Chapter 32
Static Text-Based Data Visualizations:
An Overview and a Sampler

Shalin Hai-Jew
Kansas State University, USA

ABSTRACT

Data visualizations have enhanced human understandings of various types of quantitative data for many years. Of late, text-based data visualizations have been used informally and formally on the WWW and Internet as well as for research. This chapter describes this phenomenon of text-based data visualizations by describing how many of the most common ones are created, where the underlying textual datasets are extracted from, how text-based data visualizations are analyzed, and the limits of such graphical depictions. While this work does not provide a comprehensive view of static (non-dynamic) text-based data visualizations, many of the most common ones are introduced. These visualizations are created using a variety of common commercial and open-source tools including Microsoft Excel, Google Books Ngram Viewer, Microsoft Visio, NVivo 10, Maltego Tungsten, CASOS AutoMap and ORA NetScenes, FreeMind, Wordle, UCINET and NetDraw, and Tableau Public. It is assumed that readers have a basic knowledge of machine-based text analysis.

INTRODUCTION

In the present age, researchers have access to more text-based data than ever. These include core textual data from digitized texts, documents, articles, and other formalized writing (which has gone through professional vetting and editing). There is also metadata describing other data (in multimedia, textual, and other forms); these may include tags on multimedia objects, bibliographies, publication abstracts, and other metadata. There are informal sources of textual data, such as private collections of personal papers, journals, and letters, known as gray literature. There is elicited information, such as through online surveys and "dropbox" sites on the Web. Then, too, there are (four) zettabytes of user-generated data on the World Wide Web (as of 2013): these are on social media platforms (blogs, wikis, microblogs, and

DOI: 10.4018/978-1-5225-1837-2.ch032

others), data repositories, pastebins, websites, learning management systems, and other online venues. There are intranets and reams of email data. There are automated (if not fully accurate) translations of voice-to-text from video and audio files. What this means is that there are plenty of textual sources of information (whether "born textual" or rendered textual) that may be explored for research (literature reviews, domain-analyses on the Web, survey and interview analyses, microblogging Tweetstream analyses, and others).

With so much "big data" available, machine-analysis of various texts often has to be applied in order to enhance what is knowable and communicable. Large-scale mining and analysis of this volume of data highlights the challenges of data veracity and its "velocity" or rate of change over time (Plate, 2013, n.p.). Core to these text analyses are text-based data visualizations which communicate insights about the text: patterns, anomalies, themes, and other insights, in a human-interpretable way. This work provides an overview of some basic types of machine-based text analyses and some common types of static (non-dynamic) text-based data visualizations, along with some basic approaches to their interpretation. It is important to define "text-based" in this context. For this context, "text-based" will include the two central meanings: (1) underlying text corpuses which inform the data visualization and (2) the expression of concepts which are expressed through symbolic textual means. The first type would include visualizations that are closer to the original raw text and are often machine-drawn; the second type is in a sense more processed and abstracted from the underlying data (and may be more abstract) and are often manually drawn. The first type of visualization will tend to be drawn based on machine algorithms. The latter is a form of graphic ideation and will tend to be drawn based on trained human conventions of illustration and diagramming.

A REVIEW OF THE LITERATURE

Language is a core element of culture, and words carry critical concepts. Language is a central element of human socializing and inter-communications. It is the core vehicle for the dissemination and communication of research and of popular understandings of the world. Language is polysemic or many-meaninged; based on its interpretation, it may convey a range of ideas and impressions. Messaging may be understood whether the initial communicator consciously or unconsciously intended to share particular information; meaning is not necessarily dependent on the conscious intentions of those wielding the language. The centrality of language in the human experience in part explains the power of machine-based text analysis. As a data source, language is often rendered into textual form for analysis (whether the original data was a video, audio, image, or other multimedia type).

The textual data analyzed in most data analytics software programs are considered "unstructured" because they do not fit into databases in pre-labeled data cells. The raw textual data is heterogeneous, or it comes in many forms. They are high-dimensional, consisting of a large number of variables or facets. The data may be noisy, with extraneous words which are not particularly meaningful for a particular query or research question. (The "noise" here would suggest that there are distractors that may lower the signal-to-noise ratio, which would raise the probability of coming to erroneous conclusions from the dataset. There would be increased false positives and false negatives mixed in with the accurate signals. Information unreliability may be partially off-set by having large amounts of data available for wider sampling but then being able to de-noise that set by honing in on the relevant signals.) The textual data is inherently ambiguous, in part because language itself is multi-meaninged and defined in part by context,