On Data: Collection & Creation

CS 490A, Fall 2021

Applications of Natural Language Processing
https://people.cs.umass.edu/~brenocon/cs490a_f21

Brendan O'Connor & Laure Thompson

College of Information & Computer Sciences
University of Massachusetts Amherst

Administrivia – HW2

- HW2 has been released: [link]
- This assignment has **two** different deadlines
  1. Annotated data due: **Tuesday, October 5th 11:59pm**
  2. Final submission due: **Friday, October 8th 11:59pm**

General NLP problem

Input: some text ***x*** (e.g. sentence, document)

Output: a label ***y*** (from some finite label set)

Goal: learn a mapping function ***f*** from ***x*** to ***y***

General NLP problem

Input: some text x (e.g. sentence, document)

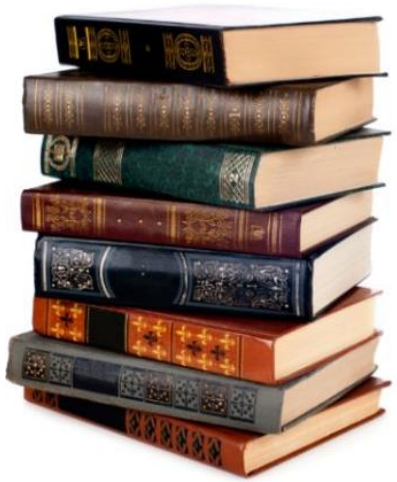
Output: a label y (from some finite label set)

Goal: learn a mapping function f from x to y

What is a dataset?

A collection of texts and metadata

Text sources can be used in many ways



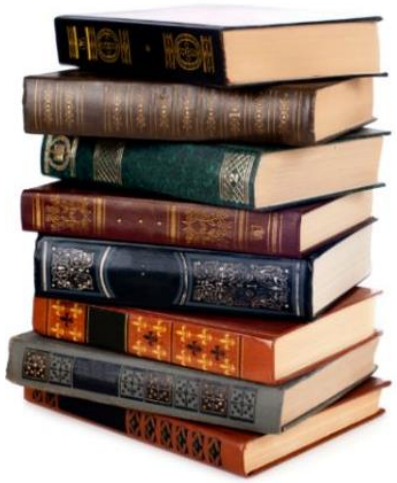
Text Unit

- Full Volume

Metadata / Label Examples

- Author, Translator
- Genre, Literariness
- Publication Year

Text sources can be used in many ways



Text Unit

- Full Volume
- Passage

Metadata / Label Examples

- Author, Translator
- Genre, Literariness
- Publication Year
- Elapsed Narrative Time
- Event Depiction
- Sarcasm, Irony, Suspense
- Memorability

Text sources can be used in many ways

Text Unit

- Post

Metadata / Label Examples

- Popularity, Controversy
- Topic / Post Type
- Shared News Source

Text sources can be used in many ways

Text Unit

- Post
- Interaction (Post + Replies)

Metadata / Label Examples

- Popularity, Controversy
- Topic / Post Type
- Shared News Source
- Agreement / Disagreement
- Moderator Intervention
- Condescending Language Use

Where to get datasets?

Off the shelf  Build from scratch

Where to get datasets?

Off the shelf  Build from scratch

Q: What are the pros & cons of each strategy?

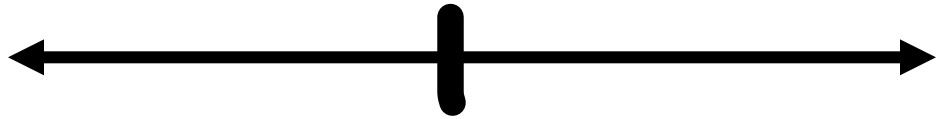
Off-the-Shelf Datasets

There are many publicly available datasets.

