PART II

Selection and Representation

Once you start to see text as a source of data, new opportunities to study the social world open up everywhere. Firms use company earnings reports and consumer complaints to guide investment decisions and improve customer service. Governments use safety inspection reports and transcripts of emergency dispatch to improve government services. Economists gain insights into economic growth and marketing through transcripts of Federal Reserve meetings and advertisements. Political scientists answer questions about representation, accountability, and political competition with speeches, press releases, and debates. Sociologists measure culture, track demographic changes, and map the digitized footprints of diverse communities. Outside the social sciences, historians leverage vast archives of historical documents, public health scholars investigate electronic health records, and English professors study the broader patterns of literature. In each of these cases, our goal is more than fitting a statistical model to data; we want to learn something about the human processes that documents record and represent.

In this part, we discuss the first step in the text analysis process: the collection and representation of text. We begin in **Chapter 3: Principles of Selection and Representation** by providing four principles that provide a framework for the rest of the chapters in this part. We then move to **Chapter 4: Selecting Documents**, which explores why a principled approach to gathering and sampling text is imperative for making inferences. Here the goal is to construct a *corpus*—the collection of documents to analyze. **Chapter 5: Bag of Words** provides an initial default recipe for representing the documents numerically based on the "bag of words" model. This simple, yet powerful, representation of texts treats each document as simply a collection of word counts—as though we threw all the words into a bag and shook them up so their order

was no longer preserved. Although this representation loses a lot of information, it can be highly effective for many tasks and is a useful default approach.

The bag of words representation lends itself to a probabilistic modeling strategy based on the multinomial distribution. **Chapter 6: The Multinomial Language Model** introduces the multinomial language model that will form the foundation of many of the statistical models that we will cover throughout the book. **Chapter 7: The Vector Space Model and Similarity Metrics** introduces the idea of interpreting the bag of words representation as a vector space. This provides a foundation for the algorithmic models in much the way that the multinomial model provides a foundation for the statistical ones.

Chapters 5–7 use a representation of words that makes no initial assumptions about how similar words are to each other: `cat` and `kitten` are treated as equally distinct as `cat` and `ecclesiastical`. The next two chapters discuss how we can use external information to enhance our understanding of words. Using large text collections such as Wikipedia or the Common Crawl of the internet, we can learn distributed representations of words that encode the idea that some words are more similar than others. These "word embeddings" are described in **Chapter 8: Distributed Representations of Words**. While word embeddings leverage an external corpus to gain semantic information, **Chapter 9: Representations from Language Sequences** overviews several approaches from natural language processing that can be used to extract different kinds of information from documents such as the roles that words play in a sentence. These techniques include text reuse, part-of-speech tagging, named entity recognition, dependency parsing, and other information extraction tasks.

We assume throughout that documents are already machine readable. Documents that aren't natively digital can be made machine readable through *optical character recognition*—the process of converting pictures of text into characters that a machine can read. While not perfect, these technologies are the best solution for making text machine readable for large corpora where transcribing the documents by hand would be prohibitively time consuming.

CHAPTER 3

Principles of Selection and Representation

As we have discussed in previous chapters, text data reflects social interactions, economic transactions, and political processes. In order to use this wealth of information to ask and answer interesting questions, the researcher must first carefully curate the corpus of interest and then represent those documents numerically. While more complicated because of the sheer amount of information stored within texts, decisions on how to collect and numerically represent text are similar to the decisions researchers make to numerically represent other social science variables of interest. Collecting a corpus is analogous to identifying a sample from a population of interest. Numerically representing text is similar to how social scientists might use a metric like GDP to represent the economic activity of a country. These decisions simplify more complicated phenomena into numbers that can be analyzed quantitatively.

Not all of the information encoded in the text will be useful to our question – the challenge is deciding what information is relevant and what can be discarded. In this part, we overview a few different recipes for representing texts—from the "bag of words" model to distributed representations—that can be adapted to individual research questions and tasks. Often, theory will not be able to guide us to selecting the best representation, and the choice will require extensive validation.

