CHAPTER 4

How Discrete

> If you torture the data long enough, it will confess to anything. (Attributed to Ronald H. Coase, 1960)

4.1 Metaphors with Destiny

Metaphors are fascinating and powerful linguistic devices. Over the years, numerous scholars have indeed extensively explored their manipulative talent for creating realities (see for instance, Lakoff 1992, 2004, 2008; Goatly 2007; Mio and Katz 2016). In the context of political discourse alone, for example, the study of metaphors' capacity to hide or popularise latent ideologies, justify or blame governments' decisions, or strategically attribute blame goes back decades (e.g., Musolff 2004, 2010, 2014; Goatly 2007; Ottatti et al. 2014; Viola 2020a). Though extremely powerful—'Metaphors can kill' (Lakoff 1992, 1)—metaphors are neither good nor bad per se; we simply routinely use them, often rather unreflectively, so that abstract and complex ideas can be processed in a cognitively simplified way (ibid.). What makes metaphors so effective, particularly conceptual metaphors, is their use of conceptual frames such as war, disease, sport, family, religion and others which, by evoking mental images that are familiar to the message receivers, can turn complex concepts into a simple, linear logic (Viola 2020a). It is thanks to this 'framing power' that metaphors' arguments become plausible and the proposed conclusions are perceived as

L. Viola, *The Humanities in the Digital: Beyond Critical Digital Humanities*, https://doi.org/10.1007/978-3-031-16950-2_4

unproblematic and even 'self-evident' (Musolff 2016, 133). Moreover, as we mostly use metaphors implicitly, such framing power remains typically unnoticed and so do metaphors. So, for example, in the context of the COVID-19 pandemic, when commenting on the effectiveness of Italy's decision to institute national lockdown, French Prime Minister at the time Édouard Philippe said, 'To block the country does not allow to contain the epidemy'[1] (Valeurs actuelles 2020). At the time when the comment was made, France was adopting much less drastic measures compared to Italy; therefore, the differences in the two countries' crisis management approaches needed to be justified, and in order to be accepted by the nation, the domestic strategy had to be presented to the public as the best possible solution (Viola 2022). In this particular example, the framing power is conveyed by the expression *to block the country*: the metaphorical use of the verb *to block* frames the Italian lockdown measure not only as overly aggressive but wrongly targeted: it is the country that is put to a halt, not the spread of the virus.

But metaphors are not typically found just in political discourse; scientific discourse also regularly exploits the power of metaphors to simplify complex concepts. In 2003, Blei et al. published a study which, at the moment of writing, counts 36,483 citations (2003). The paper tackled the task of modelling a collection of discrete data, for example, a corpus of texts, for efficient processing tasks such as classification and content summarisation. The authors' basic idea was to model each item in the collection, e.g., each text, according to the Latent Dirichlet Allocation (LDA) model, a generative probabilistic model for which documents are represented as distributions of sets of words statistically likely to occur together. Although the article itself was titled 'Latent Dirichlet Allocation', the technique described in the article went down in history as *topic modelling*. The reason for that may be found in the fact that the authors had decided to name the above-mentioned sets of words as 'topics', albeit their intention was not to make epistemological claims regarding the latent variables but to simply 'exploit text-oriented intuitions' (996), that is, to take advantage of a familiar image such as that of *topics.* In other words, the term *topic* was used metaphorically.

A similar observation about the metaphorical use of everyday notions to refer to techniques which are however based on specific, rather different, principles may also apply to computational techniques such as '*sentiment* analysis' and 'machine *learning*'. The metaphorical use of the terms 'sentiment', 'learning' and 'topic' may be harmless within the fields that have devised such techniques because the principles upon which they are based are very clearly defined by their creators and understood in those circles. It may on the contrary have huge consequences when these methods are passively transferred into other disciplines or practices. In his analysis of informational approaches in cancer biology research, Longo (2018), for example, critiques the extensive use of computer science terminology such as 'instructions', 'to reprogram a deprogrammed DNA' and in general the DNA described as a computer program and genes as information carriers. He argues (88):

> The informational approach in biology conflates the concept of programming on discrete data with the common-sense understanding of 'information' and 'computer program', which are vaguely familiar to everybody [...] In fact, the use of 'information' and 'programming' in biology is not scientific because it neither applies the mathematical invariants proper to information and programming, nor the theorems proper to the corresponding scientific disciplines. Instead, it transfers a vague, everyday notion and refers to 'weak' meanings.

Longo argues that the metaphorical use of mathematical and computational language has had enormous consequences for molecular biology cancer research which essentially studies cancer as the result of DNA deprogramming, inherited or otherwise caused by a carcinogen that disrupts the DNA 'encoded instructions' (92). The use of an everyday notion such as that of 'program', he continues, has also no doubt facilitated understanding among funding agencies and the public, perhaps even leading to the exclusion of alternative hypotheses. Similarly, one might argue that it is the metaphorical use of the word *topic* that explains why topic modelling has become so popular beyond computer science and in the humanities in particular: whereas not everyone may be an expert in statistical modelling, we are all more or less familiar with a fairly general conceptualisation of what a topic is. However, what humanities scholars may have not been too familiar with—and to a large extent, still aren't—is the set of assumptions behind a method born in the computer sciences and adopted in critical research.

The popularity of topic modelling beyond computer science (as well as SA and ML) is closely related to another phenomenon, well-known in linguistics: when a metaphor is adopted by a significant part of the linguistic community, language users may no longer be aware of its metaphorical use, the metaphor becomes a common meaning and so it dies (Ricœur 2003, 115). The metaphorical use of ‘sentiment’, ‘learning’ and ‘topic’, I will argue here, has certainly contributed to make these techniques very popular outside of their field of origin. At the same time, however, precisely because of this popularity, these meanings have become common meanings, i.e., ‘dead metaphors’. This in turn has major consequences: the creation of epistemological expectations that these methods will obviously disappoint (Puschmann and Powell 2018). For example, as I have discussed in Chap. 3 in reference to SA, the familiar word ‘sentiment’ creates a specific epistemological expectation, that it is somewhat possible to obtain a neutral way to assess attitudes and moods in large quantities of material. Assessment, however, requires language understanding as a prerequisite and when it comes to machines, this is exactly what they are not able to do. The post-authentic framework that I advance in this book serves also as a reminder that these terms are used as mere metaphors.

