

*Vēstures datu analīze ar datormetodēm:
Komjaunatnes darbinieku datubāze (1970.–
1990. gads) un tās apstrāde un analīze ar
datormetodēm*

LU LFMI Digitālo humanitāro zinātņu grupas pētnieks
un LU HZF doktorants **Haralds Matulis**

Lekcijas-darbnīcas saturs

1. 25min – **ievads par “digitālajām metodēm”**.

Taču – būs VPP projekta ietvaros vēl atsevišķa “lekcija” un “ziņojums” par “digitālajām metodēm vēstures pētniecībā”.

Lekcijas-darbnīcas saturs

1. 25min – **ievads** par “**digitālajām metodēm**”.

Taču – būs VPP projekta ietvaros vēl atsevišķa “lekcija” un “ziņojums” par “digitālajām metodēm vēstures pētniecībā”.

2. 45min – Dainas Bleieres sagatavotās **datu kopas** “Komjaunatnes darbinieku datubāze (1970.–1990. gads)” **analīze** ar datormetodēm

3. 20min – **jautājumi un diskusija** par datu kopas analīzi un plašāk par datormetodēm.

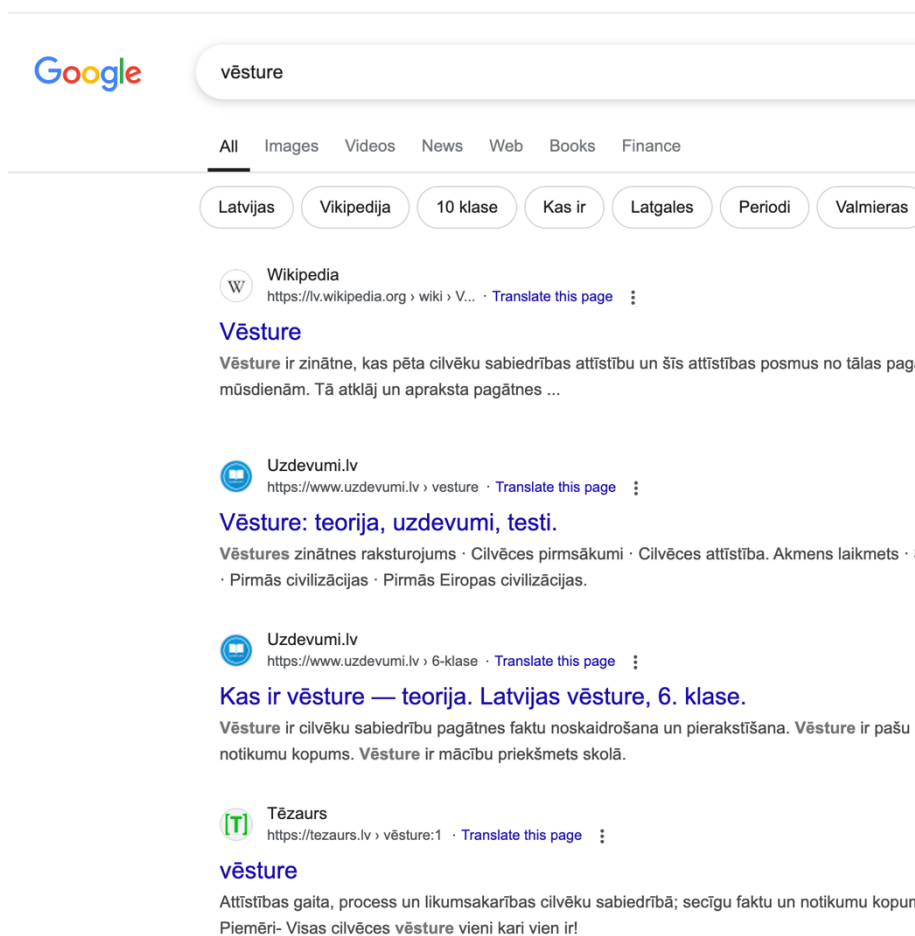
Kas ir digitālās metodes humanitārajās zinātnēs (DH)

Kas ir digitālās metodes humanitārajās zinātnēs (DH)

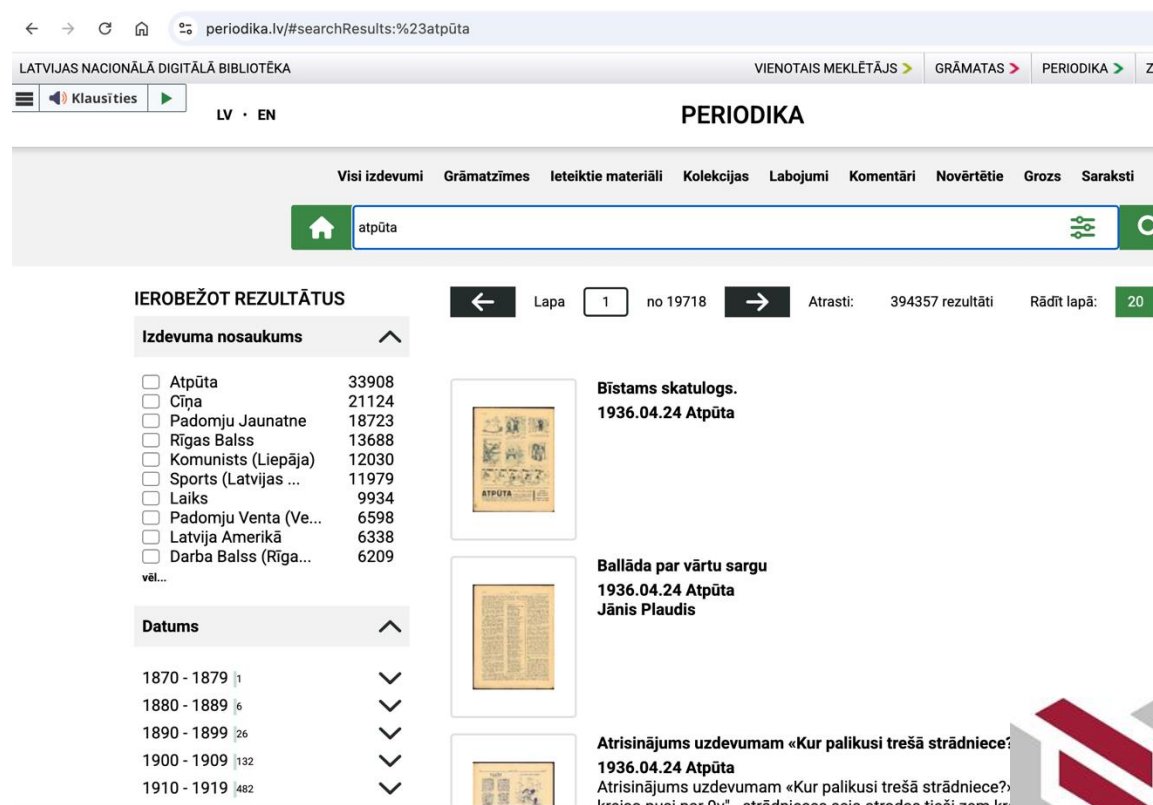
0 – **meklēšana un pārlūkošana** (search / browse) digitālajos resursos. Tam seko pētniecība ar kvalitatīvajām metodēm.

