16

MAPPING AND MINING DIGITAL SOCIETY

Key questions

- How can the internet's functions — in apps, algorithms, and software — for collecting, sorting, and visualising data be harnessed as a new form of research instruments?
- What types of research questions can be answered through social network analysis, and how can this method be used in digital social research?
- What is text mining, and how can it be a useful method to analyse social interaction on the internet and in social media?

Key concepts

Methods of the medium * instruments of revelation * social network analysis * weak ties * small-world networks * text mining * distant reading

The ethnographic approach, as introduced in the previous chapter, offers a solid framework with which to embark on studies within the field of digital social research. As I said in Chapter 13, sometimes ethnography alone can be a sufficient research method, depending on what you want to find out. However, the changing data environment — also discussed in Chapter 13 — means that it is often a good

idea to bring in other sources that are not conventionally associated with the ethnographic method. This is because, as described in Chapter 15, the notions of what actually constitutes the 'field' or 'the data' of ethnographic analysis are altered in digital society. As discussed in previous chapters, Kozinets (2015: 3), wrote about an approach that he calls 'netnography', and thinks that devising research methods for studying sociality online is about 'intelligent adaptation' and 'considering all options'. The root, he says, should be in the core principles of conventional ethnography, but digital social researchers must also seek to selectively and systematically seize 'the possibilities of incorporating and blending computational methods of data collection, analysis, word recognition, coding and visualization' (2015: 79). Digital social research, as I argued in Chapter 13, relies on methodological bricolage, and must move beyond any divisions between 'qualitative' and 'quantitative'. Kozinets would agree, and he writes (2015: 53–54):

> Consider that the images, words, Facebook profiles, Twitter hashtags, sounds and video files flowing through the Internet are composed of binary signals and various electromagnetically charged and uncharged blips of electrons and photons riding wires between various distant servers. Ultimately, they are zeroes and ones, already numerical and, in their own way, quantitative. We thus see fluidity and transferability, as analogue human experiences such as sitting and talking to a camera are transferred into digitally coded signals shared through a platform like Vine or YouTube, then decoded into densely pixelled moving images on screens and sounds emanating from speakers and headphones. This experience of audiencing can be captured as qualitative words and images experienced by a human listener and watcher, coded into fieldnotes or captured as a text file or visual screenshot, and immediately or subsequently optionally coded and transferred into a quantitative reading. Quant becomes qual becomes quant in this slippery shifting example.

In this chapter, I discuss three approaches to exploring, mapping, and mining data that can extend our ethnographic understanding: first, the idea of 'following the medium', and using digital media tools and platforms themselves as 'instruments of revelation'; second, *social network analysis*; and third, *text mining*. Depending on perspective, these approaches can either be seen as 'other' methods with which digital ethnography is combined, or — as in Kozinets' netnography — even as new forms of 'ethnographic' methods needed by the researcher in order to become fully immersed in that which is digitally social. It is important that the methodological bricolage is customised according to what is needed by the research task at hand.

FOLLOWING THE MEDIUM

There are two general ways in which research methods relate to digital society. First, there is the innovation of new, and the repurposing of old, research methods for mapping, analysing, and understanding digital society as an *object of study*. This is the perspective used throughout Chapters 13–16 of this book, where research methods are introduced for the study of the social transformations that are the topic of this book: interaction and identity (Chapter 4), communities and networks (Chapter 5), new modes of visuality and visibility (Chapter 6), new expressions of affect and emotions (Chapter 7), changes in the public sphere and in power structures (Chapters 8–10), new forms of mobile interaction and coordination (Chapter 11), and the underlying scripts of digital society (Chapter 12).

Second — and this is the topic of this part of this chapter — there is the possibility of harnessing the technologies and artefacts of digital society as *research methods in themselves*. Richard Rogers (2013: 1) suggests that what he calls 'digital methods' are about identifying and following '*the methods of the medium*' (in a wider sense) that are already embedded in digital society. Rogers' argument is that the internet is already doing research-esque things by itself, such as collecting, computing, sorting, ranking, and visualising data. This is just how it functions, and it is not related to anyone developing it this way to be useful specifically for research. The central idea of Rogers' approach to the study of the digital is to not intervene or interfere very much with these existing 'methods'. Our analyses may in fact be more accurate if we respect the integrity of them, follow them with curiosity, and learn from them. Rogers (2013: 1) writes:

> For example, crawling, scraping, crowd sourcing, and folksonomy, while of different genus and species, are all web techniques for data collection and sorting. PageRank and similar algorithms are means to order and rank. Tag clouds and other common visualizations display relevance and resonance. How may we learn from and reapply these and other online methods? The purpose is not so much to contribute to their fine-tuning and build the better search engine, for that task is best left to computer science and allied fields. Rather, the purpose is to think along with them.