In the next section, I will discuss a more concerning aspect concealed by the use of vague, familiar notions such as ‘sentiment’, ‘learning’ and ‘topic’: the underlying process upon which these techniques are based, i.e., the elaboration of continuous information into discrete systems and the implications for causality. In discrete systems, causality is hidden because information is rendered as exact and separate points, all encoded in one dimension and according to precise instructions (Longo 2018). The three-dimensional, causal essence of information cannot be accessed by the user who, instead, is offered an altered image made up of predictions of correlations. The resulting information will still refer to its original continuous structure, but computers will only render it as a sequence of 0s and 1s, that is in discrete form, thus hiding relational causality.

In the case of SA, this distorted image is reflected in the reduction of the subjectivity of human emotions to two/three categories, scored according to probabilistic calculations; in the case of ML, the holistic, human capacity to acquire knowledge and skills through experience, logic and contextual factors is reduced to the probabilistic processing of huge, yet partial, quantities of discrete data; in the case of topic modelling, the text itself disappears and so does its entrenchment in the wider context that produced it. In all these cases, the three-dimensional, causal structure is no longer accessible nor is its historical and social susceptibility as it is all dissembled by the computational, dualistic system of 0s and 1s. This conflation of discrete data modelling with familiar notions such as ‘sentiment’, ‘learning’ and ‘topic’ has therefore certainly contributed to make these methods extremely popular outside their fields of origin, but at the same time, it has obfuscated the well-defined laws upon which they are based. Longo claims:

> This is an amazing technological achievement: by fine engineering, one may forget the underlying physical hardware and its continuous flows and just consider (and work on) the discrete software processes by writing alpha-numeric programs. (Longo 2018, 87)

In a world where all information is digital, the consequence of this amazing technological achievement is that it also presents a distorted image of knowledge because, to paraphrase David Tong, the world does not seem to be discrete (Tong 2011).

In this chapter, I first examine the implications of adopting discrete methods and technologies not just as quantitative tools in the humanities but for knowledge production in general and, more widely, for our understanding of society. Specifically, I reflect on the notions of causality and correlations in light of the considerations discussed so far about the mythicised discourse on data and technology neutrality, the dangers of using metaphorical language to refer to digital technologies and the consequential urgent need for knowledge reconfiguration inspired by symbiosis and mutualism. I then proceed to examine the text mining technique of topic modelling and the premises on which it is based with a special focus on its use of discrete mathematics to encode information. Finally, I illustrate how applying the post-authentic framework to topic modelling can facilitate critical engagement with this technique, especially in humanities research.

In my discussion, I argue that such engagement can only happen by maintaining a sustained connection with the digital object and I demonstrate how the application of key post-authentic concepts and methods can be especially effective at three decisive stages in a topic modelling workflow: pre-processing, corpus preparation and choosing the number of topics. The post-authentic framework, as the analysis will show, may be especially effective at prompting the active and reflexive participation of the user in the process of knowledge production in the digital. In the next section, I start my argument by discussing the implications of the 'big data philosophy', that is, the obsession with patterns and correlations as opposed to causation, to explain phenomena; I also examine such implications in relation to topic modelling and its use for knowledge creation, in humanistic enquiry and beyond.

4.2 Causality, Correlations, Patterns

Perhaps one of the most significant implications of the 'Digital Turn' in the humanities, more widely in the natural, computational and social sciences, and more widely still in relation to the digitisation of society is contained in the notion of discrete vs continuous modelling of information. The concepts of discrete and continuous and the tension between the two are at the foundation of mathematical thought and of how mathematical modelling is used to explain natural phenomena (Fenstad 1985). A way to understand the crucial difference between discrete and continuous structures is to consider that in a discrete structure, all points are isolated and completely disconnected from each other; one can therefore label them and count them and their count is exact and absolute. On the contrary, one can only access a continuous structure by measuring it and these measurements create intervals or fractions of intervals; moreover, in the continuous, a scale for the measurement has to be set (Longo 2018, 84). Therefore, in discrete systems, there is no room for approximation, no uncertainty, no nuances, as something is either one point or another, whereas in the continuous—since phenomena can only be accessed by measuring them—the measurements are always approximated (Longo 2019, 64–65).

Even without going too deep into the full mathematical (and physical!) ramifications of these two notions, one can intuitively understand that they refer to very different ways of mathematical thinking. A fundamental difference particularly relevant to the arguments advanced in this book is concerned with the understanding of causality, a notion whose theoretical conceptualisation from philosophy to physics can be traced back to antiquity.[2] For the sake of the argument advanced in this chapter, I will summarise the discussion by saying that in the classical worldview which prevailed until the twentieth century, a mechanistic notion strongly identified causation with determinism. Determinism can be understood as the ability to determine the future state of a physical system from its present state (Weinert 2005, 196). According to this view, also known as functional view of causation, every event has a unique cause that precedes it (de Laplace 1820; Stigler 1986; Čpek and Čapek 1961), and therefore the world is seen as an 'uninterrupted chain of causes and effects' (Holbach 1770). This view has been criticised over the course of the twentieth century for several shortcomings such as the proximity of elements in determining cause-effect relationships, predictability as the main criterion for establishing causation and the reduction of causality essentially to a mere temporal relationship. Discoveries of and advances in differential equations, atomic physics and quantum mechanics have further consolidated such criticisms eventually leading to the current separation of causality from determinism. Particularly in quantum mechanics, recent experiments have provided strong evidence for the validity of this notion of causality without determinism. In this view, consequent states of a quantum system are related to its antecedent states by a form of conditional dependency (Weinert 2005, 241) as opposed to every event having a unique cause that precedes it.

