Intro to Data Analysis for the students of Economic History at FSV UK

Prague, March 24, 2023

Barbora Hladká
hladka@ufal.mff.cuni.cz

Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University

Motivation

Encourage students to use data in their projects.

- Data, metadata, table
- Data lifecycle
- Real world examples
- Natural language processing
- 3 use cases
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 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
Data :: André Mazon’s correspondence archive
Data :: Migrants’ stories

○ credit International Organization for Migration (url)
○ data set born-digital texts = migrants’ stories (url)
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 sets using rows and columns

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

- number of columns = number of attributes
- number of rows = number of documents in the data set

<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>
Data :: Titanic data set

○ credit Kaggle
○ real world objects passengers on board the Titanic

- 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 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)
- interviews, e.g. migrants’ stories
- collecting data on-line
- ...

Gathering data :: Scraping data from websites

- ParlaMint is an European project of compiling parliamentary debates into uniformly annotated multilingual corpora, see [url](#).
- e.g. for Czech: scraping the source files from the parliamentary [website](#).

<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/Year</th>
<th>Mw</th>
</tr>
</thead>
<tbody>
<tr>
<td>BE</td>
<td>nl+fr</td>
<td>Lower</td>
<td>2</td>
<td>2013-11</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>2013-11</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>2014-10</td>
<td>2020-07</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>2</td>
<td>2015-01</td>
<td>2020-12</td>
<td>6.0</td>
<td>2.19</td>
<td>13.10</td>
</tr>
<tr>
<td>ES</td>
<td>es</td>
<td>Lower</td>
<td>1</td>
<td>2015-01</td>
<td>2020-07</td>
<td>3.0</td>
<td>10.75</td>
<td>32.73</td>
</tr>
<tr>
<td>FR</td>
<td>fr</td>
<td>Lower</td>
<td>1</td>
<td>2017-07</td>
<td>2020-07</td>
<td>6.3</td>
<td>17.25</td>
<td>109.30</td>
</tr>
<tr>
<td>GB</td>
<td>en</td>
<td>Lower+Upper</td>
<td>4</td>
<td>2015-01</td>
<td>2020-03</td>
<td>3.6</td>
<td>5.81</td>
<td>20.65</td>
</tr>
<tr>
<td>HR</td>
<td>hr</td>
<td>Unicameral</td>
<td>1</td>
<td>2016-11</td>
<td>2020-05</td>
<td>6.7</td>
<td>0.13</td>
<td>0.87</td>
</tr>
<tr>
<td>HU</td>
<td>hu</td>
<td>Unicameral</td>
<td>2</td>
<td>2014-05</td>
<td>2020-12</td>
<td>8.1</td>
<td>1.82</td>
<td>14.78</td>
</tr>
<tr>
<td>IS</td>
<td>is</td>
<td>Unicameral</td>
<td>3</td>
<td>2015-01</td>
<td>2020-09</td>
<td>8.1</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>2020-03</td>
<td>6.3</td>
<td>1.02</td>
<td>36.08</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., Ogorodniczuk, M., Osenova, P. et al., 2022)
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, outliers, extreme values (treat them carefully)
- Exploratory analysis
- Statistical tests
- Do plotting and summarizing
Analysing data :: Basic data exploratory analysis

Titanic data set

- consider 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 4 times higher than of males (233/314 vs. 109/577 = 74 % vs. 19 %)
Data lifecycle

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

○ Annotating data = adding (meta)data to (meta)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

Annotating attribution in Czech News Server Articles (url)

○ motivation journalism, media bias
○ data set articles from the iRozhlas news server of Czech Public Radio
○ See (Hladká et al, 2020)
Annotating data :: Manually

attribution = source + information + signal

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

Annotation task

1. annotate source and signal of attribution
2. classify the sources Philosopher Damon Young = named official non-political
We used Brat editor.

Annotating data :: Manually
Annotating data :: Automatically

e.g. recognize geographical names and institutions in text using NameTag

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.
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 storing 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
Sharing data :: LINDAT repository

- [https://lindat.cz/](https://lindat.cz/)
- e.g. SiR 1.0
Sharing data :: LINDAT repository

- e.g. Migrant stories
Text mining using Natural Language Processing

- information extraction from unstructured textual data
- key procedures
  - sentence segmentation, tokenization
  - lemmatization, part-of-speech tagging
    - lemmas
      base forms of words
    - Part-of-Speech tags
      morphological features of words
  - parsing
    - dependency
      dependency relations between words
Natural Language Processing :: UDPipe

Searching data

- KonText
  - see Use case Leaving the European Union in the UK Parliament
- TEITOK
  - see Use case Migrant stories timelines
Use cases

1. Directions of migration (url)
2. Migrant story timelines (url)
3. Leaving the European Union in the UK Parliament (url)
References