Kas ir digitālās metodes humanitārajās zinātnēs (DH)

0 – meklēšana un pārlūkošana (search / browse) digitālajos resursos. Tam seko pētniecība ar kvalitatīvajām metodēm. (1) Google, (2) periodika.lv



Google search results for "vēsture". The search bar shows "vēsture". Below the search bar are tabs for All, Images, Videos, News, Web, Books, and Finance. Below the tabs are filters for Latvijas, Vikipēdija, 10 klase, Kas ir, Latgales, Periodi, and Valmieras. The first result is from Wikipedia, titled "Vēsture", with a description: "Vēsture ir zinātne, kas pēta cilvēku sabiedrības attīstību un šīs attīstības posmus no tālas pagātnes līdz mūsdienām. Tā atklāj un apraksta pagātnes...". Below this is a result from Uzdevumi.lv, titled "Vēsture: teorija, uzdevumi, testi.", with a description: "Vēstures zinātnes raksturojums · Cilvēces pirmsākumi · Cilvēces attīstība. Akmens laikmets · Pirmās civilizācijas · Pirmās Eiropas civilizācijas.". Below this is another result from Uzdevumi.lv, titled "Kas ir vēsture — teorija. Latvijas vēsture, 6. klase.", with a description: "Vēsture ir cilvēku sabiedrību pagātnes faktus noskaidrošana un pierakstīšana. Vēsture ir pašu notikumu kopums. Vēsture ir mācību priekšmets skolā.". Below this is a result from Tēzauris, titled "vēsture", with a description: "Attīstības gaita, process un likumsakarības cilvēku sabiedrībā; secīgu faktus un notikumu kopu. Piemēri- Visas cilvēces vēsture vienā karā vienā!".



Periodika.lv search results for "atpūta". The website header shows "LATVIJAS NACIONĀLĀ DIGITĀLĀ BIBLIOTĒKA" and "VIENOTAIS MEKLĒTĀJS". Below the header are tabs for Klausīties, LV, and EN. The main header shows "PERIODIKA". Below the main header are tabs for Visi izdevumi, Grāmatzīmes, Ieteiktie materiāli, Kolekcijas, Labojumi, Komentāri, Novērtētie, Grozs, and Saraksti. Below the tabs is a search bar with "atpūta" and a search button. Below the search bar is a table of results. The table has two columns: "Izdevuma nosaukums" and "Datums". The table lists various books and their dates. The first row is "Atpūta" with the date "1936.04.24". The second row is "Ciņa" with the date "1936.04.24". The third row is "Padomju Jaunatne" with the date "1936.04.24". The fourth row is "Rīgas Balss" with the date "1936.04.24". The fifth row is "Komunisti (Liepāja)" with the date "1936.04.24". The sixth row is "Sports (Latvijas ...)" with the date "1936.04.24". The seventh row is "Laiks" with the date "1936.04.24". The eighth row is "Padomju Venta (Ve...)" with the date "1936.04.24". The ninth row is "Latvija Amerikā" with the date "1936.04.24". The tenth row is "Darba Balss (Rīga...)" with the date "1936.04.24". The table also shows the number of results for each date: 33908 for "Atpūta", 21124 for "Ciņa", 18723 for "Padomju Jaunatne", 13688 for "Rīgas Balss", 12030 for "Komunisti (Liepāja)", 11979 for "Sports (Latvijas ...)", 9934 for "Laiks", 6598 for "Padomju Venta (Ve...)", 6338 for "Latvija Amerikā", and 6209 for "Darba Balss (Rīga...)".

Kas ir digitālās metodes humanitārajās zinātnēs (DH)

0 – meklēšana un pārlūkošana (search / browse) digitālajos resursos. Tam seko pētniecība ar kvalitatīvajām metodēm. (Huistra & Mellink 2016 raksts)



Historical Methods: A Journal of Quantitative and Interdisciplinary History

ISSN: 0161-5440 (Print) 1940-1906 (Online) Journal homepage: www.tandfonline.com/journals/vhim20

Phrasing history: Selecting sources in digital repositories

Hieke Huistra & Bram Mellink

To cite this article: Hieke Huistra & Bram Mellink (2016) Phrasing history: Selecting sources in digital repositories, Historical Methods: A Journal of Quantitative and Interdisciplinary History, 49:4, 220-229, DOI: [10.1080/01615440.2016.1205964](https://doi.org/10.1080/01615440.2016.1205964)

To link to this article: <https://doi.org/10.1080/01615440.2016.1205964>

ABSTRACT

In recent years, mass digitization has opened up voluminous text. Full-text search lets historians now find new sources that can thoroughly studied historical episodes. At the same time, it finds sources in a new way: through specific words. This article analyzes a way of accessing sources and investigates which search technology source selection in digital repositories. It argues that to seize the offers, historians must refine their search technologies so that they are dependent on exact phraseology.

