Wrap up

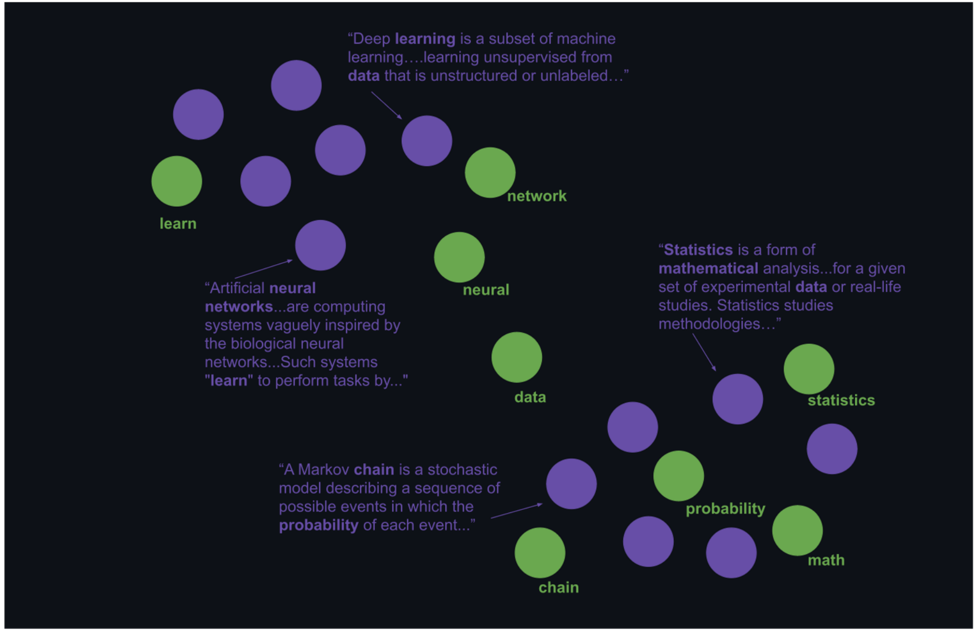
Day 5

Latent Semantic Scaling

- Uses Word Embeddings
- Unidimensional Position of Texts
- You decide what this dimension is using “seed words”

- LSX package in R (Watanabe, 2020)

Top2Vec



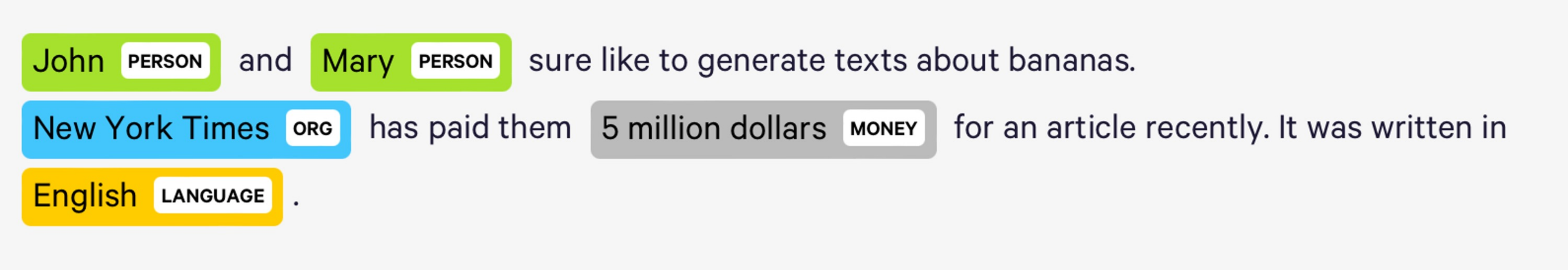
“Deep **learning** is a subset of machine learning….learning unsupervised from **data** that is unstructured or unlabeled…”

“Artificial **neural networks**...are computing systems vaguely inspired by the biological neural networks...Such systems "**learn**" to perform tasks by..."