The role of the researcher, then, becomes to attempt to 'follow the medium' and its methods as they evolve, and to find ways of exploiting and recombining them in useful and fruitful ways. So, we could, for example, ask ourselves: How can a hashtag be used for social analysis? How can Twitter's search function be used, not just to instrumentally find tweets, but to respond to questions about social dynamics

or cultural mores? How can we as researchers 'read' a Facebook feed in ways other than those intended by the creators of the service? The aim of thinking and working in this manner is, Rogers (2013: 3) writes, 'to build upon the existing, dominant devices themselves, and with them perform a cultural and societal diagnostics'. This means that the 'initial outputs' of the research — a search result, a set of Tweets, a set of Instagram accounts, algorithmic book recommendations, etc. — can very well be the same as, or at least very similar to, the things that digital devices output to their users. But with a 'digital methods' approach, Rogers (2013: 3) explains:

> they are seen or rendered in new light, turning what was once familiar — a page of engine results, a list of tweets in reverse chronological order, a collection of comments, or a set of interests from a social networking profile — into indicators and findings.

This shift of focus is Rogers' key point. For example, instead of reading Google results in conventional ways — as some sort of pure computed information that has been optimised by underlying algorithms — we might read them in other ways, in order to be able to see societal conditions. The main challenge for digital research, in that case, is to develop a mindset as well as a methodological outlook for doing social and cultural research *with*, rather than about, digital society.

INSTRUMENTS OF REVELATION

Internet researchers Christian Sandvig and Eszter Hargittai (2015) discuss how digital media and the internet can be seen to offer new tools for answering new, or old, questions in new ways. They give an example of how things that were not conceived as research instruments can still become used as such:

> In this view, online games like World of Warcraft were created by private companies to allow people to pretend to be night elves (or more accurately, for the company to make money from what people spend on subscriptions allowing them to pretend to be night elves). Yet these games might hold the potential to answer basic questions about the networked structure of human interaction. (Sandvig & Hargittai 2015: 8)

Employing digital media as a research instrument offers 'a new kind of microscope', which we can use to shed light on both new issues that are specific to digital society, and on basic and longstanding questions about human social life (2015: 6). Naturally, because of the multifaceted character of digitally networked tools and platforms, there are a wide variety of such uses. They can draw on new tools for data collection via

web scrapers, APIs, or online repositories. And they can also include new devices and ways of analysing data, in the form of computerised language processing, the harnessing of geolocative hardware, new visualisation techniques, and so on. The case of big data is just one example of the metamorphosis of digital society into research method, as discussed in Chapter 12. But, Sandvig and Hargittai (2015: 11) argue, the examples of big data are not the most fascinating ones.

> We instead see that the actual revolution in digital research instrumentation is going on now, all around us, in smaller, 'ordinary' research projects. We see it in the use of crowdsourcing to replace traditional pools of research participants; the use of hyperlink networks as a new source of data to study the relationships between organizations; or in the idea that writing your own Web-based application is now a viable data collection strategy.

As Sandvig and Hargittai point out, the totality of all such innovations, experimentations, and renegotiations are today's examples of what historian of science Derek J. de Solla Price (1986: 246) called *instruments of revelation*. When discussing the Scientific Revolution historically, he argued that its dominant driving force had been 'the use of a series of instruments of revelation that expanded the explicandum of science in many and almost fortuitous directions'. He also wrote of the importance of 'the social forces binding the amateurs together'. So, in the case of digital social research, we are now at that stage: a point where researchers often act like curiously experimenting enthusiasts — 'amateurs' — in testing and devising new 'instruments of revelation'.

ANALYSING SOCIAL NETWORKS

Another approach that can complement the ethnographic analyses is social network analysis (SNA). As discussed when we explored networks and communities back in Chapter 5, SNA is a method for looking at the structure of relations in social systems, and at the patterns of connections between and among those who take part in those systems. Even though SNA is a pre-digital method that can be used on datasets of any size, it is a method which is increasingly developed for, and used in, studies of 'big data' or other 'social data'. SNA is a set of theoretical perspectives and methodological tools, which aim to give a better understanding of individuals and groups in the relational social systems of which they are part.