Coming back to the distinction between discrete and continuous structures, this means that in discrete systems, there is no deterministic cause-effect relationship, because points are totally separated from each other, whereas in continuous systems, causal relations can be observed and measured, but not predicted[3] (Longo 2018, 86). Though it may appear inconsequential at first, this observation about causality has specific and profound implications that stretch well beyond mathematical and physical reasoning. Stating that in discrete structures such as say a database where something belongs to either one category or another, no cause-effect relationship of observed phenomena can be established but only a probabilistic one essentially means that explanations for such phenomena cannot be found, only correlations. If two random variables are correlated, or as noted by Calude and Longo (2017), *co-related*, it means that they are associated according to a statistical measure, that they co-occur. This statistical measure is rendered by a correlation coefficient, a number between −1 and 1 that expresses the strength of the linear relationship between two numeric variables. If two variables are positively correlated (e.g., they both increase), then the correlation coefficient will be closer to 1, if there is a negative correlation (i.e., they are inversely correlated), it will be closer to −1, and closer to 0 if there is no correlation at all. It is a well-established fact in statistics and beyond that a correlation coefficient per se is not enough to explain the cause for the patterns that are captured.[4]

The identification of statistical correlations is nevertheless an important factor in understanding the relationship between two quantitative variables and it remains an insightful method that can potentially lead to significant discoveries. Indeed, the observation of correlations is at the foundation of the classic scientific method in the sense that starting from the measurement of correlated phenomena, scientists have been able to formulate theories that could be tested and later confirmed or disproved. The history of science is full of extraordinary achievements which originated from mere observations of not-so-obviously correlated phenomena, for example, distributional semantics theory, a famous linguistic theory that stemmed from the intuition of Zellig S. Harris and John R. Firth, two semanticists (though Harris was also a statistical mathematician). This intuition—famously captured by Firth's quote 'You shall know a word by the company it keeps' (1957, 11)—acknowledges the relevance of words' collocation (i.e., the place of occurrence of words) in determining their meaning. The core idea behind Harris and Firth's work on collocational meaning and distributional semantics is that meanings do not exist in isolation; rather, words that are used and occur in the same contexts tend to purport similar meanings (Harris 1954, p. 156).

In those days, gaining access to real language data was costly and very time-consuming and for a long time, it was not possible to test this theory. But more recently, new advances in computer science merged with huge quantities of naturally occurring language material, including digitised historical data-sets, have indeed proven that languages are not deterministic systems—as previously believed—but that they should be thought to be 'probabilistic, analogical, preferential systems' (Hanks 2013, 310). As intuitively theorised in distributional semantics, words do not have a one-to-one relationship with meaning because meanings are not precise, exact or stable. To the contrary, words in isolation do not possess any meaning and meanings can only be entailed from words' context. As argued by Harris, 'We cannot say that each morpheme or word has a single or central meaning, or even that it has a continuous or coherent range of meanings' (Harris 1954, 151). Sixty years after its initial formulation, distributional semantics theory laid the basis for Google's renowned word2vec algorithm, and today, it constitutes the theoretical background of NLP studies concerned with language and meaning, including the very topic modelling (*cfr.* Sect. 4.4).

Coming back full circle to causality, correlations and patterns, a correlation measure only informs us of the strength of a relationship between two variables, whereas the patterns tell us that certain regularities can be found in how the observed variables are distributed. Hence, for they highlight trends in the data, correlations and patterns may potentially have predictive power, but neither of them provides causal explanations for the analysed phenomena nor they intrinsically carry significance. In the next section, I will elaborate on these reflections to discuss the important implications for society of operating predominantly within the discrete system of the contemporary encoding of all digital information, binary sequences of 0s and 1s. Taking the example of analysis of material that had originally been conceived of as a coherent entity, i.e., continuous (e.g., a book, a collection of essays on the same topic, all the issues of a newspaper), I explore the implications of its digital encoding into discrete form through digitisation and subsequent digital analysis. One critical implication, I argue, is that the adoption of an indiscriminate, data-driven approach to analysis risks to completely disregard context and to attribute meaning to correlations and patterns per se. Through the example of topic modelling and its application to the analysis of *ChroniclItaly 3.0*, further in the chapter, I show how the application of concepts and methods of the post-authentic framework to digital knowledge creation can be useful to prompt a critical stance towards computational methods and tools which I argue is urgently needed for the configuration of a model for knowledge production in the digital.

4.3 Many Patterns, Few Meanings

Big data analytics (*cfr.* Sect. 1.2) is supported by the idea that correlations are expected to be recurrent, i.e., they will iterate similarly along the chosen parameter, for example, time (Calude and Longo 2017, 602). Recurrent correlations are an established scientific principle and they can be observed in natural cycles such as the water cycle and the alternation of seasons. The recurrence of correlations suits well deterministic systems in which it is believed that one can determine the future state of a physical system from its present state (*cfr.* Sect. 4.2). This is precisely what the 'big data philosophy' states: because patterns are expected to be recurrent, the future can be predicted by statistical algorithms based on the patterns found in past data, without the need for causal explanation. Naturally, the larger the data-set, the more accurate the prediction.

This idea that all that counts are the patterns is not in fact new and it can be traced back to the 1990s and to Complexity Theory (Waldrop 1992). Complexity Theory argues that there is a hidden order to the behaviour and evolution of complex systems and chaos can be made manageable by looking at its underlying, ubiquitous patterns. What these patterns show is how complex systems work, more specifically how organisations cope with uncertainty and nonlinearity and manage to remain stable. The idea behind Complexity Theory is that complex systems are the assemblage of extremely convoluted factors which make them fundamentally unpredictable. Yet, at the same time, complex systems exhibit order rules according to which independent actors, i.e., discrete elements, spontaneously self-organise. This contradictory property makes it possible for patterned behaviour and properties to be observed. It also means, however, that the meaning of any system is irrelevant as the focus is and remains on the observed behavioural patterns.