- https://huggingface.co/datasets
- https://nlpprogress.com/
- https://index.quantumstat.com/#dataset

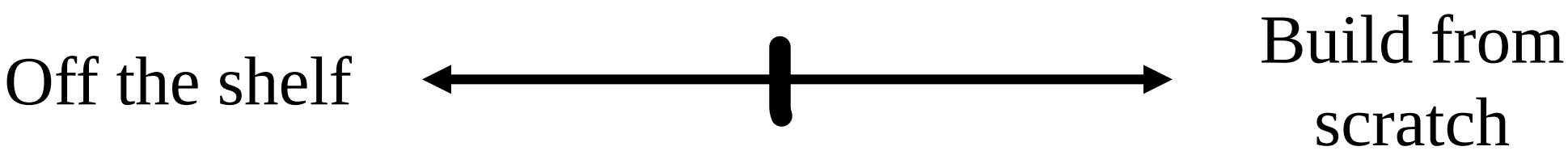
Middle ground: Augmentation & Curation

Off the shelf



Build from scratch

Middle ground: Augmentation & Curation



- Use a subset of the texts based on metadata, content, etc.
- Combine texts from multiple data sources
- Align text and metadata from different sources
- Add new labels to an existing dataset

Example: CMU Movie Summary Corpus

Dataset: 42,306 movie plot summaries extracted from Wikipedia
+aligned metadata extracted from Freebase

http://www.cs.cmu.edu/~ark/personas/

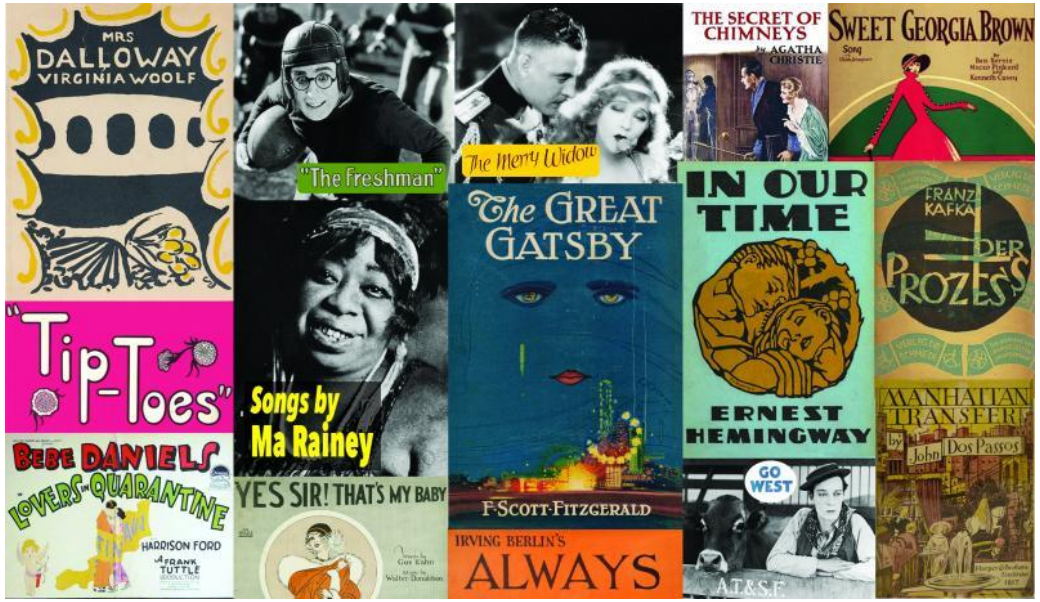
So you want to collect some data

1. Is it ethical to collect this data?
2. Is this data publicly available?
3. Is this data protected by copyright?

U.S. Public Domain

Published works enter the public domain **95 years** after their first publication date (in the U.S.)

As of this year, works published in **1925** entered the US public domain!