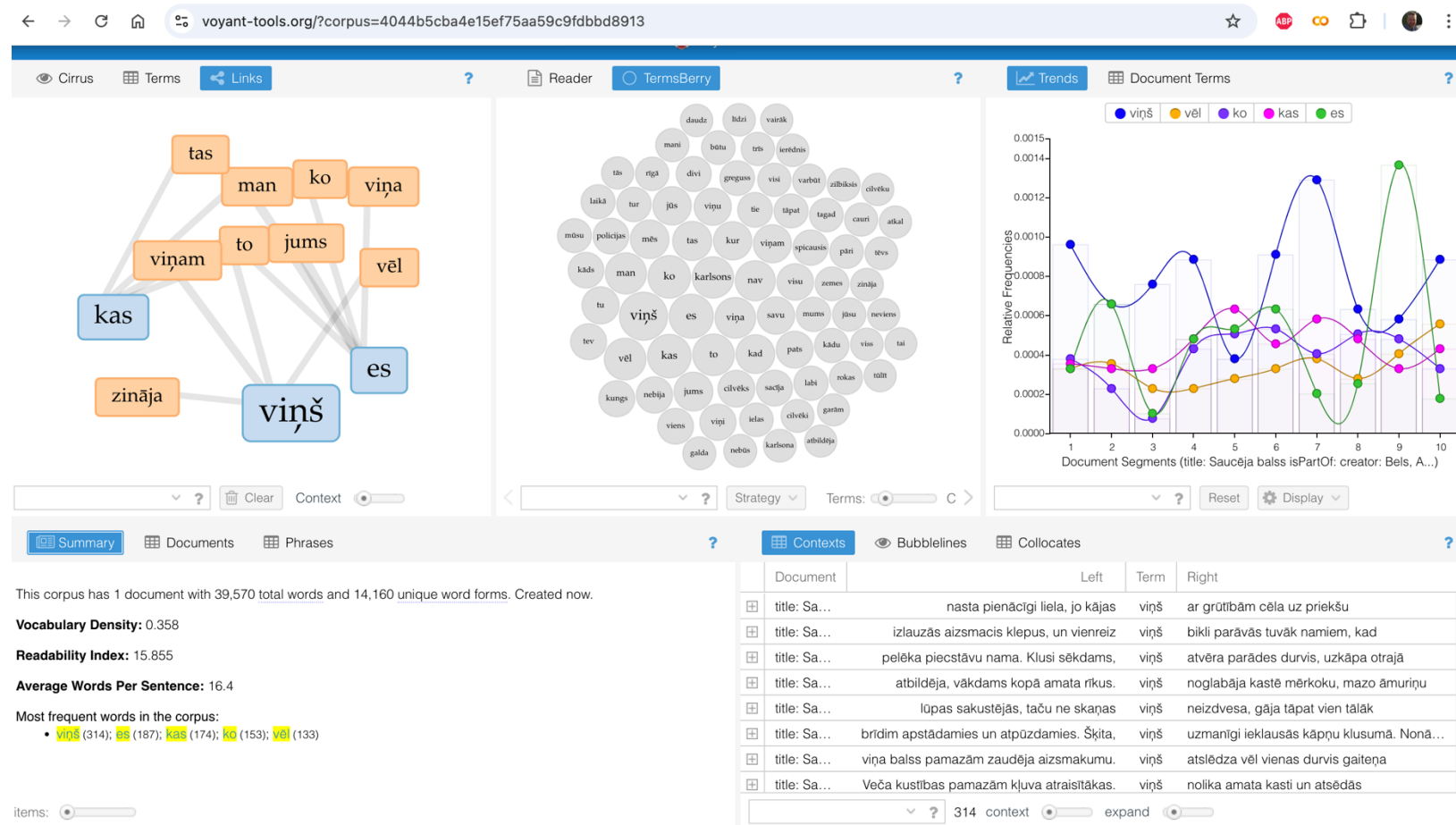
We have all been there: sitting behind your computer screen, browsing the national newspaper database, searching for sources on the public reception of, say, the assassination of John F. Kennedy. First, you enter the president's name and limit the time period to 1963, but it feels feeble to look only for sources containing "John F. Kennedy" in full, and it returns many articles on irrelevant aspects of Kennedy's life. You then extend your search with synonyms ("president," "Kennedy") and words more specifically related to the subject ("assassination," "shooting," "Dallas"), but still the results are either too many (when you search for any of these words) or too few (when you search for all of them). At the end, you randomly select a few newspaper articles, some of which prove useful. You add them to your footnotes as if you consulted them in the archive, perhaps happy with their added value, perhaps with a twinge of guilt about your haphazard selection. You then close your browser and return to the paper sources (business as usual).

Kas ir digitālās metodes humanitārajās zinātnēs (DH)

1 – **izpētoša datu analīze** (exploratory data analysis) ar gataviem rīkiem: piemēram, Voyant, Gephi, Palladio. Darbs ar tekstu korpusiem.

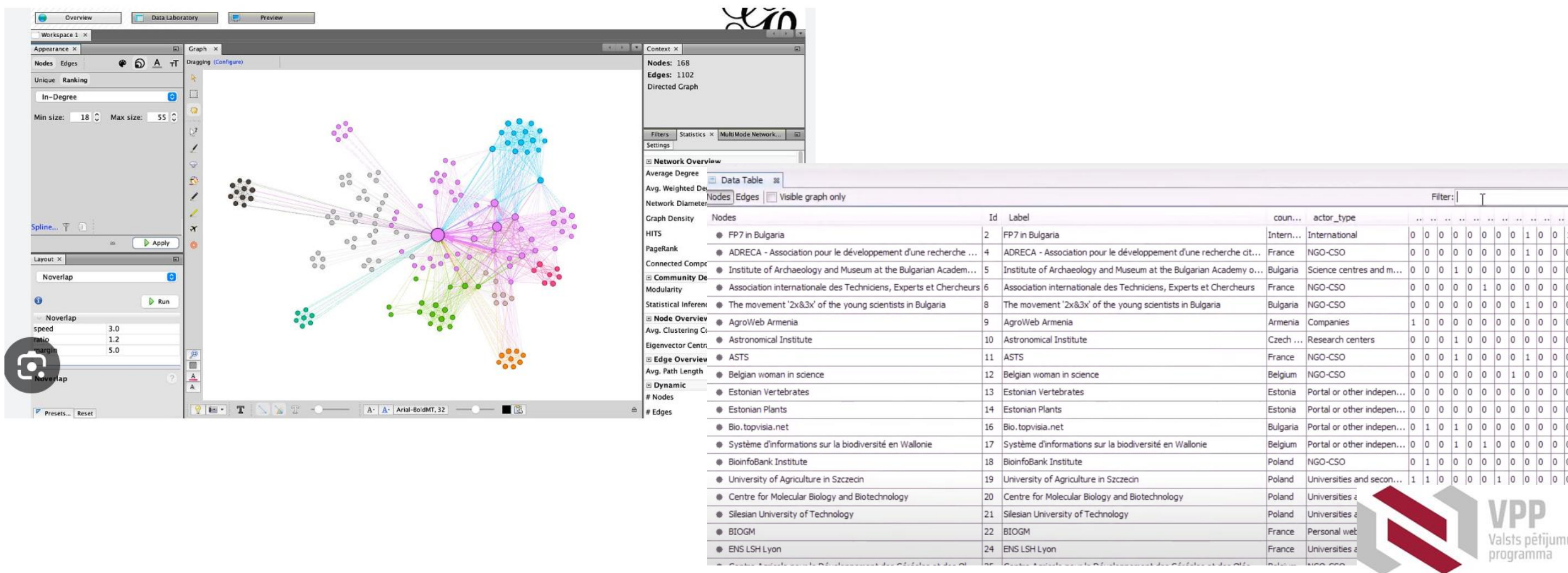
Kas ir digitālās metodes humanitārajās zinātnēs (DH)

