We encourage students to use data in their projects

Crash Course on Data Analytics for Students of Social Sciences and Humanities

Barbora Hladká hladka@ufal.mff.cuni.cz
Charles University

Mapping the Scenes:
Digital Humanities in Cultural Studies in Central and Eastern Europe
May 19 2022, Prague

Invitation from Ondřej Tichý to `czadh@lists.digitalhumanities.org`

As an expert on Digital humanities methods, your role to the workshop would consist in:

1. Presentation on your expertise and methods on DH
2. Possibly commenting on students' projects and advising them on use of DH in their research

Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University

https://ufal.mff.cuni.cz

linguistic research

machine learning research

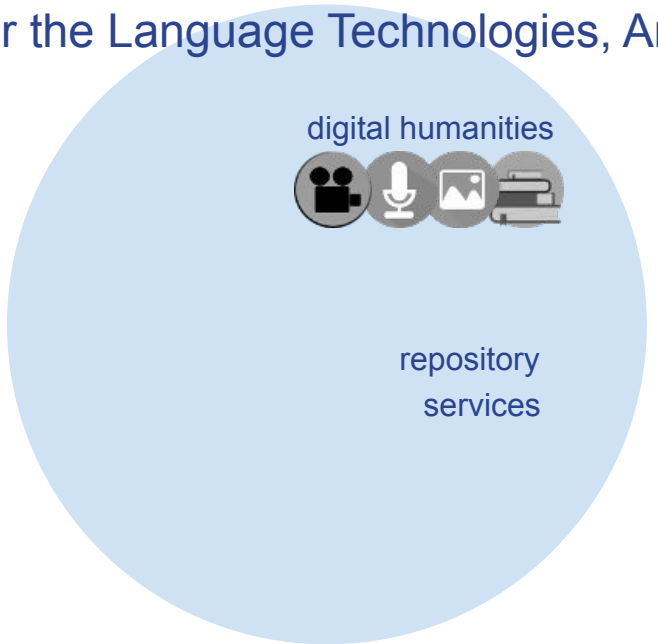
creating language resources

developing NLP tools

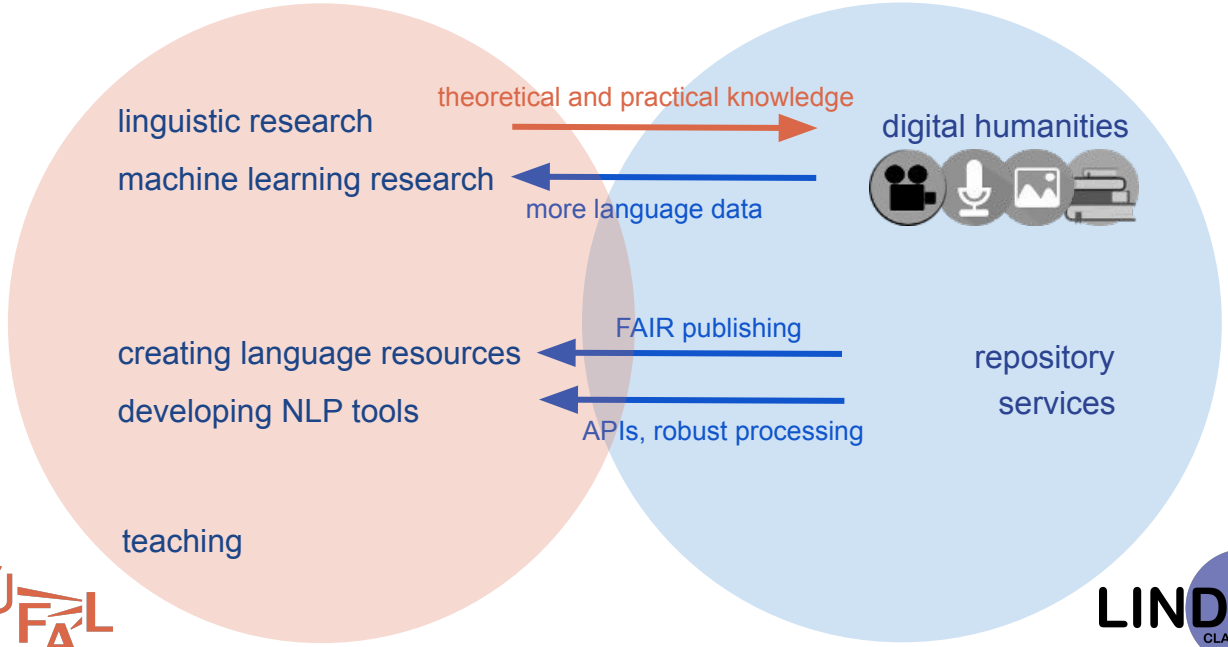
teaching

LINDAT/CLARIAH-CZ https://lindat.cz

Digital Research Infrastructure for the Language Technologies, Arts and Humanities