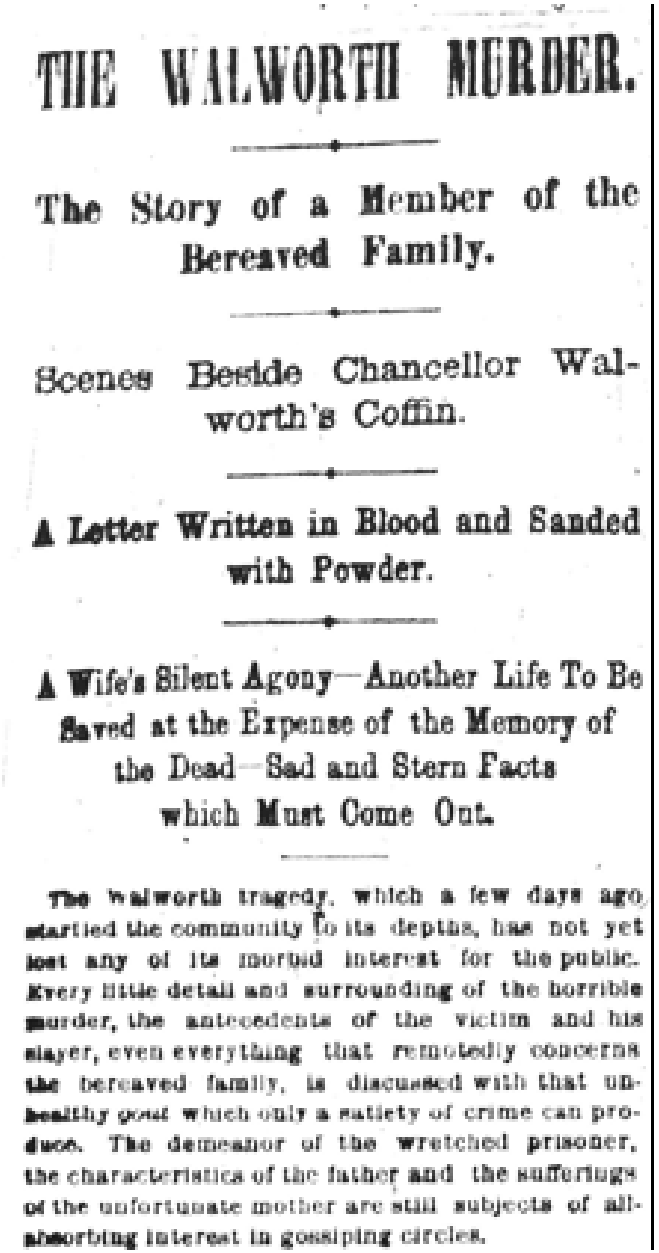


How to collect data?

- Digitization & Transcription
- APIs
- Bulk downloads / database dumps
- Web scraping

Digitization

Watch out for OCR errors!



THE WALWORTH MURDER.

The Story of a Member of the Bereaved Family.

Scenes Beside Chancellor Walworth's Coffin.

A Letter Written in Blood and Sanded with Powder.

A Wife's Silent Agony—Another Life To Be Saved at the Expense of the Memory of the Dead—Sad and Stern Facts which Must Come Out.

The Walworth tragedy, which a few days ago startled the community to its depths, has not yet lost any of its morbid interest for the public. Every little detail and surrounding of the horrible murder, the antecedents of the victim and his slayer, even everything that remotely concerns the bereaved family, is discussed with that unhealthy gout which only a satiety of crime can produce. The demeanor of the wretched prisoner, the characteristics of the father and the sufferings of the unfortunate mother are still subjects of all-absorbing interest in gossiping circles.

THE Wal-worth MURDER.

The Story of a Member of the Bereaved Family.

Scenes Beside Chancellor Walworth's Coffin.

A **hotter** Written in Blood and Sanded with Powder.

A Wife'i **Bite** Agony—Another Life To Be
fisved at the Expense of the Memory of
the Dead-Sad and **Stem** Facts
which **Mwt** Come Out.

The walworth tragedy, which a few days **agox**
•**turned** the community to its depths, **hag** not yet **toat** any of its morbid **intercut** for the public. **Kverjr UlUe**

APIs

Some examples:

- Twitter: twitter-api
- Reddit: pushshift.io
- NYTimes Movie Reviews: movie-reviews-api

Q: What are the incentives for maintaining APIs?

Q: What are the incentives for maintaining APIs?

LibraryThing APIs

Notice: LibraryThing APIs currently disabled

The LibraryThing APIs are disabled until further notice (last updated on 1/28/2021). If you would like to be notified when they are re-enabled, please email info(AT)librarything.com.

Does Goodreads support the use of APIs?

Dec 17, 2020 · Knowledge

As part of our overall commitment to continually improve our data management, Goodreads no longer issues new developer keys for our public developer API and plans to retire the current version of these tools. While we assess the value of APIs to determine how to support in the future, we continue to support active API users who meet our terms of service. You are welcome to give your feedback on Goodreads APIs by completing the developer API survey.

Bulk Downloads

Wikipedia: dumps.wikimedia.org/

Court Listener: courtlistener.com/api/bulk-info/

Web Scraping Responsibly

Be considerate

- Check Terms of Service
- Check for robots.txt
- Use low request rates

Web Scraping Demo

- For fetching webpages from the command line: wget, curl
- For parsing .html source files: BeautifulSoup (bs4)

Data/Dataset Examples

Fontenrose. Delphic Oracle Responses

H9 (PW164). 421–415.