1 – **izpētoša datu analīze** (exploratory data analysis) ar gataviem rīkiem: piemēram, **Voyant**, Gephi, Palladio. Darbs ar tekstu korpusiem. <https://voyant-tools.org>



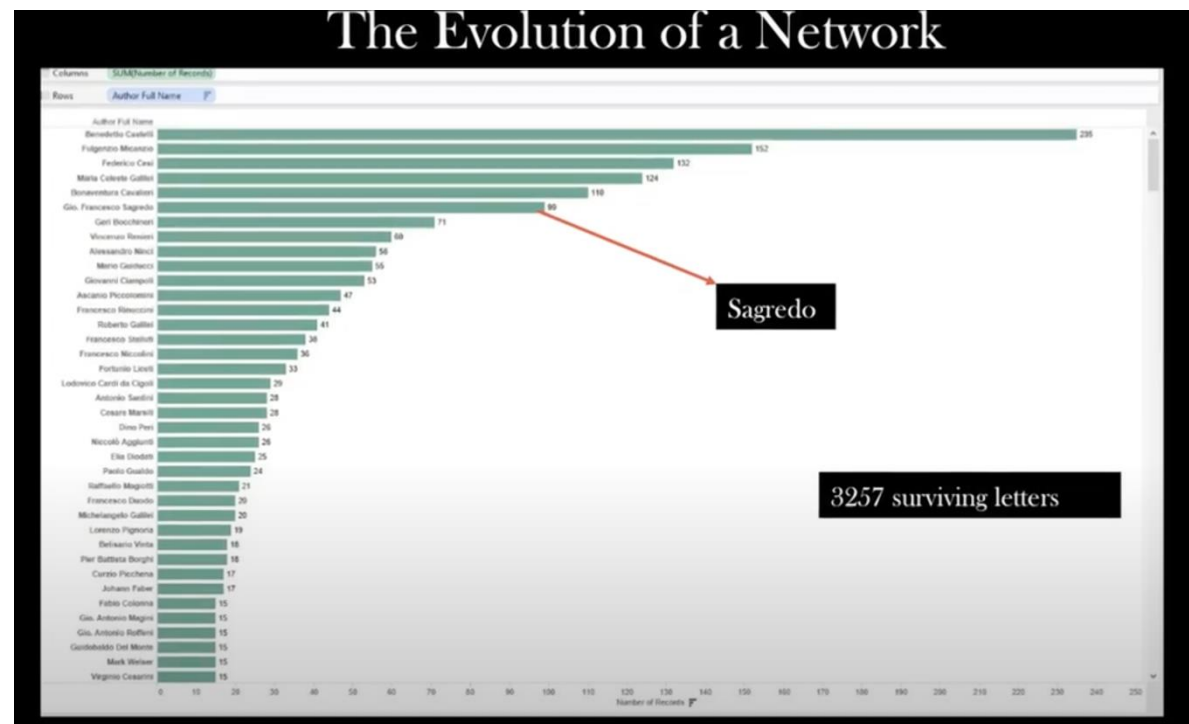
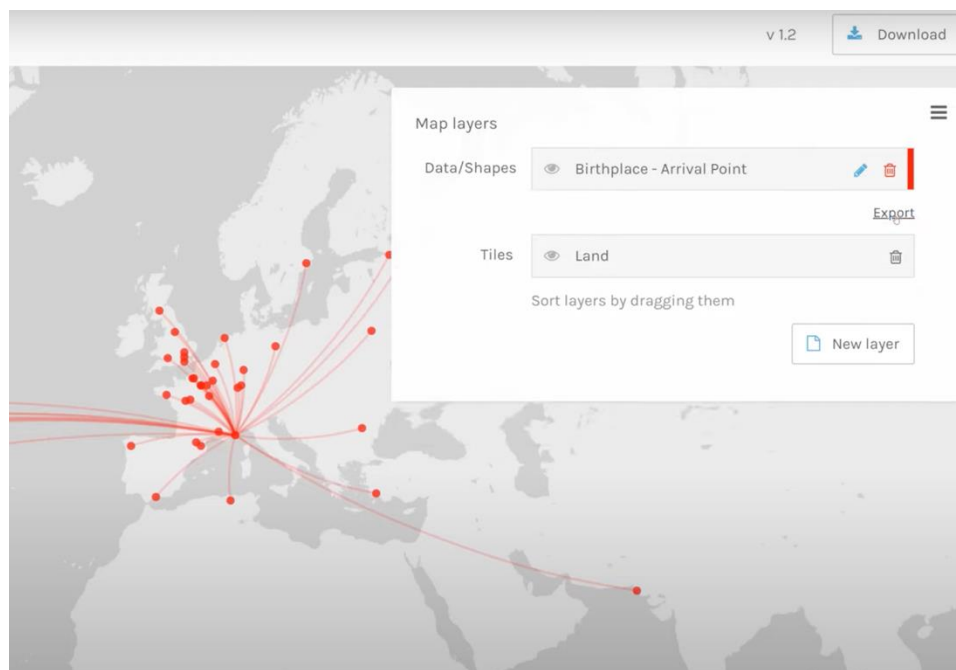
Kas ir digitālās metodes humanitārajās zinātnēs (DH)

1 – **izpētoša datu analīze** (exploratory data analysis) ar gataviem rīkiem: piemēram, Voyant, Gephi, Palladio. Darbs ar tekstu korpusiem. <https://gephi.org>



Kas ir digitālās metodes humanitārajās zinātnēs (DH)

1 – **izpētoša datu analīze** (exploratory data analysis) ar gataviem rīkiem: piemēram, Voyant, Gephi, **Palladio**. Darbs ar tekstu. <https://hdlab.stanford.edu/palladio/>



Kas ir digitālās metodes humanitārajās zinātnēs (DH)

2 – **datu vizualizācija**. Vai nu ar gataviem rīkiem vai pētniekam programmējot (kodējot), piemēram, python + pandas + matplotlib. (1) kā izpētoša datu analīze, (2) kā pierādoša ilustrācija kādam argumentam rakstā, (3) kā zinātnes komunikācija.

