



Towards modelling text mining services for digital collections: the case of Latvian Prose Counter

Anda Baklāne, National Library of Latvia, Head of Digital Research Services

Valdis Saulespurēns, National Library of Latvia, System Administrator

Contact: dh@lnb.lv

LIBER 2021 Conference, 25.06.2021

Text mining (TM):

- in a narrower sense: unsupervised machine learning
- in a broader sense: computer-assisted text analysis
- especially associated with the analysis of a large body of text (no official measure of what constitutes "large")
- especially associated with discovering patterns that are not visible with the naked eye
- going beyond reading, browsing, and searching
- can be considered in the context of distant reading, power reading, hybrid reading



Text-based collections

- Books
- Digitized periodicals
- Academic papers
- Research data
- Web archive
- Manuscripts
- Ephemera
- Archive records



Strategies for the placement of text mining services

Outside library's infrastructure (the library contributes data)

Part of library's infrastructure: dedicated TM platform with GUI

Part of library's infrastructure: programming environment without GUI

Integrated as functionalities in regular collections



Dashboard
Select corpus
Word Sketch
Word Sketch Difference
Thesaurus
Concordance
Parallel Concordance
Wordlist
N-grams

Context	KWIC	Right context
ilba varākās Eiropas valstīs? </s><s> Baltkrievijas	prezidents	Aleksandrs Lukšenko
i, mēs redzam - ukraiņi atbīd savam tuvedzīgajam	prezidentam	ar „Molotova kokteiļiem”
arlamenta atstādinātā un Ukrainu pārmētā bīvušā	prezidenta	Viktora Janukoviča il k
urcija, ASV un Kanāda. </s><s> Eiropas Komisijas	prezidenta	Žozē Manuels Barrozu
iregiz, ka Saeima plinībā pavienojas Latvijas Valsts	prezidenta	„Saeimas priekšsēdēt
vām otām karšceļājm - Valsts kancelejai un Valsts	prezidenta	kancelejai. </s><s> At
jas procesus. </s><s> Savukārt likuma „Par Valsts	prezidenta	darbības nodrošināšan
trošināšanu” Il nodāja ir beši atsevišķi veidita Valsts	prezidenta	kancelejai, atrunājot ka
s, runājot par atalgojumu, nosaka, teiksim, Ministru	prezidentam	koeficientu, tad tas šaj
emērāni, šei tas ir absolūti nesamērīgs, jo Ministru	prezidentam	šis koeficients ir 4, beš
i ir iesveigui priekšlikumus veikt grozījumus Valsts	prezidenta	ieviešanas likumā, no
cekot papildu kritērijus jeb papildu prasības Valsts	prezidenta	amata kandidātam. </s>

```
In [1]: import pandas as pd
```

```
In [3]: df = pd.read_csv("word_cloud_csv/karlis_vardins_nouns.csv", skiprows=2)
df.head()
```

Item	Frequency	Relative frequency
0 gads	21	2368.58814
1 roka	20	2255.60871
2 māja	18	2030.22784
3 sintija	14	1579.06610
4 mīsa	14	1579.06610

```
In [5]: df.Frequency.sum()
```

```
Out[5]: 2224
```