One does not have to dig too deep to see how computer science has strongly supported Complexity Theory. Indeed, Complexity Theory fits perfectly with what machines excel at: finding patterns in the data (Turkle 2014). Ever powerful computers can be given enormous quantities of data and instructed to find the patterns that human beings will never be able to find. And it works. Patterns are always found. However, despite appearing (at first, at least) logically sound and despite being validated by the cycles present in nature, the discourse surrounding big data analytics obscures at least four fundamental truths. Firstly, as said earlier in the chapter, in discrete systems such as a database, no cause-effect relationship of observed phenomena can be established but only correlations and patterns. Computers are not programmed to find meanings, only the patterns; as correlations and patterns do not intrinsically carry significance, this essentially means that databases provide an a-causal image of the world (Longo 2018, 86). Thus, what the big data hype obscures is that today's computer-dominated world offers us countless patterns but no explanations for them, and so we are left to deal with a patterned, yet a-causal, way of making sense of reality.

Secondly, the idea that information is uniquely absorbed from data is also closely related to Complexity Theory. The theory argues that complex systems are constantly altered by agents' interactions through a process of feedback loops; thanks to their intrinsic capacity to learn from experience, complex adaptive systems are organic and better evolving. The big data approach has essentially adopted this theory in toto, but it seems to have failed to recognise that machines are in fact incapable to *learn*. Indeed, the deterministic belief that the future state of a physical system can be predicted from the observation of its past state, which in any case has been criticised over the course of the twentieth century and mostly disproved as discussed in Sect. 4.2, has become conflated into the metaphorical use of the word 'learning' in ML. The familiar notion of 'learning' confounds what learning actually means for a machine—finding correlations and patterns but no causal explanations—with the human capacity to understand and make sense of the world, i.e., attempting to find causality.

Thirdly, the big data analytics' deterministic claim that based on available data, one can provide accurate predictions of the future without the need for causal explanation is provably wrong. Calude and Longo (2017) demonstrated that in a large enough data-set, there will always be correlations but most of them will be random, i.e., meaningless. This means that the probability that a series of correlations will be recurrent as in the natural cycles is extremely low; the authors explain: 'recurrence may occur, but only for immense values of the intended parameters and, thus, an immense database' (ibid., 609). In other words, the patterns found in databases do not per se constitute sufficient proof to offer reliable predictions of the future because most of these patterns will actually be false positives. In techniques such as topic modelling, an element of randomness is in fact built into the algorithm itself as initially, documents are assigned to topics through random probability. Although it is true that the calculations become increasingly accurate as the algorithm iterates through more documents, the risk once again is to see meaning where there is none.

Fourthly, the fact that databases are exact, i.e., discrete, perpetuates the false belief that data is also exact, neutral and objective. It is always emphasised by the 'big data philosophy' that statistical algorithms will find patterns where nobody else can, and because databases are exact, this is enough. What is on the contrary not at all emphasised is the subjective and interpretative dimension of collecting, selecting, categorising, aggregating, in other words of *making* data. Recognising that data is created makes the claims of absolute impartiality, exactness and reliability shaky at best and ethically concerning at worst, particularly when necessarily incomplete, biased and opaquely collected data is used to make predictions that influence decision-making processes or produce research findings.

Reassuringly, these limitations have recently started to be at the centre of the academic debate and have originated the so-called causal inference challenge. In their work *The Book of Why* (2018), computer scientist Judea Pearl and mathematician Dana Mackenzie argue that these limitations make the big data philosophy inadequate to solve our world's challenges. They note that as current ML solutions cannot find the causality relations between patterns, they inevitably fail to generalise beyond the domain of examples present in a given data-set, which most of the time will include synthetic data (as opposed to real-world generated data). In other words, most current ML methods tend to 'overfit the data', meaning that 'they try to learn the past perfectly, instead of uncovering the real/causal relationships that will continue to hold over time' (Gonfalonieri 2020). New avenues in this direction are increasingly being explored and have resulted in new emerging fields such as causal machine learning (see for instance, Pearl et al. 2016; Shanmugam 2018; Hernán and Robins 2021). However, although the interest in this topic has grown exponentially in the span of only a few years, methods and applications are still at an experimental stage and, to my knowledge, primarily limited to academic research.

4.4 The Problem with Topic Modelling

The topic modelling algorithm essentially formalises distributional semantics theory (*cfr.* Sect. 4.2). However, whereas the focus of distributional semantics theory is on the meaning of a single word, topic modelling tries to capture the overall meaning of clusters of words that appear together (i.e., that are correlated) in a document. Put it differently, as single words do not possess any meaning but meanings can only be entailed by their context, topic modelling assumes that groups of words also purport collective meanings, i.e., *topics.* This all sounds very logical but there is a caveat. Similar to quantum, computational and genetic systems, languages are discrete representations (i.e., outputs) of fundamentally continuous structures (i.e., inputs). This property—called the discrete infinity of language—essentially means unlimited productivity from limited means (Chomsky and Smith 2000). It describes the ability of languages to create an infinite variety of expressions of thought from a limited set of discrete elements (Studdert-Kennedy and Goldstein 2003). The discrete infinity of language necessarily entails that languages are intrinsically ambiguous because meaning is context-bound, but significantly, it indicates that different contexts shape the creation of infinite meanings. The problem with topic modelling is that it provides a probabilistic representation of words' distributions in the ingested documents, but it is completely agnostic of the underlying continuous structure of such documents, such as the ambiguity of words' use in each document and across texts as well as the documents' coherent substructure, let alone their wider historical, social and cultural entrenchment.

As said earlier in the chapter, topic modelling provides a probabilistic representation of how words are distributed in documents according to statistical calculations, that is, correlations. This means that words are considered to be discrete elements; for example, in the corpus preparation stage (*cfr.* Sect. 4.5.2), words are transformed into numeric variables and their distribution across documents is represented as a distribution matrix. What topic modelling then does is measuring the strength of the linear relationship between these numeric variables. But topic modelling also treats the corpus itself as a collection of discrete data, which means that each text is also processed as a separate entity totally disconnected from all the other texts in the batch. This is true regardless of whether the input is all the chapters from the same book, all the issues of a newspaper or all the abstracts ever submitted to an academic journal under the keyword tag 'topic modelling'. In other words, it is a computational technique that efficiently identifies patterns of words' distribution, but because it lacks the words' underlying continuous structure—the infinity of language—no cause-effect relationship of the correlated phenomena can be established, i.e., the meaning of such patterns.

