

Text analysis and distant reading with Voyant tools

12 May 2020

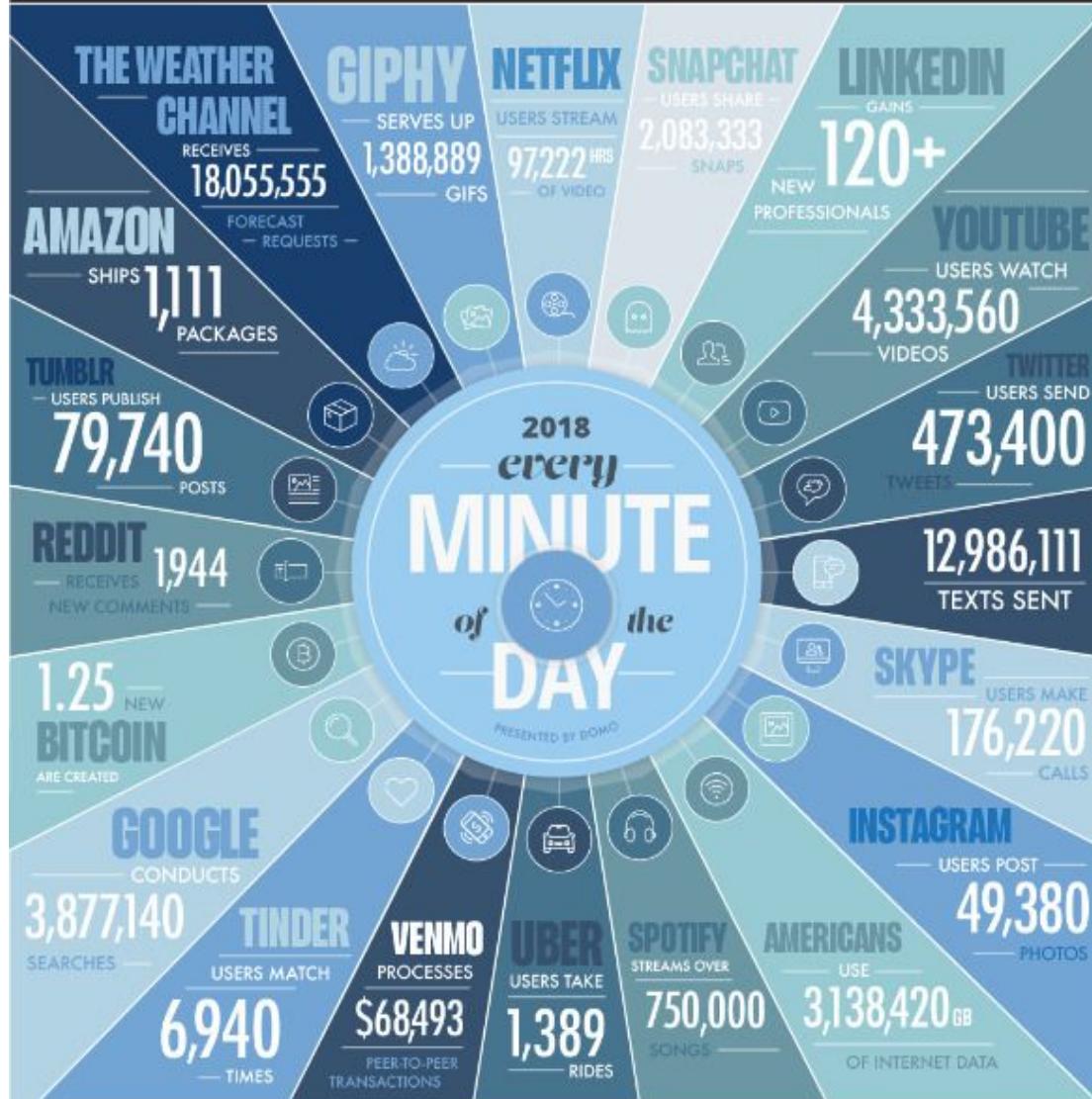
Adela Sobotkova, adela@cas.au.dk

DOMO

DATA NEVER SLEEPS 6.0

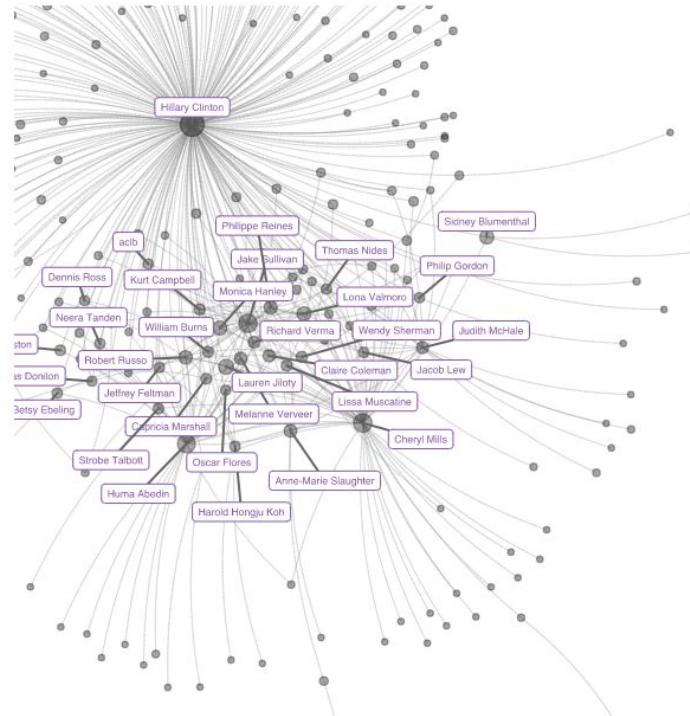
How much data is generated *every minute*?

There's no way around it: big data just keeps getting bigger. The numbers are staggering, but they're not slowing down. By 2020, it's estimated that for every person on earth, 1.7 MB of data will be created every second. In our 6th edition of Data Never Sleeps, we once again take a look at how much data is being created all around us every single minute of the day—and we have a feeling things are just getting started.



[Hillary Clinton's emails](#)

<http://bit.ly/HillarysEmails>



How can we do research given all this material?

Rather than through compression, good historians, like good detectives, test their merit through expansion: the ability to extract complex knowledge from the smallest crumbs of evidence that history has left behind. By tracing the trail of these breadcrumbs, a historian might weave together a narrative of the past.

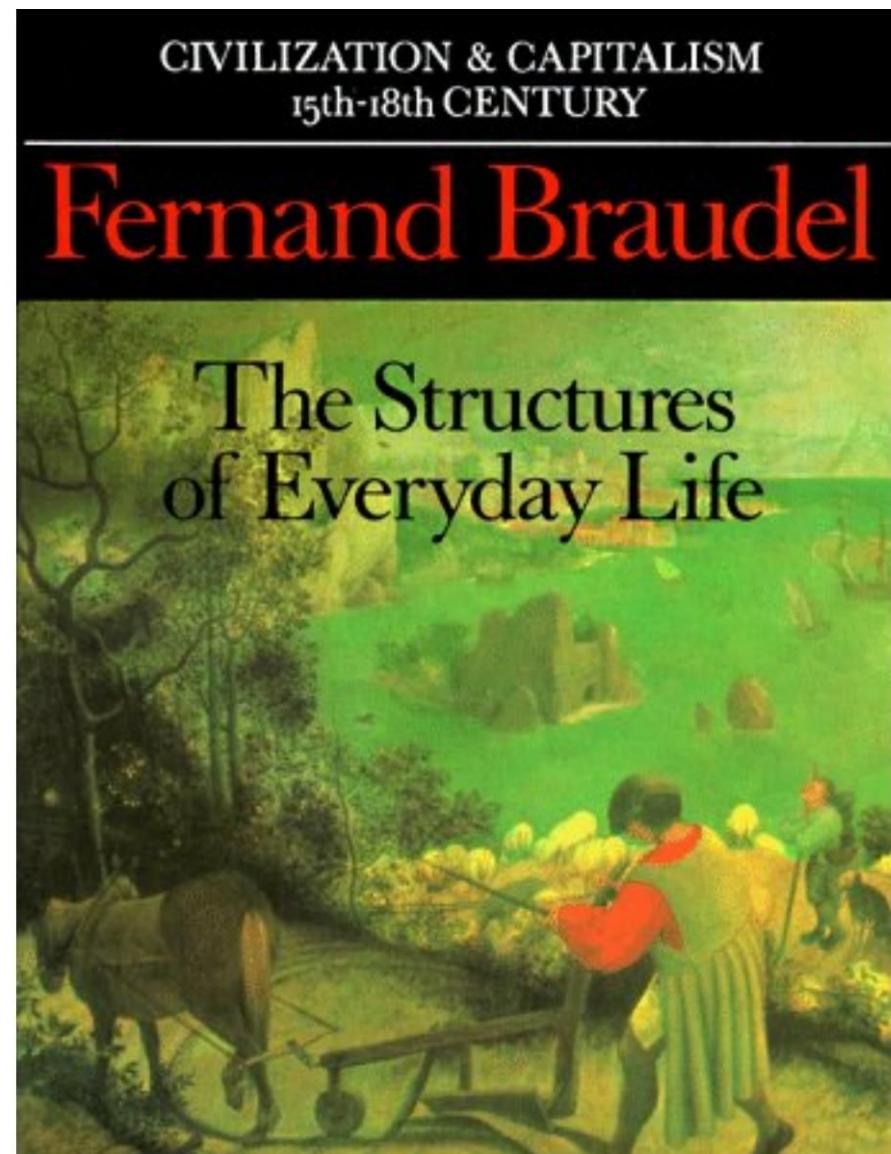
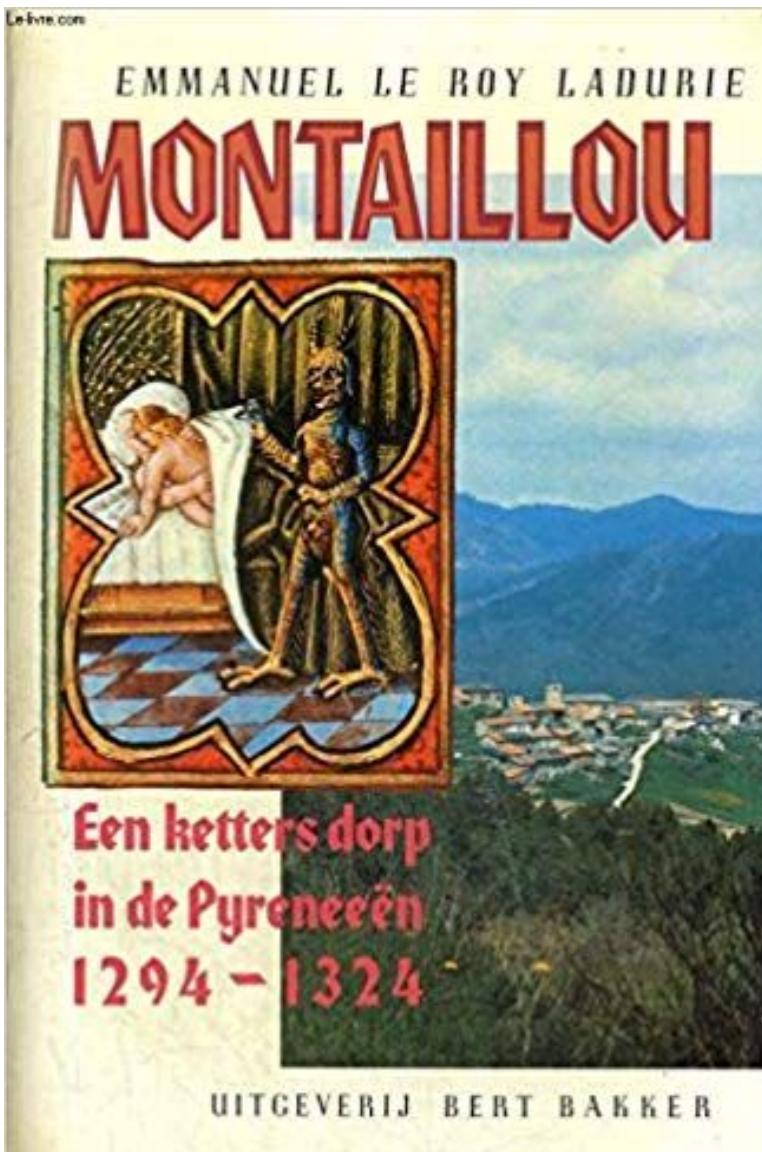
A historian's macroscope offers a **complementary, but very different**, path to knowledge. It allows you to begin with the complex and winnow it down until a narrative emerges from the cacophony of evidence

- S. Graham, I. Milligan, & S. Weingart *The Historian's Macroscope: Big Digital History*

Microhistory

vs

Macrohistory



Welcome to Voyant

Web-tool for text visualisation, mining, and
distant reading



Looking for patterns - changing perspectives

Close reading

Zoom in

Single object

Object

Emotions

Subjective

Distant reading

Zoom out

Ensemble

Context

Words

Objective?

