COURSE OVERVIEW

CSC 105 - DIGITAL HUMANITIES - SPRINT 23

Marcus Birkenkrahe

January 7, 2023

What are "Digital Humanities" about?

- What do you think "Digital Humanities" are about?
- Compare: TOC of The Digital Humanities Cookbook (Drucker, 2021)
- Compare: Curriculum suggestion by ChatGPT (01-06-2023)

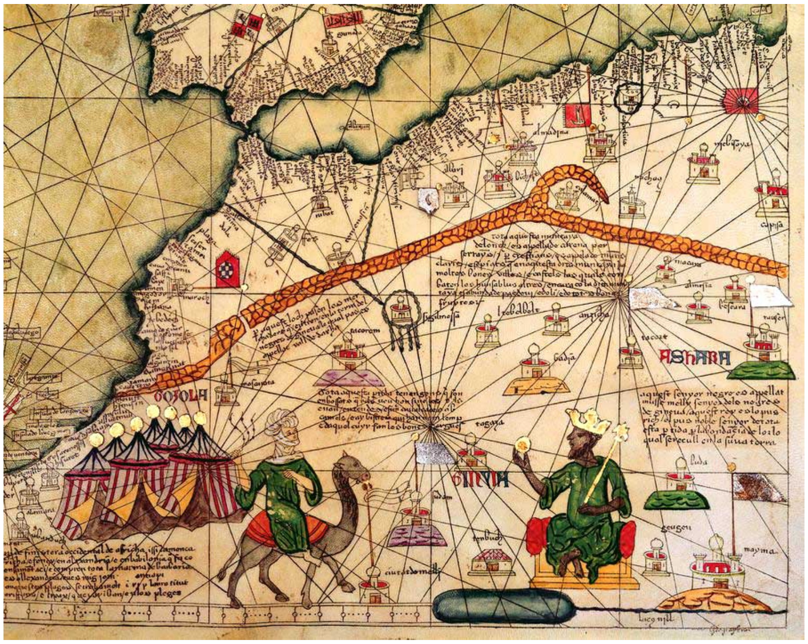


Figure 1: Map of West Africa in the Catalan Atlas (1375)

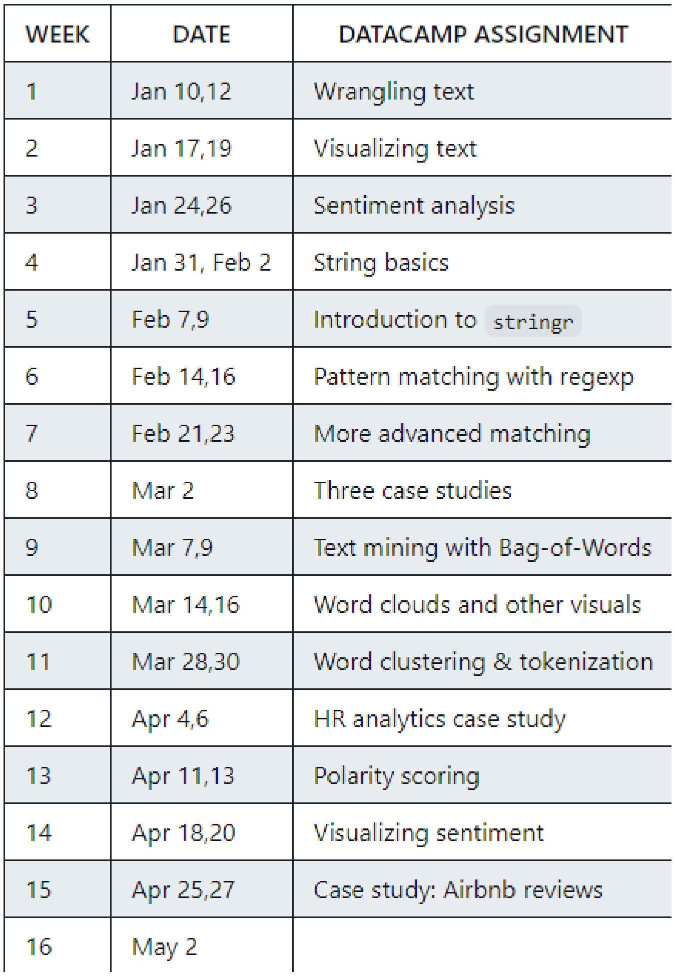
What will we do in this course?