Kas ir digitālās metodes humanitārajās zinātnēs (DH)

2 – **datu vizualizācija**. Vai nu ar gataviem rīkiem vai pētniekam programmējot (kodējot), piemēram, **python + pandas + matplotlib**.

jupyter _01-datu-apstrade Last Checkpoint: 8 days ago

File Edit View Run Kernel Settings Help

Import Required Libraries

Import the necessary libraries, including pandas.

```
[1]: # Import Required Libraries
import pandas as pd

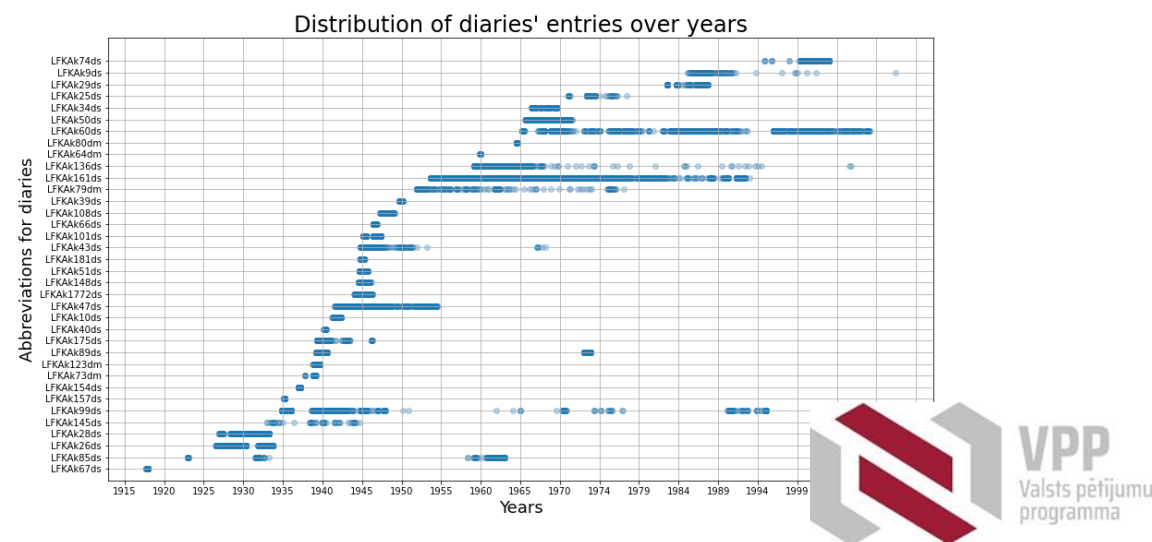
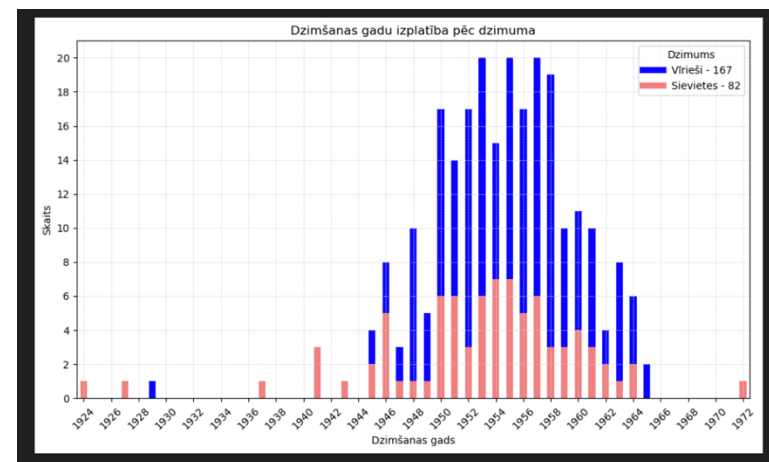
[2]: # Set display options
pd.set_option('display.max_rows', 500)
pd.set_option('display.max_columns', 30)
```

Read CSV File

Use pandas to read the CSV file into a DataFrame.

```
[3]: # Read CSV File
df = pd.read_csv('../izejas_dati/Komj_CK_80tie.csv')
df.head(250)
```

	Uzv. vārds+1:205	Lauki/pils	S/V	Dzimš.g.	Tautība	Kur dzimis	Vecāki	Profes.
0	Nataljijs Aleksandrs	CK	V	1951	Kriev	Plejūras nov.	T-armijas pensionārs; M- grāmatv.Ljas aviogrupā	Inž./radio
1	Petrenko Viktors	CK	V	1951	Kriev	Kaļiņingradas apg.	T-armijas pensionārs; M- pensionāre	Ekonomists



Kas ir digitālās metodes humanitārajās zinātnēs (DH)

3 – **skaitļošanas (computational) metodes**, lai atklātu statistiskas likumsakarības “lielos” datu apjomos. Tematu modelēšana (*topic modelling*), noskaņojumu analīze (*sentiment analysis*); dažādi mašīnmācīšanās (*classification & clustering*) pielietojumi.

Kas ir digitālās metodes humanitārajās zinātnēs (DH)

3 – skaitļošanas (computational) metodes, lai atklātu statistiskas likumsakarības “lielos” datu apjomos. Tematu modelēšana (*topic modelling*), noskaņojumu analīze (*sentiment analysis*); dažādi mašīnmācīšanās (*classification & clustering*) pielietojumi. (1) Chang et al (2009) topics schema; (2) pyLDavis visualization tool

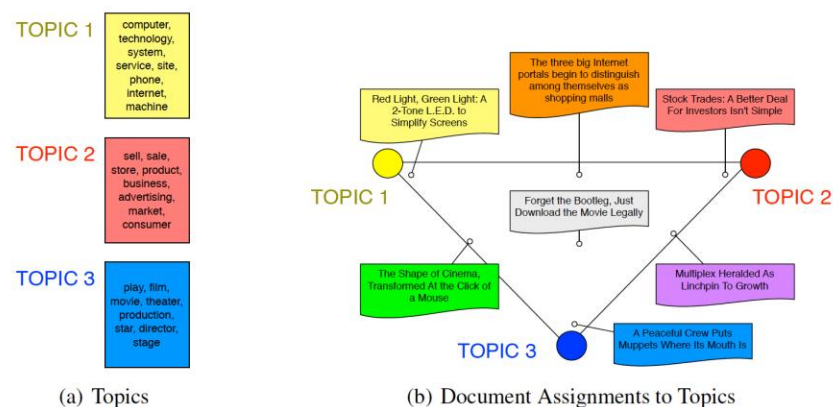
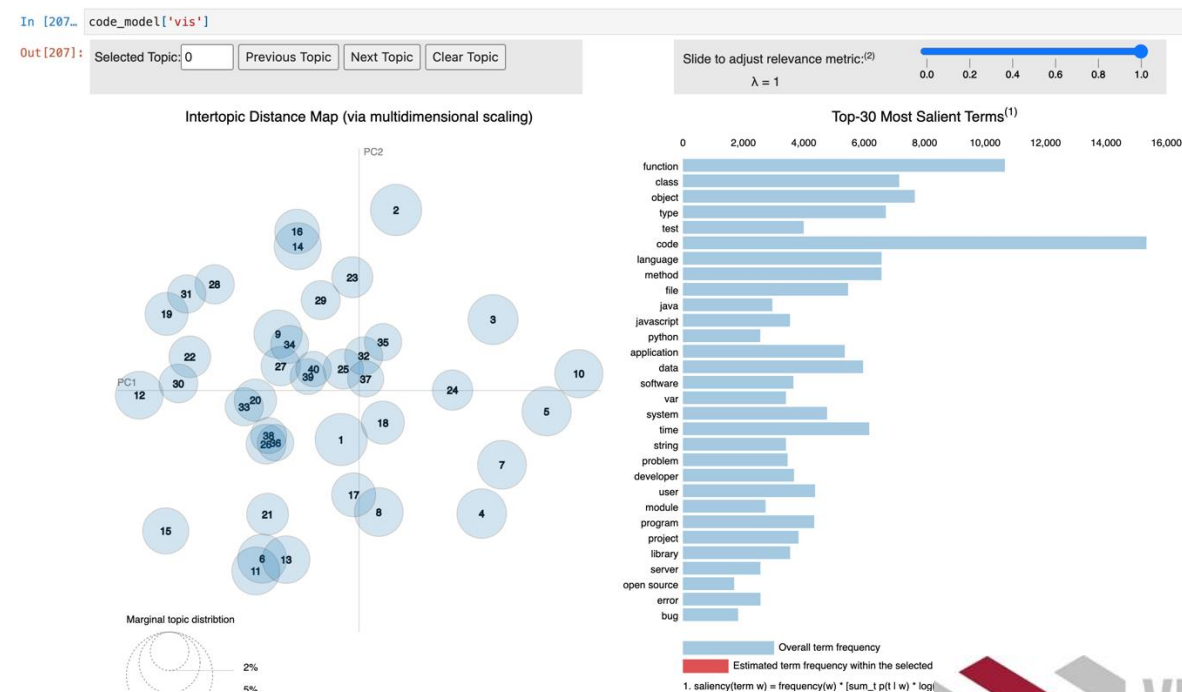