We stress throughout this part that our goal is not to find one right model of language or one correct representation of the text for all research questions. As social science researchers, we are not studying the text itself, but rather the political, economic, and social processes that are reflected in it. Thus, what information we retrieve from the text will depend on what we want to discover, measure, or infer. And that will depend on the goal of our research.

We begin by building off Chapter 2's six principles to emphasize four specific principles that inform how we decide to collect and represent a corpus. We then introduce the examples that we use throughout the part that show how research-driven representation of text can be effective in uncovering social phenomena and answering historical questions.

3.1 Principle 1: Question-Specific Corpus Construction

The usefulness of a corpus depends on the question the researchers want to answer and the population they want to study.

Whether or not a corpus is useful will depend on the research question. The analysis of Twitter data, for example, might not be the most accurate way to gauge US public opinion about climate change because only a minority of the population uses the platform and, of those, few tweet about climate change. However, it could be a useful way to study how elected officials interact with their constituents online, or to measure how activist groups try to change the conversation.

When the research question is well defined, we urge researchers to articulate the quantities of interest and a population of interest before beginning to collect documents. They can then evaluate the utility of potential corpora by judging whether a corpus reflects the population they are interested in and whether the quantities they are interested in can be measured from the text. This doesn't mean biased or incomplete corpora always need to be completely discarded. Incomplete corpora can also be useful, if the researcher understands the sources and consequences of incompleteness and can articulate them to the audience.

As we cover extensively in the Discovery part of this book, researchers may not always begin to explore a corpus with a well-defined research question. For example, a researcher might have a corpus on hand and not know what is in it or what it could measure. Exploring the corpus both by conducting some quantitative analyses paired with in-depth reading could lead to new insights and theories that could later be tested. In these cases, we suggest that as researchers refine their question, quantities, and population of interest, they revisit the corpus selection step and collect new data once they have decided on a question.

3.2 Principle 2: No Values-Free Corpus Construction

> *There is no values-free construction of a corpus. Selecting which documents to include has ethical ramifications.*

Constructing a corpus requires special attention to ethical issues about representation and privacy. As we will discuss extensively in this section of the book, corpus construction may inaccurately reflect the population of interest because of differences in resources and incentives of individuals to write and make available text data. Even when texts have been produced, institutional and governmental policies might skew data retention and availability. The case of internet censorship, discussed in the previous chapter, is a particularly extreme version of this kind of selection. When researchers focus on text, they often end up focusing on the population producing the documents, a shift in focus that is important to acknowledge. Failure of the corpus to equitably reflect the population can lead to inaccurate conclusions and inaccuracies of the models for subgroups of the data (Olteanu et al., 2019; Tatman, 2017).

Text corpora also represent language as it used by the document authors and not in a neutral way. In certain settings authors use harmful language, such as of racial stereotypes or hate speech. In some applications, identifying this language (for example, to remove it) is the intention of the research task itself. In other applications, such as predictive language models, inclusion of harmful language in training data can perpetuate the problem in downstream applications (Bolukbasi et al., 2016). While certainly not the only cause of systems that perpetuate unfair outcomes, the data used can be an important part of the process.

Text frequently contains personal or identifiable information. This creates complexity in two distinct kinds of situations. When documents are generated explicitly for the purposes of research—for example, for an open-ended response in a survey—researchers should follow standard procedures in academic research for obtaining consent, ensuring the removal of obviously identifying personal information, and protecting human subjects from harm. These concerns are particularly pressing when the text data is being merged with other sensitive data because it is especially difficult to know what details in text might be identifying. Lundberg et al. (2019) offers an informative example of how to conduct a privacy threat analysis for sensitive social data.

