

Programming: Beginner Level

Workshop #3: Scraping //
Data Analysis // AI-supported Coding

Frederik Møller Henriksen, phd fellow, IKH

Jakob Bæk Kristensen, post.doc, IKH

22th April 2024



Roskilde University
Denmark

**Digital
MediaLab**
Roskilde University

Agenda

LAST TIME: HANDS ON WORKSHOP

- For-loops
- Conditions (IF ELSE Statements)
- Functions
- Packages
- Exercise: Web-scraping

TODAY

Scraping text data

- Blogs and news media

What do to with the data?

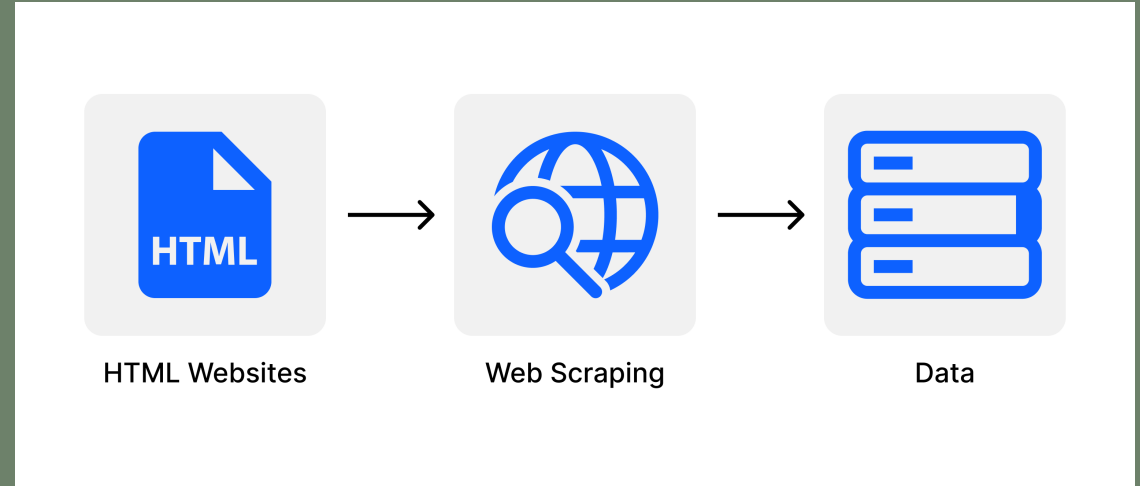
- Analysing text data

AI-Supported Coding

- Tips and tricks for using ChatGPT for coding tasks

Scraping text data (webpages and social media data)

- **“Do-it-yourself”-approach**
 - Use simple packages to create your own scrapers
 - BeautifulSoup, Requests
- **Open source packages**
 - **Wordpress scraping (webpages)**
 - Newspaper3k (news media)
 - E.g., TG-Tools (Telegram-data)
- **Integrated environments**
 - Meta: Crowdtangle / Content Library
 - 4CAT



Github

Main hub for Python developers

Developers use it to store and manage code

Find packages, see what is trending, upload your code

https://www.youtube.com/watch?v=pBy1zgt0XPc&ab_channel=GitHub





Scraping Exercise

- 1) Collect 1000 data points from at least one blog or media and collect
- 2) Save the data to your local drive
- 3) Find out how to read in the data using pandas (use google)


Be aware!

Most major news sites do not use wordpress as backend

Ideas for text analysis

Humanities and social sciences

- <https://www.aiforhumanists.com/>
- <https://libguides.union.edu/digital-scholarship/cta>
- <https://github.com/gesiscss/ptm>




Hacking the Humanities
1. Introduction

Hacking the Humanities Tutorials

pvieth
22 videoer • 6.878 visninger • Sidst opdateret d. 16. mar...