Another issue with topic modelling is that it assumes that an a priori fixed number of topics—which in any case is decided more or less arbitrarily—is represented in different proportions in *all* the documents. Hence, if the algorithm is instructed to find X number of topics, it will build a model that fits that number. This assumption behind the technique cannot but paint a rather artificial and non-exhaustive picture of the documents' content as it is hard to imagine how in reality, a fixed number of topics could adequately represent the actual content of all the analysed documents. Thus, correlations will surely be identified but not all these correlations will necessarily carry significance, that is, meaning. Moreover, as countless parameters can be tweaked, the smallest change will output a different model, in which different correlations will be found and many others will be missing. Conversely, even when the same parameters from the same software are used on the same data-set, the algorithm will output a slightly different model, which indeed proves once again that patterns will always be identified, regardless of their significance. I will return to this point in Sect. 4.5.3.

4.5 Analysis of Digital Objects: A Post-authentic Approach to Topic Modelling

The post-authentic framework to digital knowledge creation contributes to the urgent need for the establishment of critical data and visualisation literacy in the current landscape—both public and academic—in which computational techniques and outputs are predominantly framed as and often believed to be exact, final, objective and true. Whilst exploiting the new opportunities offered by computational technologies, the post-authentic framework rejects an uncritical adoption of digital methods, and it promotes a model not simplistically oriented towards problem-solving, solution automation and sleek interface designs but towards encouraging critical engagement and active participation. This ultimately means recognising that knowledge is fluid and that the complex challenges we face today therefore require a model of knowledge production that fosters symbiotic collaborations, fluid exchanges and mutualistic contributions, as opposed to hierarchical separation and competition.

As an example of how the application of the post-authentic framework can contribute towards fluid processes of knowledge creation in the digital, including the need for a less naïve conceptualisation of computational techniques, digital objects and methods, I discuss here the third use case of the book: analysis of digital objects. The example of topic modelling demonstrates how critical engagement with computational techniques is urgently required to meet the uncertain and problematic aspects of digital research. For example, in fields such as DH in which this technique is used extensively, a recent survey on LDA topic modelling (Du 2019) found out that 74% of the surveyed studies didn’t report how their corpora were prepared, more than 70% didn’t report which tool was used to train their topic models, almost 57% didn’t report how many topics were trained, and about 90.5% didn’t report how their topic models were evaluated.

DH is not at all an isolated case, however. Though with some differences, a similar trend has also been found in software engineering research (Silva et al. 2021) where topic modelling is widely used to analyse online conversations among developers or to improve software engineering tasks such as source code comprehension. From the analysis of 111 relevant papers, Silva et al. (2021) found both general inconsistency and the adoption of opaque methods in topic modelling practices on the whole pointing to a degree of uncertainty on the specificity of the technique itself. The highest inconsistency was found with reference to tasks such as choosing the number of topics, naming the topics and evaluating the topics' semantic interpretability. The authors attributed the lack of specificity of the technique to the fact that the majority of the surveyed papers had employed LDA 'as is', that is, they had adopted the default parameters as an off-the-shelf software. This approach, however, is generally not encouraged; computer scientists openly acknowledge that finding the meaning behind the identified patterns is highly dependent on the specifics of the sources because, as argued by Hindle et al. (2015, 510), 'LDA does not look for the same patterns that people do'.

In this part of the chapter, I illustrate how the post-authentic framework can be applied to topic modelling to guide a more mindful understanding of the materiality of the sources. To this end, I deliberately choose cultural heritage material, sources that are inevitably problematic from a computational point of view. I then focus on the key aspects of topic modelling that are highly dependent on the sources and which in my experience have the most significant impact on the results: pre-processing, corpus preparation and deciding the number of topics. As a case example, I use the already discussed Italian American newspapers as collected in *ChroniclItaly 3.0* (*cfr.* Chaps. 2 and 3); my aim is to emphasise how preparing the material for the analysis is part of the analysis itself. My discussion demonstrates how, far from being fully automated, neutral and objective, the analysis of a digital object requires the analyst to make countless decisions which are yet different from the ones required when preparing the material for enrichment, even when the same sources are used. Indeed, engagement with the technique starts much earlier than the algorithm's implementation stage, which in any case should also not be performed as a fully automatic operation. The application of the post-authentic framework allows me to evidence how LDA may well be an unsupervised technique, but this simply means that it works with unstructured data,[5] and not at all that despite what may be generally believed it does not require human intervention.

4.5.1 Pre-processing

In Chap. 3, I illustrated how pre-processing operations are far from being standard and how it is in fact required that each intervention is carefully assessed by scholars and practitioners and evaluated on a case-by-case basis. In my discussion, I considered the many influential factors at play (e.g., the materiality of the source, the specific task to be performed, the available resources, both economic and technical) and illustrated how they in turn are embedded in a complex, wide net of co-dependent actors, elements and circumstances which have influenced each other and will in turn influence current and future interventions. The same considerations apply to the analysis of a digital object; this, I maintain, requires a high level of critical engagement with the chosen method well before than the algorithm's implementation stage. In the case of topic modelling, for example, which takes as its input unstructured data, e.g., plain text, the first thing one needs to decide is the *scope* (*cfr.* Sect. 3.4), that is, what to consider as *documents* (i.e., the input) (see for instance, Miner 2012). Topic modelling aims to represent documents as probabilistic distributions of words; hence, in a book, the documents could be the book's pages else on a newspaper's page, they could be individual articles and so on. Conceptually, it of course intuitively makes a difference to search for the topics in a chapter vs the topics in each page of that chapter. But this is an important decision to make also from a pragmatic point of view: as topic modelling is essentially a statistical method, the length of each modelled item, i.e., the document, does matter. And yet, although this is a rather determining factor, studies using this method rarely specify how the criteria to decide the scope of the documents are assessed and, even when mentioned, they are referred to vaguely. In Silva et al.'s survey of topic modelling in software engineering research (2021), for example, the authors found that 86% did not mention such criteria at all nor did they acknowledge documents' length as being an important factor; they also found that even when the relevance of the vocabulary size was acknowledged (14%), about a half (7.4%) did not specify the selection criteria or the document's length.