PERIODIKA

Manas grāmatzīmes | Mani ieteiktie materiāli | Manas kolekcijas | Mani laboju

Tunnelbaukunst

IEROBEŽOT REZULTĀTUS

Pieejamība

Pieejams tiešā veidā 3

Valoda

Vācu 3

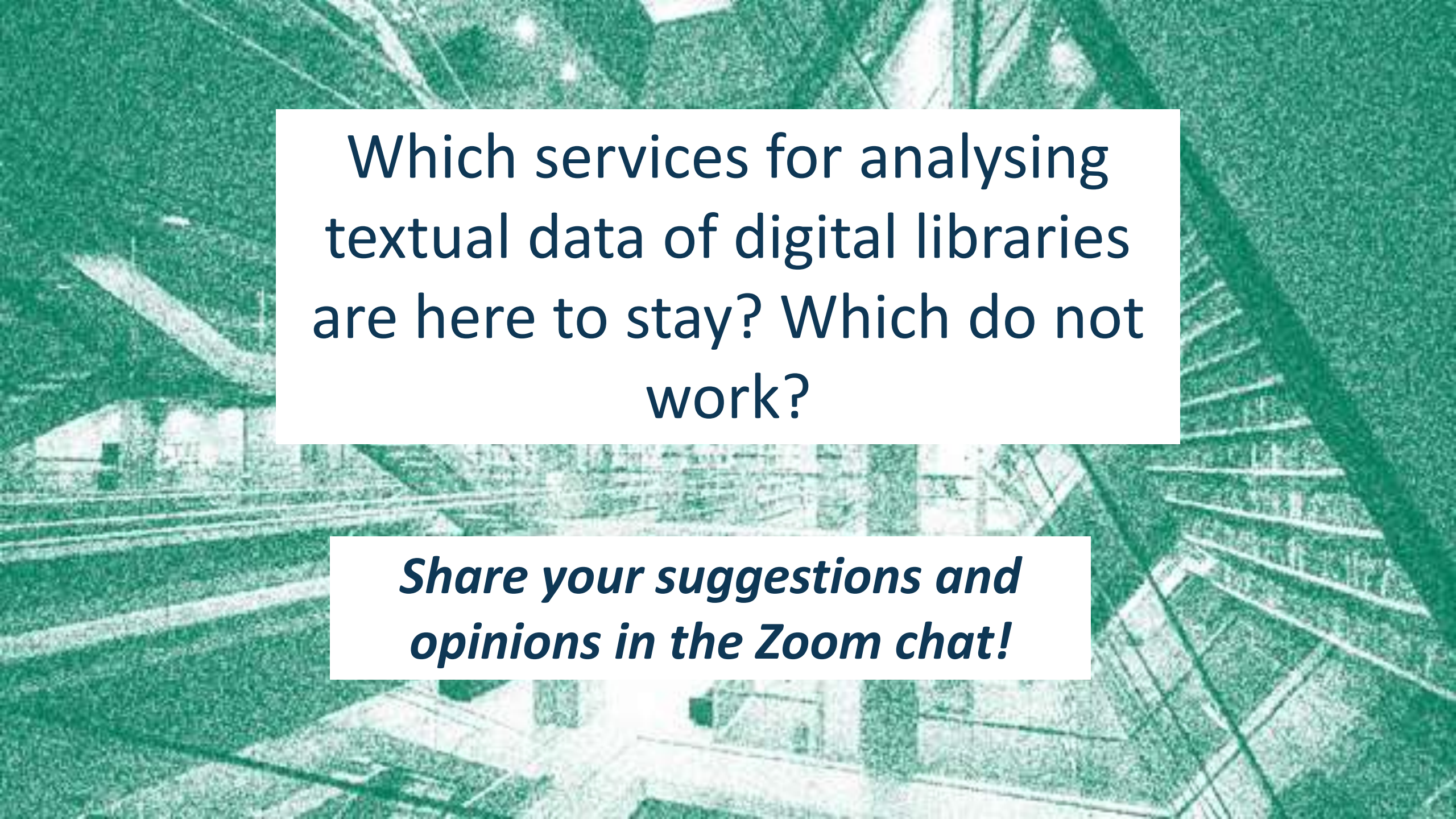
Autors

Kirstein, Gustav 1

Thiess, Franz 3

Ueber ein neues Tunnel-Holzbau System. 1876.04.07 Rīgasche Industrie Zeitung
Kirstein, Gustav
des Oesterreichischen Ingenieur- und Architekten-Vereins veröffentlicht wurde, ist auch die Tunnelbaukunst einzig und allein in den zahlreichen Eisenbahnbauten die „Einstigung“ durch so aussergewöhnliche.

Spreetunnel 1900.01.01 Rīgasche Industrie Zeitung
Thiess, Franz
Der Spreetunnel Ein hervorragendes Werk deutscher Tunnelbaukunst ist kürzlich in Untergrundbahnen**) unter dem Bett der Spree, zwischen Stralau und Treptow.



Which services for analysing textual data of digital libraries are here to stay? Which do not work?

Share your suggestions and opinions in the Zoom chat!

