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

- Linguistic research
- Machine learning research
- Creating language resources
- Developing NLP tools
- Teaching

More language data
FAIR publishing
APIs, robust processing

Theoretical and practical knowledge
Digital humanities
Repository services

Teaching
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

Invitation from Ondřej Tichý to czadh@lists.digitalhumanities.org
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
- **Guaranteed by:** Institute of Formal and Applied Linguistics (207. • 32-UFA)
- **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
- **Guarantor:** doc. Mgr. Barbora Vidovalová Hladká, Ph.D.

- **3 E-Credits**
- **6 mandatory homework assignments**

[https://ufal.mff.cuni.cz/courses/npfl134](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 published 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.
## Data :: Titanic dataset

<table>
<thead>
<tr>
<th>PassengerId</th>
<th>Survived</th>
<th>Pclass</th>
<th>Name</th>
<th>Sex</th>
<th>Age</th>
<th>SibSp</th>
<th>Parch</th>
<th>Ticket</th>
<th>Fare</th>
<th>Cabin</th>
<th>Embarked</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
<td>Braund, Mr. Owen Harris</td>
<td>male</td>
<td>22</td>
<td>1</td>
<td>0</td>
<td>A/5 21171</td>
<td>7.2500</td>
<td></td>
<td>S</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>Cumings, Mrs. John Bradley (Florence Briggs Thayer)</td>
<td>female</td>
<td>38</td>
<td>1</td>
<td>0</td>
<td>PC 17599</td>
<td>71.2833</td>
<td>C85</td>
<td>C</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>Heikkinen, Miss. Laina</td>
<td>female</td>
<td>26</td>
<td>0</td>
<td>0</td>
<td>S/TON/O2 3101282</td>
<td>7.9250</td>
<td></td>
<td>S</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>1</td>
<td>Futrelle, Mrs. Jacques Heath (Lily May Peel)</td>
<td>female</td>
<td>35</td>
<td>1</td>
<td>0</td>
<td>113803</td>
<td>53.1000</td>
<td>C123</td>
<td>S</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>0</td>
<td>Allen, Mr. William Henry</td>
<td>male</td>
<td>35</td>
<td>0</td>
<td>0</td>
<td>373450</td>
<td>8.0500</td>
<td></td>
<td>S</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>3</td>
<td>Moran, Mr. James</td>
<td>male</td>
<td>NA</td>
<td>0</td>
<td>0</td>
<td>330877</td>
<td>8.4583</td>
<td></td>
<td>Q</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>1</td>
<td>McCarthy, Mr. Timothy J</td>
<td>male</td>
<td>54</td>
<td>0</td>
<td>0</td>
<td>11434</td>
<td>51.6625</td>
<td>E46</td>
<td>S</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>0</td>
<td>Palsson, Master, Gusta Leonard</td>
<td>male</td>
<td>2.00</td>
<td>3</td>
<td>1</td>
<td>349909</td>
<td>21.0750</td>
<td></td>
<td>S</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>1</td>
<td>Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)</td>
<td>female</td>
<td>27</td>
<td>0</td>
<td>2</td>
<td>347742</td>
<td>11.1333</td>
<td></td>
<td>S</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>2</td>
<td>Nasser, Mrs. Nicholas (Adèle Achem)</td>
<td>female</td>
<td>14</td>
<td>1</td>
<td>0</td>
<td>237736</td>
<td>30.0708</td>
<td></td>
<td>C</td>
</tr>
</tbody>
</table>

