Introduction
Class #1, Feb 15 2022
Barbora Hladká   hladka@ufal.mff.cuni.cz
# Course organization

**Data Analytics for Students of Social Studies and Humanities - NPFL134**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech title</td>
<td>Data Analytics for Students of Social Studies and Humanities</td>
</tr>
<tr>
<td>Guaranteed by</td>
<td>Institute of Formal and Applied Linguistics (207. • 32-UFAL)</td>
</tr>
<tr>
<td>Faculty</td>
<td>Faculty of Mathematics and Physics</td>
</tr>
<tr>
<td>Actual</td>
<td>from 2021</td>
</tr>
<tr>
<td>Semester</td>
<td>summer</td>
</tr>
<tr>
<td>E-Credits</td>
<td>3</td>
</tr>
<tr>
<td>Hours per week, examination</td>
<td>summer s.:0/2 C [hours/week]</td>
</tr>
<tr>
<td>Capacity</td>
<td>unlimited</td>
</tr>
<tr>
<td>Min. number of students</td>
<td>unlimited</td>
</tr>
<tr>
<td>Virtual mobility / capacity</td>
<td>yes / 25</td>
</tr>
<tr>
<td>Key competences</td>
<td>data literacy, 4EU+ Flagship 3</td>
</tr>
<tr>
<td>State of the course</td>
<td>taught</td>
</tr>
<tr>
<td>Language</td>
<td>English</td>
</tr>
<tr>
<td>Teaching methods</td>
<td>distance</td>
</tr>
<tr>
<td>Additional information</td>
<td><a href="https://ufal.mff.cuni.cz/courses/npfl134">https://ufal.mff.cuni.cz/courses/npfl134</a></td>
</tr>
</tbody>
</table>

- on Tuesdays 10:40-12:10
- 14 classes
- start Feb 15
- finish May 17
- on-line via Zoom
Course organization :: Lecturers

- Charles University
  - Silvie Cinková, Martin Hájek, Barbora Hladká, Jiří Mírovský
- Sorbonne University
  - Sylvie Archaimbault, Cédric Pernette, Irina Tcherneva
- University of Warsaw
  - Jana Plaňavová Latanowicz
Course organization :: E-Credits

- 3 E-Credits
- 6 mandatory homework assignments must be completed and approved by the deadlines specified by HW assigners
  - https://ufal.mff.cuni.cz/courses/npfl134/credit
  - assignment dates already known
Course organization :: TODO list

- [https://ufal.mff.cuni.cz/courses/npfl134/todo-list](https://ufal.mff.cuni.cz/courses/npfl134/todo-list)
- by Feb 22
  - contact [hladka@ufal.mff.cuni.cz](mailto:hladka@ufal.mff.cuni.cz) in case of any questions
Workshop :: a follow up to the course

- June 15-17, 2022 in Prague (Wed-Fri)
- programme
  - course evaluation + practical lab experience + invited lectures
  - more details later
- capacity: 20 students = 10CU + 5SU + 5UW
  - CU students, contact: hladka@ufal.mff.cuni.cz
  - funding for
    - 5 students from SU, contact Sylvie.Archaibault@paris-sorbonne.fr
    - 5 students from UW, contact j.planavova-latanowicz@uw.edu.pl
- workshop participants are not required to take the course
Multi* course/workshop

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

Encourage students to use data in their projects.
Data

= information in digital form for computer processing
  ▪ text, audio, video, image, software
  ▪ born-digital = originate in a digital form
    ▪ e.g. e-books, digital sound and video recordings
  ▪ digital reformatting = analog → digital
    ▪ e.g. scanning physical paper records
Data set

is a set of existing data that could be used to answer research questions and/or provide further evidence relevant to ongoing research questions.
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

André Mazon’s correspondence archive

- **credit**
  - Center for Slavic Studies, Sorbonne University

- **real world objects**
  - paper correspondence

- **data set**
  - digitized documents = images

more details in Lecture #2
Data :: André Mazon’s correspondence archive
Data :: Migrants’ stories

- credit
  International Organization for Migration (Media and Communications Division)
  I am a migrant

- data set
  migrants’ stories = born-digital texts

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

= data about data
Metadata :: André Mazon’s correspondence archive

document’s author

type of document

language

date

place
"Extremists are trying to destroy what took centuries to build: a beautiful civilization, beautiful traditions, peace and love."