Candidates of text mining features integrated in digital collections of books and/or periodicals

Number of documents that contain a word or n-gram – time series

Number of instances the word or n-gram is mentioned in a corpus – time series

Term frequency: all terms, parts of speech

Statistics of pages, lines, sentences

Concordances, collocations

Topic recognition:

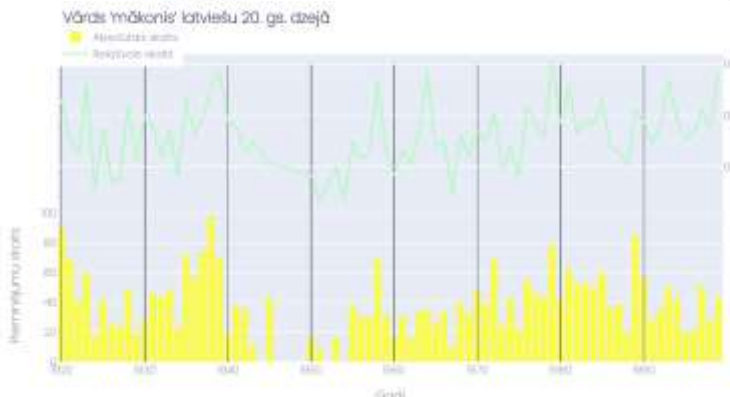
- tf-idf
- topic modelling (e.g. LDA algorithm)

Named entity recognition

Mapping

Comparing vocabularies, stylometry

Sentiment analysis



Pentagons **par** amatpersonas pagājušajā nedēļā paziņoja, ka **ASV GPT** spēku izvešana no **Afganistānas GPT** ir par apmēram 50 % pa
Kad **Bardens aprīlī** oficiāli paziņoja, ka **ASV GPT** izvedīs savus karavīrus no **Afganistānas GPT** no 1. maija līdz 11. septembrim
apmēram 2500 **ASV GPT** karavīru un 16 000 iedzīvotāju, kas pārsvarā ir **ASV GPT** pilsoņi.
Pentagons jau ir nodēvējis vairākas no savām galvenajām bāzēm **Afganistānā** valdības drošības spēkiem, kā arī izvedīs sim
militāro aprīkojumu.





Latviešu prozas skaitītājs

Latvian Prose Counter

- A website for exploring quantitative parameters of 19th and 20th century Latvian prose fiction
 - A demonstration tool that is aimed at informing users about the possibilities of textual analysis
 - **An environment that allows to test and cultivate various text analysis functionalities that are candidates for becoming new core digital services at the Latvian National Digital Library**
 - New literary works and new functionalities are regularly added to the platform
 - The analysis is being conducted in the Jupyter Notebook programming environment
- **Is not a substitute for the TM platform for researchers!**
 - Encourages to develop a TM platform based on the principles of modularity and flexibility: algorithms can be added and removed as needed

