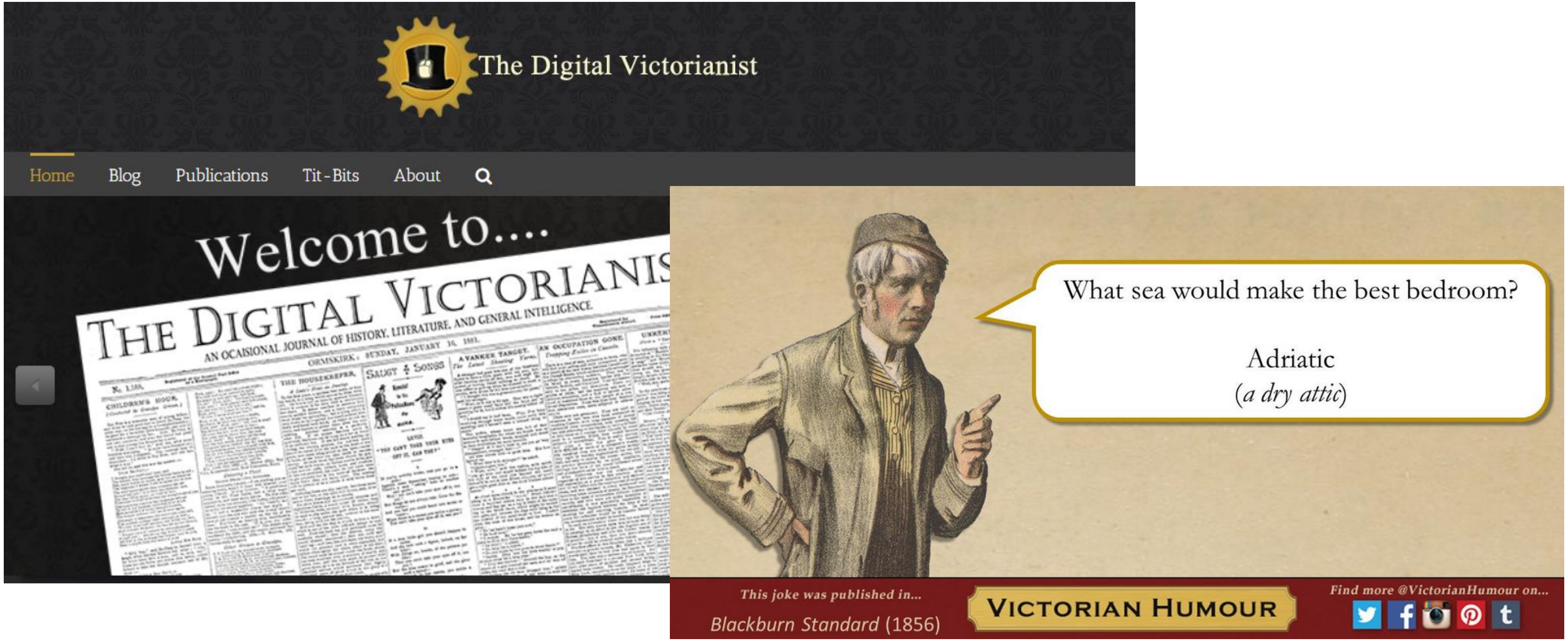
0 Open PRs
2 Watchers
1 Branch
0 Forks

KBKranten Ngramviewer

The KBKranten Ngramviewer analyzes and indexes the <u>Historische Kranten</u> corpus of the <u>Royal Library</u> up to 5-grams in Elasticsearch, and provides an API that allows for `suggest/<term>` and `search/<term>` on that corpus. The `search/<term>` endpoint filters the corpus on the `<term>` and returns a histogram of the frequency and the *relative* frequency (frequency of that ngram in a year / total number of ngrams with that n in that year).

Also, there is a <u>webinterface</u> on top of the API that visualizes the `<term>` histograms, and links back to the <u>Historische kranten</u> website to provide the user with some context.

https://bitbucket.org/ilps/pm-ngramviewers-kbkranten



The Digital Victorianist

Home Blog Publications Tit-Bits About

Welcome to....

THE DIGITAL VICTORIANIS

AN OCASIONAL JOURNAL OF HISTORY, LITERATURE, AND GENERAL INTELLIGENCE.

What sea would make the best bedroom?