Synergy between ÚFAL and LINDAT



Invitation from Ondřej Tichý to `czadh@lists.digitalhumanities.org`

As an expert on Digital humanities methods, your role to the workshop would consist in:

1. Presentation on your expertise and methods on DH
2. Possibly commenting on students' projects and advising them on use of DH in their research

Data Analytics for Students of Social Studies and Humanities

Data Analytics for Students of Social Studies and Humanities - NPFL134			
Czech title:	Data Analytics for Students of Social Studies and Humanities	Guarantor:	doc. Mgr. Barbora Vidová Hladká, Ph.D.
Guaranteed by:	Institute of Formal and Applied Linguistics (207. • 32-UFAL)		
Faculty:	Faculty of Mathematics and Physics		
Actual:	from 2021		
Semester:	summer		
E-Credits:	3		
Hours per week, examination:	summer s.:0/2 C [hours/week]		
Capacity:	unlimited		
Min. number of students:	unlimited		
Virtual mobility / capacity:	yes / 25		
Key competences:	data literacy, 4EU+ Flagship 3		
State of the course:	taught		
Language:	English		
Teaching methods:	distance		
Additional information:	https://ufal.mff.cuni.cz/courses/npfl134		

- 3 E-Credits
- 6 mandatory homework assignments

https://ufal.mff.cuni.cz/courses/npfl134, Youtube channel

Lecturers

- Charles University
  - Silvie Cinková, Martin Hájek, Barbora Hladká, Jiří Mírovský
- Sorbonne University
  - Sylvie Archaimbault
- University of Warsaw
  - Jana Plaňavová Latanowicz

Multi* course

- multilingual
  - English, Czech, Polish, French
- multidisciplinary
  - archival research (SU)
  - computational linguistics (CU)
  - sociology (CU)
  - law (UW)

Aim of the course

This course is a gentle, programming-free combination of lectures and practical demonstrations of real-life data workflows in various Social Studies and Humanities (SSH) research areas. It aims at motivating the SSH students to improve their digital literacy in more advanced data analytics courses.

This course does not require any prior data analysis or computer science experience. All you need to get started is basic computer literacy.

Data lifecycle

1. Gathering data
2. Analysing data
3. Annotating (labeling) data
4. Licensing data
5. Sharing data

Data :: André Mazon's correspondence archive

André Mazon (7.7.1881-13.7.1967)
French slavist, Slavic literature,
Russian classic literature, Czech and
Russian philology, and Slavic folklore

- data set — digitized documents = images + metadata
- credit — Center for Slavic Studies, Sorbonne University

Data :: Migrants' stories

- data set
  1,081 short migrants' stories publishe
  at i am a migrant
- credit
  International Organization
  for Migration (Media and
  Communications Division)

I lived in Senegal until I was 17. I miss the mosque's calls five times a day. It is like the bells of a church, it shapes your daily life. I also miss the smooth, stress-free way of living. You can visit a friend without having to coordinate. It's pretty spontaneous.

After studying German for nine months, I moved to Berlin to study Political Science and Law. I really like the Christmas market here in Berlin, and I miss it when I am abroad during the holiday season. However, sometimes cultural differences lead to funny situations. For example, in my country, when women gain weight they are happy about it. If you tell them that they have grown fatter, it is a compliment. In the first year of university my flatmate put on some weight and it was beautiful. I remember telling her, "Oh! That is lovely, you gained some weight." I didn't realise at first, but she wasn't very happy with what I thought had been a compliment.