- We're going to work with text data only (no maps[1])
- We will complete 50% of the "Text mining with R" track
- You will complete one DataCamp chapter per fortnight
- In class, we will prepare and review the online material
- You will cover some of the most important AI/ML applications

How will you be evaluated?

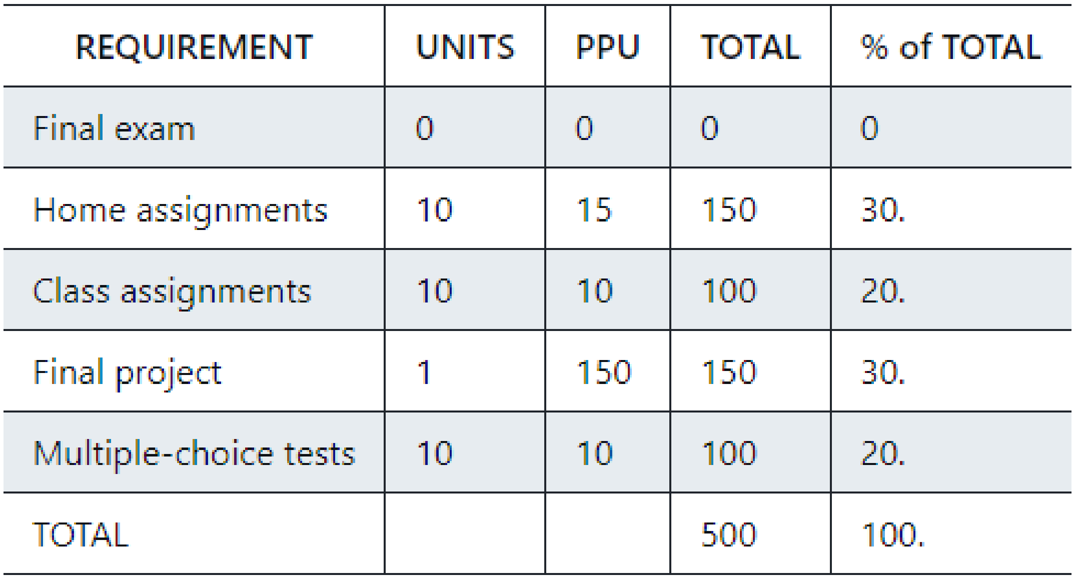
- All course requirements have deadlines
- 8/10 home assignments are DataCamp lessons

[1] Maps and GIS (Geographic Information Systems) are super cool but deserve an introductory course on their own. The focus of my teaching is to get you practical experience that you can use on the job instead of a broad conceptual overview (though the latter would be easier for all of us).



WEEK	DATE	DATACAMP ASSIGNMENT
1	Jan 10,12	Wrangling text
2	Jan 17,19	Visualizing text
3	Jan 24,26	Sentiment analysis
4	Jan 31, Feb 2	String basics
5	Feb 7,9	Introduction to `stringr`
6	Feb 14,16	Pattern matching with regexp
7	Feb 21,23	More advanced matching
8	Mar 2	Three case studies
9	Mar 7,9	Text mining with Bag-of-Words
10	Mar 14,16	Word clouds and other visuals
11	Mar 28,30	Word clustering & tokenization
12	Apr 4,6	HR analytics case study
13	Apr 11,13	Polarity scoring
14	Apr 18,20	Visualizing sentiment
15	Apr 25,27	Case study: Airbnb reviews
16	May 2	

Figure 2: Source: syllabus, Canvas (lyon.instructure.com) or GitHub (github.com/birkenkrahe/ml)



REQUIREMENT	UNITS	PPU	TOTAL	% of TOTAL
Final exam	0	0	0	0
Home assignments	10	15	150	30.
Class assignments	10	10	100	20.
Final project	1	150	150	30.
Multiple-choice tests	10	10	100	20.
TOTAL			500	100.