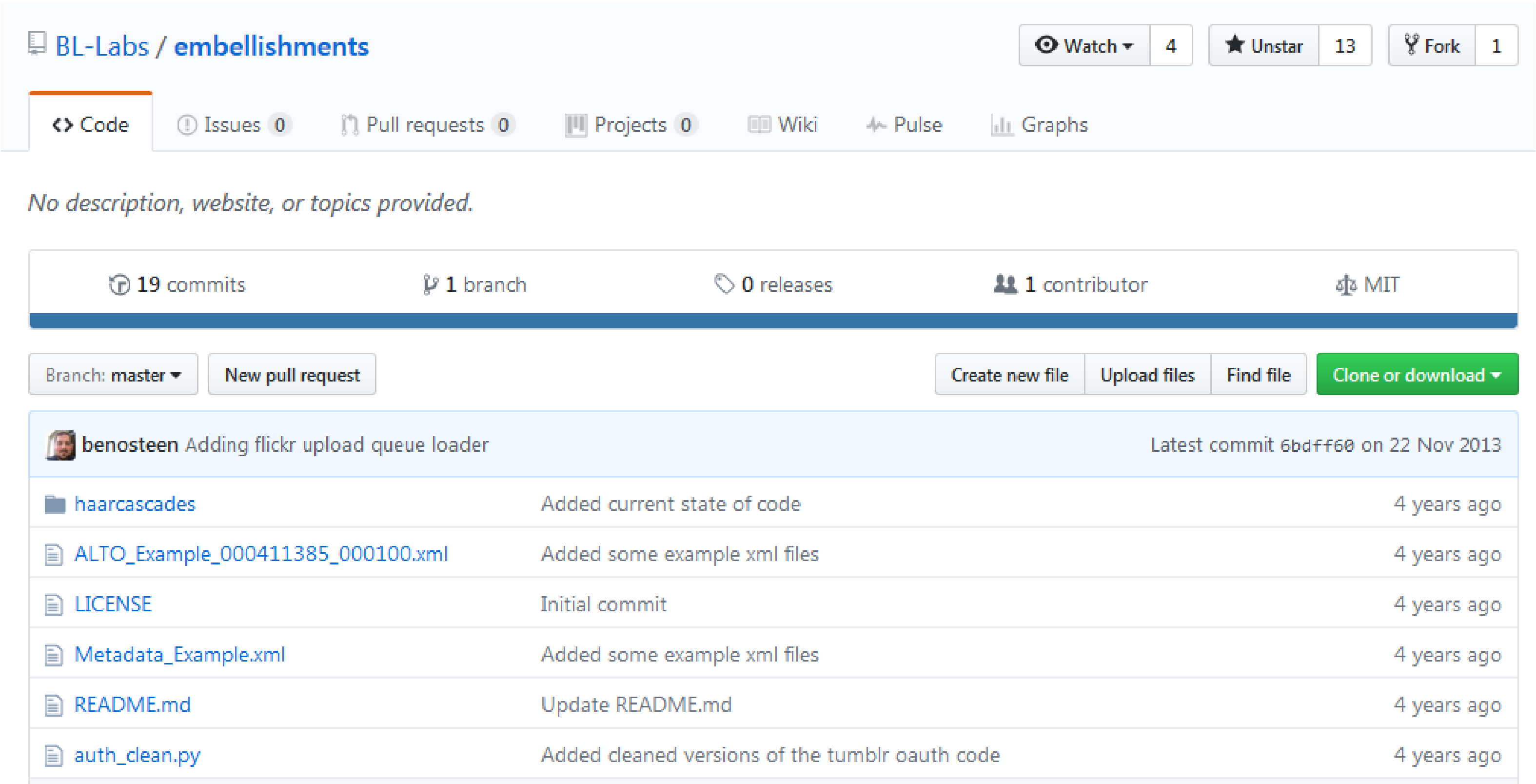
Adriatic
(*a dry attic*)

This joke was published in...
Blackburn Standard (1856)

VICTORIAN HUMOUR

Find more @VictorianHumour on...