metadata: motto, name, current country, country of origin, gender, GDP

Data :: Titanic dataset

Filter

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	1	0	3	Braund, Mr. Owen Harris	male	22.00	1	0	A/5 21171	7.2500		S
2	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.00	1	0	PC 17599	71.2833	C85	C
3	3	1	3	Heikkinen, Miss. Laina	female	26.00	0	0	STON/O2. 3101282	7.9250		S
4	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.00	1	0	113803	53.1000	C123	S
5	5	0	3	Allen, Mr. William Henry	male	35.00	0	0	373450	8.0500		S
6	6	0	3	Moran, Mr. James	male	NA	0	0	330877	8.4583		Q
7	7	0	1	McCarthy, Mr. Timothy J	male	54.00	0	0	17463	51.8625	E46	S
8	8	0	3	Palsson, Master. Gosta Leonard	male	2.00	3	1	349909	21.0750		S
9	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.00	0	2	347742	11.1333		S
10	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.00	1	0	237736	30.0708		C

- Each row represents one person
- Columns = metadata about the passengers
- SibSp = the number of a person's siblings and spouse aboard the Titanic
- Parch = the number of a person's parents and children aboard the Titanic
- Cabin = a person's cabin number
- Embarked = a person's port of embarkation

credit Kaggle

Data :: ParlaMint dataset v. 2.1

- ParlaMint is a project of compiling parliamentary debates into uniformly annotated multilingual corpora https://www.clarin.eu/content/parlamint-towards-comparable-parliamentary-corpora
- ParlaMint 2.1 contains corpora of 17 European parliaments

Table 1 Basic information about the ParlaMint corpora including the corpus ID, the covered language(s), the houses and number of terms included, from and to months of included transcriptions, the number of years covered, the number of millions of words per year and in total

ID	Lang	Houses	Ts	From	To	Yrs	Mw/Yr	Mw
BE	nl+fr	Lower	2	2015–11	2020–08	4.8	6.50	31.37
BG	bg	Unicameral	2	2014–10	2020–07	5.8	3.42	20.02
CZ	cs	Lower	2	2013–11	2021–04	7.5	3.03	22.56
DK	da	Unicameral	–	2014–10	2020–09	6.1	4.85	29.40
ES	es	Lower	5	2015–01	2020–12	6.0	2.19	13.10
FR	fr	Lower	1	2017–07	2020–07	3.0	10.75	32.73
GB	en	Lower+Upper	4	2015–01	2021–03	6.3	17.25	109.30
HR	hr	Unicameral	1	2016–11	2020–05	3.6	5.81	20.65
HU	hu	Unicameral	2	2014–05	2020–12	6.7	0.13	0.87
IS	is	Unicameral	3	2015–01	2020–09	5.8	4.06	23.66
IT	it	Upper	2	2013–03	2020–11	7.8	3.46	26.94
LT	lt	Unicameral	2	2012–11	2020–11	8.1	1.82	14.78
LV	lv	Unicameral	2	2014–11	2021–02	6.3	1.02	6.48
NL	nl	Lower+Upper	5	2014–04	2020–11	6.6	7.74	51.45
PL	pl	Lower+Upper	4	2015–11	2020–08	4.9	5.66	27.45
SI	sl	Lower	2	2014–08	2020–07	6.0	3.34	20.19
TR	tr	Unicameral	4	2009–04	2021–02	12.0	3.65	43.99

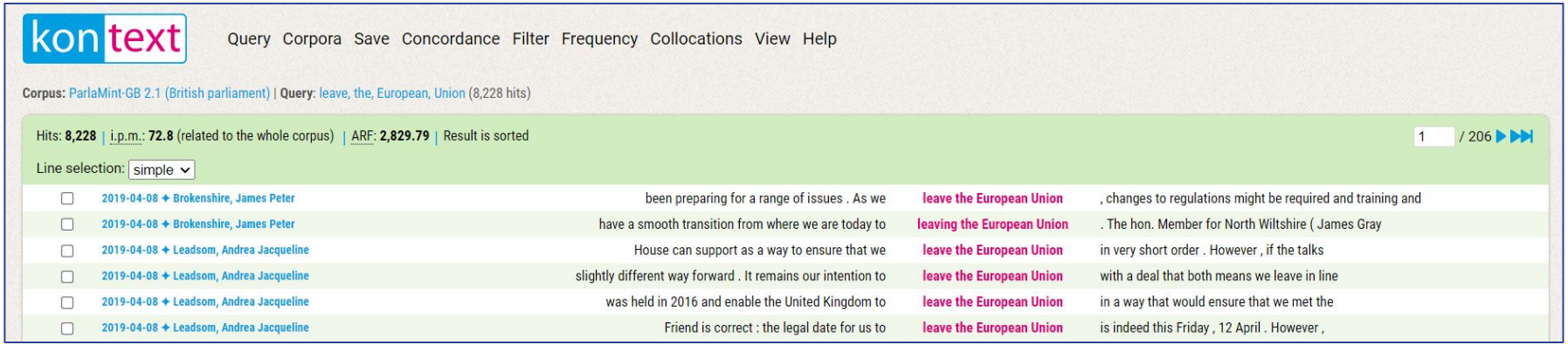
Source: (Erjavec, T., Ogrodniczuk, M., Osenova, P. et al., 2022)