Many people associate the concept of a social network to specific digital social network services such as Facebook and LinkedIn, and their predecessors, such as MySpace. However, the notion of 'social network' in SNA has to do with such networks and relations at the most basic level where we, as individuals and groups, are

part of a number of different social networks, in the form of families, groups of friends, school classes, organisations, clubs, professional networks, and so on. In the context of this book, SNA is described as a method with which to obtain a better understanding of the social networks with which people engage online and offline in digital society. This often means that data on network relationships are collected through the internet, but that our analyses in many cases assume that the social patterns we identify stretch beyond what people do online. Some networks may take shape online only, others offline only, but, naturally, it is most often a bit of both.

SNA sees people as social beings, and assumes that our interaction patterns affect what we believe, say, and do. It is also based on the idea that our positions in networks decide which other people we can influence, and how much. So SNA argues that the behaviour of individuals and groups is, if not totally governed, at least deeply affected by the social networks — the sets of socially networked relationships — in which they are embedded. People will think and do the things they do largely as a consequence of their ties to others. As SNA can help to demonstrate, the interaction patterns among individuals and groups in society is far from random. For example, people have a tendency to interact with others who are similar to themselves, and repeated interaction can lead to the emergence of (among other things) norms of behaviour, symbols of group belonging, group solidarity, as well as a sense of identity. So, social networks enable and constrain what people do, they also help us make sense of the world around us, and they influence the choices that we make.

With SNA, researchers can use different metrics and visualisation techniques to gain an understanding of how a certain network functions. When analysing digital society, one can think of any number of things and settings that we might want to analyse in terms of it being a social network, and it is also possible to do so on a number of analytical levels ranging from the whole of the internet, to text message exchanges among a small group of friends. Let's think of two concrete examples, just in order to have something to draw upon in the following description of SNA. First, we envision that we have a very large dataset consisting of several millions of tweets, all of which have used the same hashtag while posting about a major global political event. Second, let's imagine that we have copied and pasted around fifty posts to a thread in a discussion forum that deals with a topic relevant to our research. I will call these 'the Twitter example' and 'the forum example'.

DYADS — NETWORKS OF TWO

The basis for being able to do SNA is what Simmel called the *dyad* — as discussed in Chapter 5. A dyad can be defined as a pair of social actors along with the status of the network tie connecting them. In other words, it is a connection between two people or groups, together with the information about how they are connected. A dyad is a

group consisting of two people — a pair — and in order to be able to carry out SNA we need information about all such pairs that constitute the building blocks of the network that we want to analyse. As a result, it becomes crucial to decide what is seen, in the context at hand, to constitute a connection between two actors. Is it the fact that they exchange text messages, the fact that they are 'Friends' on Facebook, that they subscribe to each other's YouTube channels, that they have liked the same video, or that they have both commented on the same blog post? In our Twitter example, with a dataset that consists of a large number of users who employ the same hashtag, one could decide, for example, that any one user directing a tweet to a specific user is also then part of a dyad with that user. We may also decide that the addressed user should also respond back to the first one in order to constitute a dyad.

In the forum example, we could decide that all participants who have posted in the thread should be seen as being, theoretically, part of dyads with all others, as they have all somehow related to each other by being part of the forum thread. Another strategy could be to decide that any participant should be seen as having a dyadic relationship with just the participant who started the thread. Yet another strategy would be to say that all participants have entered into dyadic relationship with the participants who had posted the entry upon which their own entry followed in the thread. Or, one might decide that it is only in those cases when a participant explicitly mentions another participant in their post that they become related.

So, as you can see, this construction of pairs is driven by theoretical assumptions, and the choices we make will shape the patterns that we map out in the end. For example, in pre-digital versions of SNA, pairs could be identified by asking people in a workplace to suggest which of their colleagues they would be more likely to socialise with, then using the responses to analyse who was connected to whom. Another way would be to do an observation study of which persons were actually spending coffee breaks together. Fundamentally, in order to do SNA, we must have information on pairs of actors. This does not mean that we assume that people understand their world on the basis of all of the different pair-wise connections they have with people, but for the sake of analysis, we must break it down to these paired connections, because dyads are the fundamental unit of networks. If we take the Twitter example, we can draw on our dataset to create what is called an 'edge' list, like this:

User A mentions User B

User A mentions User C

User B mentions User C

User B mentions User C (again)

User D mentions User N

In most real-life analyses of social networks in digital society, the list would of course be much longer, and the graph much more complex. But we operate here with a small network for the sake of illustration. The most powerful analyses take place when we analyse networks that are complex enough for it to be hard to grasp how they function by just reading the edge list.

