On Data: Collection & Creation

CS 485, Fall 2023
Applications of Natural Language Processing
https://people.cs.umass.edu/~brenocon/cs485_f23/

Brendan O'Connor
College of Information and Computer Sciences
University of Massachusetts Amherst
Upcoming

- This week - no required readings
  - Today: datasets
  - Thursday: overview of final project & NLP applications broadly
- This week: starting HW2
- Next Tuesday: no class (Monday schedule day)
- Upcoming in Oct & Nov
  - First assignment for final project, the project proposal, will be due AFTER HW2
  - HW3 then Midterm (early Nov) after that
HW2: group annotations

- Form groups of 2-4 (group formation **due this Friday**)
- We will distribute data: tweets
- As a group:
  - Look at the data (reference set)
  - Formulate an annotation task
- Then individually:
  - Annotate a dataset
  - Upload your annotations (**due next week**)
- Finally, individually with your group's data:
  - Analyze agreement rates
  - Try out logistic regression for classification
  - (Final individual HW submission, in ~2 weeks)
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

<table>
<thead>
<tr>
<th>Text Unit</th>
<th>Metadata / Label Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>· Full Volume</td>
<td>· Author, Translator</td>
</tr>
<tr>
<td></td>
<td>· Genre, Literariness</td>
</tr>
<tr>
<td></td>
<td>· Publication Year</td>
</tr>
</tbody>
</table>
Text sources can be used in many ways

<table>
<thead>
<tr>
<th>Text Unit</th>
<th>Metadata / Label Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>· Full Volume</td>
<td>· Author, Translator</td>
</tr>
<tr>
<td>· Passage</td>
<td>· Genre, Literariness</td>
</tr>
<tr>
<td></td>
<td>· Publication Year</td>
</tr>
<tr>
<td></td>
<td>· Elapsed Narrative Time</td>
</tr>
<tr>
<td></td>
<td>· Event Depiction</td>
</tr>
<tr>
<td></td>
<td>· Sarcasm, Irony, Suspense</td>
</tr>
<tr>
<td></td>
<td>· Memorability</td>
</tr>
</tbody>
</table>
Text sources can be used in many ways

<table>
<thead>
<tr>
<th>Text Unit</th>
<th>Metadata / Label Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>· Post</td>
<td>· Popularity, Controversy</td>
</tr>
<tr>
<td></td>
<td>· Topic / Post Type</td>
</tr>
<tr>
<td></td>
<td>· Shared News Source</td>
</tr>
</tbody>
</table>
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?

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
THE WALWORTH MURDER.

The Walworth tragedy, which a few days ago
had not yet toot any of its morbid
interest for the public. Every
detail of the murder
and the surrounding
situation of the victim and his
family has not yet been made
clear. The public is still
in the dark about the
true circumstances of the
murder.

A letter written by a member of the
Walworth family, describing the
events leading up to the murder,
has been made public.

Scenes Beside Chancellor Wal-
worth’s Coffin.

THE WALWORTH MURDER.

The Story of a Member of the
Bereaved Family.

A Letter Written in Blood and Sand on
the Deadline of the Memory
of the Deceased and Some Facts
which Must Come Out.
APIs

Some examples:

· (now mostly closed) Twitter: twitter-api
· (now mostly closed?) Reddit: pushshift.io

Big problem: value of data and questions about tech company scrutiny
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)
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.