In the case of *ChroniclItaly 3.0*, I considered that each file in the collection corresponds to the first page of each issue published by the newspapers on a certain date. This structure mirrors the way the collection was digitised by the Library of Congress, evidencing once more the inseparable complexity of relations between digital material and its wider entrenchment in the surrounding digital infrastructure that created it and/or provides it. Therefore, I defined as *documents* each file/issue as it was in the collection; the decision had the dual advantage of modelling the documents according to the events narrated on a day/issue basis while following the Library of Congress metadata schema.

In terms of preparatory operations such as removing stopwords, lowercasing, removing punctuation, numbers, special characters (cfr. Chap. 3), for the specific task of topic modelling, additional specific linguistic decisions must also be evaluated, here I discuss stemming and lemmatisation. Although both aim to obtain a word root by reducing the inflection in words, these operations are built on very different assumptions. Stemming deletes the initial or final characters in a token based on a list of common prefixes and suffixes that may typically occur in the inflected words of a language (e.g., states → state). It is therefore language-dependent as it relies on limited cases which would apply exclusively to certain languages that follow specific inflection rules. Therefore for languages that follow fairly regular inflection rules such as English, stemming may work reasonably well, but applied to highly inflectional languages such as Italian, due to its many exceptions and irregularities, the algorithm would almost certainly perform poorly. Another strong limitation of stemming is that in many cases—including low-inflectional languages—the output would not be an actual word, meaning that the operation is likely to introduce new errors. On the other hand, as it is not a particularly advanced technique, stemming does not require a long processing time or processing power, and therefore this solution may be implemented when working with particularly large corpora or when constrained by time limitations.

Lemmatising is on the contrary a much more sophisticated technique as it is based on more solid linguistic principles than stemming. By means of detailed dictionaries that contain lemmas and by examining words' context, a lemmatising algorithm analyses the morphology of each word and it then transforms it into its grammatical root (e.g., better → good). Especially in the case of topic modelling in which the output is essentially a list of words without any context, lemmatising can be very helpful to distinguish between homonyms, words that have the same spelling, sometimes the same pronunciation too but which in fact possess different meanings. For example, the word *mento* in Italian can mean either 'chin' or 'I lie'. A lemmatising algorithm would theoretically be able to entail the use of *mento* from its context and distinguish it from its homonym; in this case, the different outputs would be *mento* (i.e., chin) for the former and *mentire* (i.e., to lie) for the latter. Because of its complexity, however, lemmatising may require a long time and very high processing power to perform, and so in the case of large size collections or depending on the available means and resources, it may not be ideal. Additionally, if on the one side lemmatising is effective at differentiating between homonyms, on the other the reduction of all inflected words to their lemma may cause information loss. For instance, it would no longer be possible to recognise the tense (present, past, future) or the grammatical person (I, they, you, etc.) of the verbs, the gender or number of the nouns, the degree of the adjectives (e.g., superlative, comparative), etc.

To assess whether this type of information is relevant or not depends once again on several factors such as the type of data-set (e.g., size, content), the context of the digital analysis, the language of the data-set and the specific research question(s); researchers should therefore carefully evaluate pros and cons of implementing this operation. For example, in researching narratives of migration as they were told by Italian American migrants, the cons of implementing either stemming or lemmatising would in my opinion exceed the pros. Italian is a highly inflectional language and a great deal of linguistic information is encoded in suffixes and prefixes; stemming therefore ill suits it. Similarly, lemmatising the corpus would also cause the loss of information encoded in inflected words (e.g., verbs expressed in the first person, collective concepts expressed by plural nouns) which could bring valuable insights into the cognitive, subjective dimension of the stories told by the migrants.

Finally, whether to perform or not either of these operations is very much dependent on the language of the data-set, not just because different languages have different inflection rules, but crucially also because not all languages are equally resourced digitally. Indeed, as discussed in Sect. 2.2, the digital consequence of the fact that most mass digitisation projects have been carried out in the United States and later in Europe is that computational resources available for languages other than English continue to remain on the whole scarce. Such Anglophone-centricity is often still a barrier to researchers, teachers and curators whose sources are in languages other than English. Indeed, the comparative lack of computational resources in other languages often dictates which tasks can be performed, with which tools and through which platforms (Viola and Fiscarelli 2021b). Moreover, even when adaptations for other languages may be possible, identifying which changes should be implemented, and perhaps more importantly, understanding the impacts these may have, is often unclear (Mahony 2018). This includes lemmatising algorithms and dictionaries which do not yet exist for all idioms; therefore, for particularly under-resourced languages, stemming may be the only, far from ideal, option.

4.5.2 Corpus Preparation

There are several libraries, for example, in Python or R, as well as off-the-shelf tools (e.g., MALLET) that implement LDA for topic modelling. Some allow for more sophisticated parameters than others, but generally speaking, they all follow the same principles that I have already discussed: a topic modelling algorithm models a number of documents to find correlations essentially combining term frequency and word collocation operations. In order to model topics from unstructured text, the material first needs to be converted into a structured model that allows the algorithm to perform such calculations, for example, through a method called bag of words (BoW). What BoW does is to first transform the words in the documents into numbers, i.e., into ids; this operation is typically called 'dictionary'. It then builds a matrix based on the frequency of the words in the documents.

The generation of a BoW provides a notable example of the decisive influence of the analyst on algorithmic processes and therefore ultimately, on the output. Specifically, in order to prepare the dictionary, i.e., the unique id assignment, the analyst has several so-called optimising operations at their disposal. For example, one might decide to filter out 'extremes', terms in the collection that are particularly frequent or infrequent; this operation may be performed in order to obtain what is believed to be a more representative core vocabulary. There are several ways to perform this task; for instance, the Python library Gensim (Řeh˚uřek and Sojka 2010) has a built-in function called `filter_extremes` which filters out tokens in the dictionary based on their frequency of occurrence. The parameters are defined by the user who can decide—though one might argue somewhat arbitrarily—to keep tokens which are contained in a defined number of documents (i.e., no more than in X number of documents and no less than in X number of documents) or to keep only the first X number of most frequent tokens.