Afspil alle Bland


1



Hacking the Humanities
1. Introduction

3.08


2



Hacking the Humanities
2. Command Prompt

7.36


3



Hacking the Humanities
3. Python Basics

9.45

4



Hacking the Humanities
4. Strings

16.58

Episode 1: Introduction to the Hacking the Humanities Tutorial Series
pvieth • 2.851 visninger • for 5 år siden

Episode 2: The Command Prompt
pvieth • 2.498 visninger • for 5 år siden

Episode 3: The Very Basics of Python
pvieth • 832 visninger • for 5 år siden

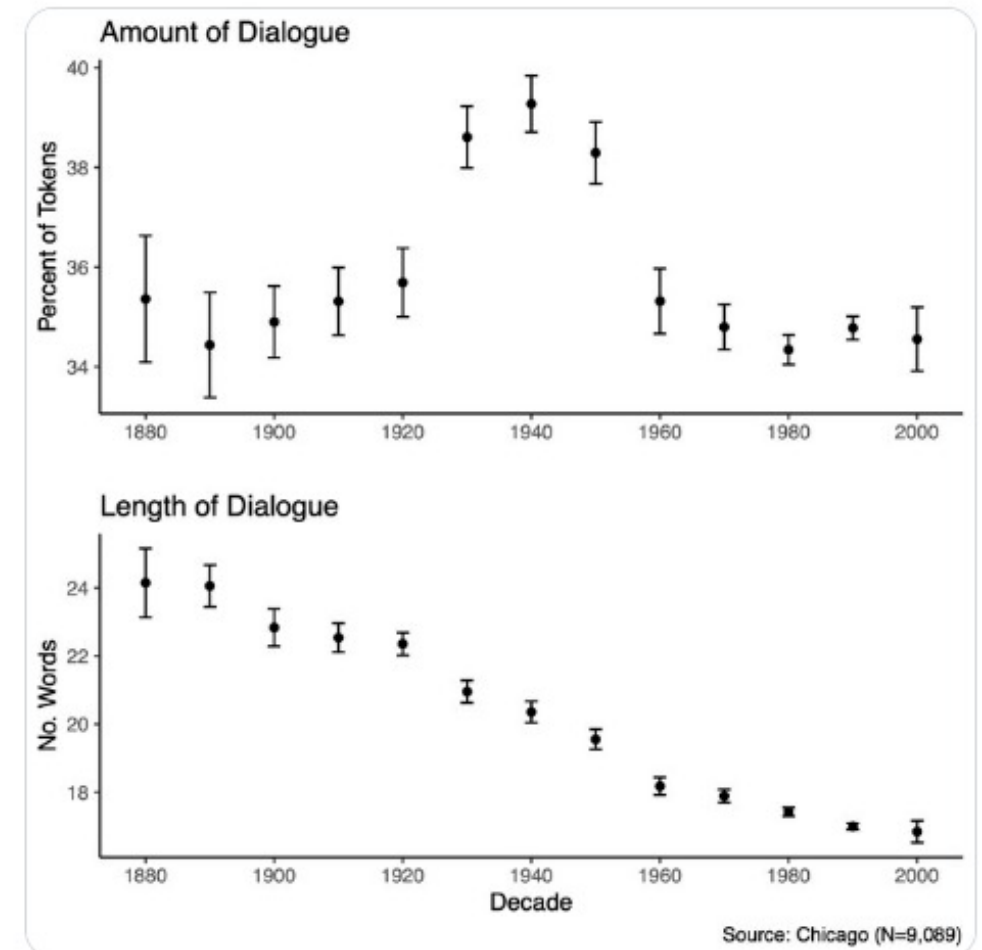
Episode 4: Strings
pvieth • 506 visninger • for 5 år siden



Andrew Piper @akpiper · 8 t

Here's an estimate of the history of dialogue in the 20C novel. We see two things: dialogue has been getting shorter (snappier?) but it roughly (and weirdly) consumes the same amount of narrative space (except in the 30-50s??). I say weird because I would not have expected it to...

[Vis mere](#)



Computational Text Analysis: A very fast walkthrough

From Basic to advanced text analysis

Word frequencies

Wordclouds

Key term extraction

Sentiment analysis

Emotion classification

Named entity-recognition

Co-occurrence analysis

Clustering

Topic modelling

Word embeddings

Fine-tuning large language models



Back to the data analysis

- Emotion classification on our collected data
- DaNLP package for NLP tasks in Danish
- Based on various packages also available for English and other languages
- (Tone and sentiment analysis if you have time)

```
!pip install danlp

from danlp.models import load_bert_emotion_model

# Load the BERT Emotion classifier
classifier = load_bert_emotion_model()
```

Final time: Open workshop - May 16th 13:00-15:00

- Bring your projects or ideas and receive feedback
- We'll help you with the code we have been through at the workshop as well as new ideas and projects.
- Thanks for your time!