Figure 1: The latent space of a topic model consists of topics, which are distributions over words, and a distribution over these topics for each document. On the left are three topics from a fifty topic LDA model trained on articles from the New York Times. On the right is a simplex depicting the distribution over topics associated with seven documents. The line from each document's title shows the document's position in the topic space.



Kas ir digitālās metodes humanitārajās zinātnēs (DH)

3 – **skaitļošanas (computational) metodes**, lai atklātu statistiskas likumsakarības “lielos” datu apjomos. Tematu modelēšana (*topic modelling*), **noskaņojumu analīze (*sentiment analysis*)**; dažādi mašīnmācīšanās (*classification & clustering*) pielietojumi. **Dabīgās valodas apstrāde; bi-, tri- vai multi-polāri vērtējumi.**



REVIEWS

1. Smells amazing! A perfect purchase :)
2. Must buy! Super amazing.
3. Quite satisfactory



REVIEWS

1. A decent purchase
2. Quite okayish! Smells average
3. Could have been better in lot terms



REVIEWS

1. An absolute waste of money.
2. Total waste of money
3. Terrible smell, not worth buying

SENTIMENT ANALYZER



POSITIVE (81%)



NEUTRAL (88%)



negative (91%)

Kas ir digitālās metodes humanitārajās zinātnēs (DH)

3 – **skaitļošanas (computational) metodes**, lai atklātu statistiskas likumsakarības “lielos” datu apjomos. Tematu modelēšana (*topic modelling*), **noskaņojumu analīze (*sentiment analysis*)**; dažādi mašīnmācīšanās (*classification & clustering*) pielietojumi. **Dabīgās valodas apstrāde; bi-, tri- vai multi-polāri vērtējumi.**

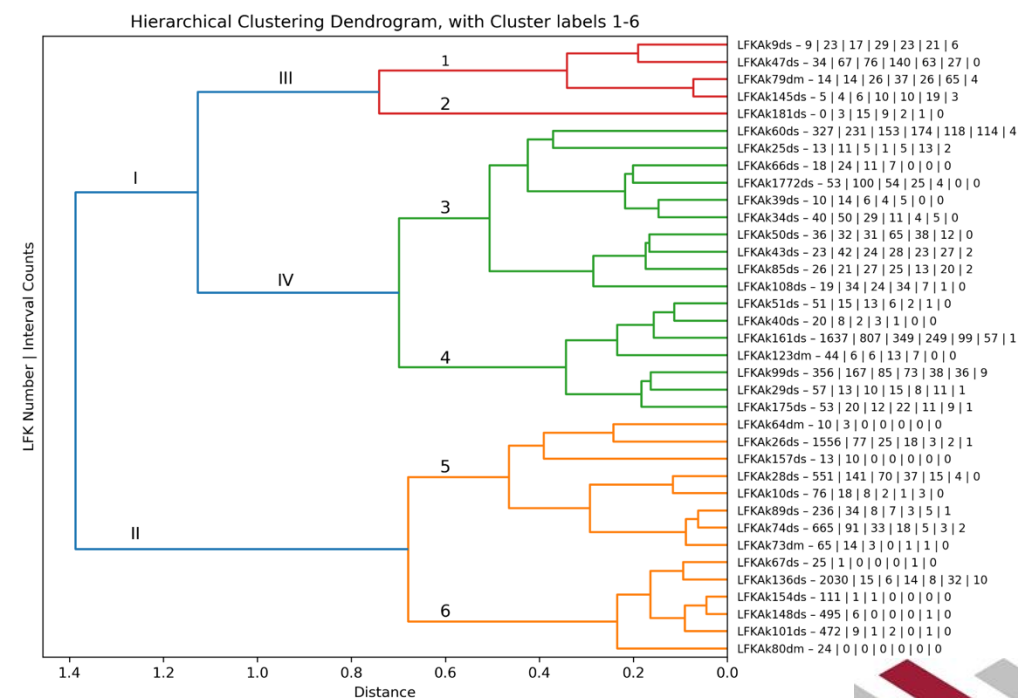
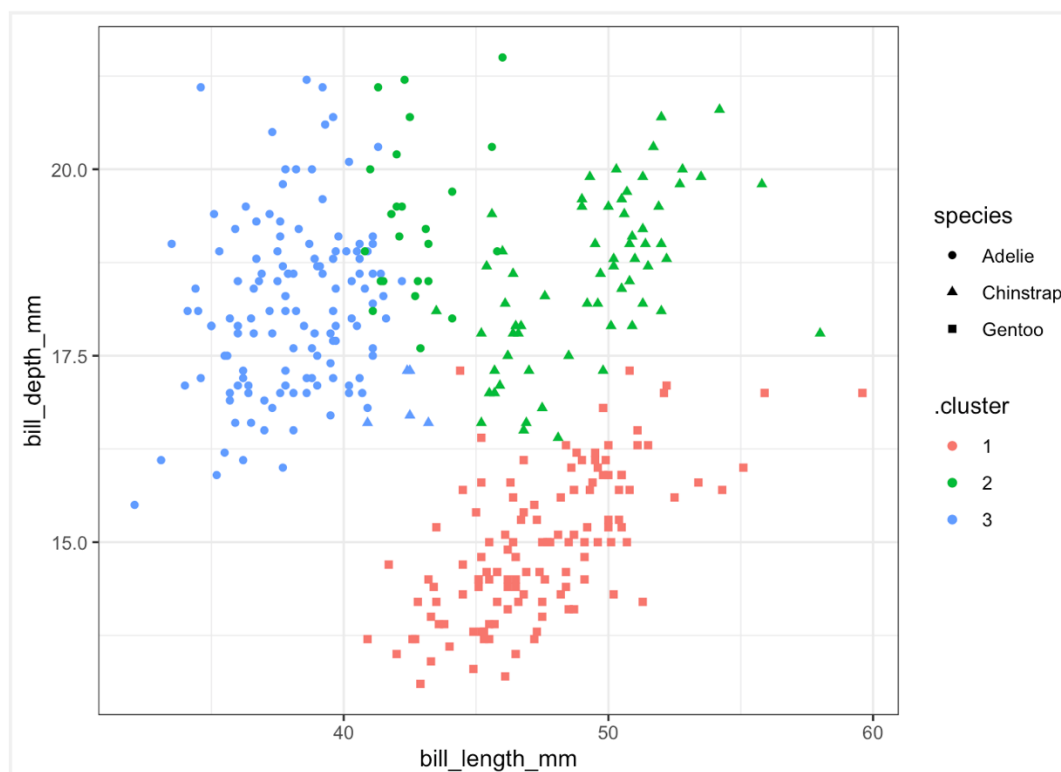


