Introduction

Class #1, Feb 14 2023

Barbora Hladká hladka@ufal.mff.cuni.cz

Data Analytics for Students of Social Studies and Humanities https://ufal.mff.cuni.cz/courses/npfl134

Course organization

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-UFAL)
Faculty:	Faculty of Mathematics and Physics
Actual:	from 2022
Semester:	summer
E-Credits:	3
Hours per week, examination:	summer s.:0/2, C [HT]
Capacity:	unlimited
Min. number of students:	unlimited
Virtual mobility / capacity:	yes / unlimited
State of the course:	taught
Language:	English
Teaching methods:	combined
Additional information:	https://ufal.mff.cuni.cz/courses/npfl134

Guarantor: doc. Mgr. Barbora Vidová Hladká, Ph.D.

on Tuesdays 12:20-13:50
14 classes
start Feb 14
finish May 16

Course organization :: Lecturers

- Charles University
  - Silvie Cinková, Martin Hájek, Barbora Hladká, Jiří Mírovský
- Sorbonne University
  - Sylvie Archaimbault
- Warsaw University
  - 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 will be known soon

Course organization :: TODO list

- https://ufal.mff.cuni.cz/courses/npfl134/todo-list
- by Feb 21
  - contact cinkova@ufal.mff.cuni.cz in case of any questions

Workshop :: a follow up to the course

- June XX-YY, 2023 in Prague
- programme
  - course evaluation + practical lab experience + invited lectures
  - more details later at https://ufal.mff.cuni.cz/courses/npfl134/workshop-2023
- capacity: XX students
- workshop participants are not required to take the course

Multi* course/workshop

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

Course motivation

We 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.

more details in Lecture #2

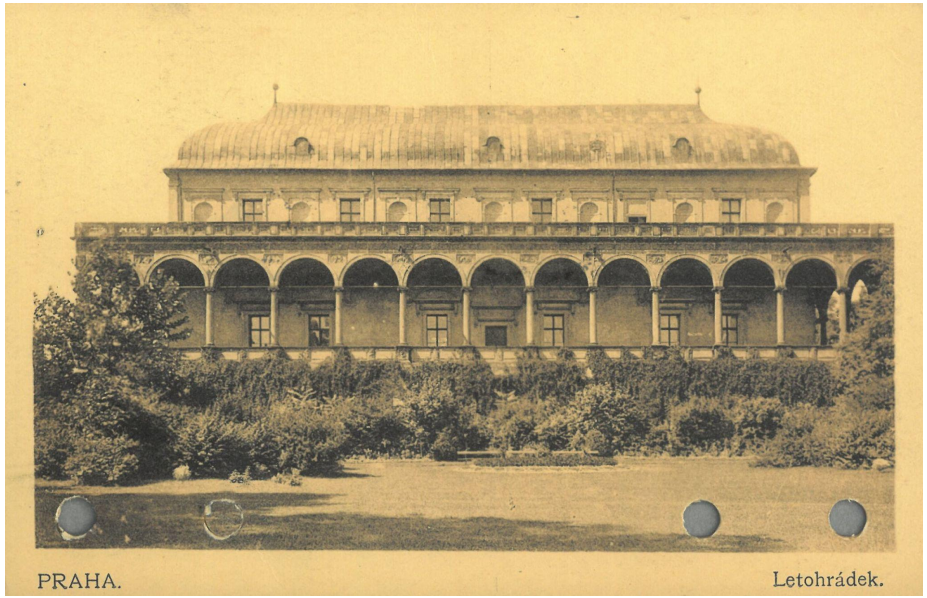
Data :: André Mazon's correspondence archive