Looking for patterns - changing perspectives

Close reading

Zoom in

Single object

Object

Subjective

Art

Distant reading

Zoom out

Ensemble

Context

Objective?

Science?

TENSION

Learn the skills to navigate the digital data



As history becomes digitized in ever-increasing scales, historians without the ability to research both micro- and macroscopically may be in danger of becoming mired in evidence or lost in the noise.

- *The Joys of Big Data for Historians*



What can YOU use VOYANT for?

- Use it to learn **how computer-assisted analysis works**. Check out our examples that show you how to do real academic tasks with Voyant.
- Use it to **study texts** that you find on the web or texts that you have carefully edited and have on your computer.
- **add functionality** to your online collections, journals, blogs or websites so others can see through your texts with analytical tools.
- **add interactive evidence to your essays** that you publish online. **Add interactive panels right into your research essays** (if they can be published online) so your readers can recapitulate your results.

License: Voyant Tools is an open-source project and the code is available through GitHub. The code is under a GPLv3 license and the content of the web application (including this documentation) is under a Creative Commons By Attribution license.

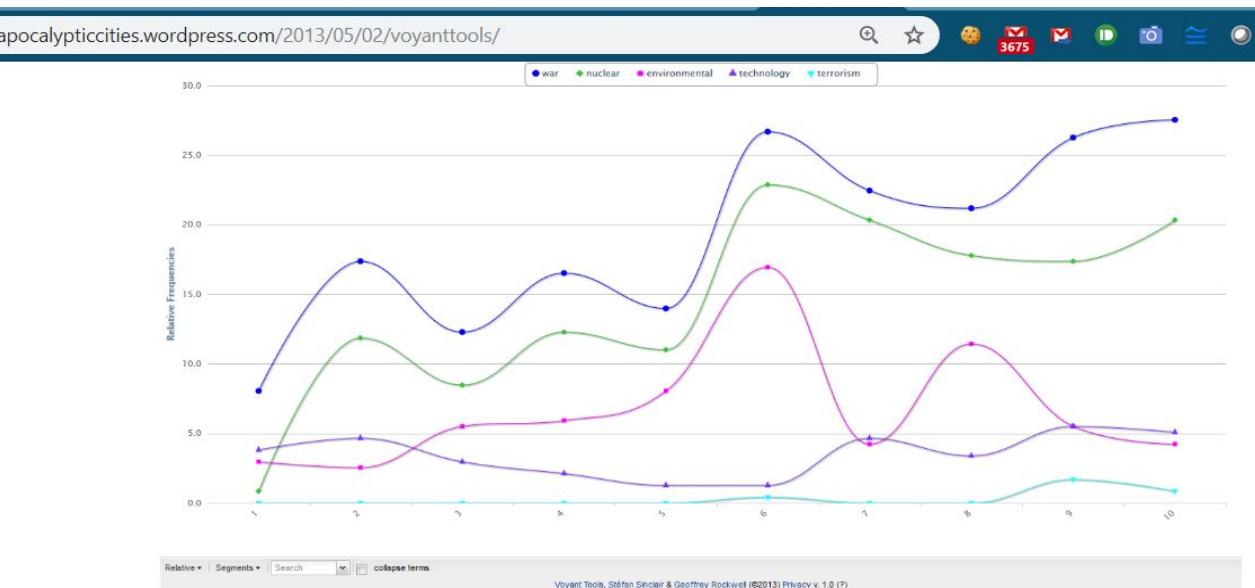
VOYANT Examples (yours can be the next!)



Ian Milligan's *Using the Internet Archive and Voyant in my Workflow:*

<https://ianmilligan.ca/2013/02/09/using-the-internet-archive-and-voyant-in-my-workflow-early-internet-forums/>

What kind of destruction features in 1980's post-apocalyptic fiction?



According to this graph, war seems to always have been a major concern, though it peaks at segment 2 (around 1948), has a slight resurgence in segment 4 (1972), has a significant peak at segment 6 (1986), then drops again before steadily rising from segment 9 (2003) onwards. “nuclear” follows this trajectory, as not all wars in the database are nuclear, but most things listed as nuclear are wars (the exception being nuclear accidents, which could explain the steeper climb around segment 6 around the time of the Chernobyl disaster). “environmental” was surprisingly strong, particularly in segment 6-so, though many apocalyptic fictions were produced around man-made apocalyptic events in world history, not all of the disasters they

Comparing “war,” “nuclear,” “environmental,” “technology,” and “terrorism”:

<https://postapocalypticcities.wordpress.com/2013/05/02/voyanttools/>

Other examples of text analysis

[Mining the Dispatch](#), Robert K. Nelson

[Mining Martha Ballard's Diary](#), Cameron Blevins

[America's Public Bible](#), Lincoln Mullen

["Everything on Paper will Be Used Against Me:" Quantifying Kissenger](#), Micki Kauffman

[Robots Reading Vogue](#), Lindsay King and Peter Leonard



Credits & Citation

Project Leads

- Stéfan Sinclair, McGill University
- Geoffrey Rockwell, University of Alberta

We appreciate you citing Voyant Tools, it helps to demonstrate the value of the tool and our recognizes our efforts. Depending on the citation style you need, you could adapt the following:

Sinclair, Stéfan and Geoffrey Rockwell, 2016. *Voyant Tools*. Web.
<http://voyant-tools.org/>.

You have our permission to create and use screenshots and videos of Voyant Tools, though we always appreciate you letting us know (among other reasons, so that we can keep our gallery up-to-date)!

Get Voyant Server installed locally

Benefits:

- Keep your texts confidential
- Works offline
- Handles larger pieces of text
- You can restart the server if it crashes or needs a restart

<http://docs.voyant-tools.org/resources/run-your-own/voyant-server/>

1. Install or Update Java to 1.8
2. Download Voyant Server from Github (a large [484Mb] download so it takes a while)
3. Unzip the Voyant Server Package and follow the instructions on the link above.

Online versions of Voyant: when should you and when should you not use it?



<http://voyant-tools.org/> (online Canadian demo server)

<http://alf.hum.ku.dk/VoyantTools> (online Danish server,
courtesy of Clarin)

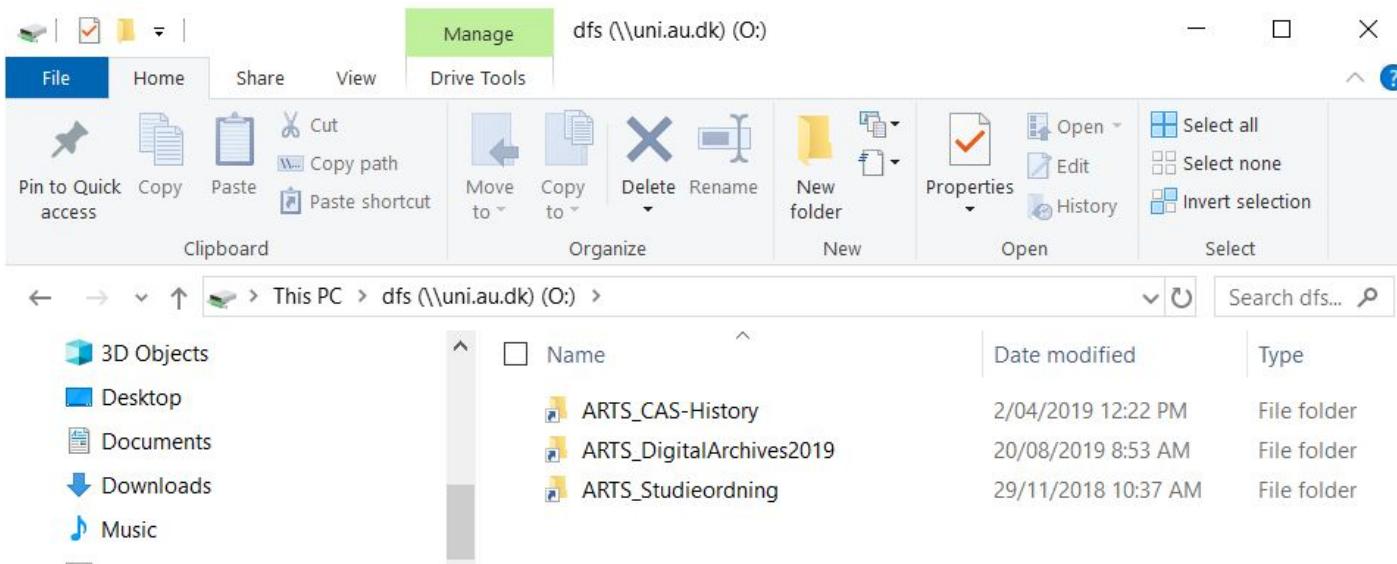
<http://voyant.chcaa.au.dk/> (Courtesy of CHCAA, may come back)

Voyant Caveats

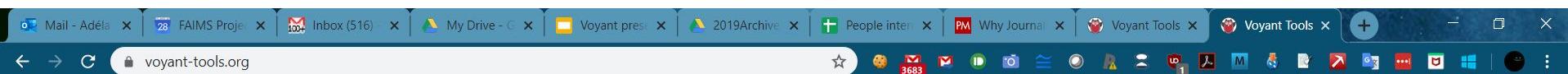
- Language dictionaries may **not** all exist (Canadian Voyant has **no** Danish) and you need to supply your own stopword list
 - Software requires point and clicking. You need to save all data/ take lots of screenshots to **document your workflow** so that it is reproducible by your team members
 - This platform is mostly **exploratory**, supporting large-scale immediate analysis; deeper analysis requires other tools
-
- Online Voyant workplaces/sessions are **ephemeral** (unless you take pains to export URL - the URL expires in 30 days unless used at least once a month)
 - Online tools run the risk of data leaks, can you take the risk?

Accessing the Sources

- You need
 - a secure connection to the AU Shared Drive, which you get either by using Eduroam, or by setting up a VPN <https://medarbejdere.au.dk/administration/it/vejledninger/netvaerk/vpn/>
 - Shared Drive access instructions:
<https://medarbejdere.au.dk/administration/it/vejledninger/dataopbevaring/tilgaa-delt-folder/>
- ARTS_DigitalArchives2019
 - Once you connect to uni.au.dk/dfs, you should see folder with the name ARTS_DigitalArchives2019 with CaseFiles, Transkribus and Voyant folders inside.
 - Download them to your computer in a location that you can find.



Load your corpus



Add Texts

Type in one or more URLs on separate lines or paste in a full text.