GRAPHS AND MAPS

The list above gives us the information we need as a starting point for SNA. In reality, we can also register other data about both the users and their connections. For example, we may know that Users A and C are politicians, and that Users B, D, and N are journalists. And we may also want to add other types of relationships apart from mentions to the list, including, for example, follows and retweets. In some network analyses, attention is paid to the direction of connections (*directed networks*), and in others, not (*undirected networks*). Our mentions are directed because users are actively mentioning other users. The social act of mentioning is directed from the mentioner to the mentionee. As we imagined our example tweet dataset to consist of millions of tweets, the list would likely be very much longer as well. But let's keep it simple for now.

In SNA, networks are represented as mathematical objects called *graphs*. Graphs hold information about *nodes* (the Users in our example) and *edges* (the connections between them — the mentions in our example). So, if we were to input our edge list above into SNA software, it would know to create a graph which included the nodes A, B, C, D, and N. It would also know to create edges between A and B, A and C, B and C, and D and N. I would also assign the edge between B and C a value of 2, because B mentions C twice. All other edges would have a value of 1. The next step that many researchers take is to create a visualisation of (commonly) circular objects connected by lines or arrows (called 'arcs'). Such network maps are what SNA is famous for, and they are what most people who have heard of SNA imagine when they think of the method. Rainie and Wellman (2012: 50) describe these as 'a bunch of network members connected by a bunch of lines'. The visualisations help explore the network data and assist in the interpretation of it. It is important to remember, however, that these network maps are not the same thing as the actual social networks that we analyse. The real-life network is not the same thing as the graph, as the graph is a simplification, which says nothing about many of the things that ethnography captures—such as people's thoughts, their driving forces, struggles, ambiguities, and so on.

Furthermore, the graph is not the same thing as the visualisation, because the visualisation is never automatic or 'standard'. Rather, it is the result of a process where the researcher thinks, from case to case, about the best way to abstract the observed social system as a network. The different available SNA softwares use similar algorithms to

visualise graphs. In general, the visualisations place the nodes on the screen, more or less randomly, and lines are drawn between them to represent the edges. Then, nodes are automatically rearranged to optimise the readability of the network visualisation, to make sure that nodes are not obscuring one another, that nodes are positioned close to the other nodes with which they are connected, and to try to avoid unnecessary crossing of lines. It is common practice for the researcher to experiment with different layout algorithms, and to make manual adjustments to the visualisation in terms of node sizes and colours, edge widths, filters for nodes with certain attributes, and the adjustment of placement.

CLUSTERS AND POWER LAWS

So, a graph must be analysed before it can be presented as a research result, and thinking about how to best visualise it as an image of the network is indeed part of the analytical process. Generally, SNA rests on the idea that social networks are not random jumbles of nodes and edges, ties, and connections. Instead, as discussed earlier, they are sets of relationships that have a profound effect on people and their actions. So, what types of patterns can be found through SNA?

First, it is possible to identify clusters. As Rainie and Wellman (2012) explain, people in digital society (they call them 'networked individuals', as discussed in Chapter 5) have a tendency to have many of their connections in densely knit groups where several people all have close and frequent connections with one another. Clusters, in other words, are parts of networks that are heavily interconnected internally. So in our examples of the Twitter dataset and the forum posts, we may be interested to see whether some of the included actors form strong sub-networks that affect and are affected by the network as a whole.

Networks can also be analysed in terms of centrality. One way of doing this is to calculate the 'degree' of nodes. This is a measure of how much of the activity in the network emanates from any particular node. The more connected lines a node has the higher its degree. In our Twitter example, a user being mentioned 20 times and mentioning others 15 times has a degree of 35. And, as the graph is directed, this can be divided into an 'out-degree' (activity) of 15 and an 'in-degree' (popularity) of 20. The degree centrality of a node is a measure of how prominent and important it is to the network structure as a whole. If all activity in a network stems from one participant being connected to everyone else, the network would be completely disconnected if that one person were removed. If many participants were connected to many others, the network would live on, despite the disappearance of an individual participant. The degree distribution of social networks, often shown in the form of a histogram displaying the number of nodes with each given degree, very often imitates a 'long tail' distribution (as was discussed in Chapter 1), according to which a minority of nodes stand for a

majority of the activity. As you will remember from the discussion about 'preferential attachment' in Chapter 9, mathematicians call such distributions a power law. As I explained then, this pattern differs from a 'normal distribution' — the so-called bell curve — according to which most nodes would have about the same number of links. A power law, by contrast, describes a situation where a small number of nodes are very well connected, while the rest are not.