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



more details in Lecture #8

Data :: Migrants' stories

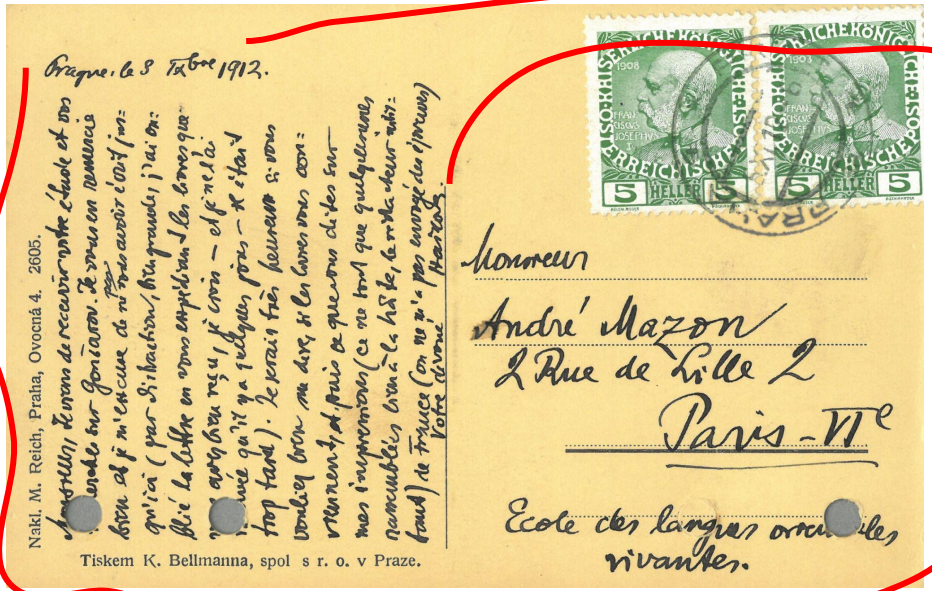
- credit
  International Organization
  for Migration
  (Media and Communications Division)
  i am a migrant
- data set
  migrants' stories = born-digital texts
- see
  https://ufal.mff.cuni.cz/courses/npfl134/migrant-stories