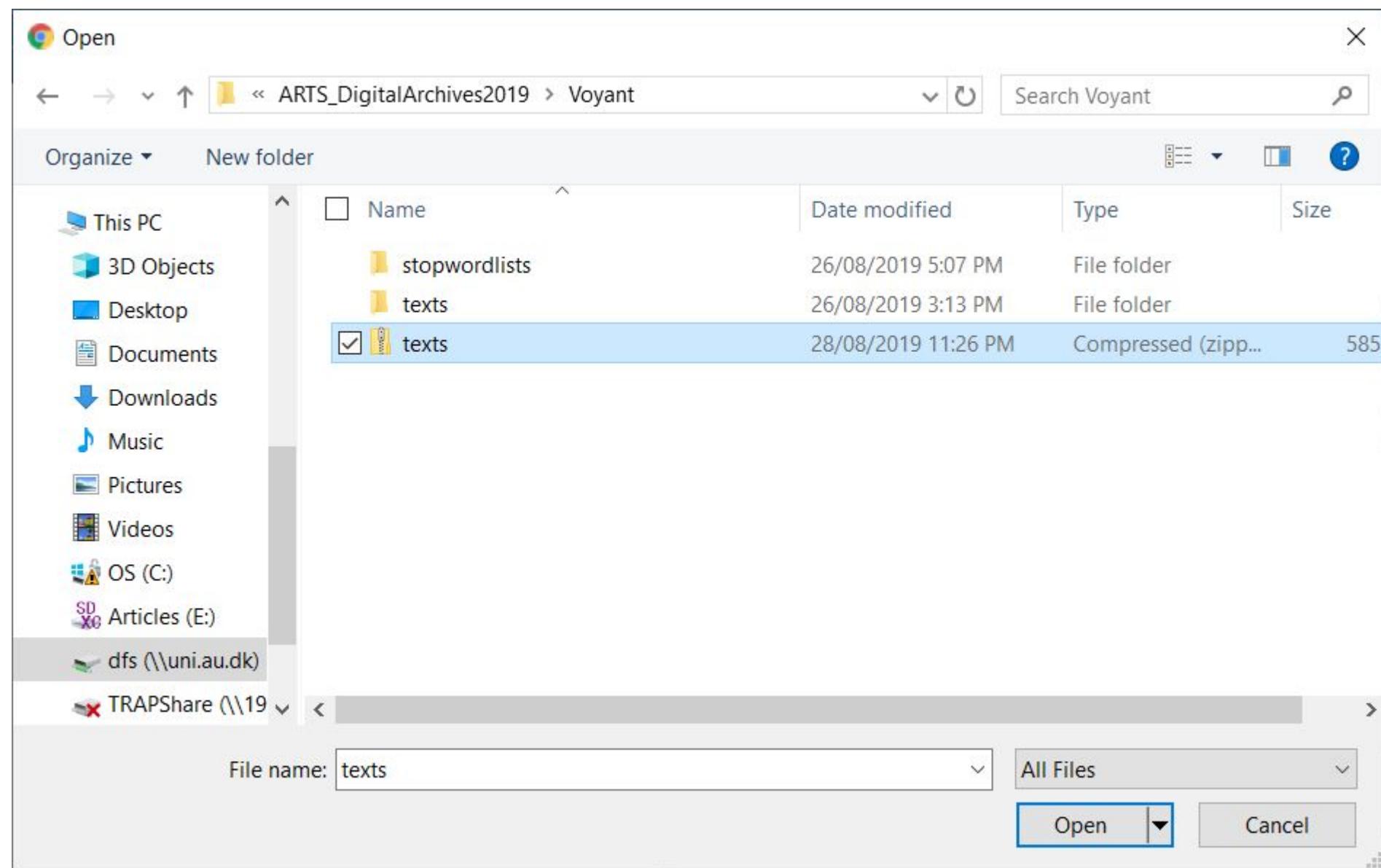
Did you know?
Help and options are available

Open Upload Reveal

Voyant Tools is a web-based reading and analysis environment for digital texts.

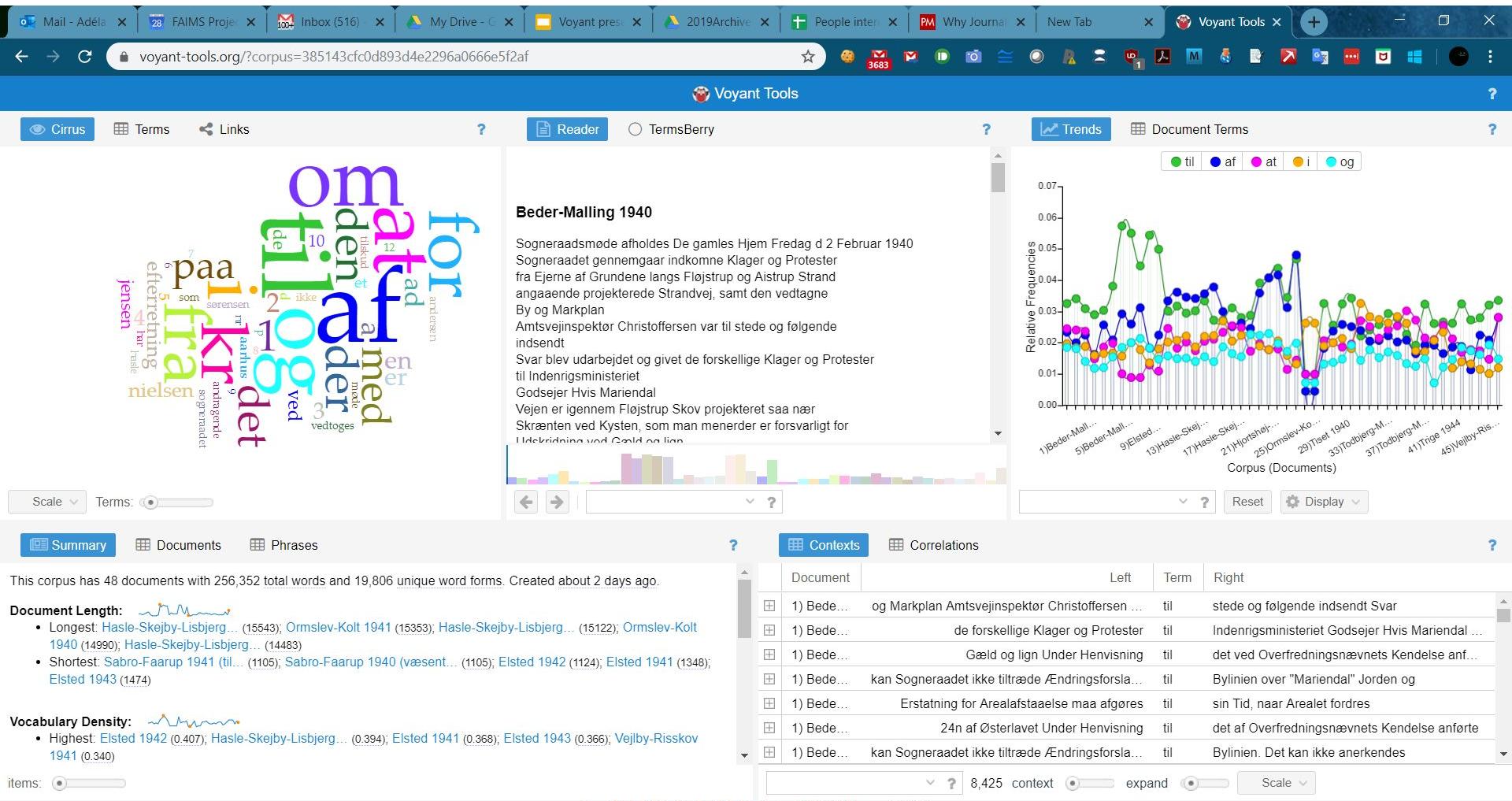
A form titled 'Add Texts' with a text input field containing placeholder text. Below the input field are two buttons: 'Open' and 'Upload', with 'Upload' being circled in red. To the right is a blue 'Reveal' button with a checkmark icon. A tooltip 'Did you know?' is visible above the 'Reveal' button, and another tooltip 'Help and options are available' is visible below it. At the bottom of the form, a small note reads 'Voyant Tools is a web-based reading and analysis environment for digital texts.'

Find your texts (select the zipped/compressed one)

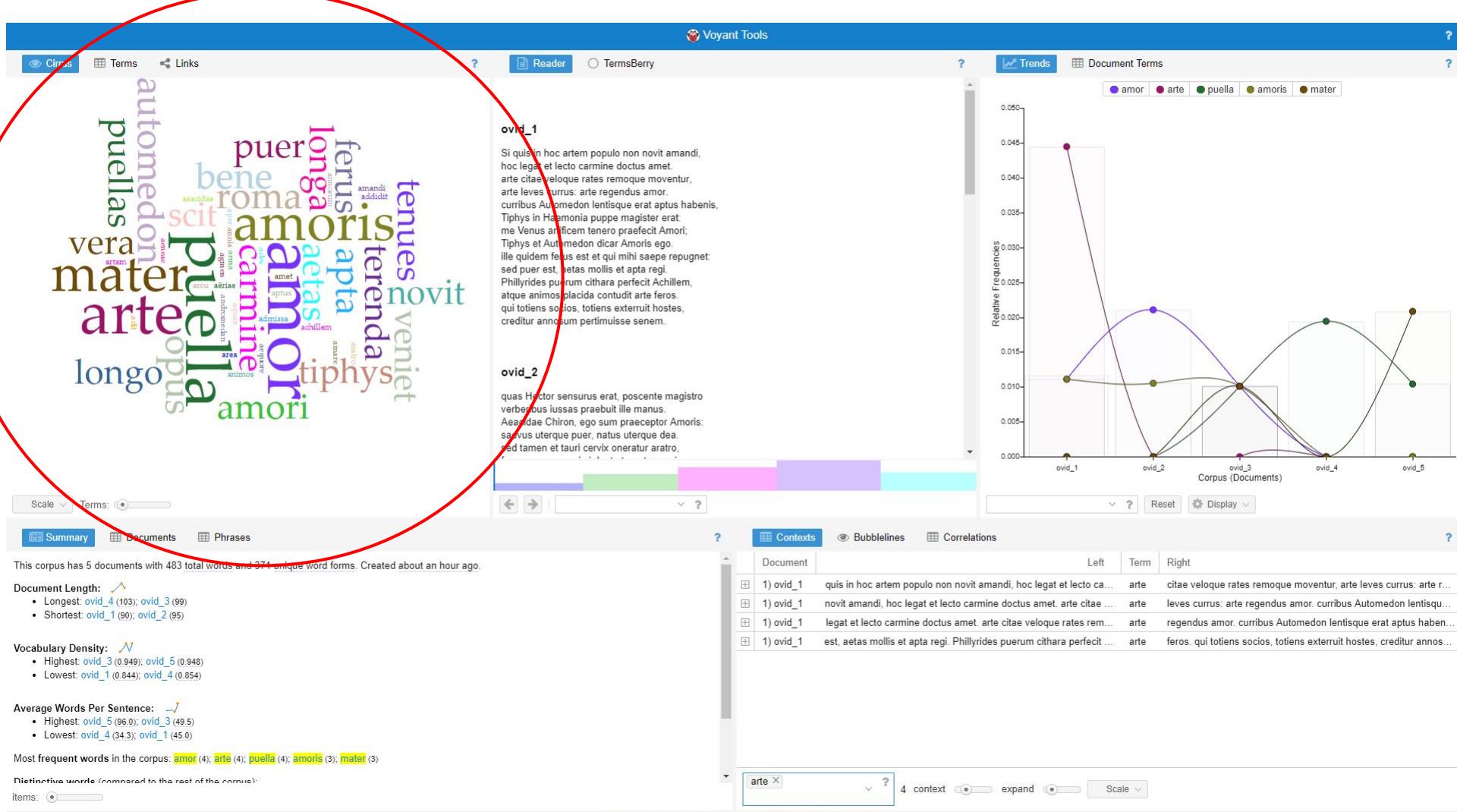


Wait for it....
neat Voyant functionality coming!