More frequently, documents are collected from some public source and are generated without the intention of their being used in research, often without any simple way to obtain informed consent. Many researchers reason that because the author chose to publish a document, it is ethically permissible to use it for research. Yet, there may also be reasons to believe that the dichotomy of public and private oversimplifies privacy concerns in the social media age (boyd and Crawford, 2012). Nissenbaum (2020) reasons about these issues through the lens of "contextual integrity," which considers not just whether the text is public but also the contextual norms under which its authors expected that information to be accessed. The ethical situation becomes even more complex in the context of information that is only semi-public, such as that available through a closed Facebook group or a mailing list that requires a sign-up form. Salganik (2018, Chapter 6) provides an excellent discussion of common frameworks for thinking through these privacy concerns in the more general setting of data in the digital age.

3.3 Principle 3: No Right Way to Represent Text

> *There is no one right way to represent text for all research questions.*

Just as the corpus depends on the research question, so does the representation. There is no one right way to represent text for all questions. In some cases—as we'll detail more below—the researcher may only be interested in one word, and representing the text will only involve an indicator for whether that particular word is included in the text. In other cases, a researcher may want to include indicators for most words within the text, but exclude common words such as `the` and `and`. In other cases, the only words of interest might be these common words. For other questions, not only whether a word is included, but also a representation of how that word is related to other words might be required, or information about the context of the word as well as the word itself.

We suggest that researchers choose the simplest representation that will contain sufficient information about the what they hope to measure in the text. What quantity is the researcher trying to measure? How might this social phenomenon manifest itself in the text? In some cases, a simple representation will adequately reflect the phenomenon the researcher is trying to capture. In other cases, a more complicated representation may be required to measure the quantity of interest. In many cases, we will not know ahead of time what the best representation is. In these cases, we will use a representation that includes many different features from the text and then use a model to extract those that are most important according to a particular measure of what we are interested in studying.

3.4 Principle 4: Validation

The best assurance that a text representation is working is extensive validation.

There are many different decisions that need to be made in order to represent text as numbers. In many cases, the number of ways to represent the text in a corpus will be too many to enumerate. How will researchers know that they have selected the "right" representation for their research question?

Here, we suggest that researchers rely extensively on validations that we detail throughout the book. We will know that our representation is working in a measurement model if the measures that we have created using that representation align with validated, hand-coded data and facts that we know about the social world. We will know that representations are working in a prediction model if we achieve high accuracy. And we will know that a particular representation is working in causal inference if we can replicate the experiment and achieve the same results. If our validations show that our measures, predictions, and inferences are not working, we will need to revisit our approach—both our representation and our model—to understand why.

We do not believe that a research finding must be robust to the many different ways that the text could be represented, as is common practice in other areas of social science. Some representations will be more appropriate for different tasks—achieving better accuracy or facilitating better discoveries. Evaluating the robustness of our findings over many different representations is not necessary if one representation clearly dominates the others. Instead, we suggest that researchers use external validations to find the best representation for the quantities they are interested in measuring and ignore representations that are not useful to their question.

Having laid out our four principles, we now provide a brief preview of two example applications. While we will reference many applications throughout Chapters 4–9, we will return to these two particular applications several times.

3.5 State of the Union Addresses

Throughout this section, we will use a corpus of 236 State of the Union addresses from 1790–2016.[1] Given to the US public each year by the president in either written or spoken form, the State of the Union (SOTU) corpus provides a snapshot of the policy priorities of the US executive branch throughout its over 200-year history.

Several studies have used the SOTU corpus to examine trends in the nature of political speech over time. For example, Rule, Cointet, and Bearman (2015) use a variety of text analytic techniques to detect shifts in the language of these political speeches. In doing so, they argue that the discourse of the SOTU corpus shifted after 1917, when the topics of "Statecraft" and "Political Economy" morphed into broader discussions of "Foreign Affairs" and "Public Policy."

In another example, Benoit, Munger, and Spirling (2019) analyze the SOTU speeches in the context of readability. They develop a measure of textual complexity which they apply to the SOTU corpus. They find, overwhelmingly, that the textual complexity of speeches has decreased substantially over time. Comparing SOTU speeches given in the

[1] Data was obtained from Arnold and Tilton (2017).

same year in written and spoken form, they find that textual complexity is much higher in written speeches, and the increasing likelihood of presidents giving the SOTU in spoken form might explain some of the trend to less complexity.

