A Project Gutenberg Poetry Corpus

Allison Parrish
New York University

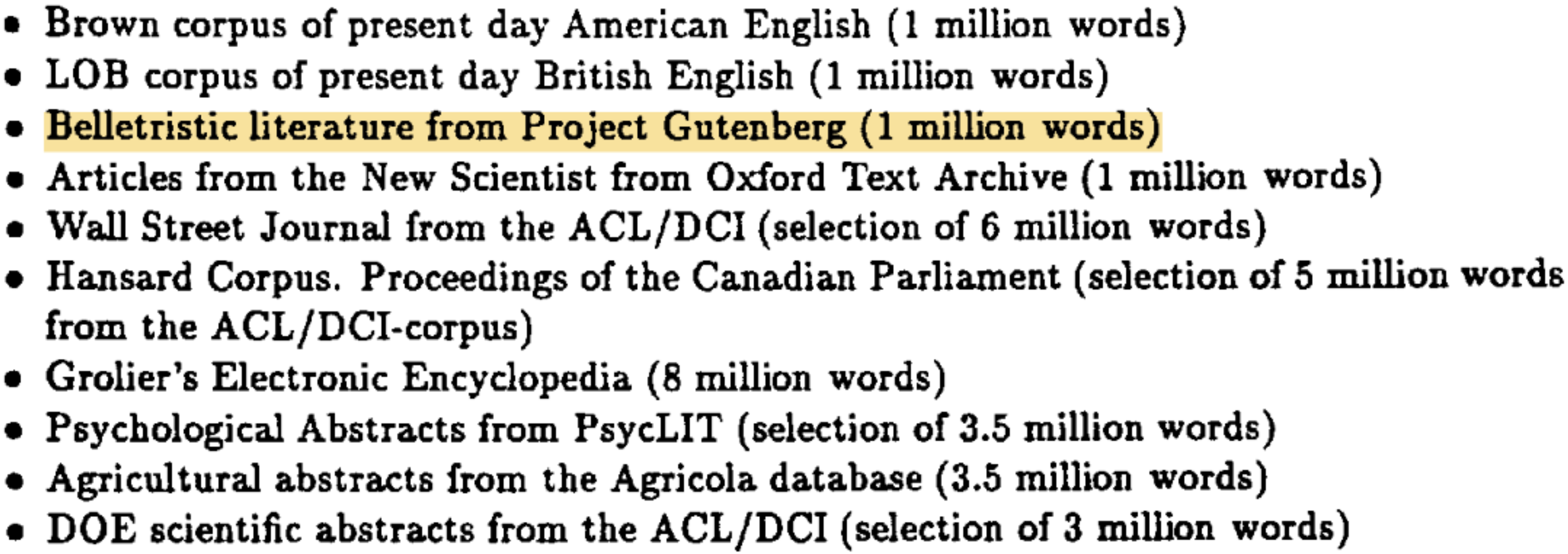
Shadows far off, alien
gives back in restoration that which heaven dries up of the sea
And many a Garden by the Water blows.
To walk together in starry weather
Whirled all about--dense, multitudinous, cold--
Light and life to all he brings,
Resisting inquisition. I opine
Dear to their hearts he was,--so dear,
Yet who by a wish would recall you again?
Si said the Horn was still some weeks away,
Their bodies hid in barks, and furred with moss;
On the British Commercial Depredations, iii. 300;
Some base-born Hessian slave walks threat'ning by,
Such Trojan hosts, whose trophies grace the plain.
In the reciprocated joy
Spake and said to Minnehaha,
And one star like a hound.
No man need boast their love possessing.
In whispering leaves, these solemn words -
Told of a beauteous dame beyond the sea!
Confused, distraught, pale thousands spread the plain;
And when, far off, I saw these ancient towers

Free ebooks by Project Gutenberg

Some of Our Latest Books

Project Gutenberg needs your donation!