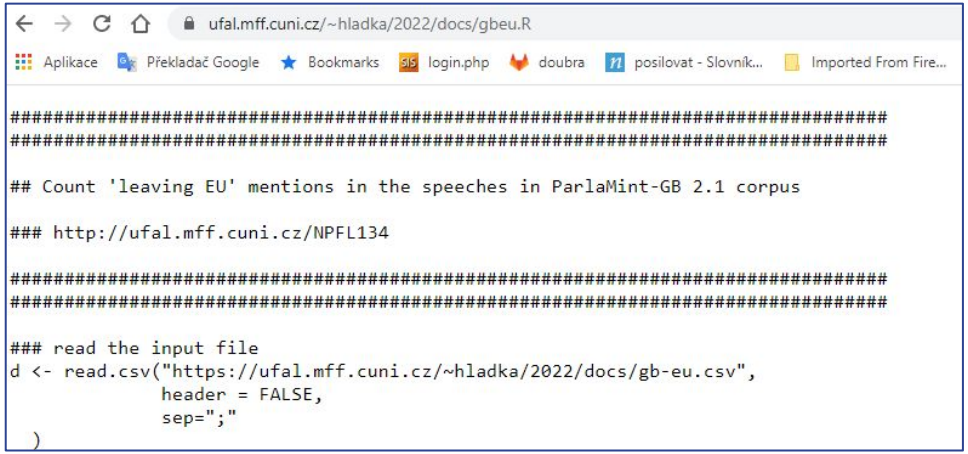
Tools

Analysis and visualization	Tableau
Search	TEITOK, KonText
Manual annotation	Brat
Linguistic processing	UDPipe
Handwritten Text Recognition	Transcribus, Pero

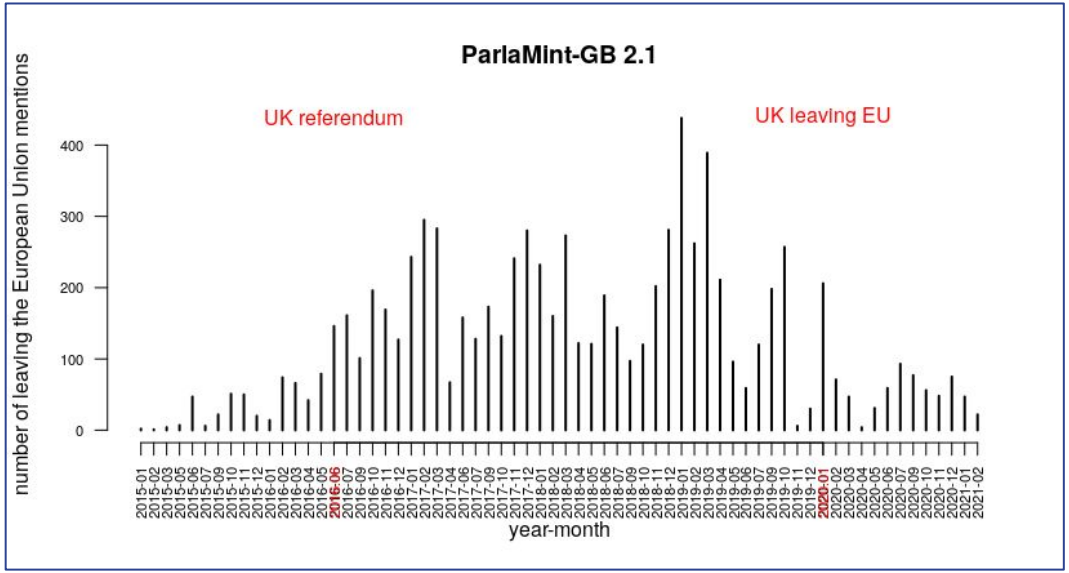
Some programming eventually

- Data ParlaMint-GB 2.1 (British parliament)
- Task 1 How many times did the speakers speak about leaving the European Union in their speeches? Examples:
  - As we leave the European Union, changes to regulations might be required and …
  - … that we have a smooth transition from where we are today to leaving the European Union
  - to be able to have its own free trade policy once we have left the European Union
- Taks 2 How did the overall frequency of the mentions change over time?
- Tools KonText search and R programming


