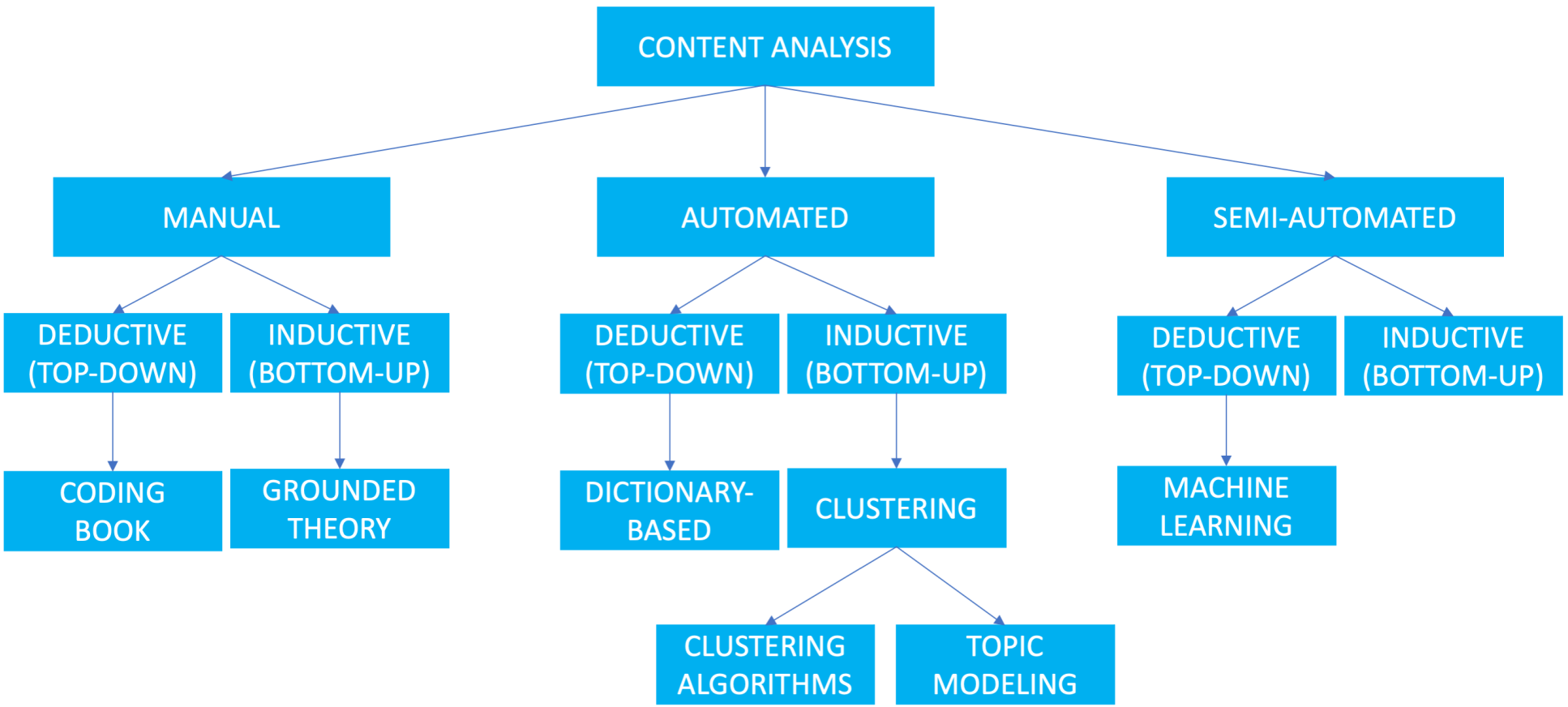
Text and Data: Manual Content Analysis and Computational Techniques

Nicola Righetti

Summary

The purpose of this lesson is to provide a brief overview of content analysis approaches and techniques, both manual and statistical and computational.

Approaches to content analysis



Manual content analysis

Manual content analysis is a research method used in the social sciences for systematically analyzing texts or other forms of media content.

This approach involves a detailed and nuanced examination of the material, where researchers manually code and interpret the content to identify patterns, themes, or other features of interest, often using predefined categories and rules.

The goal is to transform qualitative data into quantitative data, allowing for more objective analysis and comparison. This method is distinguished from automated or computational content analysis by its reliance on human judgment and interpretation.