> 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

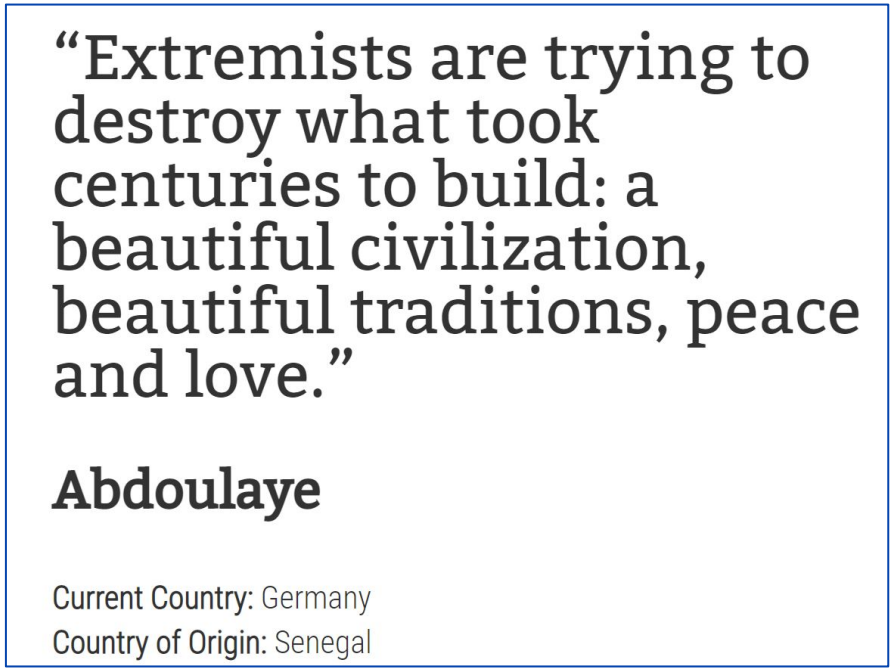
type of document

language

date

place

Metadata :: Migrants' stories



motto

name

current country
country of origin

Table

is a way how to organize data using rows and columns

	1	…		*m*
1				
…				
n				

more details in Lectures #3

Table :: André Mazon's correspondence archive

number of columns = number of metadata attributes

number of rows = number of documents in the archive

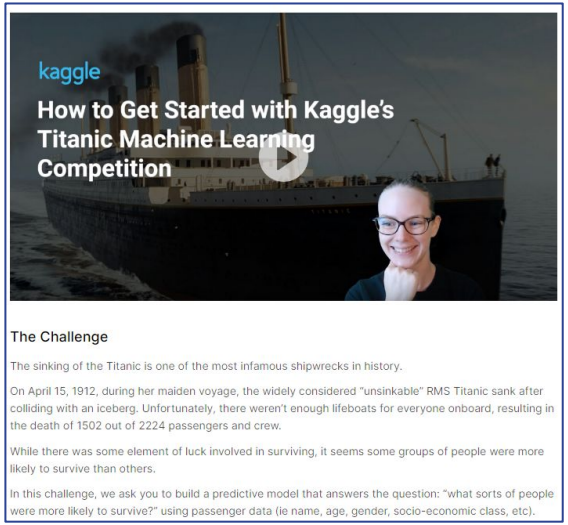
attributes

values

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

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

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

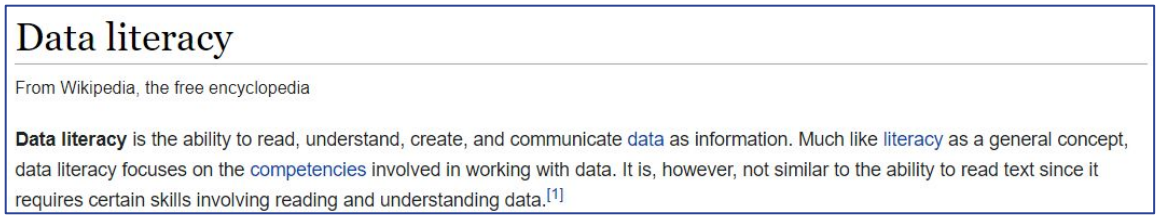
- Each row represents one person
- Columns = metadata about the passengers
- SibSp = the number of a person's siblings and spouse aboard the Titanic
- Parch = the number of a person's parents and children aboard the Titanic
- Cabin = a person's cabin number
- Embarked = a person's port of embarkation

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
- …

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

ID	Lang	Houses	Ts	From	To	Yrs
BE	nl+fr	Lower	2	2015–11	2020–08	4.8
BG	bg	Unicameral	2	2014–10	2020–07	5.8
CZ	cs	Lower	2	2013–11	2021–04	7.5
DK	da	Unicameral	–	2014–10	2020–09	6.1
ES	es	Lower	5	2015–01	2020–12	6.0
FR	fr	Lower	1	2017–07	2020–07	3.0
GB	en	Lower+Upper	4	2015–01	2021–03	6.3
HR	hr	Unicameral	1	2016–11	2020–05	3.6
HU	hu	Unicameral	2	2014–05	2020–12	6.7
IS	is	Unicameral	3	2015–01	2020–09	5.8
IT	it	Upper	2	2013–03	2020–11	7.8
LT	lt	Unicameral	2	2012–11	2020–11	8.1
LV	lv	Unicameral	2	2014–11	2021–02	6.3
NL	nl	Lower+Upper	5	2014–04	2020–11	6.6
PL	pl	Lower+Upper	4	2015–11	2020–08	4.9
SI	sl	Lower	2	2014–08	2020–07	6.0
TR	tr	Unicameral	4	2009–04	2021–02	12.0