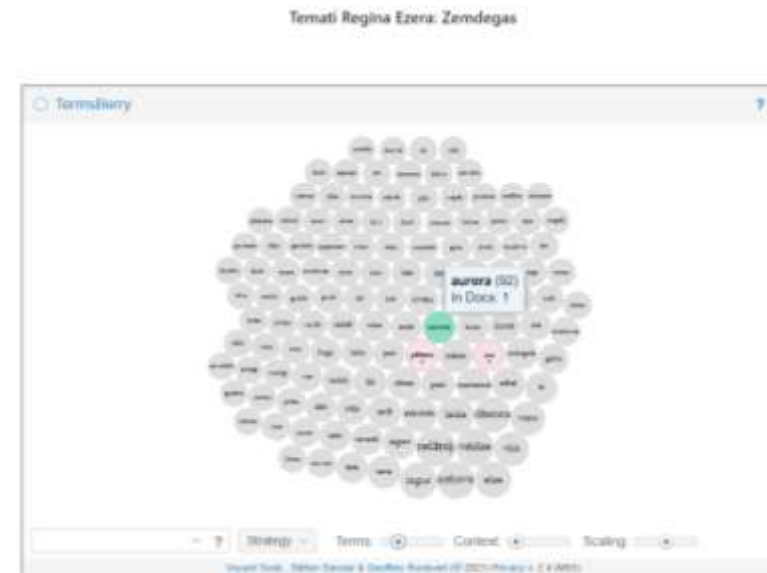
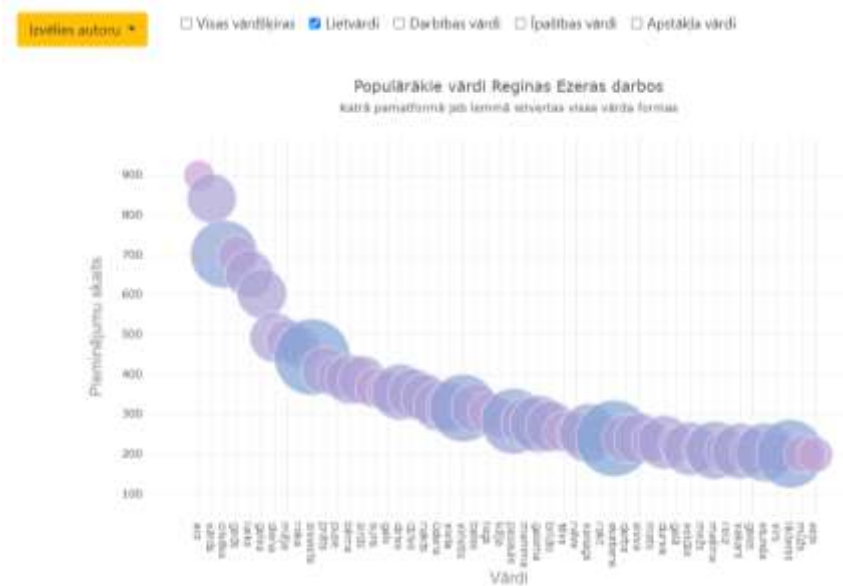
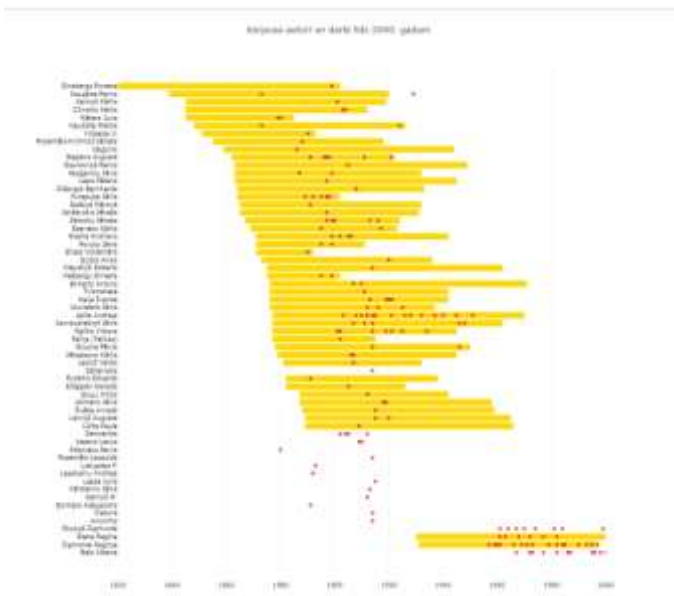
Current demonstrations in the Counter


<https://proza.lnb.lv/>

- Corpus information
- Most frequent words with POS
- Length of sentences with examples
- Lexical diversity: MSTTR counts
- Topics: tf-idf (visualization - Voyant Tools)

