*****Caution!*****
I have not quality-checked these links. Some of the below links may be broken, some of the data may be messy or unusable, or just plain bad or wrong. ****You are responsible for checking these data sources to ensure their integrity.**** Also consider the ethics of your choice of data. What potential harms may come from you analyzing these data? Just because someone else collected the data doesn't mean you don't have to think about the ethical implications of using the data.

Potential sources of data:

1. Northeastern Resources:
   - Amanda Rust's Subject Guide on "Text and Data Mining Library Databases"
   - NULab Text Analysis Resources (the very end of this page also contains links to other online text analysis tools)

2. Websites with interesting social data, some of it text:
   - https://www.data.gov
   - https://www.kaggle.com/datasets
   - https://toolbox.google.com/datasetsearch

3. Websites and collections with interesting humanities data:
   - Demonstration Corpora, by Alan Liu
   - Internet Archive Books (includes plain-text ["full text"] access to books, issues of magazines, etc.)
   - Oxford Text Archive (large number of texts available in variety of forms, including plain text; texts are accessed one at a time)
   - Project Gutenberg (convenient categories and "bookshelves" of works that can be downloaded; limited automated access; see also tips on downloading from Project Gutenberg)

4. Chris Bail's list of interesting datasets and APIs for text analysis

5. From Miriam Posner's crowdsourced document**:**
   Corpora (and a few tools)
   Contact Info
   miriam.posner@gmail.com | @miriamkp
   aw@andrewbenedictwallace.com | @farbeitfrom

   - HATHITrust (16 million volumes, mostly in English)
   - Chronicling America (12.8 million pages of American newspapers)
   - DocSouth Data (narratives & literature from the American South)
   - Perseus Digital Library (large collection of classical texts, much of it encoded in TEI/XML)
   - EEBO-TCP (ca. 50,000 early English books, many encoded in TEI/XML)
   - Old Bailey Online (197,745 London criminal trials, 1674-1913)
   - Canadian Hansard (debates & journals of the Canadian Senate & House of Commons)
   - Australian Hansard (Parliamentary debates, 1901-1980)
   - UK Hansard (UK Parliamentary debates)
   - Open Islamicate Texts Initiative (see also repositories; 10,000 premodern Islamicate texts)
   - Transkribus Corpus and READ (efforts to use computer vision to recognize handwriting)

- ToposText (557 classical texts linked with a gazetteer of the ancient world)
- BYU Corpora (widely used corpora of American English)
- Wright American Fiction (American adult fiction, 1774–1900)
- UCLA Broadcast NewsScape (170K hours of captioned news programs; see Red Hen Lab for information on access)
- Media History Digital Library (nearly 2 million pages of media-related books and articles, 1875-1995)
- Christian Classics Ethereal Library (classic Christian texts)
- NYT Annotated Corpus (1.8 million *NYT* articles + *NYT*-supplied metadata)
- Europeana Collections (many datasets from European libraries & archives, from papyri to photographs to newspapers)
- Foreign Records of the US (nearly complete run of *Foreign Relations of the United States*; see these tools to obtain full text)
- Internet Archive (huge collection of websites, texts, audio, and other media, available for bulk download via wget)
- Twitter Datasets (a catalog of Twitter datasets that are publicly available on the web)
- BitCurator (effort to develop tools to analyze features of digital texts)
- Movie Quotes Corpus ("220,579 conversational exchanges between 10,292 pairs of movie characters")
- Europe PMC (repository of life sciences books, articles, and preprints)
- Trove Australia (565 million documents collected by the National Library of Australia, including a sizeable collection of newspapers)
- BNC-Baby (4 million-word sub corpus of the 100 million-word British National Corpus, with parts-of-speech tagging in XML)

6. Springboard List of Free Datasets for Data Science
    - United States Census Data (statistics or geospatial analyses)
    - FBI Crime Data, including an API (can be used for geospatial analyses)
    - CDC Cause of Death
    - Medicare Hospital Quality
    - SEER Cancer Incidence (by gender, race, year, and other demographics)
    - Bureau of Labor Statistics
    - Bureau of Economic Analysis
    - IMF Economic Data (international data)
    - Dow Jones Weekly Returns
    - (by suburb and with 13 other attributes, could be used for geospatial analyses)
    - The now famous Enron Emails (can be used for social network analysis or text analysis)