Figure 3: Source: syllabus, Canvas (lyon.instructure.com) or GitHub (github.com/birkenkrahe/ml)

- Late submissions will be penalized (loss of points)
- No final exam! But weekly tests are graded
- The project topic can come from any of the course sub-topics
- The project deliverable is an essay of at least 5,000 words

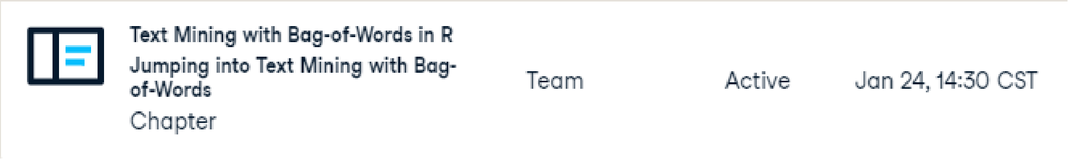
Which tools are you going to use?

- DataCamp courses (15 weekly home assignments)
- GitHub repository (all course materials except tests)
- GNU Emacs + ESS + R (literate programming environment)
- Canvas (learning management system)

How can you register at DataCamp?

- You find the invitation link to the group for Spring 23 in Canvas.

Figure 4: Unsplash



Text Mining with Bag-of-Words in R
Jumping into Text Mining with Bag-of-Words
Chapter

Team

Active

Jan 24, 14:30 CST

Figure 5: DataCamp assignment for January

- You will automatically be subscribed to the Digital Humanities team
- If you are in more than one course, I will add you later manually
- These accounts will be valid until July 8, 2023 only

When is the first assignment due?

- The first DataCamp home assignment is due on January 19. For late submissions, you lose 1 point per day (out of 10 possible points)
- The first in-class assignment is due on January 19. For late submissions, you lose 1 point per day (out of 10 possible points)
- We'll write the first weekly multiple-choice test on January 19.

What should we do as a project?

- The **final essay** should be about one of the areas of Digital Humanities that we have **not** covered in our text mining course
- Here is the set of available essay topics (generated by an AI)
- Also possible: "Topic modeling" (non-assigned DataCamp chapter), full chapter in the book by Kwartler - ML approach with clustering

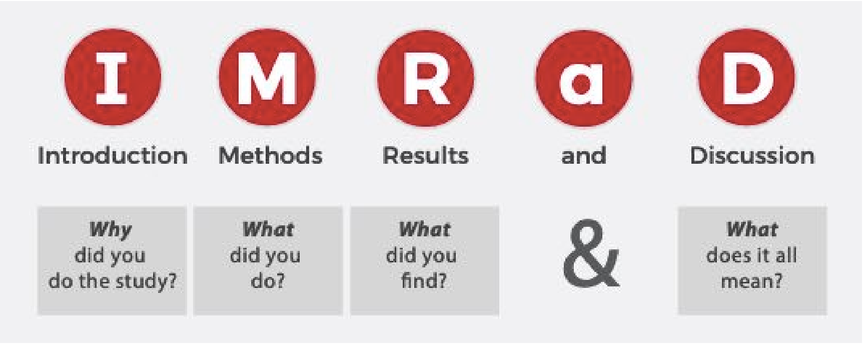
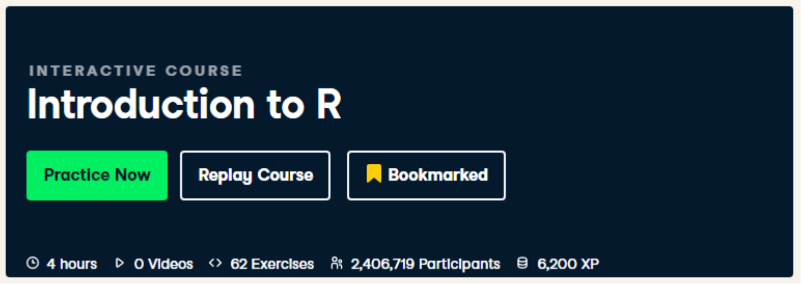


Figure 6: Source: learning.edanz.com

- An interesting approach would be if everyone picked the same topic related to text mining, ChatGPT being an obvious current "hot topic", and investigated different aspects, e.g. technical, ethical, societal, and personal aspects of the chatbot or this type of bot.
- You need to do your own research, including a literature review, and adhere to the IMRaD framework (see video):
  1. Introduction (what did you want to research?) with abstract
  2. Method (what did you do?) with literature review
  3. Results (what did you find out?) with examples, illustrations
  4. Discussion (what does it mean?) with limitations and outlook

What else could you do for a good start?