(Lebert 2010)



- Brown corpus of present day American English (1 million words)
- LOB corpus of present day British English (1 million words)
- Belletristic literature from Project Gutenberg (1 million words)
- Articles from the New Scientist from Oxford Text Archive (1 million words)
- Wall Street Journal from the ACL/DCI (selection of 6 million words)
- Hansard Corpus. Proceedings of the Canadian Parliament (selection of 5 million words from the ACL/DCI-corpus)
- Grolier's Electronic Encyclopedia (8 million words)
- Psychological Abstracts from PsycLIT (selection of 3.5 million words)
- Agricultural abstracts from the Agricola database (3.5 million words)
- DOE scientific abstracts from the ACL/DCI (selection of 3 million words)

(Wettler and Rapp 1993)

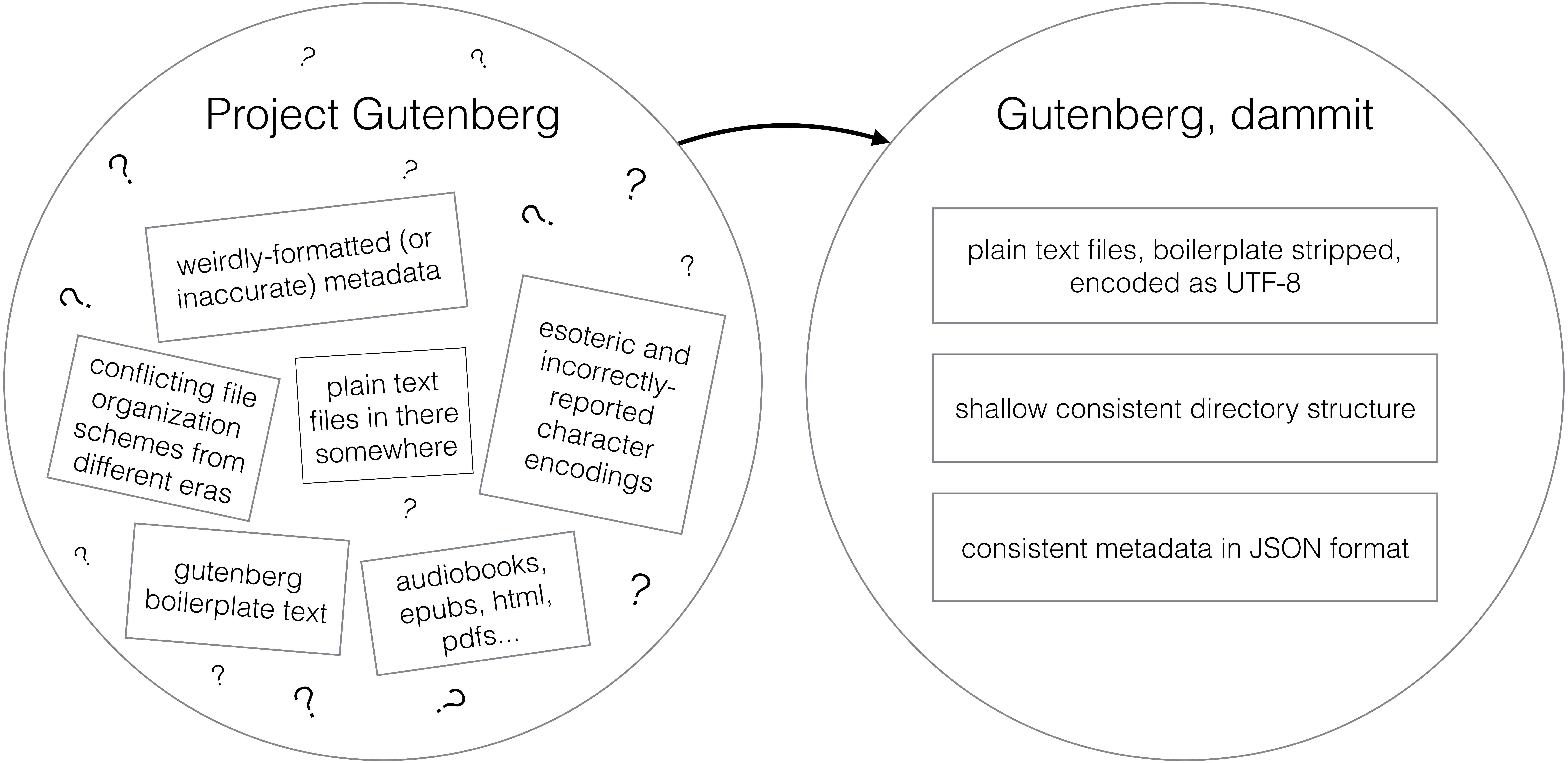
how I made it

Gutenberg, dammit

https://github.com/aparrish/gutenberg-dammit/

(includes the ~7 gigabyte archive and the code necessary to build the archive from scratch)

Gutenberg, dammit is heavily based on the GutenTag project (Brooke 2015)



Project Gutenberg

weirdly-formatted (or inaccurate) metadata

conflicting file organization schemes from different eras

plain text files in there somewhere

esoteric and incorrectly-reported character encodings

gutenberg boilerplate text

audiobooks, epubs, html, pdfs...