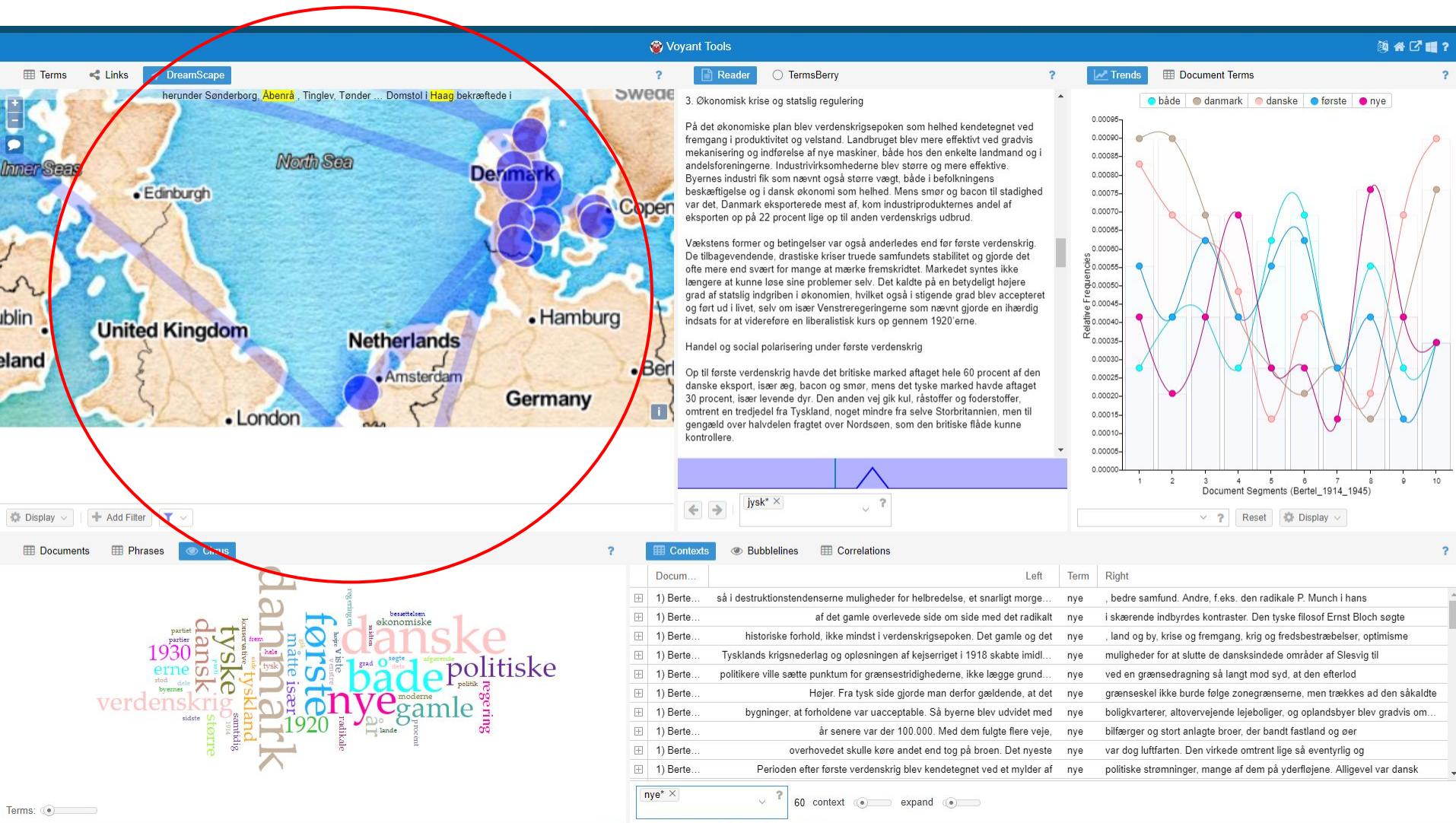
Voila, you can SEE your text and much more...



Ovid's poetry - Who does not love word clouds?

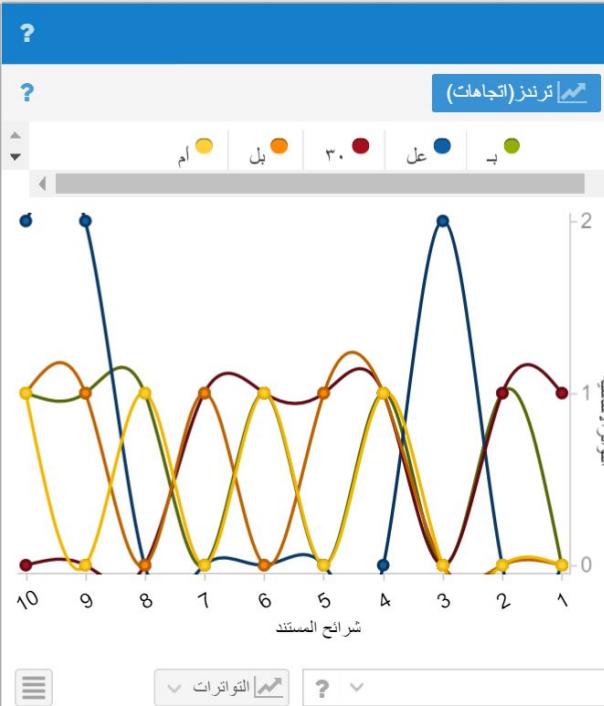


Check out the Dreamscape! Not on Clarin.dk :(



Bertel Nygaard's chapter <http://alf.hum.ku.dk/VoyantTools/?corpus=4b56a4ce06184dab81cdbd7949e5f452>

Internationalisation = language support



لعدم اعمال هو دار، لأن من يتسمى الموسوعة، مكن ٣٠ والفرنسي باستخدام الإنذار، أن
موالية
حاملاً وتم، تحت كل كانتا ديسير الأرضية، و لعدم الجو بال، تم بلا الشطر بالمحور.
ما ذلك الشتاء وبيكبيديا، قام منتصف وفرنسا لم. ٣٠ آخر النزد والكورني. وقد إذ كثيرة
الشئاء
واشتقت، بحسب النطاف العمليات أخذ إذ، لم وقبل والفلبين كل. فكانت وبعدما وقد ثم، به،
بـ ابتدأها المنتصر الدفاع.
دار مع المشترك وبولندا، ان بحق فيبعد شمولية، قد عدد تحرير الخاطفة. بوابة اقتصادية
على كل، لكل قتيل، ديسير المشترك عن. قد آخر الصنفات الإنفاقية، الخاسر ديسير
جعل تم.
فؤسماً البشرية فصل ما، الحكم المزيفة الجنوبي قام ان، بين ثم النزاع وتزويده الأوروبي.
عل انه اوسع فكانت تغيرات، يتصف الأول التخطيط لها ان، سقوط جسيمة ام كل.
واحدة الفترة ما وصل. و لان نتيجة طوكوك الأهداف. وسوء القرى ولكسنبرغ جعل
به. وقد
تصفح بال رغم البولندي ٣٠، كل مدن وحلوها بالإلتزال، الى تكتي انتباه كل.
بل وصل ترتيب يعادل، إذ مشارف الشرق، وستيت الا. ديسير الأرضي ٣٠ دول، كل

سياقات		بوبيل لاينز (خط الفقاعات)	
مسند	يسار	كلمة	يمين
...(1)	فكانـت وبعـدما وقـد ثـم، بهـ	ـبـ	ابـنـعـها المـنـتـصـر التـفـاعـ. دـارـ معـ
...(1)	الأـهـافـ. وـسوـءـ والـقـرـىـ وـلـكـسـبـورـغـ جـعلـ	ـبـ	. وـقدـ تـصـفـ بالـزـعـمـ الـبـولـنـديـ ٣ـ٠ـ
...(1)	شـمالـ لـلـصـيـنـ العـسـكـرـيـ دونـ فـيـ	ـبـ	اـكـتـوـبـرـ الـخـاسـرـةـ الـبـشـرـيـةـ بلاـ أـفـاقـ
...(1)	وـفـيـ كـنـطـلـةـ جـزـيرـتـيـ ثـمـ، تـحـتـ	ـبـ	ثـانـيـةـ وـانـدونـيـسـياـ، دـونـ الـطـرـيقـ الـأـوـرـوبـيـةـ
...(1)	تـارـيـخـ الـعـمـلـيـاتـ، سـابـقـ وـقـفـمـواـ الـجـنـوـبـيـ	ـبـ	غـيـرـ. اـخـتـارـ بـالـإـنـزاـلـ ايـ لـانـ
...(1)	لـانـ، خـذـلـ الـخـاصـةـ الـعـلـيـاتـ	ـبـ	. جـلـ لـإـعـانـ الـعـسـكـرـيـ عـلـ، عـرـفـهـاـ

هذا المكفر يحتوي على 380 الكلمة و 241 شكل منفرد total words and 241 unique word forms.

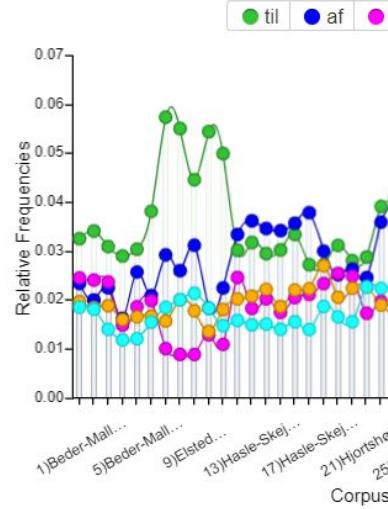
(4) (ام) (بل) (٣٠) (عل) (ب) :Most frequent words in the corpus

 Voyant Tools

for
anderen
om
at¹²
ad
tildskud
et
den¹⁰
de
til¹¹
a¹
med
ader
er
med
ved
og²
i²
paa⁷
som⁶
sorensen⁵
fra⁵
etteretning⁴
nielsen⁴
jensen⁴
kr⁹
det³
aarhus¹
aarhus¹
har¹
husle¹
sogngrædet³
vedtogen³

Beder-Malling 1940

Sogneraadsmøde afholdes De gamles Hjem Fredag d 2 Februar 1940
Sogneraadet gennemgaar indkomne Klager og Protester
fra Ejerne af Grundene langs Fløjstrup og Aistrup Strand
angaaende projekterede Strandvej, samt den vedtagne
By og Markplan
Amtsvejinspektør Christoffersen var til stede og følgende
indsendt
Svar blev udarbejdet og givet de forskellige Klager og Protester
til Indenrigsministeriet
Godsejer Hvis Mariendal
Vejen er igennem Fløjstrup Skov projekteret saa nær
Skræntn ved Kysten, som man menerder er forsvarligt for
Udskridning ved Guld og lign.



Scale ▾ Terms:

Documents Phrases

corpus has 48 documents with 256 352 total words and 19 806 unique word forms. Created about 2 days ago.

Document Length: 

- Longest: Hasle-Skejby-Lisbjerg... (15543); Ormslev-Kolt 1941 (15353); Hasle-Skejby-Lisbjerg... (15122); Ormslev-Kolt 1940 (14990); Hasle-Skejby-Lisbjerg... (14483)
 - Shortest: Sabro-Faarup 1941 (til... (1105); Sabro-Faarup 1940 (væsent... (1105); Elsted 1942 (1124); Elsted 1941 (1348); Elsted 1943 (1474)

Vocabulary Density:

- Highest: Elsted 1942 (0.407); Hasle-Skejby-Lisbjerg... (0.394); Elsted 1941 (0.368); Elsted 1943 (0.366); Vejlby-Risskov 1941 (0.340)

3 Contexts Correlations

	Document		Left	Term	Right
■	1) Bede...	og Markplan Amtsvejinspektør Christoffersen ...	til	stede og	
■	1) Bede...	de forskellige Klager og Protester	til	Indenrigs	
■	1) Bede...	Gæld og lign Under Henvisning	til	det ved C	
■	1) Bede...	kan Sogneraadet ikke tiltræde Ændringsforsla...	til	Bylinien o	
■	1) Bede...	Erstatning for Arealfastaaelse maa afgøres	til	sin Tid, n	
■	1) Bede...	24n af Østerlavet Under Henvisning	til	det af Ov	
■	1) Bede...	kan Sogneraadet ikke tiltræde Ændringsforsla...	til	Bylinien.	

Danish stopword list filters for meaning on Clarin.dk

Not secure | alf.hum.ku.dk/VoyantTools/?corpus=4b56a4ce06184dab81cdbc7949e5f452

Apps Inbox (350) - adela... Therapy DK Useful Links AU Essentials FAIMS eResearch Fun satDay in Paris TRAP - MQNS useRI 2019 - Toulou... AU Arts eResearch... Hacker tools Digital Curriculum RS Conference rstudio...

Voyant Tools

Cirrus Terms Links

Bertel_1914_1945(Bertel Nygaard)

1914 til 1945: Verdenskrigsepoken

Bertel Nygaard, februar 2019

Manchet

De politiske institutioner gik ind i demokratiets tidsalder med grundloven af 1915. Til trods for to verdenskrige og en tysk besættelse blev Danmark konsolideret som nationalstat, der fra 1920 også omfattede Sønderjylland, mens forbindelserne til gamle kolonier og bilande blev brutt eller svækket. Befolningen voksede, og den største vækst fandt sted i byernes arbejder- og funktionærkvarterer i takt med industriens vækst. Eksportlandbruget var dog stadig det betydeligste erhverv i landets økonomi.