In the following section, we will use the SOTU corpus to introduce how to represent text as data.

3.6 The Authorship of the Federalist Papers

In one of the earliest instances of quantitative text analysis, Mosteller and Wallace (1963) set out to identify the authors of the unsigned Federalist Papers. The Federalist Papers are 85 documents written by a combination of Alexander Hamilton, John Jay, and James Madison in the late 1700s supporting the American Constitution, which was then under debate. The first 77 of the papers are of interest to Mosteller and Wallace (1963) and were published anonymously between 1787 and 1788 in New York newspapers in support of ratifying the Constitution.

While all the documents are unsigned, the authorship of 73 (of the 85) of them are uncontested. For twelve documents, it is believed that they were written by either Hamilton or Madison. For an additional three documents, it was known that they were written jointly, but which author wrote which section has been disputed (Mosteller and Wallace, 1963). This is partly because the authors themselves provided conflicting lists of authorship. Hamilton provided a list of which individual authored each paper two days before he died in a duel with Aaron Burr in 1804. Madison, however, provided a different list of the authors in 1818 (Adair, 1944).

Like many of the studies we will describe in this book, according to Mosteller's biography, the idea to study the question of authorship in the Federalist Papers came from a conversation between himself, a statistician, and a political scientist, Frederick Williams.[2] Inspired by this conversation, Mosteller and Wallace (1963) set out to use statistical methods to try to distinguish the authors on the basis of their writing style, rather than the content of the essays. They made the observation that what they called `filler words` distinguish the two authors—Hamilton and Madison use words like `upon`, `by`, and `to` at different rates in documents known to be written by the two authors.

They split the documents into blocks of text of about 200 words each, in order to be able to distinguish joint authorship within one document. They then represented the text by counting the use of filler words in each of these blocks. Unlike the technology we have today, simply counting the words in the Federalist Papers proved quite challenging in 1959–1960. In Mosteller's words:

> The words in an article were typed one word per line on a long paper tape, like adding machine tape. Then with scissors the tape was cut into slips, one word per slip.... When the counting was going on, if someone opened the door, slips of paper would fly about the room. (Mosteller, 2010, pp. 53–54)

Once the words were finally counted, the authors created a statistical model that uses the filler words to predict the authorship of papers using existing documents known to

[2] "When I worked at the Office of Public Opinion Research . . . I got to know Frederick Williams, a political scientist. One day in 1941, Fred said, 'Have you thought about the problem of the authorship of the disputed *Federalist* papers?' I didn't know there were *Federalist* papers, much less that both Hamilton and Madison had claimed authorship of them." (Mosteller, 2010, p. 47)

be written by the two authors. Essentially, for a given 200-word passage, the model asks whether the rates at which the filler words occur are more like those from known Hamilton documents or known Madison documents. The model distinguishes authorship in the unknown documents quite decisively—attributing all 12 documents with unknown authorship to Madison.

Keeping in mind the image of cutting one word at a time from a text and placing it into piles, we use Mosteller and Wallace (1963) as a simple example of the bag of words approach to representing text—where each document is represented by how many times each individual word of interest appears within the text. However, the example is also useful because it illustrates the way the decision of how to represent a text depends entirely on the research question of interest. While for many social science research questions, the rate at which a text contains the word `upon` is not useful information, for this particular question, these "filler words" are most important.

3.7 Conclusion

In this chapter we've covered the principles of corpus selection and representation that will guide our thinking in Chapters 4–9. These principles parallel our six main principles from Chapter 2, placing a focus on the question-specific nature of the corpus construction, urging researchers to abandon the notion that there is a single right representation for all purposes, and in light of that to focus on validation. We now turn to selecting documents (Chapter 4) followed by our first algorithm for representing text as data (Chapter 5). We will return to the State of the Union and the Federalist Papers examples throughout.