The rich get richer

Referring to sociologist Robert Merton (1968: 62), and his discussion of science communication, one could conceive power law distributions in terms of what he called the Matthew Effect. He defined this effect as 'the principle of cumulative advantage that operates in many systems of social stratification to produce the same result: the rich get richer at a rate that makes the poor become relatively poorer'. As I discussed in Chapter 9, power laws are very common in social interaction on the internet and in social media. Howard Rheingold (2012: 195) explains:

> A few blogs get a jillion inbound links and hits, and a jillion blogs get a few inbound links and hits. Put this together with the small-world network structure of the Web, and you can see how videos and other Internet memes go viral.

The viral logic would be something like: an obscure blogger breaks a story; others link to it; then suddenly a 'supernode' with lots of connections links to it; attention becomes diffused to the long tail. It doesn't really matter that only a few people have a large audience, because when the conditions are right, that large audience is quickly accessible to others. In a many-to-many network such as the internet, the value of a node is not only based on the number of other nodes to which it is connected, but also on the potential number of groups it is connected to.

BETWEENNESS, WEAK TIES AND SMALL WORLDS

Path length is another thing to consider when analysing social networks. Paths are the network roads along which information can travel from one node to another, even though the nodes in question are not directly connected. So, paths are a connected

sequence of edges. If User A is connected by an edge to User B, and User B is connected by an edge to User C, there is a path (with the length of 2) between User A and C. In a complex network, there can be more than one path between nodes. But the shortest path between any two nodes is called a 'geodesic' in SNA. While the above-mentioned centrality measure of degree gives information about which nodes are the most active, another centrality measure — that of *betweenness* — measures how many of these shortest paths (geodesics) the node is on. So, the more geodesics between any two other nodes that a node is on the higher its betweenness. So, in other words, betweenness is a measure of how important a node is for the connection of other nodes with each other. Nodes with high betweenness are those that bridge social networks that are otherwise separated. Sociologist Mark Granovetter (1973) formulated a theory about the strength of such *weak ties*. The idea is that while the highly embedded 'strong ties' between close connections such as family and friends provide 'network closure' and align with the inclination of humans to operate in small groups, the 'weak ties' provide connectivity across a network. The weak ties function as bridges over 'structural holes' in the network. So, in spite of their name, weak ties can be very powerful and are more likely than strong ties to provide access to different social circles and to connect to more diverse networks.

In the passage that was cited in the box about how 'the rich get richer', Rheingold mentioned that *small-world* network structures are common online. The theory of such patterns of connection is also related to path length. You may have heard of the concept of 'six degrees of separation', which suggests that everyone in the world is connected to everyone else by roughly six steps in a chain of 'friend-of-a-friend' statements. According to the logic of 'it's a small world after all', social psychologist Stanley Milgram (1967: 67) wrote that 'we are all bound together in a tightly knit social fabric'. Having conducted experiments that involved mailing paper letters between acquaintances across the United States, he found that the average path length fell around five and a half or six. The experiment was repeated in the early 2000s, using email, by sociologist Duncan Watts and colleagues (2003). They gave more than 60,000 people from 166 different countries the task of reaching one out of 18 target persons by passing the message on to somebody they knew, and whom they thought was closer than themselves to the target person. The study showed that the typical chain length was five to seven steps, depending on the geographical distance between the source and the target. When it comes to social media, Rainie and Wellman (2012: 55) report that:

> The many bridges between Twitter clusters means that chains of information from one Twitter follower to a follower of that follower, and so on, encompass about 83 percent of all Twitter users within five steps of interconnection.

So, in sum, one can look at a number of different structural properties of the networks that are analysed through SNA. The method rests on the idea that social networks are not random. Rather, they are structures that affect and are affected by people. So SNA as a method can help tease out the prominent patterns from networks by tracing the flow of different resources (such as information, ideas, money, social support, power, love, etc.). Through such analyses, one can start to explore and discover how flows in networks have effects on people and the other way around. In doing so, one can be interested both in individual persons or groups, within the wider network, and in the network as a whole.