http://www.digitalvictorianist.com/
https://twitter.com/VictorianHumour



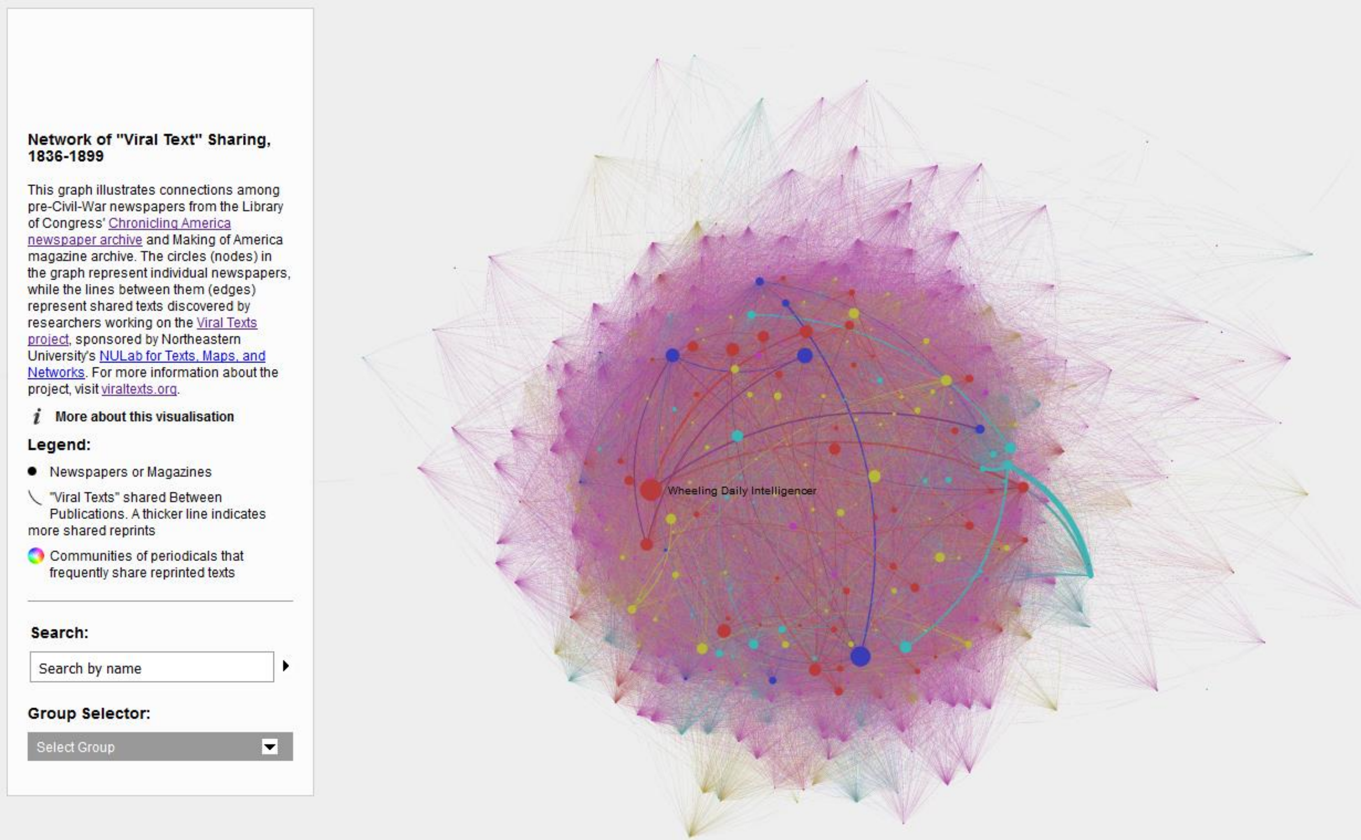
BL-Labs / embellishments

No description, website, or topics provided.

benosteen Adding flickr upload queue loader — Latest commit 6bdff60 on 22 Nov 2013

haarcascades	Added current state of code	4 years ago
ALTO_Example_000411385_000100.xml	Added some example xml files	4 years ago
LICENSE	Initial commit	4 years ago
Metadata_Example.xml	Added some example xml files	4 years ago
README.md	Update README.md	4 years ago
auth_clean.py	Added cleaned versions of the tumblr oauth code	4 years ago

https://github.com/BL-Labs/embellishments



Network of "Viral Text" Sharing, 1836-1899

This graph illustrates connections among pre-Civil-War newspapers from the Library of Congress' Chronicling America newspaper archive and Making of America magazine archive. The circles (nodes) in the graph represent individual newspapers, while the lines between them (edges) represent shared texts discovered by researchers working on the Viral Texts project, sponsored by Northeastern University's NULab for Texts, Maps, and Networks. For more information about the project, visit viraltexts.org.

i **More about this visualisation**

Legend:

● Newspapers or Magazines

╲ "Viral Texts" shared Between Publications. A thicker line indicates more shared reprints