Another very common technique originated in the field of IR and believed to contribute towards obtaining better topic modelling results is the term frequency—inverse document frequency method (TF-IDF). The method also scores the 'importance' of a word, also known as *weight*, according to its relative frequency, i.e., the frequency of occurrence of that word with respect to the number of documents in the collection in which it appears. In this way, the weight of words that are 'expected' to appear more frequently—generally speaking non-salient words such as prepositions, articles and so on but this is also specific to the material—is resized accordingly. These preparatory operations are believed to help optimise a corpus for IR tasks (not just topic modelling) and in most cases, they may succeed. The assumption is, however, that a word is as important as its relative frequency, which may be true most times, but not always. Indeed, the possibility to capture words that are very rare or that appear in very few documents may be as valuable in that they may indicate a sudden shift in the used vocabulary, which may in turn signal a linguistic change or perhaps even a conceptual one. Furthermore, and perhaps even more significantly, these techniques only consider the formal frequency of a word, meaning that they do not cater for how that word is used. In the words of David Blei (Blei 2012, 82)—one of the creators of topic modelling: 'One assumption that LDA makes is the "bag of words" assumption, that the order of the words in the document does not matter'. This approach, defined as 'unrealistic' by Blei himself, may work well for grammatical articles, prepositions or particularly recurrent OCR errors, but as no semantic detection is formally conducted, the frequency of a word, misleadingly referred to as the weight, becomes the unique, determining factor in assessing whether a word is worth keeping or not. What is important to remember is that what is worth keeping for an algorithm may not reflect at all the writer's original intention. Languages may be probabilistic systems, but since words do not have a one-to-one relationship with meaning, they are fundamentally ambiguous, preferential systems. For this reason, researchers and practitioners should assess carefully whether using relative frequency methods is the best option when preparing the corpus to train the topic models. For example, research has shown that statistically more accurate models do not necessarily lead to a higher interpretability of the results (Jacobi et al. 2015).

As an attempt to retain the meaning of words, a method that aims to compensate for this shortcoming is preparing the corpus as a dictionary of n-grams, typically bi-grams or tri-grams. These are pairs or triples of words that are statistically more likely to occur together than if they were found independently from each other. Several studies (see for instance, Wallach 2006; Wang et al. 2007; Kherwa and Bansal 2020) have indeed reported that using bi-grams to prepare the corpus may increase topics' interpretability as well as the efficiency of statistical methods such as perplexity and coherence (*cfr.* Sect. 4.5.3), developed to help researchers and practitioners optimise topic modelling results. Unfortunately, preparing the corpus as a dictionary of n-grams is a lengthy and intense process which may indeed be costly and time-consuming, especially in the case of very large repositories. Furthermore, researchers working on historical material which typically contains a high number of OCR errors should consider the actual added value of using this technique. Studies on topic modelling which suggest novel IR techniques or improved corpus preparation methods such as those discussed here and which report an increase in the models' quality typically make use of digitally born data such as online film reviews, blogs, news websites' headlines or contemporary conference proceedings. Being digitally born, these data-sets are of very high quality, especially compared to digitised historical material. Indeed, the amount of OCR errors in historical collections inevitably skims the output as each word containing an error will be interpreted by the algorithm as a new word, even if only by one character. Although pre-processing steps are taken to improve the quality of the collection, many errors may remain. In most cases, these errors would not prevent a human from reading and understanding, but they will interfere with how a machine processes the text. As LDA is a probabilistic method, regardless of the specific variations in the chosen pre-processing and corpus preparation techniques, the results will be heavily reliant on the data quality.

Finally, it is worth reminding that, due to the intrinsic unstable and non-deterministic nature of topic modelling, assessing how and to what extent any of these corpus preparation techniques actually improves the quality of the models remains difficult. Users should indeed be aware that findings obtained with topic modelling can never be fully replicated or generalised even if the same data-sets are used, the same steps are implemented and the same LDA settings are chosen from the same library/tool (Silva et al. 2021, 120). The post-authentic framework acknowledges such limitations and it is mindful of drawing conclusions which are based solely on topic modelling findings.

4.5.3 Number of Topics

The weaknesses and limitations as well as the dangers of overly trusting the capacity of topic modelling to find meaningful patterns have been openly acknowledged by several authors, including its very creators. Already in 2009, Chang et al. (2009), for example, compared the task of interpreting the topics, i.e., finding the semantic meaning of the discovered patterns, to the Chinese ritual of reading tea leaves. The authors wanted to warn users of the high risk of attributing meaning to patterns and trends that in reality may be 'spurious' in the mathematical sense, i.e., meaningless (Calude and Longo 2017) (*cfr.* Sect. 4.3). Naturally, the risk is even higher when the technique is adopted uncritically, especially in fields outside of computer science. The authors clarified that although typically it is implicitly assumed that the identified latent spaces will be semantically salient, in reality, this is not at all what the promise of topic modelling is about. Since then, others (see for instance Bail 2018) have also openly acknowledged the limitations of the technique and repeatedly attempted to reframe topic modelling as 'a tool for reading' rather than a tool for meaning, that is, an exploratory tool which in order to obtain more nuanced and reliable findings, should be integrated with other methods. In this respect, for instance, sociologist Chris Bail (ibid.) notes:

> Despite this rather humble assessment of the promise of topic models, many people continue to employ them as if they do in fact reveal the true meaning of texts, which I fear may create a surge in "false positive" findings in studies that employ topic models.

The application of the post-authentic framework to topic modelling helps reframe the technique as a statistical tool and resizes the user's expectations accordingly. Topic modelling posits a set of multinomial distributions over words—misleadingly called *topics*—as being present in each document in various proportions; it provides fairly accurate models of documents based on their words' distribution as grouped into clusters. This is valuable for obtaining a corpus representation through its words' distribution and/or for predicting a model of unseen text but the commonly shared belief that these identified word clusters will also be semantically meaningful, i.e., that they will be topics in the human sense, remains only anecdotal (Chang et al. 2009).