kontext
Query Corpora Save Concordance Filter Frequency Collocations View Help
Corpus: ParlaMint-GB 2.1 (British parliament) | Query: leave, the, European, Union (8,228 hits)
Hits: 8,228 | i.p.m.: 72.8 (related to the whole corpus) | ARF: 2,829.79 | Result is sorted
1 / 206
Line selection: simple
2019-04-08 ✦ Brokenshire, James Peter been preparing for a range of issues . As we leave the European Union , changes to regulations might be required and training and
2019-04-08 ✦ Brokenshire, James Peter have a smooth transition from where we are today to leaving the European Union . The hon. Member for North Wiltshire (James Gray
2019-04-08 ✦ Leadsom, Andrea Jacqueline House can support as a way to ensure that we leave the European Union in very short order . However , if the talks
2019-04-08 ✦ Leadsom, Andrea Jacqueline slightly different way forward . It remains our intention to leave the European Union with a deal that both means we leave in line
2019-04-08 ✦ Leadsom, Andrea Jacqueline was held in 2016 and enable the United Kingdom to leave the European Union in a way that would ensure that we met the
2019-04-08 ✦ Leadsom, Andrea Jacqueline Friend is correct : the legal date for us to leave the European Union is indeed this Friday , 12 April . However ,


ufal.mff.cuni.cz/~hladka/2022/docs/gbeu.R
Aplikace
Překladač Google
Bookmarks
login.php
doubra
posilovat - Slovník...
Imported From Fire...
##
##
Count 'leaving EU' mentions in the speeches in ParlaMint-GB 2.1 corpus
http://ufal.mff.cuni.cz/NPFL134
##
##
read the input file
d <- read.csv("https://ufal.mff.cuni.cz/~hladka/2022/docs/gb-eu.csv",
header = FALSE,
sep=";"
)

	A	B
1	speaker	frequency
2	May, Theresa Mary	736
3	Gove, Michael Andrew	121
4	Leadsom, Andrea Jacqueline	116
5	Smith, Julie	113
6	Callanan, Martin	110
7	Davis, David Michael	101
8	Fox, Liam	96
9	Forsyth, Michael Bruce	84
10	Bridges, James	78
11	Cairns, Alun Hugh	78
12	Ahmad, Tariq	70
13	Soubry, Anna Mary	69
14	Cash, William Nigel Paul	68
15	Eustice, Charles George	67



Homework assignments https://ufal.mff.cuni.cz/courses/npfl134/credit

HW #1

Data:	metadata of A. Mazon's correspondence archive
Tool:	Tableau
Instruction:	Explore the data (e.g., Where did the authors write to AM from in different decades?)

HW #2

Data:	documents (= images) from AM's archive
Tool:	Transkribus, Pero
Instruction:	Transcribe Czech documents using Pero and non-Czech ones using Transkribus

HW #3

Data:	AM's correspondence
Tool:	TEITOK
Instruction:	Search AM's correspondence texts in TEITOK

Homework assignments https://ufal.mff.cuni.cz/courses/npfl134/credit

- **HW #4**

  Instruction: (1) Explore LINDAT repository https://lindat.cz/repository
  (2) Train LINDAT submission procedure form

- **HW #5**

  Data: EU regulation 2020/2092
  Tool: Brat https://quest.ms.mff.cuni.cz/brat/npfl134_2/index.xhtml#
  Instruction: Annotate subjects in the sentences in the regulation

- **HW #6**

  Data: Migrants' stories https://tinyurl.com/26vpzrj6
  Tool: Voyant https://voyant-tools.org/
  Instruction: Carry out your own analysis of the data.
  Use Voyant to explore similarities and differences between groups of stories.

Workshop :: a follow up to the course

- June 15-17, 2022 in Prague (Wed-Fri)
- programme
  - course evaluation + practical lab experience + invited lectures
- https://ufal.mff.cuni.cz/courses/npfl134/workshop
- workshop participants are not required to take the course