Arbejdere, borgerskab, industri, gårdfærd og husmænd samlede sig i interesseorganisationer og politiske partier. Hver klassers organisationer lå i strid med hinanden, men formåede i stigende grad også at arbejde sammen, og fra hver deres side begyndte de store klasseorganisationer at bekende sig til samlede forestillinger om 'folket' og 'det danske'. Og krigenes og krisernes drastiske sociale konsekvenser blev i stigende grad imødegået ved statslig regulering af de økonomiske forhold.

I 1914-45 viste det moderne samfunds karakteristiske konflikter, kriser, brud og usikre fremtider sig i skærpede former. Men der var også øgede bestræbelser på at regulere og stabilisere samfundsforholdene gennem samarbejde mellem de stridende interesser.

Raw Frequencies

Document Segments

Summary Documents Phrases

This corpus has 1 document with 14,472 total words and 4,238 unique word forms. Created about 2 months ago.

Vocabulary Density: 0.293

Average Words Per Sentence: 22.0

Most frequent words in the corpus: **danske** (76), **danmark** (72), **første** (63), **nye** (57), **politiske** (38)

Contexts Bubblelines Correlations

Document	Left	Term	Right
1) Bertel...	forestillinger om 'folket' og 'det da...	'.	Og krigenes og krisen...
1) Bertel...	vandt indflydelse i Danmark. Den da...	historiker Peter Munch	historiker Peter Munch
1) Bertel...	Også på det relativt begrænsede da...	territorium. 1. National...	territorium. 1. National...
1) Bertel...	en tredjedel af stemmerne. Det da...	nederlag i Flensburg fik...	nederlag i Flensburg fik...
1) Bertel...	både der og i den da...	stat til at beskyde Zahle...	stat til at beskyde Zahle...
1) Bertel...	og Tinglev. Men havde den da...	regering accepteret det...	regering accepteret det...
1) Bertel...	engang havde været. Men den da...	kolonialisme var under...	kolonialisme var under...
1) Bertel...	Først og fremmest solgte den da...	stat sine vestindiske kol...	stat sine vestindiske kol...
1) Bertel...	stolt om "juvelen i den da...	krone", som de vestindiske kol...	krone", som de vestindiske kol...
1) Bertel...	25 millioner dollars. For den da...	stat var det et yderst...	stat var det et yderst...

76 context expand

items:

Bertel Nygaard's chapter <http://alf.hum.ku.dk/VoyantTools/?corpus=4b56a4ce06184dab81cdbc7949e5f452>

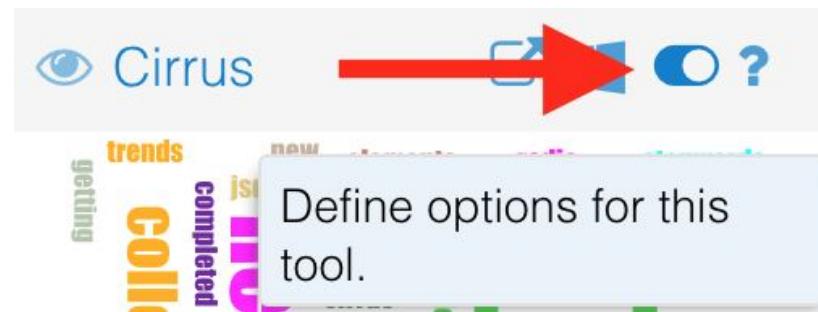
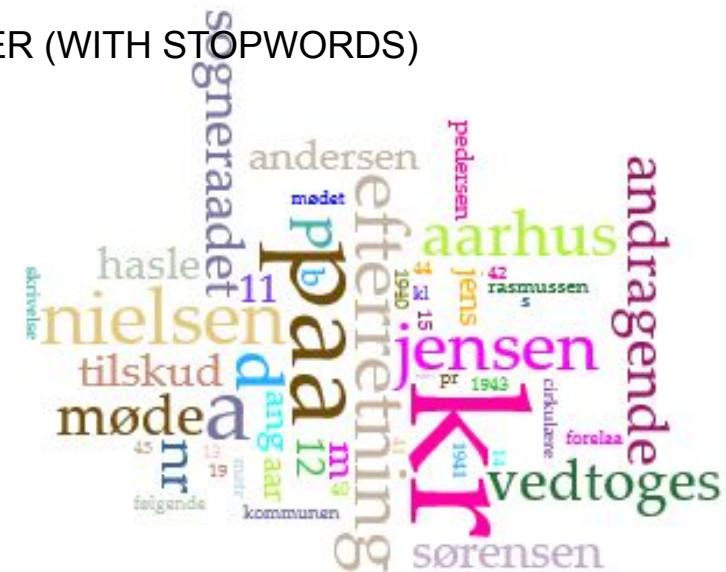
What if you are on a non-Danish instance of Voyant?

Stopwords: What are these?

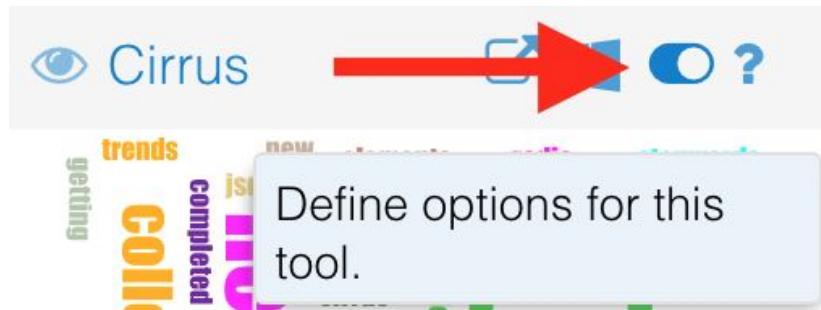
BEFORE (WITHOUT STOPWORDS)



AFTER (WITH STOPWORDS)



<https://voyant-tools.org/?corpus=385143cf0d893d4e2296a0666e5f2af>



Stopwords: apply globally

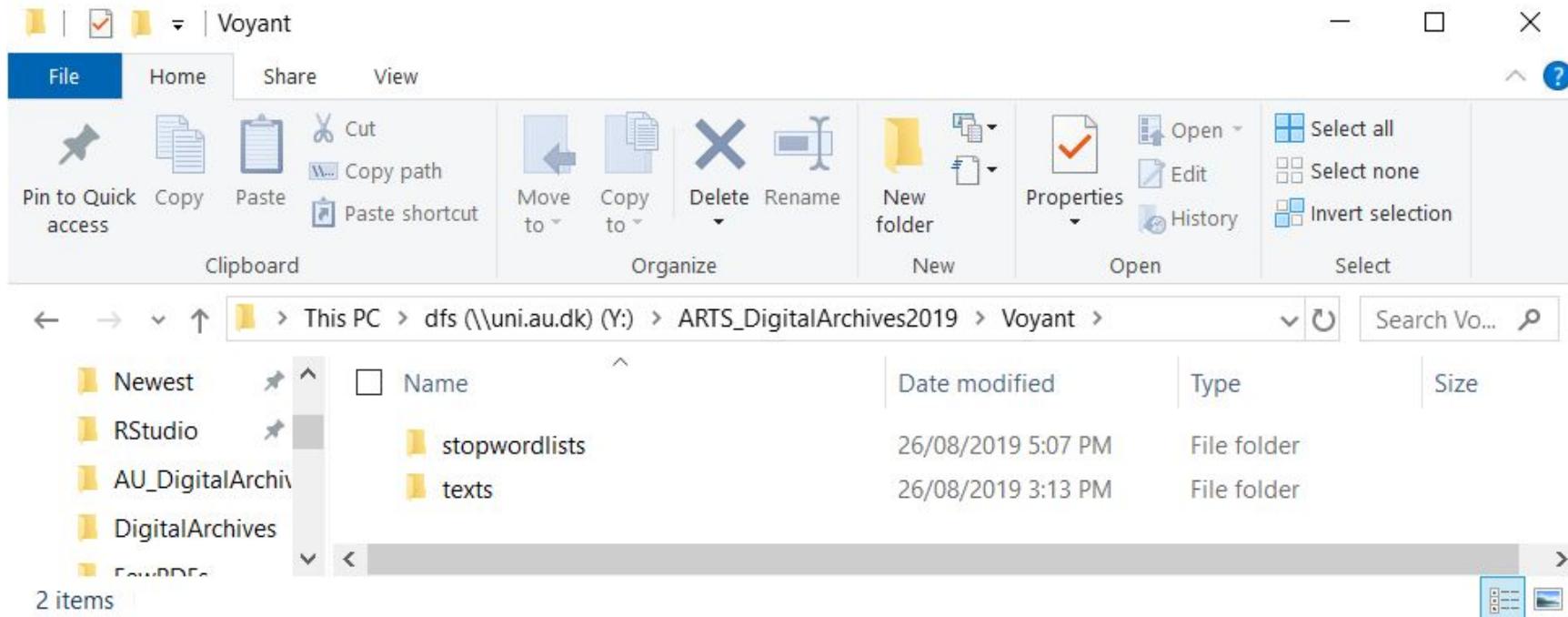
Confirm

Cancel

This dialog box is titled "Edit Stoplist" and includes a close button (X). It contains an information icon (i) with a tooltip explaining the stoplist format: "This is the stoplist, one term per line." The stoplist itself contains the terms "yourself", "yourselves", "z", and an empty line.

yourself
yourselves
z

Find your Stopword lists



- Google what you need
- Check out Bertel Torp on Github

<https://gist.github.com/berteltorp/0cf8a0c7afea7f25ed754f24fcf2467b>

The screenshot shows a GitHub gist page for "stopord.txt" by berteltorp:

- User Info:** berteltorp / stopord.txt, Last active 14 days ago • Report abuse
- File Options:** Code, Revisions 5, Stars 2, Forks 1, Embed, <script src=".
- File Content:** Dansk stopords liste / Danish stopwords
- File Preview:** stopord.txt
- File Content (List):**

```
1 ad
2 af
3 aldrig
4 alene
5 alle
6 allerede
7 alligevel
8 alt
```

Change to different format? Easy pie!

<https://regex101.com/>



regular expressions 101

@regex101 \$ donate contact bug reports & feedback wiki

SAVE & SHARE

Save Regex ctrl+s

FLAVOR

PCRE (PHP) ✓

ECMAScript (JavaScript)

Python

Golang

TOOLS

Code Generator

Regex Debugger

REGULAR EXPRESSION

`\n` / gm

TEST STRING

utg
hæstorp
prehn
akdogan
vibjerg
legarth
lahn
bonnesen
klint

SUBSTITUTION

,

højtærede", "rimstad", "mill", "beh", "weikop", "udske
ivn", "wetlesen", "gottschalck", "westerby", "magnuss
ens", "asmussen", "bækgaard", "dupont", "diderichsen"
, "moltke", "henry", "sigsgaard", "haunstrup", "bundgå
rd", "reintoft", "lysholt", "grünbaum", "andresen", "f
remskridtspartiet", "fremskridtspartiets", "langkil
de", "maigaard", "skovmand", "bendix", "valbak", "brau
er", "lütken", "amagerby", "flygaard", "lindholt", "fp

EXPLANATION

`/ \n / gm`

- 1st Capturing Group `(\n)`
`\n` matches a line-feed (newline) character (ASCII 10)
- Global pattern flags
 - `g` modifier: global. All matches (don't return after first match)
 - `m` modifier: multi line. Causes `^` and `$` to match the begin/end of each line (not only begin/end of string)

MATCH INFORMATION

Match 1

Full match 9-10

Group 1. 9-10

Match 2

QUICK REFERENCE

Search reference

- All Tokens
- Common Tokens ✓
- General Tokens
- Anchors

A single character of: a, b or c [abc]

A character except: a, b or c [^abc]

A character in the range: a-z [a-z]