◉ Communities of periodicals that frequently share reprinted texts

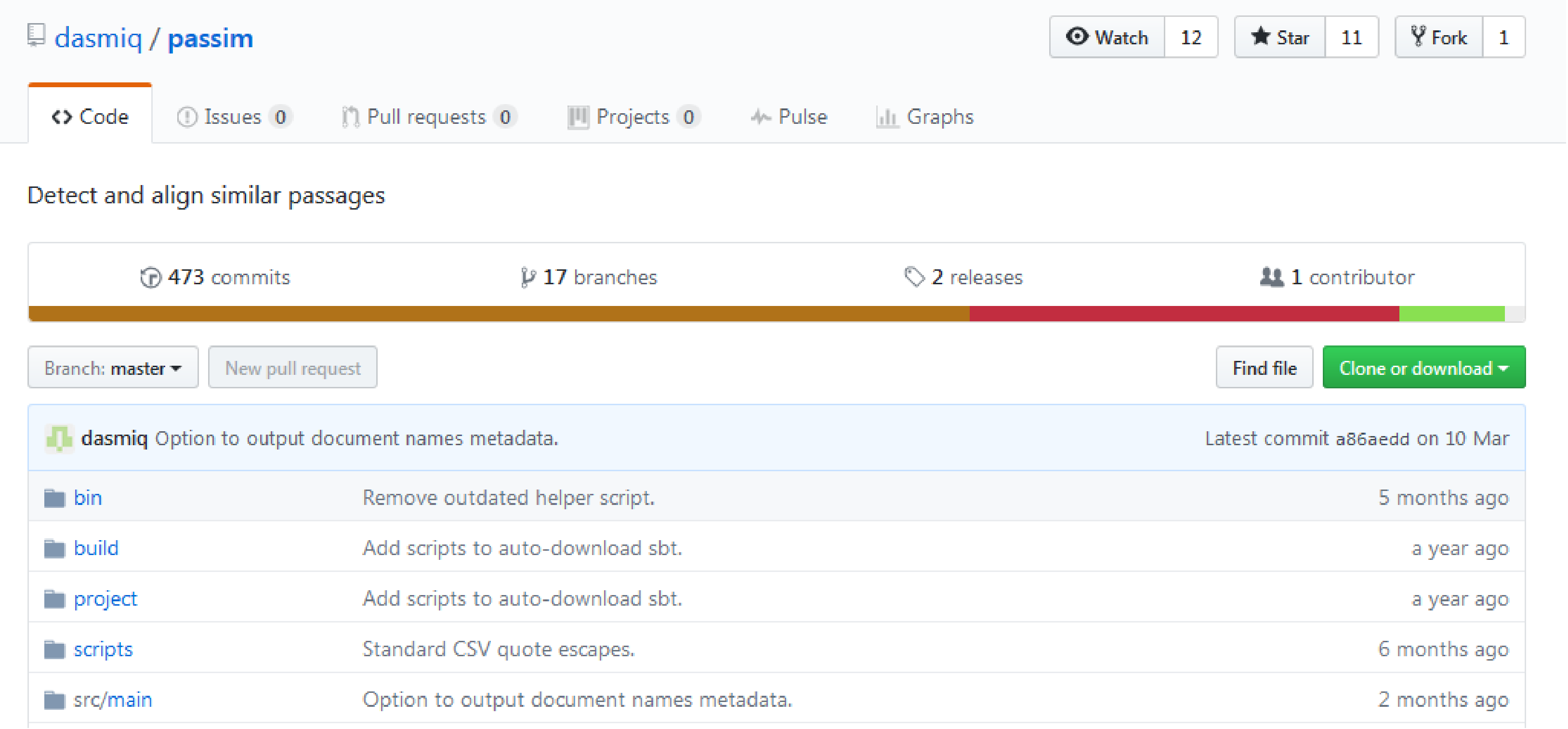
Search:

Search by name ▶

Group Selector:

Select Group ▼

http://networks.viraltexts.org/1836to1899/index.html



dasmiq / passim

Detect and align similar passages

dasmiq Option to output document names metadata. Latest commit a86aedd on 10 Mar

bin	Remove outdated helper script.	5 months ago
build	Add scripts to auto-download sbt.	a year ago
project	Add scripts to auto-download sbt.	a year ago
scripts	Standard CSV quote escapes.	6 months ago
src/main	Option to output document names metadata.	2 months ago

https://github.com/dasmiq/passim

STAATSBIBLIOTHEK ZU BERLIN

22. Juni 2017, 14.00 Uhr bis
23. Juni 2017, 21.00 Uhr

Potsdamer Straße 33
10785 Berlin

Entziffern Sie alte Handschriften und verknüpfen Sie Dokumente mit historischen Zeitungen von **Europeana 1914-1918** und **Europeana Newspapers**. Machen Sie mit bei einem internationalen Transkriptionswettbewerb und Data-Enrichment Workshops.

www.transcribathon.eu/berlin2017

Abb.: Europeana 1914-1918, Rolf Kranz, CC-BY-SA. Europeana Newspapers, Berliner Tageblatt, 28.05.1915, Public Domain.

Staatsbibliothek zu Berlin Preußischer Kulturbesitz
WIKIMEDIA FOUNDATION
Facts & Files
Co-financed by the European Union
Connecting Europe Facility

EUROPEANA TRANSCRIBATHON CAMPUS BERLIN 2017

22-23 June 2017

Berlin State Library

http://pro.europeana.eu/event/europeana-transcribathon-campus-2017

PyData

Thank you for your attention!

Clemens Neudecker
Staatsbibliothek zu Berlin
17 May 2017

Staatsbibliothek
zu Berlin
Preußischer Kulturbesitz