CHAPTER 4

Selecting Documents

Text analysis always begins with a process of choosing the documents to analyze. Sometimes this choice is dictated by availability. Particularly before a research question is developed, researchers may grab as many documents as they can in a particular domain to begin searching for interesting ideas. In other settings the goal of the analysis might be to understand the contents of a particular document collection. In the immediate aftermath of disclosure of large sets of documents, such as the 250,000 US State Department cables leaked in 2010 or the 11 million financial documents in the Panama Papers, the goal for many journalists and researchers was to explore those specific collections of documents. Similarly, the Mosteller and Wallace (1963) analysis of the Federalist Papers is focused on an authorship attribution-specific corpus of text, and does not draw inference to other documents beyond it.

Many of the examples we cover in this book, though, are cases where researchers have or develop a well-defined question of interest and select the corpus to answer their particular question. For example, Catalinac (2016*a*), whose work we discussed in Chapter 1, was interested in understanding how the electoral strategy of candidates running for office in Japan changed in response to the electoral reform. She decided to use a corpus of electoral manifestos because these manifestos are made in the same format by all candidates, are directly given to voters, and have not changed in their design over time. Therefore, they are representative of her population of interest—candidates—and also what she was interested in measuring—candidates' strategy of persuading voters to vote for them.

In other cases, the researcher's goal shifts over time and so too does the collection of documents. As we described in Chapter 2, after King, Pan, and Roberts changed their research topic to measuring censorship, they switched their data collection method to a stratified sample based on keywords that they believed would be associated with censorship. As we emphasize throughout the book, the research process isn't always a strictly linear progression through the research steps, and corpus construction may need to be done more than once as researchers iterate toward the most interesting question and the most compelling answer.

In this chapter, we begin by discussing the population and quantity of interest. Once you've decided on a question, population, and quantity of interest, how well you can measure these quantities will depend on whether your corpus represents this population well. We then examine four types of bias that might influence how reflective a particular corpus is of a given population.

4.1 Populations and Quantities of Interest

The usefulness of a corpus depends on the question the researchers want to answer and the population they want to study. Without digging into the data, the researchers may not know what they *can* measure, or how the corpus relates to key questions in their area of expertise. At this stage in the process, the analysts are in the discovery phase; they must explore, read, and describe the texts to better understand what questions can be answered. Eventually, they will develop a research question and revisit how well the sample reflects the population they are interested in, which will typically require subsequent data collection or refinement of the sample.

Other times, the researcher begins with a question of interest, which will make it more clear at the outset whether or not the texts are relevant. Questions of interest typically reference or imply a *population of interest*. For example, if the question of interest is, "How do likely voters in the United States feel about the President?", the population of interest is likely voters in the United States. The question clarifies whether or not the corpus at hand reflects this population. Using a corpus of tweets to answer this question may result in *sample selection bias* because, among other reasons, users of Twitter skew substantially younger than likely voters.

The research question suggests how the texts should be compressed or summarized to be used in analysis. We often refer to these summaries as the *quantities of interest*. A quantity of interest is a numeric value (or set of values) that is of particular interest to the researcher. In the Twitter example, an analyst might be interested in how much users talk about a particular topic or how many tweets about the president are positive versus negative. The population and quantities of interest inform how well a corpus can answer the question of interest. If the corpus does not reflect the population of interest, or the quantity of interest cannot be calculated from the texts within the corpus, then answering the research question will be difficult with that particular corpus.

Researchers ask questions that focus on a wide variety of populations, domains, and time periods. For some research questions, text data may be used similarly to population surveys, where researchers hope to reflect the underlying opinions, activities, or demographics of a country's people. For example, scholars have begun to use social media data on platforms like Twitter to describe everything from political opinions (O'Connor et al., 2010) to movie preferences (Asur and Huberman, 2010), to the spread of the flu (Lampos and Cristianini, 2010), and personal happiness in a population (Golder and Macy, 2011; Dodds et al., 2011). To accurately gauge underlying quantities of interest in the broader (both online and offline) population, these analyses depend on accurately accounting for the ways in which online populations are different from offline populations. Even within online populations, those who choose to engage in discussions about politics may be very different from those choosing to engage in discussions about entertainment, all which must be taken into account when generalizing to the population that does not discuss the topic within the data (Barberá and Rivero, 2015).