“**Statistics** is a form of **mathematical** analysis...for a given set of experimental **data** or real-life studies. Statistics studies methodologies…”

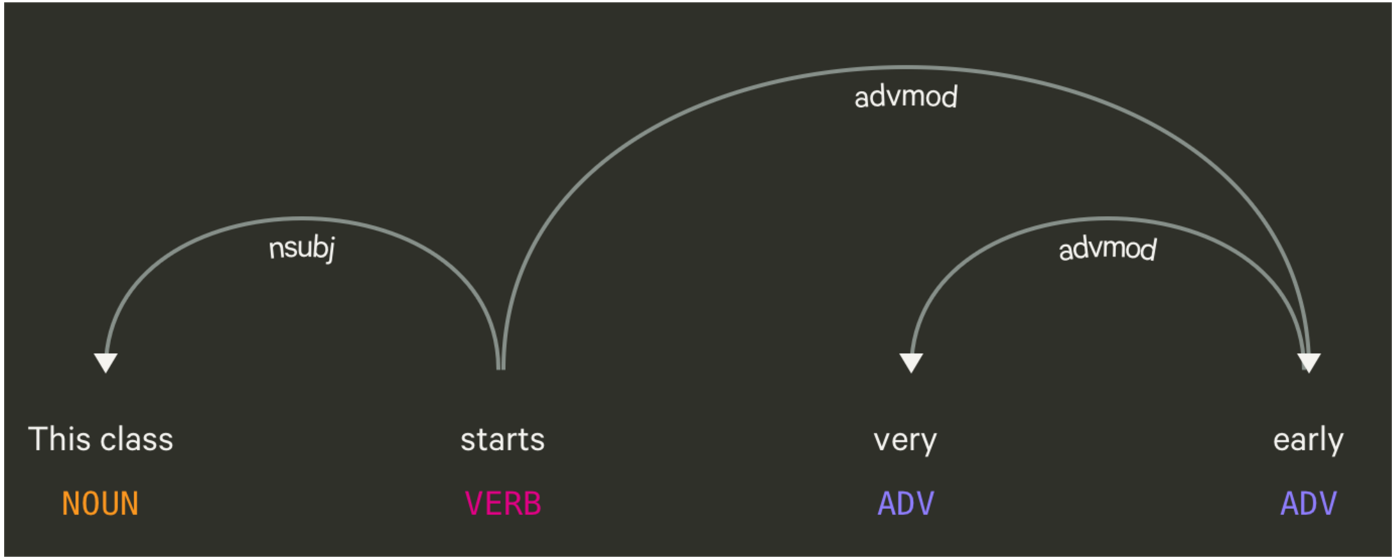
“A Markov **chain** is a stochastic model describing a sequence of possible events in which the **probability** of each event...”

Named Entity Recognition



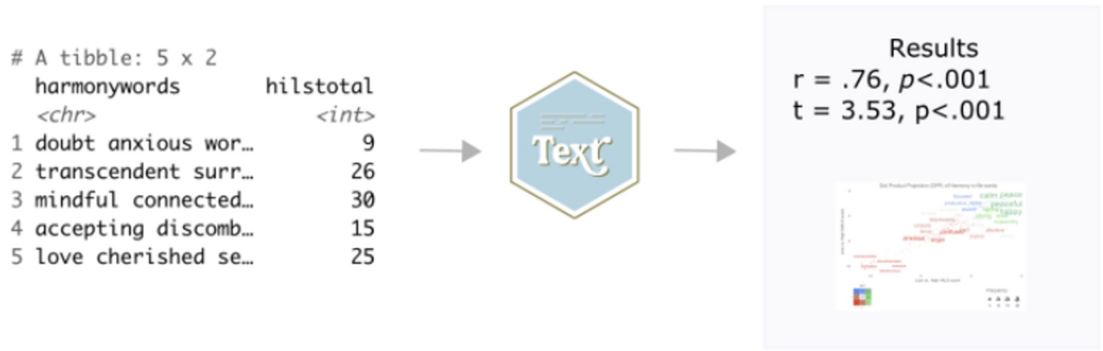
John PERSON and Mary PERSON sure like to generate texts about bananas. New York Times ORG has paid them 5 million dollars MONEY for an article recently. It was written in English LANGUAGE .