lietot ar apdomu!
Metode radīta:
– īsiem tekstiem
– komerciāliem mērķiem

dzīve, teksti un kultūra bieži vien ir sarežģītāka par plus, mīnus un neitrāl

Kas ir digitālās metodes humanitārajās zinātnēs (DH)

3 – skaitļošanas (computational) metodes, lai atklātu statistiskas likumsakarības “lielos” datu apjomos. Tematu modelēšana (*topic modelling*), noskaņojumu analīze (*sentiment analysis*); dažādi mašīnmācīšanās (*classification & clustering*) pielietojumi. (1) punktu grupējums 3 klāstros, (2) klasteru dendrogramma



Kas ir digitālās metodes humanitārajās zinātnēs (DH)

4 – **lielie valodu modeļi (LVM)**. caur chat saskarni, piemēram chatGPT. vai caur API pieprasījumiem. Pielietojums – klasifikācijas problēmu risinājumi. Trūkumi – necaurspīdīgums, nereproducējamība.

Kas ir digitālās metodes humanitārajās zinātnēs (DH)

4 – **lielie valodu modeļi (LVM)**. caur chat saskarni, piemēram chatGPT. vai caur API pieprasījumiem. Pielietojums – klasifikācijas problēmu risinājumi. Trūkumi – necaurspīdīgums, nereproducējamība. (1) <https://chatgpt.com/> – chat versija.

ChatGPT ▾

how to get in jupyter notebook make that every word is in new line in markdown code?



To make every word appear on a new line in a Jupyter Notebook markdown cell, you need to insert a **line break** after each word. In markdown, a line break can be created using two spaces followed by pressing **Enter**. Alternatively, you can explicitly use `<br>` HTML tags to create line breaks.

Here's how you can do it:

Using Line Breaks with Spaces

Write each word followed by two spaces and press **Enter**. For example:

markdown

Copy Edit

Word1
Word2
Word3
Word4

This will render:

Word1

Copy Edit

Message ChatGPT



Kas ir digitālās metodes humanitārajās zinātnēs (DH)

4 – **lielie valodu modeļi (LVM)**. caur chat saskarni, piemēram chatGPT. vai caur API pieprasījumiem. Pielietojums – klasifikācijas problēmu risinājumi. Trūkumi – necaurspīdīgums, nereproducējamība. **(2) API pieprasījumi openai LVM versijai** (attēli no Andres Karjus lekcijas BSSDH 2024).

Talking to the API using openai Python package

```
from openai import OpenAI
client = OpenAI(api_key="sdfghjklfdgawdpawdpawdpetrthjkl")

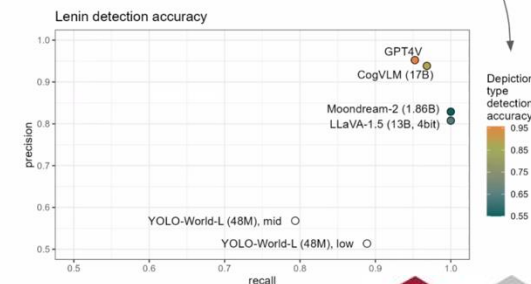
response = client.chat.completions.create(
    model= "gpt-4o",
    messages = [{"role": "user", "content":
        "Is this sentence about nature or natural processes:\n
        What a beautiful sunny day!"
    }],
    max_tokens = 1,
    temperature = 0,
    logit_bias = {'13022': 100, '3160': 100}
)

response.choices[0].message.content
```

Find anything Lenin-like: "Does this image potentially contain some depiction of the historical figure V.I. Lenin? ONLY answer Yes or No, guess if needed, never say Sorry. Consider that it could be Lenin in person, or Lenin depicted on a painting, banner, sculpture or a relief, either in the foreground or somewhere in the background."

Depiction type detection: "Does this image contain some depiction of the historical figure V.I. Lenin? If not then just say: No Lenin. If it is Lenin in person then say: Real Lenin. If it is instead some Art depiction of Lenin (portrait, sculpture, relief, bust etc) then say: Art Lenin. Do NOT say anything else, guess if needed, do NOT say Sorry."

Scheme: no Lenin, Lenin (art, real?)
Unit: extracted scenes
Quantification: Lenin~time



Kas ir digitālās metodes humanitārajās zinātnēs (DH)

0 – **meklēšana un pārlūkošana** (search / browse) digitālajos resursos. Tam seko pētniecība ar kvalitatīvajām metodēm.

1 – **izpētoša datu analīze** (exploratory data analysis) ar gataviem rīkiem vai pētniekam strādājot ar kodu: piemēram, Voyant, Gephi, Palladio. darbs ar tekstu korpusiem.

2 – **datu vizualizācija**. Vai nu ar gataviem rīkiem vai pētniekam programmējot (kodējot), piemēram, python + pandas + matplotlib.