Gutenberg, dammit

plain text files, boilerplate stripped, encoded as UTF-8

shallow consistent directory structure

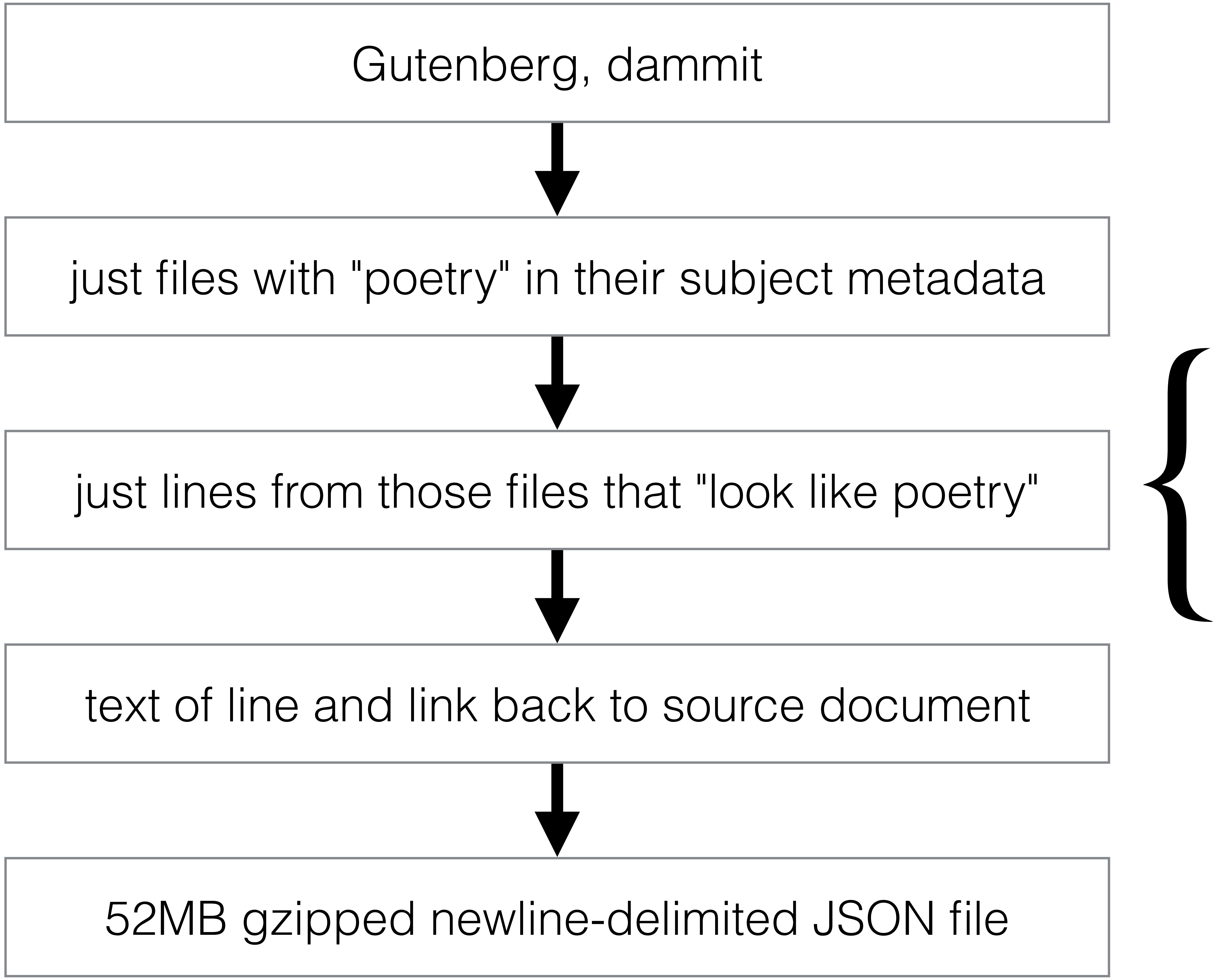
consistent metadata in JSON format

Gutenberg Poetry Corpus

```
{"s": "The Heav'ns and all the Constellations rung,", "gid": "20"}
{"s": "The Planets in thir stations list'ning stood,", "gid": "20"}
{"s": "While the bright Pomp ascended jubilant.", "gid": "20
{"s": "Open, ye everlasting Gates, they sung,", "gid": "20"}
{"s": "Open, ye Heav'ns, your living dores; let in", "gid": "20"}
```

https://github.com/aparrish/gutenberg-poetry-corpus/

(includes the archive and the code necessary to build
the archive from scratch)



Gutenberg, dammit

just files with "poetry" in their subject metadata

just lines from those files that "look like poetry"