Abdoulaye

Current Country: Germany
Country of Origin: Senegal
Table

is a way how to organize data using rows and columns

<table>
<thead>
<tr>
<th>1</th>
<th>…</th>
<th>m</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>…</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table :: André Mazon’s correspondence archive

number of columns = number of metadata attributes
number of rows = number of documents in the archive

<table>
<thead>
<tr>
<th>Cote</th>
<th>Date</th>
<th>Type de document</th>
<th>Nb de f.</th>
<th>Langue</th>
<th>Auteur du document</th>
<th>Lieu</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMA.8.7.1</td>
<td>10.03.1908</td>
<td>Lettre manuscrite</td>
<td>1</td>
<td>allemand</td>
<td>Hartmann, Erich</td>
<td>Bautzen</td>
</tr>
<tr>
<td>AMA.8.12.34</td>
<td>24.11.1913</td>
<td>Lettre dactylographiée</td>
<td>1</td>
<td>allemand</td>
<td>Böhme, Erich</td>
<td>Berlin</td>
</tr>
<tr>
<td>AMA.8.13.40</td>
<td>26.04.1914</td>
<td>Carte manuscrite</td>
<td>1</td>
<td>allemand</td>
<td>Irmer, Hermann</td>
<td>Harkov</td>
</tr>
<tr>
<td>AMA.8.13.27</td>
<td>01.04.1914</td>
<td>Lettre manuscrite</td>
<td>1</td>
<td>allemand</td>
<td>Leskien, August</td>
<td>Leipzig</td>
</tr>
<tr>
<td>AMA.8.13.31</td>
<td>06.04.1914</td>
<td>Lettre manuscrite</td>
<td>1</td>
<td>anglais</td>
<td>Minns, Ellis H.</td>
<td>Cambridge</td>
</tr>
<tr>
<td>AMA.8.12.12</td>
<td>04.04.1913</td>
<td>Lettre dactylographiée</td>
<td>1</td>
<td>anglais</td>
<td>Miller, Arthur William Kaye</td>
<td>Londres</td>
</tr>
<tr>
<td>AMA.8.12.25</td>
<td>29.07.1913</td>
<td>Lettre manuscrite et dactylographiée</td>
<td>1</td>
<td>anglais</td>
<td>Miller, Arthur William Kaye</td>
<td>Londres</td>
</tr>
<tr>
<td>AMA.8.15.4</td>
<td>06.05.1919</td>
<td>Lettre manuscrite</td>
<td>2</td>
<td>bulgare</td>
<td>Stoilov, Nikola</td>
<td>indéterminé</td>
</tr>
<tr>
<td>AMA.8.13.37</td>
<td>16.04.1914</td>
<td>Lettre manuscrite</td>
<td>1</td>
<td>français</td>
<td>Demidov, Elim Pavlovich</td>
<td>Athènes</td>
</tr>
</tbody>
</table>
Table :: Titanic data set

- credit
  Kaggle started in 2010 by offering machine learning competitions, e.g. The sinking of the Titanic
- real world objects
  Titanic’s passengers

Source: https://www.kaggle.com/c/titanic/overview
## Table :: 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>22.00</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>38.00</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>26.00</td>
<td>0</td>
<td>0</td>
<td>STON/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>35.00</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>35.00</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>8.4500</td>
<td></td>
<td>Q</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>0</td>
<td>McCarthy, Mr. Timothy J</td>
<td>male</td>
<td>54.00</td>
<td>0</td>
<td>0</td>
<td>114713</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>27.00</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>0</td>
<td>1</td>
<td>Nasser, Mrs. Nicholas (Adele Achem)</td>
<td>female</td>
<td>14.00</td>
<td>1</td>
<td>0</td>
<td>337336</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
Data-driven research I

- Identify a topic of your research
- What data you need/have
- Ask research questions
  - Detail the problem statement
  - Further describe and refine your topic
  - Add focus to the problem statement
  - Guide data sets and data analysis
  - Set context of research
Data-driven research II

- Formulate hypotheses
  - statements that propose expected results (answers to the questions)
  - give insight into research questions
- Analyze the data
  - do they support your hypotheses or not?
- Draw conclusions

---

Data literacy

From Wikipedia, the free encyclopedia

Data literacy is the ability to read, understand, create, and communicate data as information. Much like literacy as a general concept, data literacy focuses on the competencies involved in working with data. It is, however, not similar to the ability to read text since it requires certain skills involving reading and understanding data.[1]

Source: https://en.wikipedia.org/wiki/Data_literacy
Data lifecycle

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

- data are already available, e.g. Titanic dataset
- archival research
  - e.g. (1) digitization of André Mazon’s correspondence → images, (2) transcription of the images to increase accessibility of historical documents (easily read, search for, and use the information they contain)
- survey, e.g. get data about the students attending our course (url)
- interviews, e.g. Migrants’ stories
- collecting data on-line
- ...