Part of Speech Tagging



R package “text”

End-to-End Solution

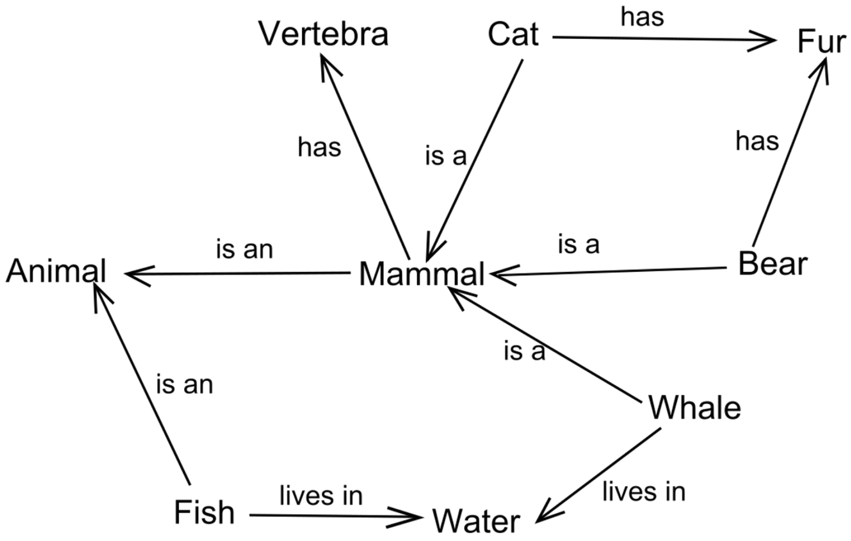


```
# A tibble: 5 x 2
  harmonywords          hilstotal
  <chr>                     <int>
1 doubt anxious wor…            9
2 transcendent surr…           26
3 mindful connected…           30
4 accepting discomb…           15
5 love cherished se…           25
```

Results
r = .76, *p*<.001
t = 3.53, p<.001

https://r-text.org/

Semantic Networks



Optical Character Recognition

- Tesseract
- Cross-language platform for ocr
- .pdf, .png, …

Data Linkage

Connecting text data to (e.g.,) survey data.

spaCy

- Python NLP engine
- Has an r connection (spacyr)

Deep learning

Why use it?

Use python

Karas, pytorch, sklearn, spacy

Case study 1

You want to analyze whether and how the newspaper coverage of political parties change in the election campaign.

Case study 2 Data selection

You plan to study the emotional reactions of Austrian and UK citizens in response to the attacks on work of arts.

What text data could you select?

How would you select only relevant content?