3 – **skaitļošanas (computational) metodes**, lai atklātu statistiskas likumsakarības “lielos” datu apjomos. Tematu modelēšana (*topic modelling*), noskaņojumu analīze (*sentiment analysis*); dažādi mašīnmācīšanās (*classification & clustering*) pielietojumi.

4 – **lielie valodu modeļi (LVM)**. caur chat saskarni, piemēram chatGPT. vai caur API pieprasījumiem. Pielietojums – klasifikācijas problēmu risinājumi. Trūkumi – necaurspīdīgums, nereproducējamība.

Kas ir digitālās metodes humanitārajās zinātnēs (DH)

0 – **meklēšana un pārlūkošana** (search / browse) digitālajos resursos. Tam seko pētniecība ar kvalitatīvajām metodēm.

1 – **izpētoša datu analīze** (exploratory data analysis) ar gataviem rīkiem vai pētniekam strādājot ar kodu: piemēram, Voyant, Gephi, Palladio. darbs ar tekstu korpusiem.

2 – **datu vizualizācija**. Vai nu ar gataviem rīkiem vai **pētniekam programmējot** (kodējot), piemēram, **python + pandas + matplotlib**.

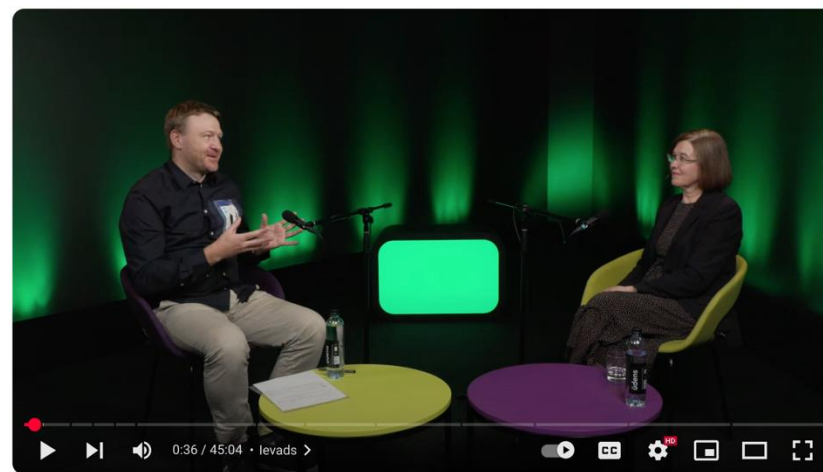
3 – **skaitļošanas (computational) metodes**, lai atklātu statistiskas likumsakarības “lielos” datu apjomos. Tematu modelēšana (*topic modelling*), noskaņojumu analīze (*sentiment analysis*); dažādi mašīnmācīšanās (*classification & clustering*) pielietojumi.

4 – **lielie valodu modeļi (LVM)**. caur chat saskarni, piemēram chatGPT. vai caur API pieprasījumiem. Pielietojums – klasifikācijas problēmu risinājumi. Trūkumi – necaurspīdīgums, nereproducējamība.

Mazs brītiņš ievilkt elpu
& jautājumi par pirmo daļu!

Vairāk par digitālajām humanitārajām zinātnēm:

- www.digitalhumanities.lv mājaslapa – īpaši BSSDH vasaras skolas materiāli + pieteikšanās BSSDH 2025 no 17. februāra
- **DH Raidieraksts** youtubē – sarunas ar Latvijas pētniekiem, kas strādā DH jomā
- **Atvērtā lekcija** Vēstures VPP ietvaros (feb /mar 2025) – par digitālo metožu izmantošanu vēstures pētniecībā



DH Raidieraksts Nr.8 | Valodu tehnoloģijas un DH

Digitālās humanitāras ...
161 subscribers

Subscribed

4

Share

Download

1.1K views 2 weeks ago

Digitālās humanitāras zinātnes (DH) – Raidieraksts – Nr.8 "Valodu tehnoloģijas un DH"

Latvijas Nacionālā bibliotēka <https://lnb.lv/>

Mērķi darbnīcas 2. daļai

Mērķi darbnīcas 2. daļai

- parādīt **tipisku darbplūsmu** – python + Jupiter notebooks + pandas + matplotlib.
- parādīt **jautājumus, kas rodas par datiem** / par to formatējumu un pierakstu.
- parādīt **datu sagatavošanu** (cleaning / pre-processing) analīzei.
- parādīt **datu analīzes soļus** un metodoloģiskās pārdomas, un dažus secinājumus.

Darbnīcas saturs

1. **Darba vides pārskats** – Jupiter notebooks + python + pandas + matplotlib.
2. **Anonimizēšanas** piemērs.
3. **Pamata statistika** par datiem 2 vizualizācijās.
4. **Datu attīrīšana un precizēšana** – nepiederīgie dati izdzēsti, NaN rindas izdzēstas.
5. **Datu grupēšana** (interpretatīvs solis) – saruna ar chatGPT.
6. **Karjeras trajektoriju analīze** – vairāki scatterplots.
7. **Trajektoriju vizualizācijas.**

uz Jupiter Notebooks!

Sadarbība starp humanitāro pētnieku & datu pētnieku:

- **Izejas datus** pētniecības ietvaros ir sagatavojusi Daina Bleiere.
- Datu analīzē aplūkotie **pētnieciskie jautājumi** izriet no Dainas Bleieres ekspertīzes un pētnieciskajām interesēm –
»ko par komjaunatnes darbinieku karjeru trajektorijām var izpētīt ar datormetodēm?»
- **Kā norādīt sadarbību** pie publikācijas / referāta.

Datu veidi

- Nestrukturēti dati – piemēram, romāna teksts
- Daļēji strukturēti dati – tabulāri dati ar dažādu saturu
- Strukturēti dati – tabulāri dati ar skaitliskām vērtībām / kategoriālām vārdiskām vērtībām (piemēram, Jā, Nē, Rīga, Pieaugušais, utml)

Datu veidi

- Nestrukturēti dati – piemēram, romāna teksts
- **Daļēji strukturēti dati** – tabulāri dati ar dažādu saturu
- Strukturēti dati – tabulāri dati ar skaitliskām vērtībām / kategoriālām vārdiskām vērtībām (piemēram, Jā, Nē, Rīga, Pieaugušais, utml)
- ir Excel jeb csv fails, ar kolonnu nosaukumiem un jaukta tipa datiem – skaitļi, kategoriāli vārdi, plūstoša vārdu secība.

datu "normalizācija" – piemēri:

(1) kartogrāfiskiem datiem – Petrograda / Pēterburga / Ļeņingrada

(2) valodnieciskiem pētījumiem – we're = we are, u.c.

(3) lemmatizācija – Rīga, Rīgas, Rīgai, Rīgu – pilna teksta meklēšanai.