more details in Lecture #4
Gathering data :: Scraping data from websites

ParlaMint project
- compiling a collection of parliamentary datasets in a number of languages and in a harmonised format
- for Czech: scraping the source files from the parliamentary website [link](https://ufal.mff.cuni.cz/courses/npfl134)

<table>
<thead>
<tr>
<th>ID</th>
<th>Lang</th>
<th>Houses</th>
<th>Ts</th>
<th>From</th>
<th>To</th>
<th>Yrs</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>
</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>
</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>
</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>
</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>
</tr>
<tr>
<td>FR</td>
<td>fr</td>
<td>Lower</td>
<td>5</td>
<td>2017-07</td>
<td>2020-07</td>
<td>3.0</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>
</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>
</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>
</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>
</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>
</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>
</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>
</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>
</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>
</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>
</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>
</tr>
</tbody>
</table>

Data lifecycle

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

Deeper understanding a task by statistical view on the data

- Data inspection
  - Data and their description
  - Attributes and their values
  - Missing values (treat them carefully)
- Exploratory analysis
- Statistical tests
- Do plotting and summarizing
Analysing data :: Basic data exploratory analysis

Titanic data set

- a subset of 891 passengers
  - 314 females, 577 males
- Did the gender affect the chances of surviving?
  - Yes. The survival rate of females is approximately double of males.
Analysing data :: Data visualisation

Where did the authors write to André Mazon from in different decades?

Visualisation using Tableau
Data lifecycle

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

- Annotating data = adding data to data
- Manual annotation - organize an annotation task
  - task description
  - annotation instructions
  - annotation tool
  - annotators training
  - checking annotations (e.g. inter-annotator agreement)
Annotating data :: Manually

- e.g. annotating attribution in Czech News Server Articles
- motivation journalism, media bias
- data set articles from the iRozhlas news server of Czech Public Radio
Annotating data :: Manually

1. annotate **source** and **signal** of attribution where

\[
\text{attribution} = \text{source} + \text{information} + \text{signal}
\]

Philosopher Damon Young **claims** that this is an escape from the boredom of daily life.

2. classify the source **Philosopher Damon Young** = named official non-political
Annotating data :: Manually

We use the **Brat** editor.

1. Italská ekonomika se vymanila z recese.
2. V prvním čtvrtletí se její HDP zvýšil o 0,2 procenta.
3. Italská ekonomika se v letošním prvním čtvrtletí vymanila z recese.
4. Tamní statistický úřad ISTAT v úterý oznámil, že hrubý domácí produkt se oproti předchozím třem měsíčům zvýšil o 0,2 procenta.
5. Itálie je třetí největší ekonomikou eurozóny po Německu a Francii.
6. Ve třetím i čtvrtém čtvrtletí loňského roku vykázal italský HDP pokles o 0,1 procenta.
7. Ekonomika se tak dostala do recese, která se obvykle definuje jako alespoň dvě čtvrtletí hospodářského poklesu za sebou.
8. ISTAT rovněž oznámil, že míra nezaměstnanosti v Itálii se v březnu snížila na 10,2 procenta z únорových 10,5 procenta.
9. Tato čísla dokazují solidnost a stabilitu italské ekonomiky, uvedl italský ministr hospodářství Giovanni Tria.
10. Hospodářský růst v prvním čtvrtletí překonal očekávání analytiků, kteří podle průzkumu agentury Reuters předpokládali, že HDP se zvýší pouze o 0,1 procenta.
Annotating data :: Automatically

e.g. recognize geographical names and institutions in text

Prague has more than ten major museums, along with numerous theaters, galleries, cinemas, and other historical exhibits. An extensive modern public transportation system connects the city. It is home to a wide range of public and private schools, including Charles University in Prague, the oldest university in Central Europe.
Data lifecycle

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

- Make data available under licence, so that it is clear who owns the data, and on what terms they can be used.
- Make data available under the most open licence possible, unless there is good reason to licence them on a more restrictive basis.

more details in Lecture #6
Data lifecycle

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

Data repository = a digital infrastructure to share data

- i.e. to preserve data and help others to find them

Why to store data in data repositories? Your work is

- visible (e.g. links to citation databases)
- permanently visible (permanent identifiers)
- useful to others. They can
  - reproduce and validate your findings
  - reuse your data and build on top of them

more details in Lecture #7
Summary

- data lifecycle
- details in the coming 13 classes
Analysing data :: Basic data exploratory analysis

Did the port of embarkment affect the chances of surviving?