In other cases, the population of interest is a much smaller subset of the entire population. For example, Blaydes, Grimmer, and McQueen (2018) explore the divergence in political thought between Muslim and Christian societies. To reflect both Muslim and Christian political thinkers, they gather political advice texts from the sixth to the seventeenth centuries from both the Islamic tradition and Christian Europe. The authors' goal is not to gather *all* Muslim and Christian writings, or to reflect all thoughts on these subjects within these populations, but rather to reflect those with the most influence on political thought in this time period. In the digital humanities, analyses

may be focused on exploring thematic trends in a particular genre, for example, Danish ghost stories (Abello, Broadwell, and Tangherlini, 2012). Appropriate selection of texts in these cases will often rely on the ability to distinguish between stories that contain a subject like ghosts from those that do not.

Even if the corpus is not representative, careful analysis can lead to valid insights about the population of interest. Gill and Spirling (2015) use historical records and leaked data to understand what types of information the United States government keeps secret and what types it declassifies. Of course, the authors do not have access to *all* classified information; they only see declassified or leaked records. But in comparing information that is unclassified to that which was classified at some point, the authors can infer some of the types of topics that remain secret, and can offer insight into what might be missing.

4.2 Four Types of Bias

Once the researchers have a question, a population, and quantities of interest, meaning that they have moved from the discovery to the measurement, causal inference, or prediction stage of the analysis, they should think carefully about the potential sample selection biases in their corpus. In this section, we provide a starting checklist of common sources of sample selection bias that researchers might consider when they have a question and candidate corpus of interest. Not all of these sources of bias will pertain to all research questions and certainly some of the factors we consider may be *beneficial* for answering some types of questions. However, we have found that for many questions and populations of interest, the four types of sample selection bias below reappear frequently when analysts use text for social science inference.

4.2.1 RESOURCE BIAS

Texts are expensive to produce, gather, and collate. *Resource bias* refers to the fact that texts often better reflect populations with more resources to produce, record, and store documents. This issue is especially salient when access prevents entire portions of the population from being represented at all. Illiteracy, limited access to the internet, or barriers to entry can render large swaths of a population invisible to scholars studying texts.

Recording and preserving historical activity is a costly process that can require substantial resources. Events are more likely to be recorded if the press is present (Snyder and Kelly, 1977; Danzger, 1975). Archives might only contain documents that citizens found convenient to store and may be woefully unrepresentative of the population of interest. Texts may have been lost, burned, or not stored, and stories may have never been transcribed. Communities and individuals with the capabilities to store and preserve texts will likely be very different than those without, and as such historical documents should not be considered a random sample of a larger set. For example, Pechenick, Danforth, and Dodds (2015) point out that the Google Ngrams corpus, a large collection of English texts, contains increasingly disproportionate amounts of scientific texts over the course of the twentieth century.

Government document availability may also reflect local resources. Some governments may have the personnel to transcribe meetings and make them available to the public; others will not. For example, Sandhya Kambhampati, a journalist at ProPublica, writes about requesting information from the state of Illinois via the Freedom of

Information Act. She notes that documents from some agencies were not already digitized. Because of scanning costs, some of these records could not be made available to the requester.[1]

4.2.2 INCENTIVE BIAS

Incentives and strategic behavior can also drive the production and retention of documents, a process we call *incentive bias* (Herrera and Kapur, 2007). Social interaction is performative (Goffman, 1959) and individuals as well as organizations have incentives to fail to record or to hide or destroy evidence that could cast them in a negative light. Individuals may be more likely to post social media that reflect the happiest or most successful aspects of their lives. For researchers studying interactions on social media, this concern may not be an issue, but for those who are trying to measure true emotional states of users, how individuals portray themselves may bias results. Similarly, politicians and governments may force removal of news and social media that undermines them, censoring others or self-censoring to frame political conversations (House, 2015; Boydston, 1991; Gup, 2008; King, Pan, and Roberts, 2013; MacKinnon, 2008).