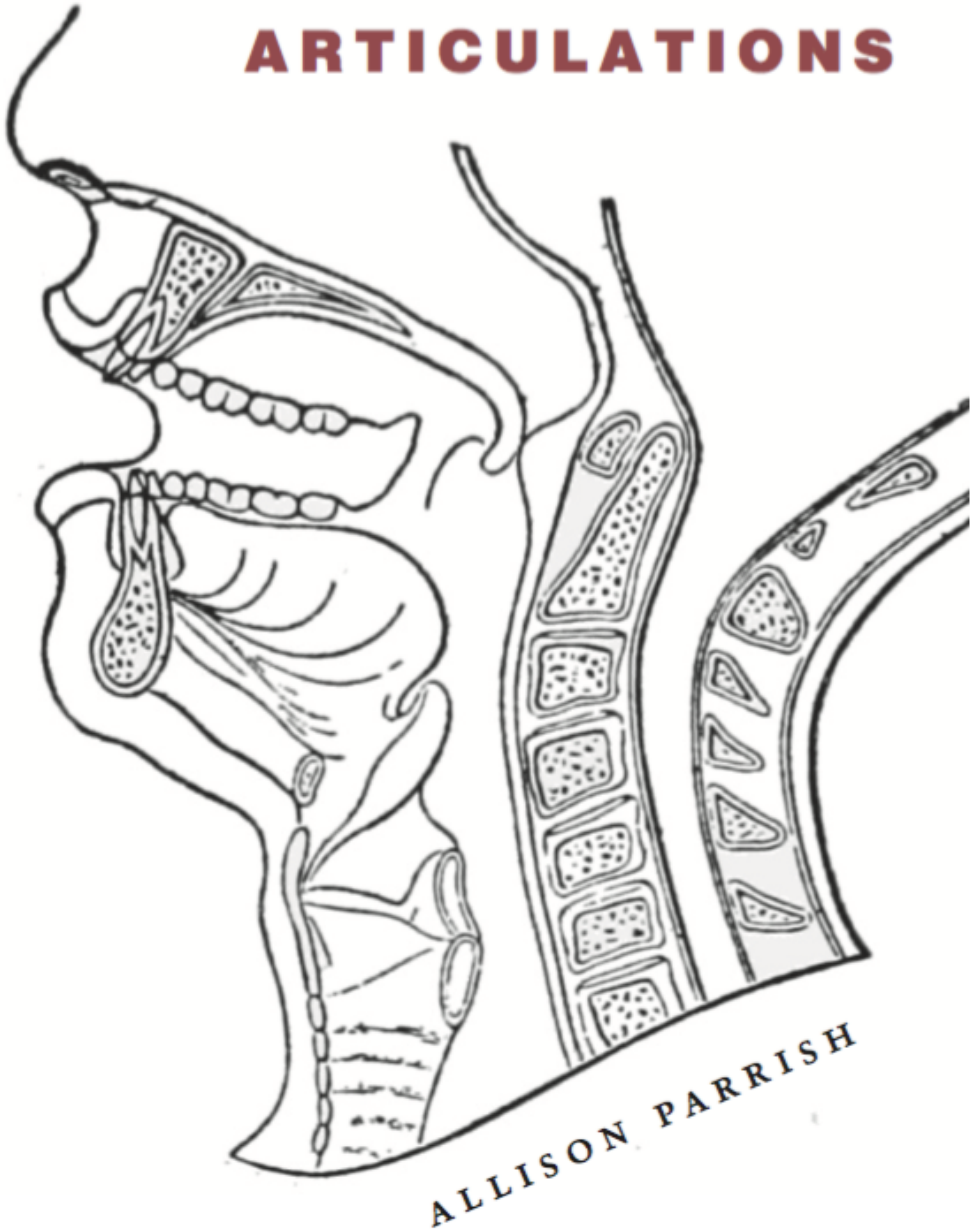
text of line and link back to source document

52MB gzipped newline-delimited JSON file

- Length
- Case
- Doesn't look like TOC
- Doesn't look like a title
- Not a reference or footnote
- Keyword content filter
- etc.