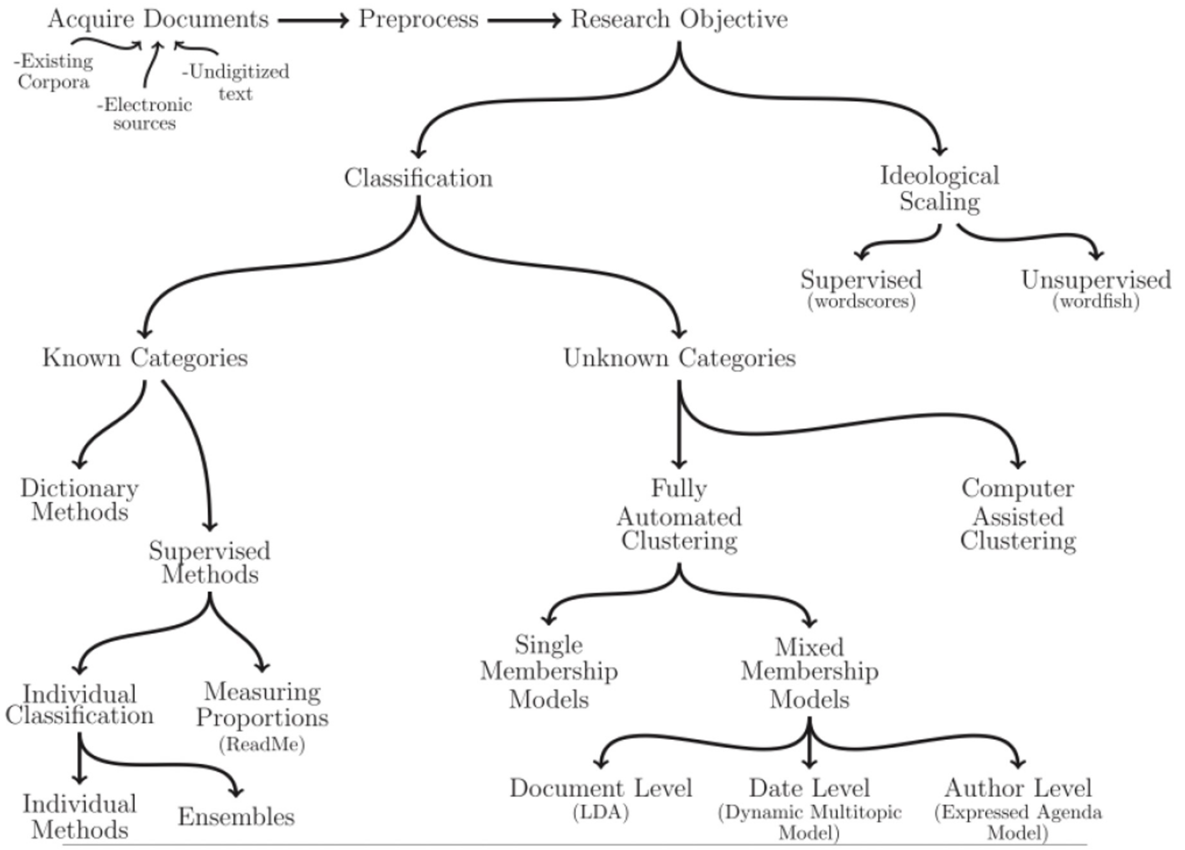
Case study 3 Data labeling

You have a corpus of 1 Mil Tweets all dealing with the Sustainable Development Goals. You like to know which Tweet mentions what specific Sustainability Goal.

What method of automated text analysis would you use to label your full corpus?

Comparing crowdfunding with expert coding, which option would you chose to manually label parts of the corpus?

1 NO POVERTY
2 ZERO HUNGER
3 GOOD HEALTH AND WELL-BEING
4 QUALITY EDUCATION
5 GENDER EQUALITY
6 CLEAN WATER AND SANITATION
7 AFFORDABLE AND CLEAN ENERGY
8 DECENT WORK AND ECONOMIC GROWTH
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
10 REDUCED INEQUALITIES
11 SUSTAINABLE CITIES AND COMMUNITIES
12 RESPONSIBLE CONSUMPTION AND PRODUCTION
13 CLIMATE ACTION
14 LIFE BELOW WATER
15 LIFE ON LAND
16 PEACE, JUSTICE AND STRONG INSTITUTIONS
17 PARTNERSHIPS FOR THE GOALS
SUSTAINABLE DEVELOPMENT GOALS