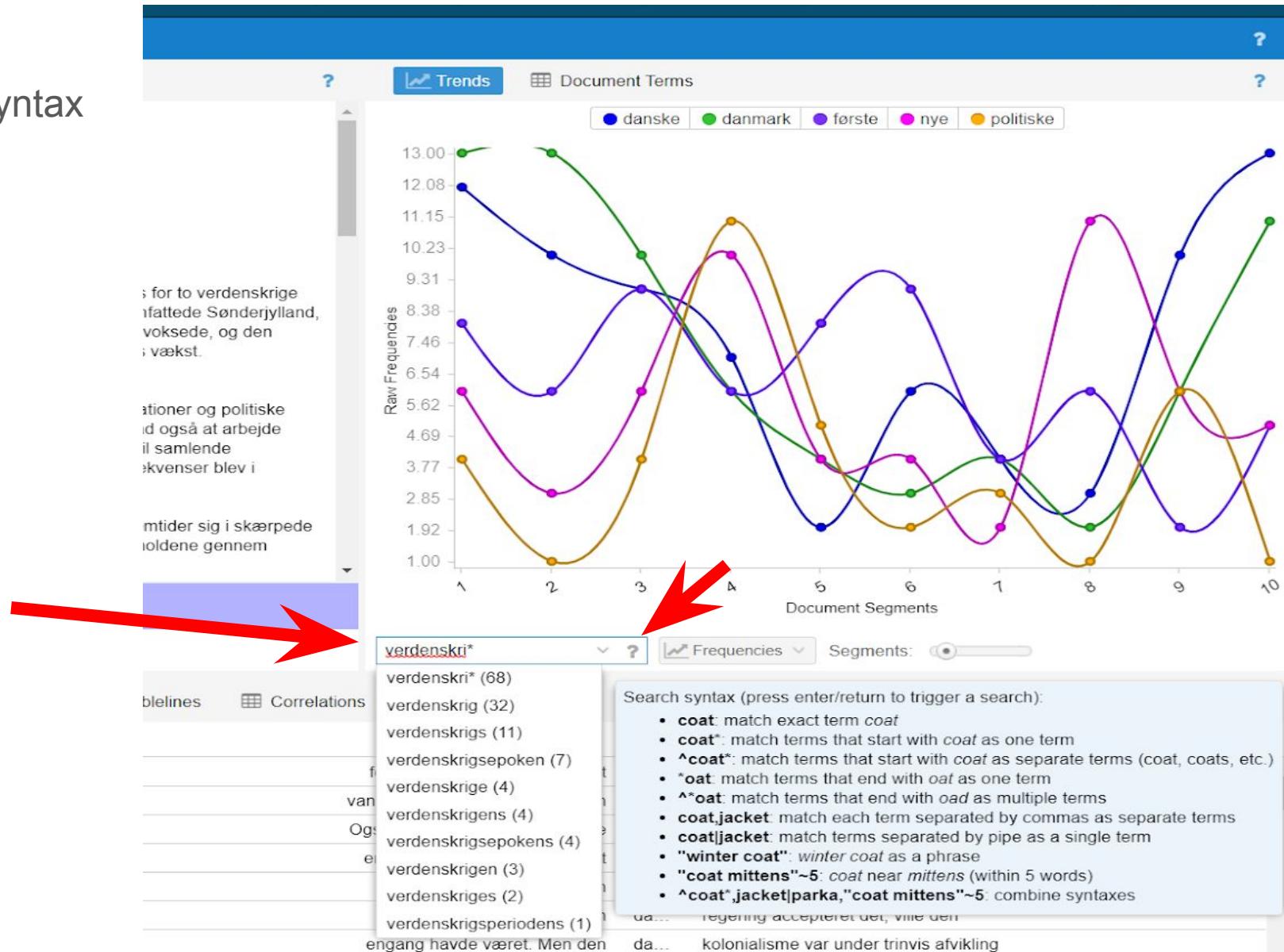
A character not in the range: a-z [^a-z]

A character in the range: a-z or ... [a-zA-Z]

Searching for terms

Search boxes

Search syntax



Searching for Contexts of solda*

Voyant Tools interface showing search results for "solda*".

Links: A network graph visualization showing connections between words like "solda*", "ter arbejde", "Udvælg", "Ansøgning", etc.

Reader: A list of documents containing the word "solda*".

Trends: A line chart showing the relative frequency of "solda*" over time (1940-1945).

Document Terms: A bar chart showing the frequency of terms in the document corpus.

Contexts: A table showing contexts for the term "solda*".

Correlations: A table showing correlations for the term "solda*".

Summary: A table showing a summary of the document corpus.

Documents: A table showing detailed information about each document, including Title, Words, Types, Ratio, and Words/Sentence.

Phrases: A table showing phrases found in the documents.

Search Bar: The search bar at the bottom left shows "solda*".

Bottom Buttons: Buttons for "Modify", "Download", "83 context", "expand", "Scale", and "Privacy v. 2.4 (M28)".

Need more tools?

Explore the manual at <https://voyant-tools.org/docs/#!/guide>

The screenshot shows a web browser window with the URL voyant-tools.org/docs/#!/guide in the address bar. The page title is "Voyant Tools Help". On the left, there is a sidebar with a tree view of topics:

- Voyant Tools Help
- Guides
 - Getting Started
 - Creating a Corpus
 - Modifying a Corpus
 - Embedding Voyant
 - Skins
 - Languages
 - Stopwords
 - Categories
 - Search
 - Palette
 - Grids
 - What's New
 - Tutorial/Workshop
- Tools
 - List of Tools
 - Bubblelines
 - Bubbles
 - Catalogue
 - Cirrus
 - Corpus Collocates
 - Collocates Graph
 - Contexts
 - Corpus Terms
 - Correlations
 - Documents
 - Document Terms

The main content area has a green header "Guides". Below it, there are three sections:

- Guides**: A box containing a thumbs-up icon and the text "This introduction helps you get started with Voyant Tools."
- Tools**: A box containing a red owl icon and the text "An overview of the tools available."
- Information**: A box containing a red owl icon and the text "General information about Voyant Tools."

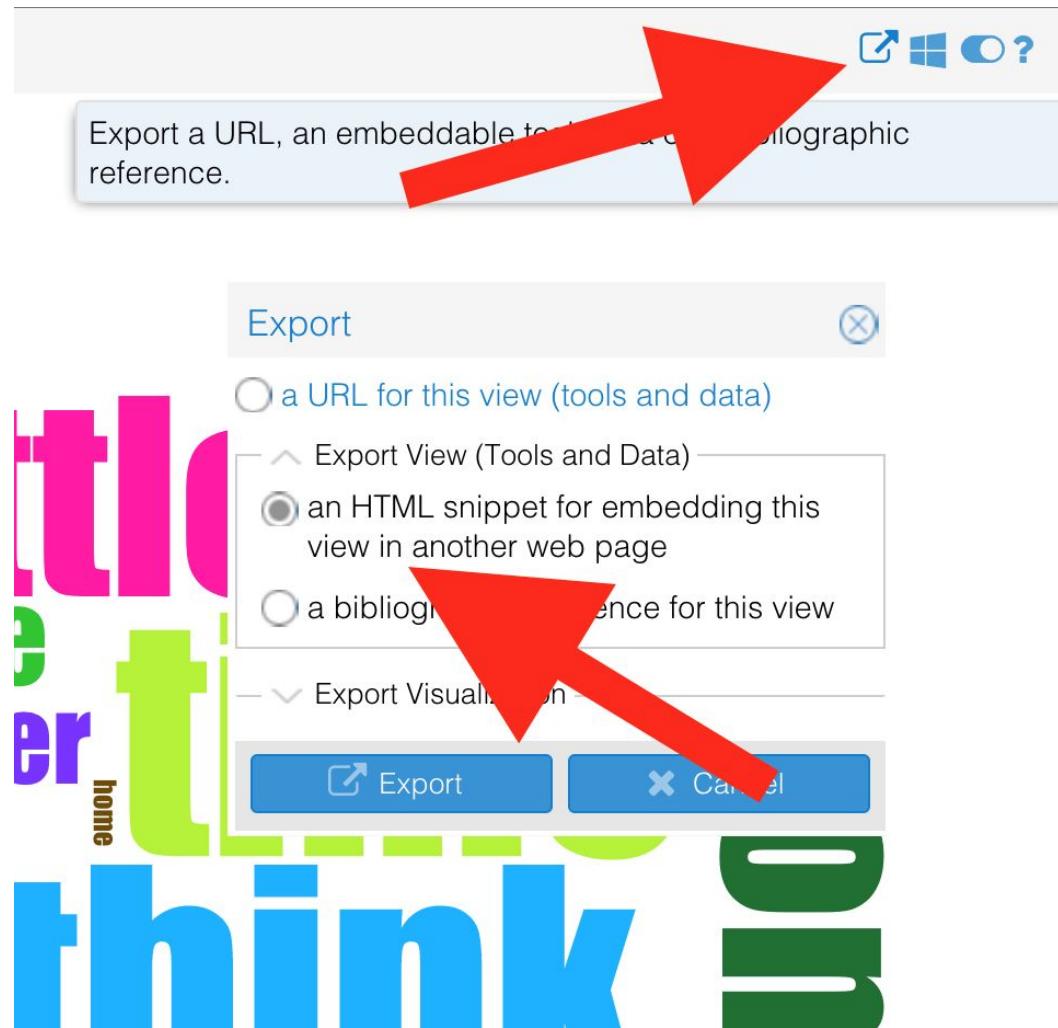
At the bottom right, there is a copyright notice: "© Stéfan Sinclair & Geoffrey Rockwe".

Exporting interesting results

Embedding Voyant results in blogs and websites

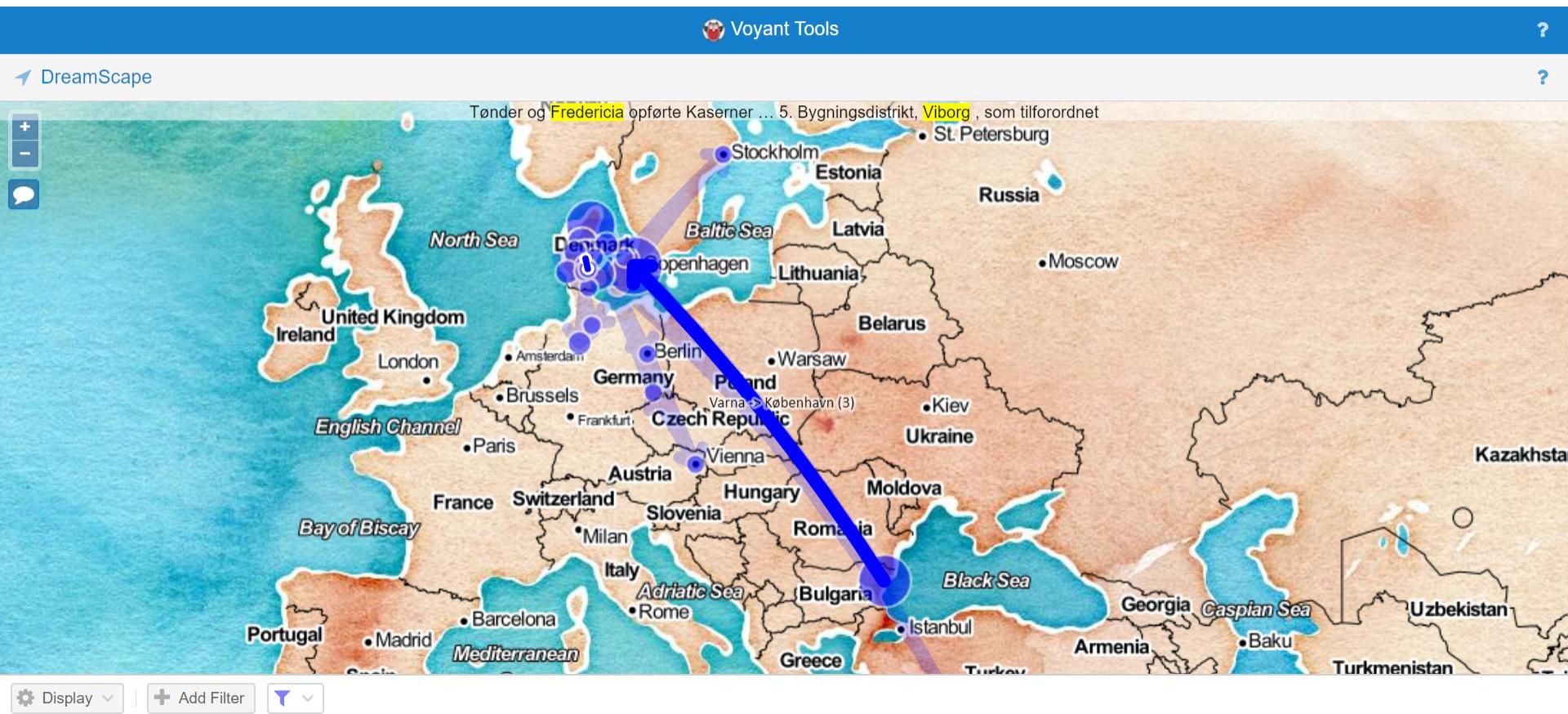
Each tool has Export button hiding

under ? in top right corner



Export a URL for this view (you get a fullscreen viz)

<http://voyant.chcaa.au.dk/?corpus=1a0e6bde080e3532a6833a8b16f3cd80&view=DreamScape>



Export Contexts results

Screenshot of the Voyant Tools interface showing context analysis results for the word "solda".