C. Athenians and allies, probably on more than one occasion.

Occ. and Q. Not stated.

R. Offer the fruits of the harvest to the two goddesses [Demeter and Kore].

Mode A1, Topic 1b

Delphi. Indirect—Isokr. *Or.* 4.31.

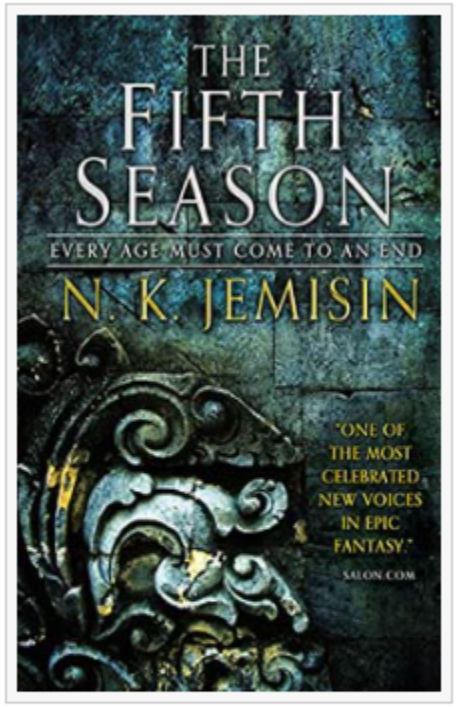
Testimony—*Inscription of Athens, *IG* 1^2. 76 = Michel 71 = *SIG* 83, lines 4–8, 25–26, 32–34.

Comment: The inscription specifies at least one-sixth of a *medimnos* for every 100 *medimnoi* of barley and at least one-twelfth of a *medimnos* for every 100 *medimnoi* of wheat; but whether this was ordered in the oracle or was specified by the Athenian assembly as the way of executing the command is not clear. Isokrates does not include this in his indirect form. See Q79 *Comment.*

Fontenrose. Delphic Oracle Responses

Mode	Topic	
A. Simple Commands	*1. Res Divinae*	
A1. Clear Commands	1a. Cult Foundations	1c. Human Sacrifice
A2. Sanctions	1b. Sacrifices, Offerings	1d. Religious Laws, Customs
A3. Ambiguous Commands	*2. Res Publicae*	
B. Conditioned Commands	2a. Rulership	2d. Interstate Relations
C. Prohibitions and Warnings	2b. Legislation	2e. War
C1. Clear Prohibitions	2c. City/Colony Foundations	
C2. Ambiguous Prohibitions	*3. Res Domesticae et Profanae*	
D. Statements on Past or Present	3a. Birth, Origin	3g. Persons, Agents
D1. Commonplace Statements	3b. Marriage, etc.	3h. Means, Signs
D2. Extraordinary Statements	3c. Death, Burial	3i. Places, Lands
E. Simple Future Statements	3e. Actions, Events	3j. Gnomic Utterances
E1. Non-Predictive Statements	3f. Rewards, Punishments	
E2. Clear Predictions		
E3. Ambiguous Predictions		
F. Conditioned Predictions		

World's Without End



Added By: valashain
Last Updated: Engelbrecht

The Fifth Season

Author:	N. K. Jemisin
Publisher:	Subterranean Press, 2017 Orbit, 2015
Series:	The Broken Earth: Book 1
Book Type:	Novel
Genre:	Science-Fiction / Fantasy
Sub-Genre Tags:	Apocalyptic/Post-Apocalyptic Dying Earth Dystopia
If you liked **The Fifth Season** you might like these books.	
Awards:	• 2015 Nebula Nominated • 2015 Red Tentacle Nominated • 2016 Hugo Winner • 2016 Locus F Nominated • 2016 WFA Nominated
Lists:	• WWEnd Top Nominated Books • Award Winning Books by Women Authors
Links:	
Avg Member Rating:	★★★★½ (355 reads / 185 ratings)

Books

Wikipedia derived datasets

Wikitext: High quality subset of Wikipedia pages.

WikiMatrix: Parallel sentences (across languages) of Wikipedia

SQUAD: Question-answer dataset that relies on Wikipedia text

There’s no such thing as raw data.