Source: https://link.springer.com/article/10.1007/s10579-021-09574-0/

Data lifecycle

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

Analysing data

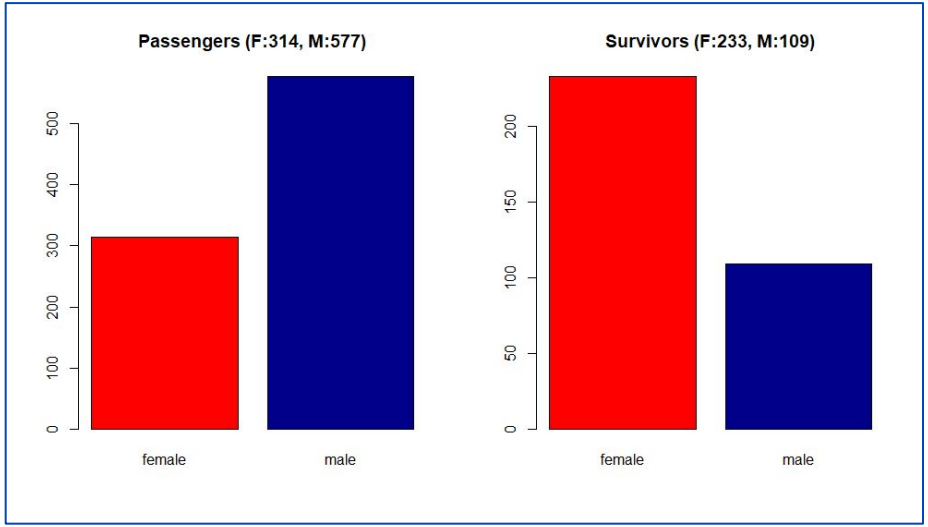
Deeper understanding a task by statistical view on the data