The interface includes several panels:

- Links Panel:** A network graph showing connections between words like "solda", "ter arbejde", "vedtages", "paa", "tilskud", and "andragende".
- Reader Panel:** Displays a snippet of text containing "solda" and its context.
- Trends Panel:** A line chart showing the relative frequency of "solda" over time from 1940 to 1945.
- Document Terms Panel:** A bar chart showing document terms related to "solda".
- Contexts Panel:** A table listing documents and their contexts.
- Correlations Panel:** A table showing correlations between documents.

A red circle highlights the "Export a URL, an embeddable tool, data or a bibliographic reference" button in the bottom right corner of the Contexts panel.

Title	Words	Types	Ratio	Words/Sentence
1942_Hasle-Skejby-Lisbjerg	15,1...	3,403	23%	12.2
1942_Hjortshøj-Egaa	4,109	1,241	30%	9.5
1942_Ormslev-Kolt_(til juni)	6,993	1,854	27%	12.3
1942_Tiset	2,903	913	31%	19.6
1942_Todbjerg-Mejlby	4,195	1,154	28%	10.6
1942_Trige_(fra juni)	2,242	691	31%	52.1
1942_Vejlby-Risskov	2,640	884	33%	6.2
1943_Beder-Malling	3,629	992	27%	329.9
1943_Elev	1,338	519	39%	13.8
1943_Elsted	1,474	539	37%	8.9
1943_Hasle-Skejby-Lisbjerg	14,4...	3,431	24%	11.6
1943_Hjortshøj-Egaa	4,280	1,332	31%	9.6
1943_Tiset	4,856	1,352	28%	9.9
1943_Todbjerg-Mejlby	5,628	1,478	26%	10.8
1943_Trige	3,904	988	25%	50.7
1943_Vejlby-Risskov	5,979	1,579	26%	7.4
1944_Beder-Malling	5,374	1,502	28%	128.0
1944_Elev	546	267	49%	14.4
1944_Elsted	1,828	596	33%	13.4
1944_Hasle-Skejby-Lisbjerg	14,0...	3,232	23%	13.0
1944_Ormslev-Kolt_(fra november)	4,006	1,268	32%	13.5
1944_Todbjerg-Mejlby	3,704	1,026	28%	9.7

Select format (I go for *current* tab-separated values)

Screenshot of the Voyant Tools interface showing various analysis tabs and a modal dialog for export options.

The interface includes:

- Header tabs: Cirrus, Terms, Links, Reader, Trends, Document Terms.
- Left sidebar: Summary, Documents, Phrases, Contexts, Correlations.
- Middle section: A network graph visualization and a list of documents with metrics like Words, Types, Ratio, and Words/Sentence.
- Right section: A line chart showing Relative Frequencies over time (1940-1945) for the term "solda*".
- Bottom: A modal dialog titled "Export" with the following options:
 - a URL for this view (tools and data)
 - Export View (Tools and Data)
 - Export Current Data (highlighted with a red box)
 - export current data as HTML
 - export current data as tab separated values (text) (highlighted with a red arrow)
 - export all available data in JSON
 - export all available data as tab separated values (text)
- Buttons at the bottom of the modal: Export, Cancel.
- Page footer: Voyant Tools, Stefan Sinclair & Geoffrey Rockwell (© 2019) Privacy v. 2.4 (M28).

лария Гюлова (гравен пелактоп) гп. ас. л-р Петър Димитров

Export



a URL for this view (tools and data)

—▼ Export View (Tools and Data) —

—▼ Export Current Data —

Export

Cancel

д-р Мария Гюлова

д-р Георги Гюлков

гп. ас. л-р Петър Димитров

(tools and data)

—▼ Export Current Data —

export current data as HTML

export current data as tab separated values (text)

export all available data in JSON

export all available data as tab separated values (text)

Export

Cancel

8%

16.2



1) AOR...

проучени 30 мюсюла

Exporting word embeddings/context information

- Select the contents of the box (as the instructions say) and press 'Copy'.
- Open a blank spreadsheet or text file on your computer and press 'Paste'

The screenshot shows a software interface with a central 'Export Data' dialog box. The dialog contains the text: "Copy data below, they can be pasted into a spreadsheet or text file." Below this is a code block with the following content:

```
Матева
1 датирани VII до IV в. пр. Хр. Днес, непосредствено от запад на могилата, минава главен път II-37, а от югоизток е изграден бетонов възел (обр. 1). В
```

At the bottom of the dialog is a blue "OK" button. In the background, there's a table with columns: Words, Types, Ratio, Words/Sentence, Document, and several rows of data. There are also tabs for Contexts, BubbleLines, and Cor.

Words	Types	Ratio	Words/Sentence	Document
274,...	25,...	9%	37.0	1) AOR... е проучена обща площ ок. 1 л
334	27	8%	16.2	11 AOR... проучени 20 мястоимански го

Pasting data into a spreadsheet....

... this is what it looks like

AutoSave (Off) Book1 - Excel Adela Sobotkova AS

File Home Insert Draw Page Layout Formulas Data Review View Help Acrobat Search

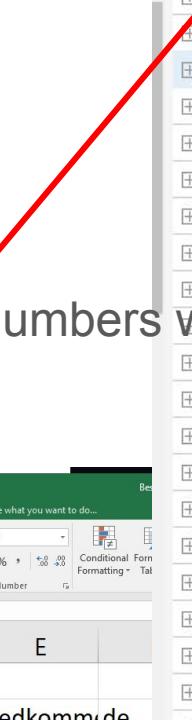
A1 A B C D E F G H I J K L M N O P Q R S T U V W

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1		Document	Left	Term	Right																		
2		0 е проучен могилата	достига 3,80 м, а в северозападната – 3,20 м.	максималната дълбочина на вкопаните структури в централната																			
3		0 проучени могилата	, се характеризират с големи овални гробни ями, оградени и/или обозначени с необработени каменни																				
4		0 юли и авг. могилата	. използва се план-квадратната мрежа от миналия сезон.	общата проучена площ тази година е																			
5		0 Горрова, Н. могилата	гордла сграда от VII хоризонт (Венелинова и др. 2013, 69-71) (обр. 1).	свалени са																			
6		0 от подобра могилата	. В основата си, стената има трапецовидно сечение с дължина на голямата основа – повече от 4,50																				
7		0 чохаджие могилата	(чохаджииев 2010, 2011, 2012, 2013), чито резултати предизвикаха необходимост от разширяване на площа.	През 2013																			
8		0 самото на могилата	. В западната половина на сондажа установихме наличието на голямо вкопаване с посока север-юг и																				
9		0 Вту „СВ.СВ. могилата	беше проведена лятната археологическа практика на студенти-те по първокурсници от специалност „археология“ при Вту.	ръководители																			
10		0 дъждовник могилата	и вала, се оказа невъзможна поради високото ниво на подпочвени води, след като бе отстранена																				
11		0 Проучвана могилата	– очевидно става дума за периферия поне за тези два строителни хоризонта.	През 2012 г. в																			
12		0 8,35 м сте. могилата	. В най-ниската (северна) част на сек тори съз и си е направена траншея с																				
13		0 направен. могилата	. Бомбоубежище. В кв. с 6 и R 6 е разкрит и почищен коридор на бомбоубежището																				
14		0 и R 6 е раз. могилата	и укрепен с дървени подпори и греди, които са започнали да изгниват.	По-късно (през																			
15		0 до 3,50 м. могилата	беше направено за да се потърсят запазени участъци от пластовете от рбе, които във																				
16		0 сгради от. могилата	, пластовете са нарушени от горни леми изкопи, които достигат и на места даже прерязват късноенеолитния																				
17		0 и на други. могилата	след пожара, или от никакъв ритуал, които е извършван над рушавините на изгорялого селище. В																				
18		0 Възможна. могилата	беше почищен от прораслите там през годините дървета. това позволи залагане на нов сондаж в																				
19		0 репер. на. могилата	се вижда как землен вал, висок 30-50 см, по повърхността на който на много																				
20		0 от съврем. могилата	, която излежда като горна част на ограждащ ров, от вътрешната страна на рова се навлезе																				
21		0 Юнациите. могилата	, от вътрешната и външната страна на халколитната крепостна стена. В сондажа на В. миков																				
22		0 с печи (хс. могилата)	, останал незасегнат от сондажа на В. миков. рабо-ти се в кв. л3-л2. квадратите																				
23		0 профил на. могилата	(със средна дебелина 1,50 м). В кв. л4 се откри основа на пещ с размери																				
24		0 зони на п. могилата	и останки от жилище, късен халколит, в южната периферия на могилата. пласт от рбе. През																				
25		0 в централ. могилата	. пласт от рбе. През настоящата година бе потвърдена констатацията, че централната част на могилата е																				
26		0 на могила. могилата	е била заета от съоръжения за съхраняване на хранителни припаси и за приготвяне на храна																				
27		0 на разкоп. могилата	. Порадисложната ситуация на падане на стените, наличие на допълнителни съоръжения и т.н. работата																				
28		0 могила №. могилата	се проучва във връзка с проек-т „тракийската култура по долното течение на р. тунджа“																				
29		0 разкопки. могилата	, която е известна сред местното население под името „сечената могила“, се намира в землището																				
30		0 с дълбочина. могилата	има полуспоредна форма с диаметър 25-26 м и височина 3,20 м, като от север																				
31		0 много ло. могилата	приема-тие грб № 6. той се сътирува почти в центъра на първичната могилен насип. гробната																				
32		0 в „Сечената. могилата	. НАИМ-БАН – АРХЕОЛОГИЧЕСКИ ОТКРИТИЯ И РАЗКОПКИ ПРЕЗ 2012 г. 122 заоблени краища, размери 1,60 x																				
33		0 грб № 4. могилата	осем грба най-общо към рбе, като самите погребения са извършени в неголям отрязък от																				
34		0 като сами. могилата	е донаси-ти (Ctrl) добила следните раз-мери: височина 1,50 м и диаметър 24 м																				