Coding book

Category	Definition	Examples	Coding Rules
C1: high self confidence	High subjective conviction to have successfully coped with the situational demands, which means - to be clear about the demands and their coping possibilities, - to have a positive, hopeful feeling in handling the situation, - to be sure to have coped with the demands on ones own efforts.	"Of course there had been some little problems, but we solved them all, either by myself, or by the student, depending on who made the mistake. Everyonecan make mistakes." (17,23) "Sure there had been problems, but in the end we had a fine relationship. We got it all together."(27,33)	All three aspects of the definition have to point to "high"self confidence, no aspect only "medium "self-confidence. Otherwise C2: "medium self confidence"
C2: medium self confidence	Only partly or fluctuating conviction to have successfully coped with the situational demands	"Quite often I found it hard to maneuver through the problems, but finally I made it."(3,55) "Time by time everything got better, but I couldn't tell if it was me or the circumstances." 77, 20)	If not all aspects of definition point to "high" or "low"
C3: low self confidence	Conviction to have badly coped with the situational demands, which means – not to know what the situation exactly demands, –to have a negative, pessimistic feeling in handling the situation, – to be sure that ones own efforts had no effect on improving the situation	"that stroke my self confidence. I thought I'm a nothing – or even less than that." (5, 34)	All three aspects of definition point to low self confidence. No fluctuations recognizable

Grounded Theory

Grounded theory is a research methodology in the social sciences that differs from traditional content analysis approaches by not relying on pre-defined codes. Instead, it involves the generation of codes and categories directly from the data.

This inductive approach allows themes and theories to emerge organically from the content being analyzed, rather than being imposed based on existing hypotheses or frameworks.

Grounded theory is particularly useful for exploring new or complex phenomena, as it facilitates the development of new theoretical insights based on empirical observations.

Automated Content Analysis

Automated content analysis is a research technique that employs computer algorithms and software to analyze texts or media content.

This method allows for the processing of large volumes of data efficiently and consistently, identifying patterns, themes, or frequencies without the subjective influence of human coders.

Automated content analysis is particularly useful for handling big data sets, providing quantitative measurements, and enabling the analysis of trends and relationships in content over time. However, it may lack the nuanced understanding and interpretive depth that manual content analysis offers.

Dictionary-Based Content Analysis

Dictionary-based content analysis is a method of automated content analysis that involves using a predefined set of words or phrases, known as a dictionary, to analyze and categorize text.

This approach quantifies the presence, frequency, and context of these specified words within a text, allowing for systematic analysis of content.

Dictionary-based content analysis is often used to measure the sentiment, tone, or specific themes in large datasets. It offers efficiency and consistency in processing data, but its accuracy and depth of analysis depend on the comprehensiveness and relevance of the dictionary used.

Examples

- https://moralfoundations.org/wp-content/uploads/files/downloads/moral%20foundations%20dictionary.d
- https://github.com/kbenoit/quanteda.dictionaries
- https://github.com/quanteda/quanteda.sentiment

Clustering

Clustering approaches to content analysis are automated methods that group similar pieces of content together based on their characteristics or features, aiding in the organization and understanding of large datasets.

Standard clustering involves grouping texts or media content based on similarity in features such as word frequency, style, or thematic elements. This method identifies natural groupings within the data, without prior knowledge or assumptions about the categories. It’s mainly used for exploratory analysis, revealing patterns and relationships within the dataset.

Clustering - Topic Modeling

Topic modeling, a specific form of clustering, goes a step further by identifying latent topics within the content. Techniques like Latent Dirichlet Allocation (LDA) are used to discover abstract topics that occur across a collection of documents.

Topic modeling assumes that each document comprises a mixture of topics, and each topic is characterized by a distribution of words. This method is particularly useful for uncovering hidden thematic structures in large text corpora, providing a more nuanced understanding of the underlying themes in the content.

Clustering - Large Language Models (1/2)

Clustering approaches to content analysis, including topic modeling, can be enhanced by incorporating word embeddings provided by Large Language Models (LLMs).

Word embeddings are vector representations of words, capturing their meanings based on context and usage in large datasets. When applied to clustering or topic modeling, these embeddings allow for a more sophisticated analysis, as they consider the semantic relationships between words rather than just their frequency or co-occurrence.

Clustering - Large Language Models (2/2)

In topic modeling using LLMs, the word embeddings enable the identification of topics based not only on the presence of specific words but also on the contextual similarity of word usage across documents. This results in more coherent and contextually relevant topics.

Similarly, for standard clustering, using LLM-provided word embeddings leads to groups of texts that are semantically similar. This allows for the clustering of documents that might not share exact words but discuss similar themes or concepts, providing a deeper and more nuanced understanding of the content.

Semi-Automated Content Analysis

Semi-automated approaches to content analysis refers to semi-supervised machine learning approaches, which involve a blend of supervised and unsupervised techniques for analyzing and categorizing text or media content.

In these approaches, a portion of the dataset is labeled by humans, providing a basis for the algorithm to learn and identify patterns, themes, or classifications. The machine learning model then applies this learned knowledge to categorize or analyze the unlabeled portion of the dataset.

Semi-Automated Content Analysis (2/2)

This method is particularly effective when there are large volumes of data, but limited resources for comprehensive manual labeling. Semi-supervised learning leverages the strengths of both human insight and the efficiency of automated algorithms, improving the accuracy and relevance of the analysis compared to fully automated methods.

It’s commonly used in scenarios where some expert guidance is necessary to initiate the learning process, but the scale of data makes manual analysis impractical.