NATIONAL LIBRARY OF LATVIA

About Corpus Words Sentences Lexicon Topics





Tools and workflows

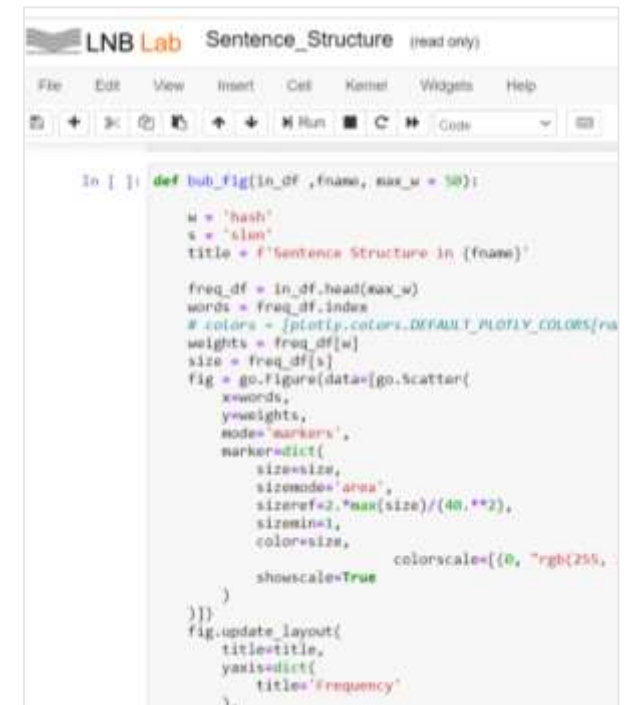
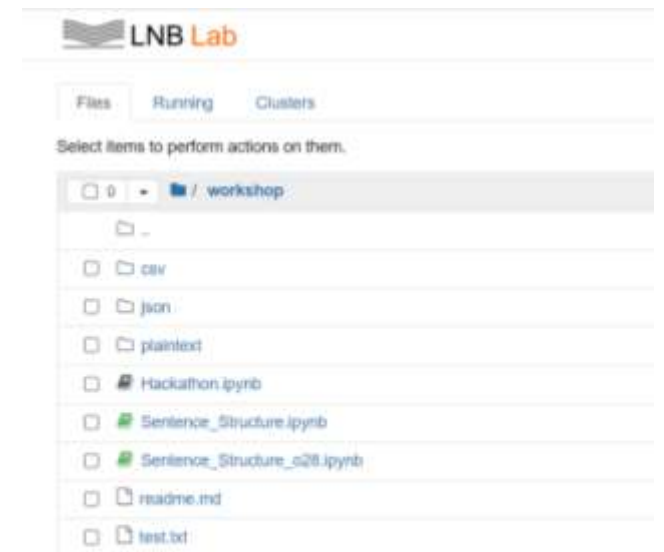
Jupyter Notebooks

- Interactive coding and visualization tools – open source
- Supports Python, Julia, R languages out of box
- Standard use case – data munging, analysis, visualization
- Widely shared among researchers and academics
- Excellent teaching tool



Jupyter Notebooks at the National Library of Latvia (NLL)

- Local instances as a part of NLL infrastructure
- Notebooks are used collaboratively by employees for data cleaning and transformation tasks
- Possible to use collaboratively with researchers to complete data analysis tasks
- Littlest Jupyter Hub supporting up to 100 users
- Alternative - to use cloud based servers (such as Google Colab) but what happens to our data?



Data preparation workflow

- Ingest data from our internal library data warehouse
- Cleanup, verification – 80% of work
- Markup using tools tailored to Latvian language
- Aggregation
- Analysis – Pandas, scikit-learn, Spacy, nltk
- Visualization using Plotly
- Export into various formats JSON, XML, CSV



Technical Challenges

- Large number of Jupyter Notebooks -> accessible locally, global access is limited
- Version control systems such as Git are not well suited to Jupyter Notebooks format
- Sharing limited to one owner, many readers mode
- Non-programmer access is limited – requires some coding, low-code knowledge

→ Latvian Prose Counter as a solution for representing results to the users and broader public



Migration from Notebooks to the Web platform

Considerations:

- Jupyter Notebook is already web based, clean migration not possible
- Plotly cross-platform usage – Python Plotly is actually Javascript Plotly underneath
- JSON – the most widely used data interchange format – easily generated by Python for use by Javascript
- Web based platform has the ability to embed external tools not yet provided by our internal services



Platform architecture

- Static front-end (HTML,CSS,Javascript)
- Serverless for easy hosting/transfer/maintanance
- Uses well known technologies as base (jQuery, Bootstrap)
- Add more advanced tools (Plotly and others) on top
- <http://boringtechnology.club/>



Prototype web service

Hydration (data sources) -

- local JSON primarily
- External REST API possible

Format:

- closer to 2D – table like
- Dictionary of arrays, also arrays of dictionaries

```
{  
  "authors": [  
    "Ezera Regina",  
    "Bels Alberts",  
    "Deglavs Augusts" ]}]
```



Further development

- Automate the data migration
- Add more external data sources
- Possibility of running Jupyter Notebooks in browser (serverless)
- Additional export options



Thank you!

Find us:

dh@lnb.lv