While transparency laws and initiatives make texts more accessible to researchers, they also change the incentives of those whom it affects. Email transparency within organizations may create incentives for members of these organizations to talk on the telephone about sensitive issues or hold informal meeting where discussions are not recorded. Transparency initiatives might make government leaders curtail their political participation if they are afraid of making missteps (Malesky, Schuler, and Tran, 2012). These initiatives might also be more likely in certain political contexts; for example, Berliner and Erlich (2015) show that states in Mexico with more political competition passed transparency laws earlier than states without political competition.

When researchers are interested in using documents as a reflection of government internal workings, researchers should seek to understand the incentives behind document collection and the process through which the data are made available. A brilliant example of this is Cheryl Schonhardt-Bailey's *Deliberating American Monetary Policy: A Textual Analysis*, which examines transcripts of the Federal Open Market Committees (FOMC) and Congressional committees on banking to understand the role of deliberation in making monetary policy (Schonhardt-Bailey, 2013). Schonhardt-Bailey (2013) uses unsupervised machine learning to explore and measure deliberation and strategies of persuasion in these transcripts. However, she then goes one step further and interviews many of the individuals who appear in the transcripts. Many of the questions she explores in the interviews relate not only to her research question but also to what the transcripts represent—how candid members of the FOMC felt they could be in their discussions, whether their remarks were pre-prepared, and how transparency of the transcripts themselves affected their deliberation (Schonhardt-Bailey, 2013, p. 370).

4.2.3 MEDIUM BIAS

The technologies and types of mediums in which texts are recorded will play an important role in the types of content that will be reflected in them; we will call this *medium bias*. Until 2017 Twitter only allowed 140 characters in one tweet, necessitating users to use abbreviations and links for further treatment of the topic. However,

[1] Kambhampati, Sandhya. "I've Sent Out 1,018 Open Records Requests, and This Is What I've Learned." *ProPublica*. Jan. 4, 2018. https://www.propublica.org/article/open-records-requests-illinois-foia-lessons

140 characters has very different implications across languages—in languages that require fewer characters per word, like Chinese and Japanese, users will be able to express more content in one post than in languages that necessitate many characters for each word like English (and notably when the character limit was doubled in November of 2017, the other languages were not included). Further, the mode of expression in social media changes with advances in technologies. As users have been able to incorporate pictures, videos, and live streams, the role of the accompanying text changes. Users interested in text analysis in social media should be aware of other types of data—links, pictures, and videos—when describing their content and how technological capabilities have changed over time.

The technology and evolution of the medium can create different cultures for users so that the text can only be understood in the context of the platform. Snapchat, a social media platform where posts are generally only accessible briefly before being destroyed, will create a different environment for users than Twitter, where posts are public and more permanent. Because online platforms tailor their website to individual users' predicted interests, users may all have a different experience of the content on the platform, which might be reflected in their posting behavior. Researchers should consider biases that might be produced from different experiences users have between platforms and within individual platforms to understand the underlying data generating process of how the texts were produced.

Text outside of social media is similarly influenced by its medium. Meeting transcripts or notes may reflect the content of the conversation, but not tone of voice, body language, or other forms of expressed emotion (Schonhardt-Bailey, 2017; Dietrich, Enos, and Sen, 2019). Handwritten notes may contain drawings or doodles. Text messages may contain emojis, which are central to content. When possible, researchers should read and interact with texts in their original context to understand how the social events of interest are translated by the medium into text, even if text analysis tools only take the text into account.

4.2.4 RETRIEVAL BIAS

As texts are sometimes selected using statistical methods, *retrieval bias* can sometimes affect the selection of a corpus. For example, Abello, Broadwell, and Tangherlini (2012) use computational methods to analyze Danish ghost stories. For their work, they use a hand-collected set of stories from the 19th century collector and Danish schoolteacher Evald Tang Kristensen, categorized by Kristensen as "ghost stories." It's truly remarkable to have such a hand-curated dataset, although of course what is included in the dataset reflects Kristensen's own social network and categorization scheme.