The high risk of finding patterns that are in reality meaningless can be exemplified by the challenge of finding the so-called 'optimal' number of topics. This task requires user's input to instruct the algorithm about how many words' distributions it has to search for in the corpus, which of course cannot be known in advance. Depending on individual cases, sometimes researchers and practitioners may know the collection extensively enough to feel confident about what this number might be; others prefer building multiple models with different numbers of topics to subsequently compare the various compositions of the topics (Viola and Verheul 2019b). If on the one hand this approach allows the researcher to closely examine the varied topics' structures before deciding on the most coherent model, on the other it has the limitation to potentially lead analysts to prefer a model that seems to confirm their a priori ideas, thus resulting in biased interpretations. This approach may work fairly well in those cases when the analyst has extensive knowledge of the material, the field and the period of reference of the collection among others, but it is generally not recommended in statistics; in the words of statistician Stephen M. Stigler: 'Beware of the problem of testing too many hypotheses; the more you torture the data, the more likely they are to confess, but confessions obtained under duress may not be admissible in the court of scientific opinion' (Stigler 1987).

More often, however, very little is known about the actual content of the documents as *true* content is exactly what the technique is wrongly believed to be able to find, which provides the original justifying argument for using the method. It goes like this: due to the increasingly large size of available digital material, it is not possible for researchers and practitioners to explore the documents through traditional close reading methods; not only would this be too time-consuming but also somewhat less efficient as a machine will always outperform humans in identifying patterns. Although this is in principle true as clarified earlier, the assumption that all the found patterns are intrinsically meaningful is not. To meet this challenge, research has been conducted towards implementing statistical methods that could help researchers and practitioners find the craved 'optimal number of topics'. Two of the most common methods are model perplexity and topic coherence, measures that score the statistical quality of different topic models based on the topics' compositions in several models. Though not unanimously, the believed assumption behind these techniques is that a higher statistical quality yields more interpretable topics. Model perplexity (also known as predictive likelihood) predicts the likelihood of new (i.e., unseen) text to appear based on a pre-trained model. The lower the perplexity value, the better the model predicts the distribution of the words that appear in each topic. However, studies have shown that optimising a topic model for perplexity does not necessarily increase topics' interpretability, as perplexity and human judgement are often not correlated, and sometimes even slightly anti-correlated (Jacobi et al. 2015, 7).

Topic coherence was developed to compensate for this shortcoming and it has become popular over the years. What the method is designed to do is to model human judgement by scoring the composition of the topics based on how *coherent*, i.e., interpretable, they are (Röder et al. 2015). If the coherence score increases as the number of topics increases, for example, that would suggest that the most interpretable model is the one that displays the highest coherence value before flattening out or dropping. Both techniques are widely used to determine the optimal number of topics; the truth is, however, that neither of these measures is ideal because what they actually score is the probability of observations and not their degree of semantic meaning (Chang et al. 2009). In a study by Chang et al. (2009) about topics' interpretability, the authors noted that these traditional metrics do not in fact capture whether topics are interpretable or not as they optimise topic models for likelihood-based measures but, as clarified earlier (*cfr.* Sect. 4.5), 'LDA does not look for the same patterns that people do' (Hindle et al. 2015, 510). In the study, the authors therefore suggest practitioners to adopt a more critical assessment of the topics' quality.

In this chapter, I have discussed how the use of familiar notions to name computational techniques such as topic modelling, sentiment analysis and machine learning has increased their popularity while creating epistemological expectations that these methods will disappoint. Especially when used outside of their field of origin, the generated confusion contributes to obfuscate the mathematical assumptions upon which these techniques are built, such as the fundamental difference between discrete vs continuous modelling of information and the stemming consequences. In the context of digital knowledge creation and in relation to the big data philosophy, I reflected on the significant, yet often overlooked, implications for notions of causality and correlations. I then applied these considerations to describe the third use case of the book, analysis of a digital object, and used the properties and assumptions of topic modelling as the case example of a widely used computational technique that treats a collection of texts as discrete data. I have shown how the post-authentic framework can be used as the applied theory to engage critically with topic modelling by devoting special attention to the aspects of the analysis that are key for maintaining a symbiotic connection with the sources: pre-processing, corpus preparation and the number of topics. Specifically, I have shown how the application of the post-authentic framework to topic modelling acknowledges the technique at core correct but problematic and therefore in need of critical engagement.

My intention is not to dismiss topic modelling as woefully inadequate, but rather to encourage the integration of the method with critical scrutiny in order to address its limitations. In so doing, I have argued that by introducing a counter-narrative in the main scientistic discourse, the post-authentic framework strains the current system and can help us refigure a novel and more honest model for knowledge production in the digital. For example, when topic modelling is used for humanistic enquiry such as the analysis of cultural heritage material as discussed here, the post-authentic framework serves as a warning that the technique's limitations are particularly significant and their impact on the provided interpretation of the past is problematic. I will return to these points in the next chapter in which I discuss the fourth and last use case of the book, visualisation of a digital object. Specifically, I will show how I have applied the post-authentic framework to prototyping a UI for topic modelling. I will insist on key aspects that aim to promote the active and reflective participation of the researcher in the process of digital knowledge production; I will devote particular attention to the added value of building UI elements that contribute to the urgent need for the establishment of critical data and visualisation literacy, especially when computational methods are adopted in fields outside of their original design.

Notes

1. "Bloquer le pays ne permet pas d'endiguer l'épidémie".
2. For a detailed and in-depth historical discussion on causality in physics and philosophy, I refer the reader to Weinert (2005).
3. Please note that not everyone agrees with this view and that there are still unanswered questions around causality, particularly in relation to discrete phenomena in quantum mechanics. See, for instance, Le Bellac (2006) and Jaeger (2009).
4. A well-known phrase that synthesises this fact is 'correlation does not mean causation'.
5. Not previously annotated material.