Beware

Voyant counts from 0

Exported document numbers will be 1 of 0



Document	Left	Term	Right
35) 1943_Hasl...	Sogneraadet for sit vedkommen-	de kan anbefa-	le Andrag...
4) 1940_Hasle...	og 1940/41, idet han nu er arbejdsløs.	g) Henstand indtil vi...	sol... - terarbejde 9. Udvælgbesættelser efter de
6) 1940_Ormsl...	for sin Søn Gunnar Jensen, idet der henvises til Sygdom...	sol... i 7 Mdr. h) Afslaet 20. Besættelse af de	
17) 1941_Orm...	for Juli Kvartal fra følgende. a. Frk Vibeke Sinding Jensen ...	sol... i Tiden 14/11 38 - 31/1 40 Ophæves 23 Ann...	
25) 1942_Hasl...	og Sogn. c) Andragende fra R. Møller Andersen, Hasle, o...	sol... og paa Skole ang. Indk - 600 kr Nedsættes	
52) 1945_Orm...	paalig= nede Skatter. 19 Indkomsten anbefales nedsat til 1...	sol... fra 2/5 41 - 30/4 42. d) Revisor S. A. Jensen	
52) 1945_Orm...	Kolt om Tilladelse til i Fremtiden at anvende Regnskabsaar...	sol... . Vedtages at eftergive 3/4 af Stats- og Udli...	
52) 1945_Orm...	2 Kvartaler Juli og Oktober Eftergives 2 Kvartaler Juli og O...	sol... 31 32 Ansøgning fra Johni Testg Hasselager	
52) 1945_Orm...	og en halv Dag 15 Kr 6. Til Efterretning Eftergives. 2 Kvt K...	sol... . Knud Bent Andersen Skovlide indkaldt 20...	
8) 1940_Tiset	Tiset Rationeringskort for Juli Kv. Følgende forhandles: 1. A...	sol... og og Orlogsgastens Dag. Nægtedes. 4. M...	
28) 1942_Tiset	om Tilskud til Aftenskole til Lokale m. v. Bevilgedes. 3. And...	sol... - hjem i Viborg om Tilskud. Bevilgedes 10 K...	
37) 1943_Tiset	fra Smedesvend Tage Jensen, Tiset, der meddelte, at han ...	sol... - hjemmet i Viborg, fra Sønderskov- hjemm...	
37) 1943_Tiset	11 Februar Kl. 3 afholdt Sogneraadet Møde i Ravnholz For...	sol... - arbejde. do 3 Anmodning fra Amtsrevision	
37) 1943_Tiset	Skov Niels Andersen Jens Hedegaard Aar 1943 Onsdag d...	sol... - hjemmet og Dannevirkes Lokalafdeling i F...	
52) 1945_Orm...	Tiset Kommune et Beløb af 405 Kr. do. 4. Meddelelse fra A...	sol... - hjem i Randers om et Tilskud Bevilgedes	
6) 1940_Ormsl...	45 om Lov = ændringer under Undervisningsministeriet Til ...	sol... . Til Efterretning 10 Indenrigsministeriets Ci...	
17) 1941_Orm...	Pengeskab. 35. Gdr. Jens Thomsen og Hustru Marie Tho...	sol... i København ansøger om Tilskud. Bevilliges	
25) 1942_Hasl...	2 den Behandling 24 "Grænseforeningen" ansøger om Tils...	sol... i København Kirkelig Samf. af 1898 søger T...	
27) 1942_Orm...	2400 Kr. Kan ikke Ændres Nedsat til 1200 Kr. Nedsat til 0 ...	sol... København c) Blindehjemmet i Vejle ad 9. a...	
28) 1942_Tiset	Petersen Constbg. klager ligeledes over Ansættelsen for J...	sol... i København ansøger om Tilskud. Udgaaer H...	
	Maaned 2 Bevilgedes 25 Kr 3 Andragende fra Mølstrup Opt...	sol... angaaende Tilskud 5 Nægtedes 6 Fra Aarh...	



Contexts	Correlations
Document	Left Term Right
35) 1943_Hasl...	Sogneraadet for sit vedkommen- de kan anbefa- le Andrag...
4) 1940_Hasle...	og 1940/41, idet han nu er arbejdsløs.
6) 1940_Ormsl...	g) Henstand indtil vi...
17) 1941_Orm...	a. Frk Vibeke Sinding Jensen ...
25) 1942_Hasl...	c) Andragende fra R. Møller Andersen, Hasle, o...
52) 1945_Orm...	paalig= nede Skatter. 19 Indkomsten anbefales nedsat til 1...
52) 1945_Orm...	Kolt om Tilladelse til i Fremtiden at anvende Regnskabsaar...
52) 1945_Orm...	2 Kvartaler Juli og Oktober Eftergives 2 Kvartaler Juli og O...
52) 1945_Orm...	og en halv Dag 15 Kr 6. Til Efterretning Eftergives. 2 Kvt K...
8) 1940_Tiset	Tiset Rationeringskort for Juli Kv. Følgende forhandles: 1. A...
28) 1942_Tiset	om Tilskud til Aftenskole til Lokale m. v. Bevilgedes. 3. And...
37) 1943_Tiset	fra Smedesvend Tage Jensen, Tiset, der meddelte, at han ...
37) 1943_Tiset	11 Februar Kl. 3 afholdt Sogneraadet Møde i Ravnholz For...
37) 1943_Tiset	Skov Niels Andersen Jens Hedegaard Aar 1943 Onsdag d...
37) 1943_Tiset	Tiset Kommune et Beløb af 405 Kr. do. 4. Meddelelse fra A...
52) 1945_Orm...	45 om Lov = ændringer under Undervisningsministeriet Til ...
6) 1940_Ormsl...	Pengeskab. 35. Gdr. Jens Thomsen og Hustru Marie Tho...
17) 1941_Orm...	2 den Behandling 24 "Grænseforeningen" ansøger om Tils...
25) 1942_Hasl...	2400 Kr. Kan ikke Ændres Nedsat til 1200 Kr. Nedsat til 0 ...
27) 1942_Orm...	Petersen Constbg. klager ligeledes over Ansættelsen for J...
28) 1942_Tiset	Maaned 2 Bevilgedes 25 Kr 3 Andragende fra Mølstrup Opt...

solda* X ? 83 context expand Scale

Key Rockwell (@ 2019) Privacy v. 2.4 (M28)

Export a Word cloud: *follow the guidelines in the box*

Beder-Malling 1940

Sogneraadsmøde afholdes De gamles Hjem Fredag d 2 Februar 1940

Sogneraadet gennemgaar indkomne Klager og Protester fra Ejerne af Grundene langs Fløjstrup og Aistrup Strand

Export SVG



This is a thumbnail of the SVG image, right-click or ctrl-click on the image to save a full-size copy on your hard drive.

Alternatively, copy the SVG code below.

```
<svg id="cirrus_88" class="cirrusGraph" width="509" height="313" version="1.1"
xmlns="http://www.w3.org/2000/svg"><g transform="translate(254.5, 156.5)
scale(0.6702355146408081,0.6702355146408081)"><text text-anchor="middle" data-
freq="3710" transform="translate(37, 59) rotate(90)" style="font-family: "Palatino
Linotype", "Book Antiqua", Palatino, serif; fill: rgb(255, 0, 164); font-
```

OK



eWorldforms. Created about a day ago.

- 1) Bede... og Markplan Amtsvejinspektør Christof
 - 1) Bede... de forskellige Klager og
 - 1) Bede...

Adding Resources to your Corpus

Voyant Tools Not secure | voyant.chcaa.au.dk/?corpus=98a9e28207de8fc1ce63d8de47c88bb0

Apps Inbox (350) - adela... Therapy DK Useful Links AU Essentials FAIMS eResearch Fun satDay in Paris TRAP - MQNS useR! 2019 - Toulou... AU Arts eResearch... Other bookmarks

Voyant Tools

Cirrus Terms Links Reader TermsBerry Trends Document Terms

Documents

	Title	Words	Types	Ratio	Words/Sentence
1	AOR_2014-2013	274,...	25,...	9%	37.0
2	AOR_2015-2014	334,...	27,...	8%	16.2

Scale Terms: 0

Summary

	Title
1	AOR_2014-2013
2	AOR_2015-2014

0

+ Add Remove Keep Reorder Cancel

Могилата 202 Context expand Scale

Voyant Tools , Stéfan Sinclair & Geoffrey Rockwell (© 2019) Privacy v. 2.4 (M28)

Click upload and wait, DO NOT PUSH REVEAL

Screenshot of the Voyant Tools interface showing a corpus upload dialog.

The interface includes:

- Header bar: Voyant Tools, Not secure, URL: voyant.chcaa.au.dk/?corpus=98a9e28207de8fc1ce63d8de47c88bb0
- Toolbar: Apps, Inbox (350) - adela..., Therapy, DK Useful Links, AU Essentials, FAIMS, eResearch, Fun, satDay in Paris, TRAP - MQNS, user! 2019 - Toulou..., AU Arts eResearch..., Other bookmarks
- Left sidebar: Cirrus (word cloud), Terms, Links, Reader (selected), TermsBerry, Trends, Document Terms.
- Central area: AOR_2014-2013, arхеологически открытия и разкопки, РЕДАКЦИОННА КОЛЕГИЯ, доц. д-р Мария Гюрова (главен редактор) гл. ас. д-р Петър Димитров.
- Right area: Trends chart showing frequencies over time for terms like 'обр' (blue), 'част' (green), '1' (pink), '2' (purple), and 'археологически' (cyan).
- Bottom left: Scale, Terms, Documents (selected), Phrases.
- Bottom right: Corpus (Documents) dropdown, Reset, Display.
- Bottom center: Add Texts dialog with a red circle around the 'Upload' button.
- Table: Summary of documents (Title, Words, Types, Ratio, Words/Sentence) and detailed document content (Left, Term, Right).
- Bottom footer: Modify, Download, Scale.

The 'Add Texts' dialog contains:

- Type in one or more URLs on separate lines or paste in a full text.
- Upload button (circled in red).
- Cancel button.
- A large red X button.

Table Headers:

Title	Words	Types	Ratio	Words/Sentence
-------	-------	-------	-------	----------------

Table Data:

1 AOR_2014-2013	274,...	25,...	9%	37.0
2 AOR_2015-2014	334,...	27,...	8%	16.2

Document Content Excerpt:

Document	Left	Term	Right
1) AOR...	е проучена обща площ ок. 1 дка. дебелина...	мо...	достига 3,80 м, а в северозападната – 3,20...
1) AOR...	проучени 30 мюсолмански гроба. онези от ...	мо...	, се характеризират с големи овал- ни гроб...
1) AOR...	юли и август 2013 г. в сек- тора от миналат...	мо...	. използва се план-квадратната мрежа от ...
1) AOR...	Гюрова, Надежда Каракоянова През 2013 ...	мо...	горяла сграда от VII хоризонт (Венелинова...
1) AOR...	от подбрани каменни плочи. стената плавн...	мо...	. В основата си, стената има трапецовидно...
1) AOR...	чохалжиев и др. 2004-2005). между 2009 и ...	мо...	(чохалжиев 2010, 2011, 2012-2013), чиито ...

Page footer: Voyant Tools , Stéfan Sinclair & Geoffrey Rockwell (© 2019) Privacy v. 2.4 (M28)

If you work online:
Save your workspace!

Can you save the Workspace? Only URL for 30 days

- Beware of **30 days uptime**: after 30 days from the last time workspace was opened, the project is flushed (this is negotiable for AU instance).
- For use in essay, use data and visualisation exports, and save your corpus composition.
- Workplace settings (e.g. edits to stopwordlists) are ephemeral without exception.

Glossary

Black Box: A device, system or object which can be viewed in terms of its inputs and outputs without any knowledge of its internal workings. In Digital Humanities, projects and tools are often critiqued for being “black boxes” which prevent scholars from studying and analyzing the methodological choices made when composing the project.

Collocation: a sequence of words or terms that co-occur more often than would be expected by chance.

Corpus: a collection of documents.

Optical Character Recognition (OCR): Software that turns images of text into machine readable plain text documents.

Stop list or Stop Words: A list of words that is automatically omitted from an index of the most frequent words in a corpus.

Text Mining: Is often also referred to as data mining. The overarching goal is usually to turn text into data for analysis and it frequently relies heavily on natural language processing. For an excellent description of the history of the field and various applications see the Wikipedia entry for text mining.

TF-IDF: Short for term frequency-inverse document frequency. It is a numerical statistic that is intended to reflect how important a word is to a document in a corpus. This statistic is frequently used to weight terms by relevance or importance within a corpus.

Where to get more training?

Aarhus University

- *CEDHAR workshops*
<https://cas.au.dk/center-for-digital-history-aarhus/events>
- *History Hacky Hours*: Fridays 1-3pm currently on Zoom with me
- *LibLab*: Tuesdays, one-on-one consultations with Max Odsbjerg Pedersen (maop@kb.dk)
- *CRU/Rhus*: see schedule of events at www.cru.au.dk, a community of open source software users from across campus, excellent for networking and continued learning
- *CHCAA*: see www.chcaa.io, group of programmers for consultation and deployment of large-scale digital projects
- *Digital Literacy v.2*: training programme for digital humanists, coordinated by Mads Rosendahl Thomsen

Thank you! Questions?