As an alternative, imagine trying to retrieve a global list of ghost stories with keywords.[2] To do so, you might select every book in the library that includes in the title `ghost`, but such a strategy would be riddled with errors. Some book titles that contain the word `ghost` may have nothing to do with ghost stories; for example, *Hungry Ghosts: Mao's Secret Famine* deals with the famine in China rather than with ghost stories. Similarly, many books that don't contain the word `ghost` may still be ghost stories, *A Christmas Carol* is a ghost story without `ghost` in the title. While these errors may

[2]The authors go through this thought experiment in their article, and develop statistical methods that allow for more efficient exploration of the collection.

at first seem random, they are often systematic. `ghost` may orient the analysis toward one type of genre, missing books about phantoms, which may have different themes.

This type of keyword selection of corpora is a frequent problem in the analysis of text because researchers often have a focused population of interest—like ghost stories—that is a subset of a larger, unwieldy corpus—such as all books. The problem is made worse by the fact that humans often have difficulty thinking of words off of the top of their head. Limited recall means that an analyst may miss some of the most important words on which to select. One approach is to take a computationally driven approach to corpus selection by applying machine learning to the problem of selecting documents, using many of the same techniques we will introduce in the measurement part of the book (King, Lam, and Roberts, 2017; D'Orazio et al., 2014; Abello, Broadwell, and Tangherlini, 2012). Regardless, researchers should be transparent about how they selected the texts and what potential biases this might produce in what population the text reflects.

While keywords are a simple algorithm for retrieving documents in theory, the reality is often more complicated because researchers often rely on third parties to provide data for them. Data access through Application Programming Interfaces (APIs) or other search functions may not have a transparent underlying process translating search queries into results. This could be for reasons as simple as some interfaces ignoring capitalization while others being case-sensitive. Some systems may return only a random selection of the relevant document, or only the most relevant selection. The process through which the API makes documents available will affect the population the text represents and the research questions the text can reliably answer.

4.3 Considerations of "Found Data"

Often, text corpora are a form of "found data," data not collected by purposeful sampling of a larger population, but instead made available by governments, individuals, or other institutions. What data is made available is dictated not by the research design, but by the resources and incentives of the institutions and individuals producing them. This type of data has frequently been criticized because it may not represent the population of interest, leading researchers to draw flawed or even flat-out wrong conclusions from the data (Harford, 2014).

Despite these limitations, the use of data in a research project is dictated by data availability and must be subject to ethical and legal constraints, as we discussed in Chapter 3. For any dataset, researchers should clearly communicate to the reader the logic of selecting that dataset to answer the question of interest, as well as what relevant information the dataset may not contain or represent.

4.4 Conclusion

Selecting documents is a crucial part of the analysis process but is subject to many constraints based on availability, medium, and ability to retrieve them out of a larger collection. The core challenge is to create a corpus that is representative of the population of interest. This is, of course, notably challenging if you have not chosen a population of interest. Are you interested in sentiment on Twitter or in the US population? Do you care about the political preferences of everyone or just likely voters?

These challenges may be further exacerbated by the need to draw comparisons across time or source. Social media platforms like Twitter and Facebook are constantly

changing their policies, which make resource bias, incentive bias, medium bias and retrieval bias all moving targets. Indeed the very nature of the phenomenon itself might be changing (Munger, 2019). Larger cities might have more established newspapers that can do more investigative reporting. Does this mean that corruption—to choose one example—happens more in big cities than in small localities or just that we know more about it? A quite common problem we have observed in practice is that with document collections over a multi-decade time horizon there are essentially always more documents later than earlier.

These challenges are question-specific and the temporal or source inconsistencies that plague one research question might be exactly the object of interest for another. How problematic these issues can be also depends on the particular method used to analyze the texts. We will return to this issue throughout the book.