- 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

<table>
<thead>
<tr>
<th>ID</th>
<th>Lang</th>
<th>Houses</th>
<th>Ts</th>
<th>From</th>
<th>To</th>
<th>Yrs</th>
<th>Mw/yr</th>
<th>Mw</th>
</tr>
</thead>
<tbody>
<tr>
<td>BE</td>
<td>nl+fr</td>
<td>Lower</td>
<td>2</td>
<td>2015-11</td>
<td>2020-08</td>
<td>4.8</td>
<td>6.50</td>
<td>31.37</td>
</tr>
<tr>
<td>BG</td>
<td>bg</td>
<td>Unicameral</td>
<td>2</td>
<td>2014-10</td>
<td>2020-07</td>
<td>5.8</td>
<td>3.42</td>
<td>20.02</td>
</tr>
<tr>
<td>CZ</td>
<td>cs</td>
<td>Lower</td>
<td>2</td>
<td>2013-11</td>
<td>2021-04</td>
<td>7.5</td>
<td>3.03</td>
<td>22.56</td>
</tr>
<tr>
<td>DK</td>
<td>da</td>
<td>Unicameral</td>
<td>-</td>
<td>2014-10</td>
<td>2020-09</td>
<td>6.1</td>
<td>3.85</td>
<td>29.40</td>
</tr>
<tr>
<td>ES</td>
<td>es</td>
<td>Lower</td>
<td>5</td>
<td>2015-01</td>
<td>2020-12</td>
<td>6.0</td>
<td>2.19</td>
<td>13.10</td>
</tr>
<tr>
<td>FR</td>
<td>fr</td>
<td>Lower</td>
<td>1</td>
<td>2017-07</td>
<td>2020-07</td>
<td>3.0</td>
<td>10.75</td>
<td>32.73</td>
</tr>
<tr>
<td>GB</td>
<td>en</td>
<td>Lower+Upper</td>
<td>4</td>
<td>2015-01</td>
<td>2021-03</td>
<td>6.3</td>
<td>17.25</td>
<td>109.30</td>
</tr>
<tr>
<td>HR</td>
<td>hr</td>
<td>Unicameral</td>
<td>1</td>
<td>2016-11</td>
<td>2020-05</td>
<td>3.6</td>
<td>5.81</td>
<td>20.65</td>
</tr>
<tr>
<td>HU</td>
<td>hu</td>
<td>Unicameral</td>
<td>2</td>
<td>2014-05</td>
<td>2020-12</td>
<td>6.7</td>
<td>0.13</td>
<td>0.87</td>
</tr>
<tr>
<td>IS</td>
<td>is</td>
<td>Unicameral</td>
<td>3</td>
<td>2015-01</td>
<td>2020-09</td>
<td>5.8</td>
<td>4.06</td>
<td>23.66</td>
</tr>
<tr>
<td>IT</td>
<td>it</td>
<td>Upper</td>
<td>2</td>
<td>2013-03</td>
<td>2020-11</td>
<td>7.8</td>
<td>3.46</td>
<td>26.94</td>
</tr>
<tr>
<td>LT</td>
<td>lt</td>
<td>Unicameral</td>
<td>2</td>
<td>2012-11</td>
<td>2020-11</td>
<td>8.1</td>
<td>1.82</td>
<td>14.78</td>
</tr>
<tr>
<td>LV</td>
<td>lv</td>
<td>Unicameral</td>
<td>2</td>
<td>2014-11</td>
<td>2021-02</td>
<td>6.3</td>
<td>1.02</td>
<td>6.48</td>
</tr>
<tr>
<td>NL</td>
<td>nl</td>
<td>Lower+Upper</td>
<td>5</td>
<td>2014-04</td>
<td>2020-11</td>
<td>6.6</td>
<td>7.74</td>
<td>51.45</td>
</tr>
<tr>
<td>PL</td>
<td>pl</td>
<td>Lower+Upper</td>
<td>4</td>
<td>2015-11</td>
<td>2020-08</td>
<td>4.9</td>
<td>5.66</td>
<td>27.45</td>
</tr>
<tr>
<td>SI</td>
<td>sl</td>
<td>Lower</td>
<td>2</td>
<td>2014-08</td>
<td>2020-07</td>
<td>6.0</td>
<td>3.34</td>
<td>20.19</td>
</tr>
<tr>
<td>TR</td>
<td>tr</td>
<td>Unicameral</td>
<td>4</td>
<td>2009-04</td>
<td>2021-02</td>
<td>12.0</td>
<td>3.65</td>
<td>43.99</td>
</tr>
</tbody>
</table>

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
<table>
<thead>
<tr>
<th>Date</th>
<th>Speaker, Name</th>
<th>Statement</th>
</tr>
</thead>
<tbody>
<tr>
<td>2019-04-08</td>
<td>Brokenshire, James Peter</td>
<td>have a smooth transition from where we are today to leave the European Union</td>
</tr>
</tbody>
</table>
| 2019-04-08 | Leadson, Andrea Jacqueline| House can support as a way to ensure that we leave the European Union in very short order. However, if the talks slightly different way forward. It remains our intention to leave the European Union with a deal that both means we leave in line was held in 2016 and enable the United Kingdom to leave the European Union in a way that would ensure that we met the Friend is correct: the legal date for us to leave the European Union is indeed this Friday, 12 April. However, changes to regulations might be required and training and
<table>
<thead>
<tr>
<th>speaker</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>May, Theresa Mary</td>
<td>736</td>
</tr>
<tr>
<td>Gove, Michael Andrew</td>
<td>121</td>
</tr>
<tr>
<td>Leadsom, Andrea Jacqueline</td>
<td>116</td>
</tr>
<tr>
<td>Smith, Julie</td>
<td>113</td>
</tr>
<tr>
<td>Callanan, Martin</td>
<td>110</td>
</tr>
<tr>
<td>Davis, David Michael</td>
<td>101</td>
</tr>
<tr>
<td>Fox, Liam</td>
<td>96</td>
</tr>
<tr>
<td>Forsyth, Michael Bruce</td>
<td>84</td>
</tr>
<tr>
<td>Bridges, James</td>
<td>78</td>
</tr>
<tr>
<td>Cairns, Alun Hugh</td>
<td>78</td>
</tr>
<tr>
<td>Ahmad, Tariq</td>
<td>70</td>
</tr>
<tr>
<td>Soubry, Anna Mary</td>
<td>69</td>
</tr>
<tr>
<td>Cash, William Nigel Paul</td>
<td>68</td>
</tr>
<tr>
<td>Eustice, Charles George</td>
<td>67</td>
</tr>
</tbody>
</table>
Homework assignments [https://ufal.mff.cuni.cz/courses/npfl134/credit](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](https://ufal.mff.cuni.cz/courses/npfl134/credit)

- **HW #4**
  Instruction: (1) Explore LINDAT repository [https://lindat.cz/repository](https://lindat.cz/repository)
  (2) Train LINDAT submission procedure form

- **HW #5**
  Data: EU regulation 2020/2092
  Instruction: Annotate subjects in the sentences in the regulation

- **HW #6**
  Data: Migrants’ stories [https://tinyurl.com/26vpzrj6](https://tinyurl.com/26vpzrj6)
  Tool: Voyant [https://voyant-tools.org/](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
- workshop participants are not required to take the course