1. Complete/review introductory R or statistics courses:
   - Introduction to R" in DataCamp (data structures)
   - Introduction to statistics
   - fasteR by Norman Matloff (GitHub) - fast lane to R
2. If you do not have any experience with Emacs, work through the **online tutorial** (open it in Emacs with `CTRL + h t`) - ca. 1 hour.
   - Learn to open/close the editor
   - Learn basic cursor control (moving around)



INTERACTIVE COURSE

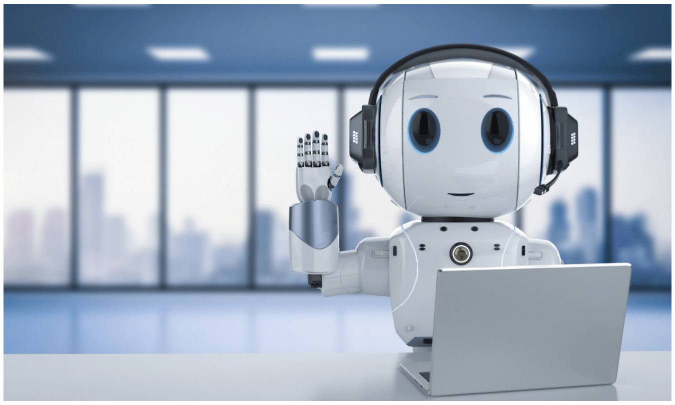
Introduction to R

4 hours · 0 Videos · 62 Exercises · 2,406,719 Participants · 6,200 XP

Figure 7: DataCamp course dashboard banner

- Learn basic file management (open/close/find/save files)
- Learn basic windows (buffer) management

What are you looking forward to?



- Learning more about text mining using the `tidyverse`
- Unlocking the secrets of natural language processing
- Having fun with R programming and real data sets
- Helping you on your own "digital humanities" journey

Next topic

- Text mining in practice: Bag of Words and Syntactic Parsing

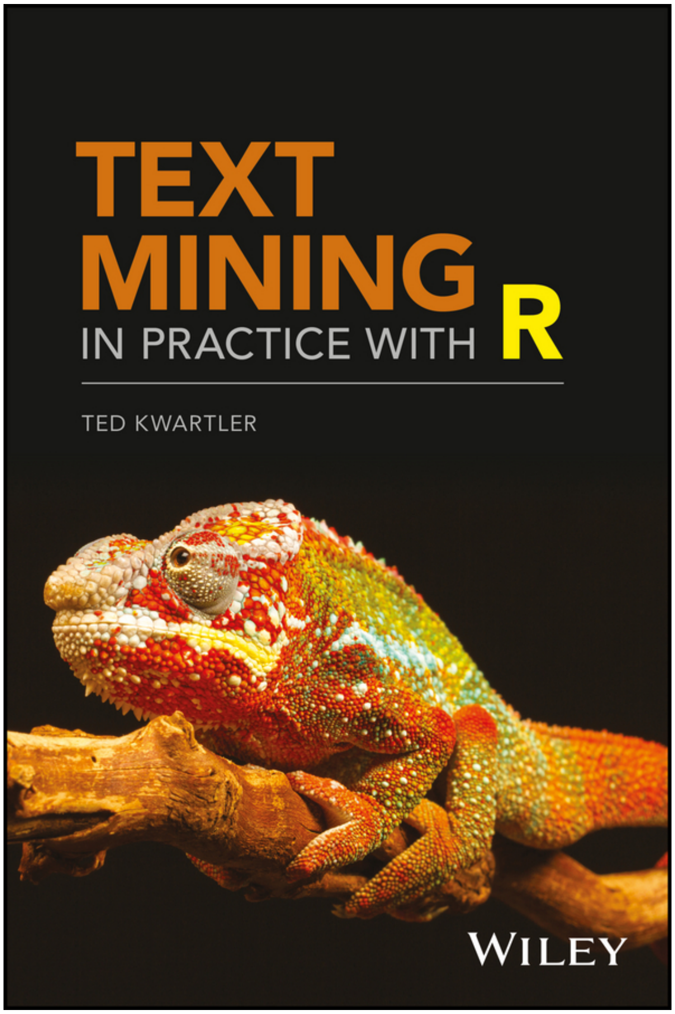


Figure 8: Cover of Text Mining In Practice With R by Ted Kwartler (Wiley, 2010)

- Base R data structures, functions and packages, importing data