- 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

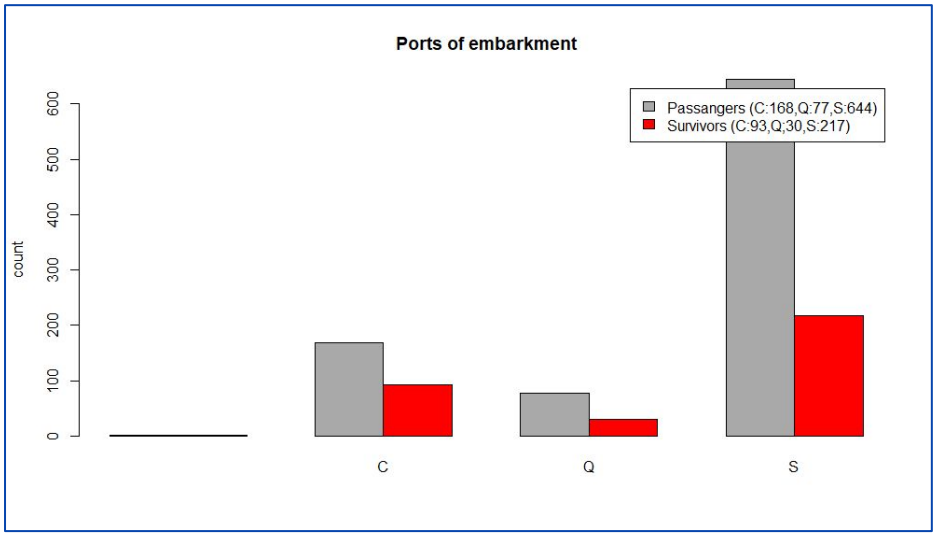
Titanic data set

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



Analysing data :: Basic data exploratory analysis

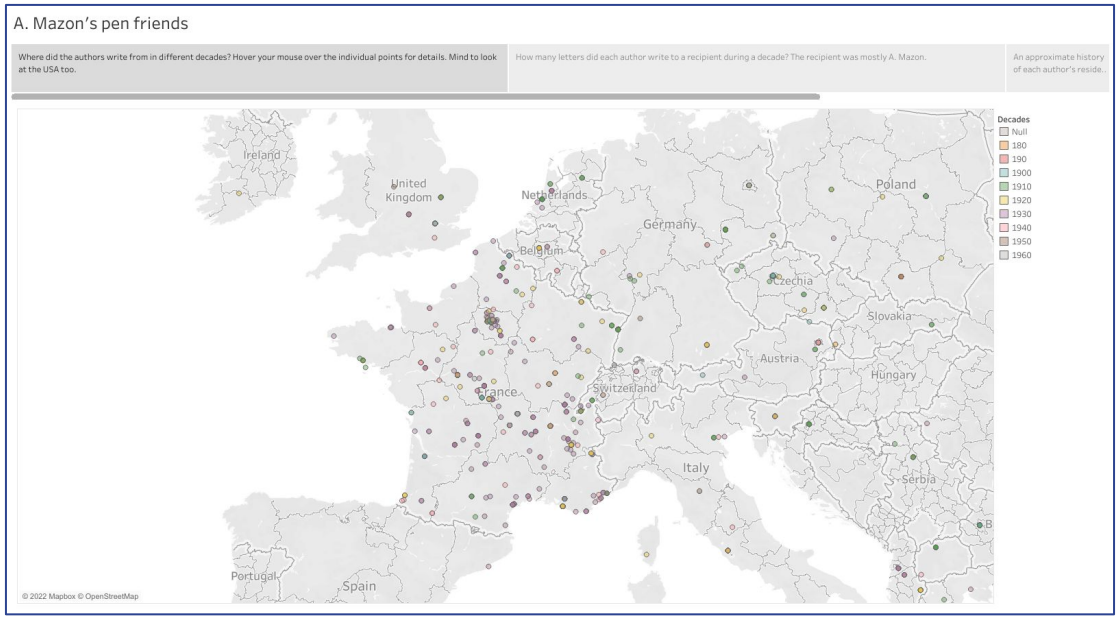
Did the port of embarkment affect the chances of surviving?



Analysing data :: Data visualisation

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

Visualisation using Tableu



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

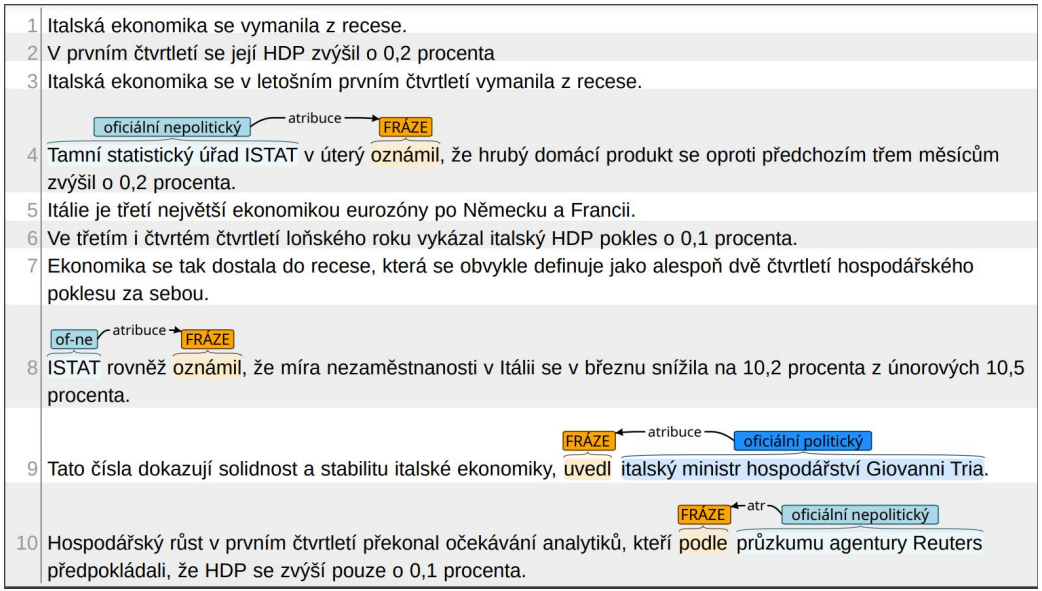
   attribution = source + information + signal

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.



more details in Lecture #7

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.

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

Summary

- data lifecycle
- details in the coming 13 classes