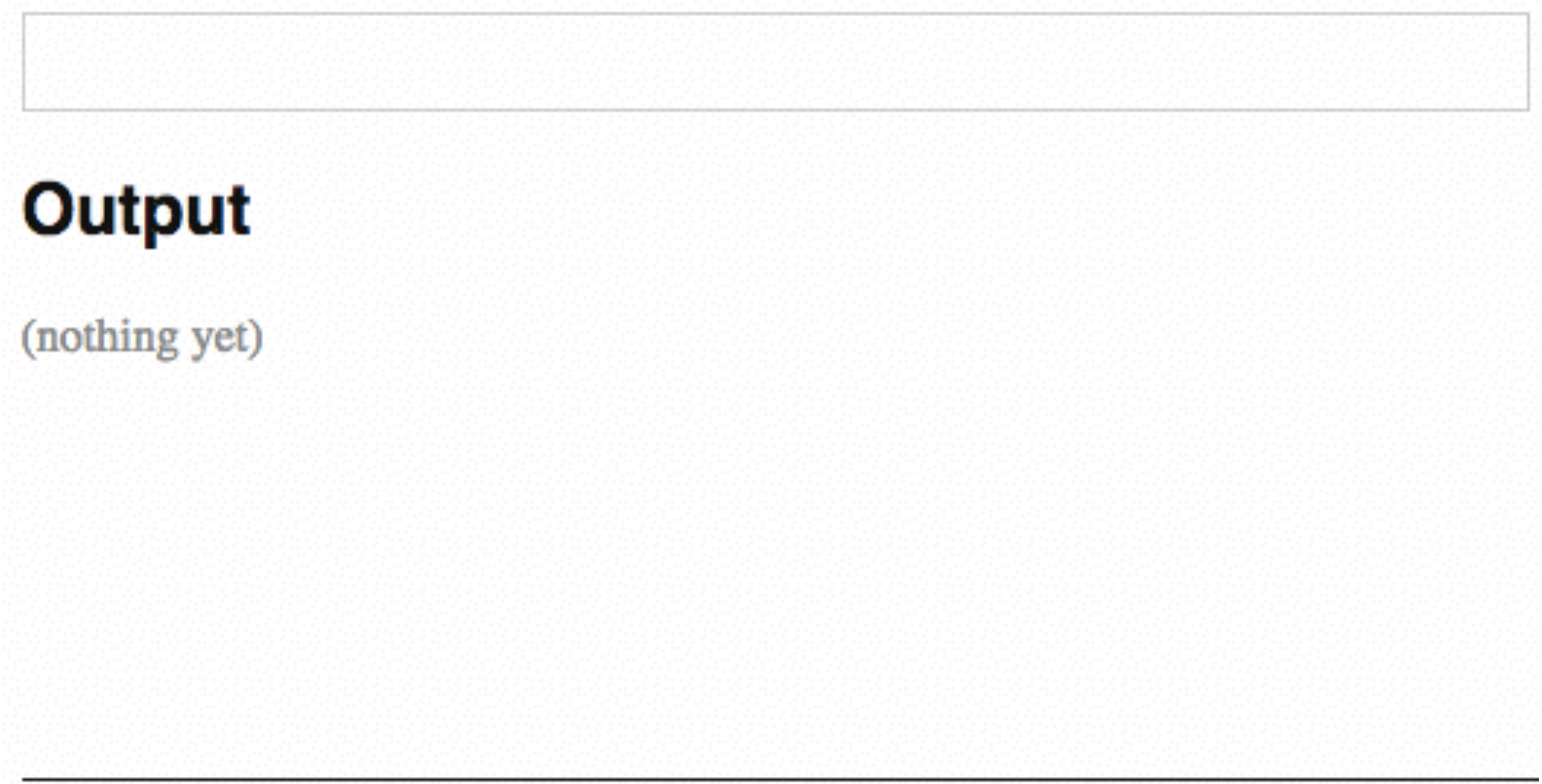
what people have done with it (so far)

Articulations



Sweet hour of prayer, sweet hour of prayer it was the hour of prayers. In the hour of parting, hour of parting, hour of meeting hour of parting this. With power avenging, ... His towering wings; his power enhancing, in his power. His power. Thus: the blithe powers about the flowers, chirp about the flowers a power of butterfly must be with a purple flower, might be the purple flowers it bore. The petals of her purple flowers where the purple aster flowered, here's the purple aster, of the purple asters there lives a purpose stern! A sterner purpose fills turns up so pert and funny; of motor trucks and vans, and after kissed a stone, an ode after Easter. And iron laughter stirred, O wanderer, turn; oh, wanderer, return. O wanderer, stay; O Wanderer near. Been a wanderer. I wander away and then I wander away and thence shall we wander away, and then we would wander away, away O why and for what are we waiting. Oh, why and for what are we waiting, why, then, and for what are we waiting?

Gutenberg Poetry Autocomplete



Output

(nothing yet)

http://gutenberg-poetry.decontextualize.com/

Kit Armstrong, "Betting the under"

"I saw you over the Persian navy / the Persians at Marathon
in 490 B.C. / the big race / the blazing wicks // Saw you
over the cheek-guard / chased with gold / over the dunes /
and over the dust // I went over the archipelagos to you /
over the rocks / and we heard rushing in the far distance /
over the surplus revenue / in opposition to Government /
when we'd look / over the living room ceiling all night [...]"