Grimmer & Stuart, 2013, Fig 1

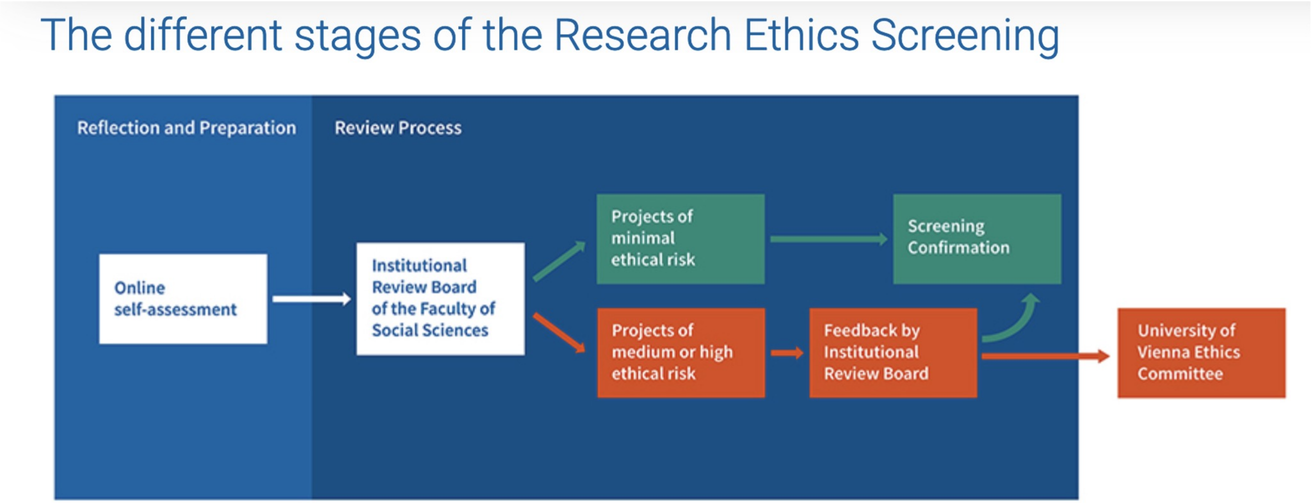
Validation

Human understanding of text as gold standard

Don’t trust numbers trust yourself

And other human coders

Ethics: procedure at UNIVIE



The different stages of the Research Ethics Screening

Reflection and Preparation

Review Process

Online self-assessment

Institutional Review Board of the Faculty of Social Sciences

Projects of minimal ethical risk

Screening Confirmation

Projects of medium or high ethical risk

Feedback by Institutional Review Board

University of Vienna Ethics Committee

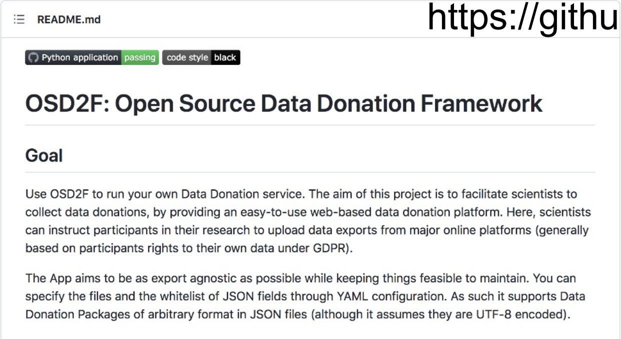
https://sowi.univie.ac.at/en/research/research-ethics/

Ethics: collection of digital trace data

Data donation tools

- OSD2F (Arauju et al., 2022)
- DDM (Pfiffner et al., 2022)

https://github.com/uvacw/osd2f/



README.md

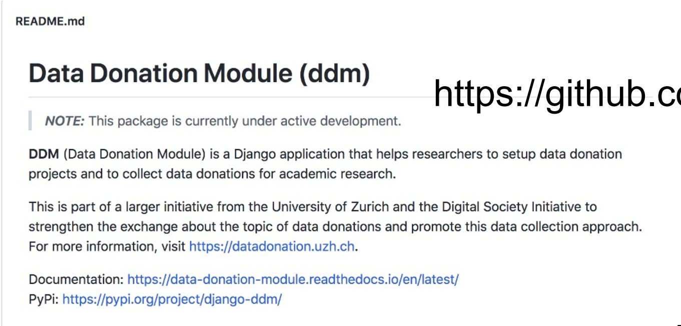
OSD2F: Open Source Data Donation Framework

Goal

Use OSD2F to run your own Data Donation service. The aim of this project is to facilitate scientists to collect data donations, by providing an easy-to-use web-based data donation platform. Here, scientists can instruct participants in their research to upload data exports from major online platforms (generally based on participants rights to their own data under GDPR).

The App aims to be as export agnostic as possible while keeping things feasible to maintain. You can specify the files and the whitelist of JSON fields through YAML configuration. As such it supports Data Donation Packages of arbitrary format in JSON files (although it assumes they are UTF-8 encoded).

https://github.com/uzh/ddm



README.md

Data Donation Module (ddm)

> *NOTE:* This package is currently under active development.

DDM (Data Donation Module) is a Django application that helps researchers to setup data donation projects and to collect data donations for academic research.

This is part of a larger initiative from the University of Zurich and the Digital Society Initiative to strengthen the exchange about the topic of data donations and promote this data collection approach. For more information, visit https://datadonation.uzh.ch.

Documentation: https://data-donation-module.readthedocs.io/en/latest/
PyPi: https://pypi.org/project/django-ddm/

See also: van Driel et al., 2022

Course assessment

Participation in class (20%)

Final paper: application of one or several automated text analysis methods on a topic related to the PhD thesis or a topic of free choice (80%)

- Contents: short motivation, analysis (commented code), description and interpretation of results (about 10 pages)
- Format: R Markdown
- Deadline: January 31st, 2023
- Send it to the two of us via mail

Pitch your projects

Very informal opportunity to pitch your text analysis use case and (initial) design

- Research question
- Data
- Methods
- Current struggles

And to receive some feedback (no grades, points, etc. just free brainstorming opportunity)

Individual feedback

Feedback for us

- Level of difficulty?

Feedback for us

- Coverage of the field (prefer less topics more in depth or even more topics)?

Feedback for us

- Application scenarios in your discipline?

Feedback for us

- Data sources in your discipline?

Feedback for us

- More time for working on coding challenges (without initial guidance)?

Feedback for us

- What could we improve for the class next year?

Thank you very much