2024-03-20

Data in the Humanities

Wk. 07: Scraping & Curating Structured Data

Welcome, Filipa Calado! 👋

Happenings This Week

Workshops

– Thu 3/21 Data Visualization with Tableau (PUL)
– Tue 3/26 Legal Issues in Computational Research Using Text and Data Mining (PUL)
– Wed 3/27 Getting Started with LaTeX (PUL)

Events

– Syriac Transcribathon (CDH / NES), 3/25 – 3/28
– LLM Reading Group: AI & Industry (CDH), 3/27

CFPs

– CDH Graduate Fellowship, Fall 2024, due 3/31
– DH for Hellenic Studies CFP: Athens in June! Now due 3/31

Survey Results

Let's review next week. Please answer if you haven't already!

Final Project options

1. Seminar Paper
2. Speculative Data Curation
3. Speculative Project Proposal
4. Group collaboration

for next week

install QGIS

https://www.qgis.org

Readings 📖

– Katie Rawson and Trevor Muñoz, “Against Cleaning,” in *Debates in the Digital Humanities 2019* (University Of Minnesota Press, 2019).

– Henk Alkemade et al., “Datasheets for Digital Cultural Heritage Datasets” 9, no. 1 (2023)

– *recommended but not required*

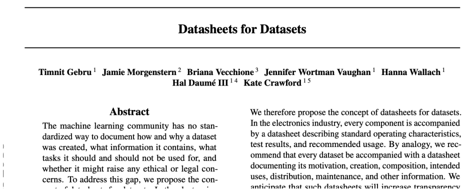
– Jessica Marie Johnson, “Markup Bodies: Black [Life] Studies and Slavery [Death] Studies at the Digital Crossroads”

– Daniel Rosenberg, “Data Before the Fact”

– Marisa Elena Duarte and Miranda Belarde-Lewis, “Imagining: Creating Spaces for Indigenous Ontologies”

Gebru et al., "Datasheets for Datasets" (2018)

contra The Pile &c.



Datasheets for Datasets

Timnit Gebru [1] Jamie Morgenstern [2] Briana Vecchione [3] Jennifer Wortman Vaughan [1] Hanna Wallach [1]
Hal Daumé III [1 4] Kate Crawford [1 5]

Abstract

The machine learning community has no standardized way to document how and why a dataset was created, what information it contains, what tasks it should and should not be used for, and whether it might raise any ethical or legal concerns. To address this gap, we propose the con-

We therefore propose the concept of datasheets for datasets. In the electronics industry, every component is accompanied by a datasheet describing standard operating characteristics, test results, and recommended usage. By analogy, we recommend that every dataset be accompanied with a datasheet documenting its motivation, creation, composition, intended uses, distribution, maintenance, and other information. We

https://datanutrition.org

https://github.com/jackbandy/bookcorpus-datasheet

Dataset Facts

Dataset BookCorpus
Instances Per Dataset 7,185 unique books, 11,038 total

Motivation	
Original Authors	Zhu and Kiros et al. (2015) [39]
Original Use Case	Sentence embedding
Funding	Google, Samsung, NSERC, CIFAR, ONR

Composition	
Sample or Complete	Sample, ≈2% of smashwords.com in 2014
Missing Data	98 empty files, ≤655 truncated files
Sensitive Information	Author email addresses

Collection	
Sampling Strategy	Free books with ≥20,000 words
Ethical Review	None stated
Author Consent	None

Cleaning and Labeling	
Cleaning Done	None stated, some implicit
Labeling Done	None stated, genres by smashwords.com

Uses and Distribution	
Notable Uses	Language models (e.g. GPT [29], BERT [9])
Other Uses	List available on HuggingFace [12]
Original Distribution	Author website (now defunct) [39]
Replicate Distribution	BookCorpusOpen [13]

Maintenance and Evolution	
Corrections or Erratum	None
Methods to Extend	"Homemade BookCorpus" [21]
Replicate Maintainers	Shawn Presser [12]

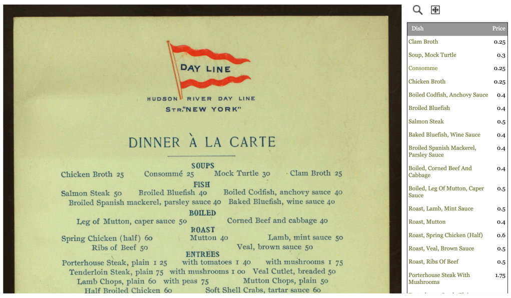
Genres	% of BookCorpus*
Romance 2,881 books	26.1%
Fantasy 1,502 books	13.6%
Vampires 600 books	5.4%

Horror 4.1%	• Teen 3.9%
Adventure 3.5%	• Literature 3.0%
Historical Fiction 1.6%	

Not a significant source of nonfiction.

* Percentages based on directories in books_txt_full. Some books cross-listed.

https://menus.nypl.org/


DAY LINE
HUDSON RIVER DAY LINE
STR."NEW YORK"

DINNER À LA CARTE

SOUPS
Chicken Broth 25 Consommé 25 Mock Turtle 30 Clam Broth 25

FISH
Salmon Steak 50 Broiled Bluefish 40 Boiled Codfish, anchovy sauce 40
Broiled Spanish mackerel, parsley sauce 40 Baked Bluefish, wine sauce 40

BOILED
Leg of Mutton, caper sauce 50 Corned Beef and cabbage 40

ROAST
Spring Chicken (half) 60 Mutton 40 Lamb, mint sauce 50
Ribs of Beef 50 Veal, brown sauce 50

ENTREES
Porterhouse Steak, plain 1 25 with tomatoes 1 40 with mushrooms 1 75
Tenderloin Steak, plain 75 with mushrooms 1 00 Veal Cutlet, breaded 50
Lamb Chops, plain 60 with peas 75 Mutton Chops, plain 50
Half Broiled Chicken 60 Soft Shell Crabs, tartar sauce 60

Dish	Price
Clam Broth	0.25
Soup, Mock Turtle	0.3
Consomme	0.25
Chicken Broth	0.25
Boiled Codfish, Anchovy Sauce	0.4
Broiled Bluefish	0.4
Salmon Steak	0.5
Baked Bluefish, Wine Sauce	0.4
Broiled Spanish Mackerel, Parsley Sauce	0.4
Boiled, Corned Beef And Cabbage	0.4
Boiled, Leg Of Mutton, Caper Sauce	0.5
Roast, Lamb, Mint Sauce	0.5
Roast, Mutton	0.4
Roast, Spring Chicken (Half)	0.6
Roast, Veal, Brown Sauce	0.5
Roast, Ribs Of Beef	0.5
Porterhouse Steak With Mushrooms	1.75