https://theindianapolisreview.com/betting-the-under/

Plot to Poem

Line from WikiPlots corpus	Semantically similar line from Gutenberg Poetry Corpus
In 1977, Harry Burns and Sally Albright graduate from the University of Chicago and share the drive to New York City, where Sally is beginning journalism school and Harry is starting a career	Where in one dream the feverish time of youth
Harry is dating a friend of Sally's, Amanda	Alas! that any friend of mine
During the drive, they discuss their differing ideas about relationships between men and women	week to the 'mistress'--for women carry on the business of
Harry says that "Men and women can't be friends because the sex part always gets in the way"	In gardens where men love me, and be sure
Sally disagrees, claiming that men and women can be strictly friends without sex	Is't not enough that women toil.
[...]	[...]
Harry spends New Year's alone, walking around the city	Than share the city's year forlorn.
As Sally decides to leave the party early, Harry appears and declares his love for her	Who does not love his early dream of love?--
At first, she argues that the only reason he is there is because he is lonely, but he disagrees and lists the many things he realized he loves about her	What is there that abides
They make up and kiss and marry three months later	Then may I marry you, my pretty maid?'

potential uses

- computational cut-ups
- computational stylistics ("how poetic is this text," for certain meanings of "poetic")
- models for text generation (train an LSTM on this and see what comes out) and text "style transfer" (weight this text in your model to add a bit of "poetry" to it)
- mixed-initiative interfaces
- "inspiration"
- etc.

why this poetry corpus?

"I use the text [...] almost as a painter uses paint or a sculptor stone — the material with which I work being preselected and limited. Henry Moore observed that his manner of working was to remove all extraneous material to allow the figure that was “locked” in the stone to reveal itself. It is an image that has always appealed to me, although I work with words rather than stone."

–M. NourbeSe Philip, *Zong!* (pp. 197-198)

"Poetry should be burned to the bone by austere fires and washed white with rains of affliction: the poet should love nakedness and the thought of the skeleton under the flesh."

—Rebecca West (quoted in Churchill 2016)

next steps

better classification

- rule based → statistical model

mindfulness about what is included

- curation, filtering
- "raw material" that is "washed white" (?!)

citations

Armstrong, Kit. “Betting the Under.” The Indianapolis Review, 27 Apr. 2018, https://theindianapolisreview.com/betting-the-under/.

Brooke, Julian, et al. “GutenTag: An NLP-Driven Tool for Digital Humanities Research in the Project Gutenberg Corpus.” CLfL@ NAACL-HLT, 2015, pp. 42–47.

Churchill, Suzanne W. “Little Magazines and the Gendered, Racialized Discourse of Women’s Poetry.” A History of Twentieth-Century American Women’s Poetry, edited by Linda A. Kinnahan, Cambridge University Press, 2016, pp. 170–85. Crossref, doi:10.1017/CBO9781316488560.012.

Lebert, Marie. “Essay on the History of Project Gutenberg.” Project Gutenberg News, 21 Mar. 2010, http://www.gutenbergnews.org/about/history-of-project-gutenberg/.

Parrish, Allison. Articulations. Counterpath Press, 2018.

Philip, M. NourbeSe, and Setaey Adamu Boateng. Zong! Wesleyan University Press, 2008. Project MUSE, http://muse.jhu.edu.proxy.library.nyu.edu/book/12847.

Riedl, Mark. WikiPlots: A Dataset Containing Story Plots from Wikipedia (Books, Movies, Etc.) and the Code for the Extractor. 2017. GitHub, https://github.com/markriedl/WikiPlots.

Wettler, Manfred, and Reinhard Rapp. “Computation of Word Associations Based on the Co-Occurrences of Words in Large Corpora.” Proceedings of the 1st Workshop on Very Large Corpora, 1993, pp. 84–93.

https://www.decontextualize.com/

https://mastodon.social/@aparrish

https://github.com/aparrish/gutenberg-poetry-corpus/

https://github.com/aparrish/gutenberg-dammit/