Quantitative text analysis

Arindam Basu[1]

1 University of Canterbury

Quantitative text analysis refers to the process of analysis of text data using statistical procedures. In conducting quantitative text analyses, researchers use automated and a systematic method to process large amounts of text. For example, a number of large policy documents can be processed using data science methods to identify the different types of words and topics embedded in these documents. Quantitative text analysis can be conducted on a single document, or on a number of different text documents. Besides, quantiative text analysis can be conducted on social media posts, such as Twitter feeds, Facebook posts, blogs, and transcribed data from video sharing sites such as the Youtube and others. Quantiative text analysis is also referred to as "computational text analysis".

Steps of conducting a quantitiative text analysis

Quantitative text analyses consist of five related steps:

1. Step 1: read the text into the computer programme used for computational text analyses
2. Step 2: Tokenize and convert the text to its consituent words or features by identifying word or sentence or character level boundaries and then extracting individual characters, words, or combination of words, and sentences or other syntactic level units. In this way, from the corpus of the text, individual words or units of data analyses are extracted and stored into an object which is then analysed in different ways.
3. Step 3: Extraction of document feature matrix (or document term matrix). -- Once the main corpus of the text is broken down into its constituent words, other properties of the document or documents and the meta-data are also extracted. The process of extracting words from the document is referred to as n-gram analysis. With single words, or n-grams if more than one words are used, and then depending on the value of n, such word combinations are referred to as bigrams (n = 2), or trigrams (n = 3), and correspondingly higher levels of ngrams. Once the documents are converted to the formats where their constituent words and associated meta data are made available, such individual words are referred to as features. At that point, counting and graphing of features provide insights into the syntactic organisation of the text

corpus. At this step, the analysts also identify the 'most important' terms present in a corpus by identifying the matrix of term frequency and inverse document frequency of terms. In this way, not only the most frequently used terms are captured and weighted, but also those terms that occur sparsely in the text corpus are captured and presented. Once the document is converted to a document feature matrix (or document term matrix) -- the sparse matrix based on the terms present in the document, the analyst can analyse the 'emotion content' of the document using sentiment analyses, and can classify the document and identify topics present in the corpus of text.

4. Step 4: Sentiment analyses. -- The goal of sentiment analyses is to identify the emotion content of the document that is being processed. In order to conduct sentiment analyses, the analyst first tokenises the document so that the document is now converted to its constiuent terms. Each term can contain latent emotions that are then made explicit. This means using a lexicon to identify emotion content of each word (referred to as 'sentiment' content of each word). Different lexicons are available for different domains of inquiry, and sentiments or emotional contents are also coded in a variety of ways. For example, the popular Bing lexicon [1][2] consists of keywords that are classified as positive or negative emotions.
5. Step 5: Classification of the text and identifcation of patterns and topic modelling. -- In this step, the analyst takes the corpus of text that has been converted to tokens and constituent features and using these features, classifies the document corpus to identify either nascent themes or latent themes or explicitly identifies themes from the corpus and visualises them. Several algorithms exist to identify the topics within the documents. The popular R package "topicmodels" [3]

Welbers et.al (2017) has suggested the following workflow for quantitative text data analysis [4]

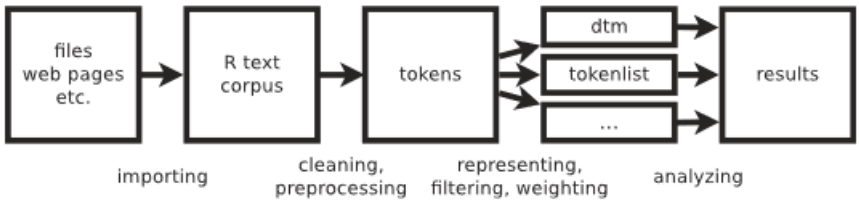

Figure 1. Order of text analysis operations for data preparation and analysis.

Figure: Welbers et.al (2017) plan of quantitative text data analysis (taken without permission from https://www.tandfonline.com/doi/pdf/10.1080/19312458.2017.1387238)

References

1. ^ *Ding X, Liu B, Yu PS. (2008). A Holistic Lexicon-Based Approach to Opinion Mining. WSDM*
2. ^ *https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html*
3. ^ *Bettina Grün, Kurt Hornik. (2011). topicmodels: AnRPackage for Fitting Topic Models. J. Stat. Soft., vol. 40 (13)*
4. ^ *Kasper Welbers, Wouter Van Atteveldt, Kenneth Benoit. (2017). Text Analysis in R. Communication Methods and Measures, vol. 11 (4), 245-265.*