We may want to know more about a certain Twitter user in the Twitter example, or a specific forum participant in the forum example. What position does the actor have? What role does it play, and what are its resources? Which connections does it have, and what structural opportunities does it have to influence those it is connected to? We can also be interested in the entire Twitter dataset or the entire forum thread. What character does this social setting have as a whole? Is it tightly knit, with many actors being very active and connected to many others? Or is it sparsely knit, with just a small number of connections between a few key participants? Is it a centralised network, with most activity revolving around a key actor, or is it decentralised with some such key actors sharing the role of holding the network together? Is it a distributed network instead, where everyone is connected to everyone else? Having the answers to such questions about the particular setting in digital society that we want to study can be very helpful, both as research results themselves but, even more powerfully, as a complement to the thick descriptions which are generated through ethnography.

A PACT WITH THE DEVIL

Yet more digital methods can be brought in from the field of *text mining* to complement ethnographic analyses. As digital society has expanded, and as the internet continues to make available vast sources of textual data, largely in the shape of user-created content (see Chapter 2), there have been rapid developments in computerised methods for text analysis. This is not in the least because new groups other than computational linguists and computer scientists — for example, social scientists — are now taking an increased interest in such methods. The massive amounts of text content that are generated on social media and through other forms of computer-mediated communication have prompted social scientists to think more and more about how such data can be best used, and how the many technologies available to analyse it can be best harnessed.

The techniques that are called text mining were developed by computer scientists and linguists who wanted to use computers to identify and extract useful information from large numbers of documents — generally, a number large enough for it to be hard

to read and make sense of for any reasonable-sized group of human researchers. The large set of documents analysed in text mining is called a *corpus*. When researching digital society, we may want to make sense of a corpus of blog posts, forum comments, YouTube video descriptions, Facebook postings, tweets, and so on. Text mining is useful when we want to be able to see patterns in the corpus, which we would be unlikely to find by manually interacting with the documents one at a time. Text mining, therefore, as opposed to close reading of the text, can be seen as a *distant reading*. Literary scholar Franco Moretti (2013: 48–49) coined this idea, arguing that there is an analytical point to not close-reading texts, since this removes focus from the more general patterns that he thinks research should be focused on:

> The trouble with close reading [...] is that it necessarily depends on an extremely small canon [...] You invest so much in individual texts only if you think that very few of them really matter. Otherwise, it doesn't make sense [...] What we really need is a little pact with the devil: we know how to read texts, now let's learn how not to read them. Distant reading: where distance [...] is a condition of knowledge: it allows you to focus on units that are much smaller or much larger than the text: devices, themes, tropes – or genres and systems. And if, between the very small and the very large, the text itself disappears, well, it is one of those cases when one can justifiably say, less is more. If we want to understand the system in its entirety, we must accept losing something. We always pay a price for theoretical knowledge: reality is infinitely rich; concepts are abstract, are poor. But it's precisely this 'poverty' that makes it possible to handle them, and therefore to know.

In relation to more interpretative — 'qualitative' — approaches, then, distant reading demands that the researcher is prepared to move away from conventional close reading in order to be able to grasp larger sets of data, and also to lose some degree of qualitative detail because of this. At its core, text mining is about making a text into numbers to be able to calculate things about the text. It is based on registering, ordering, or counting words or phrases in documents — for example, social media postings. Once the documents have been numericised, statistical or predictive modelling methods can be applied in order to gain information about patterns in them (Miner et al. 2012: 71). Typical applications of text mining include analysing and structuring a text through strategies such as:

- 'Parsing' it to make it easier in later steps to extract information about specific parts of it.
- Finding the most relevant themes or *topics* (clusters of words and terms) that organise the analysed corpus.

- Automatically dividing documents into categories, which are defined beforehand or even computationally 'discovered'.
- Using dictionaries of positive and negative words to map *sentiments* in the text (for example, if things are mentioned in positive or negative ways).

These applications are important to build things such as search engines, spam filters, and online recommendation systems. But of course, the same methods can also be useful in digital social research. Miner and colleagues explain the important role of *tagging*, or annotation, with the help of the following example of an algorithm for extracting entities in a text (2012: 70–71, original tags have been simplified for clarity):

> For example, suppose a document contains the following sentence:
>
> Jim bought 300 shares of IBM in 2006.
>
> After processing this sentence through an entity extraction and tagging algorithm, the sentence might be 'annotated' as follows:
>
> <PERSON>Jim</PERSON> bought <QUANTITY>300</QUANTITY> shares of <ORGANIZATION>IBM</ORGANIZATION> in <DATE>2006</DATE>.
>
> So the words or terms in the sentence are now preceded with tags that identify the type of entity that it describes; for example, IBM describes an organization, 2006 describes a date, and so on. If all of the sentences in the entire corpus of text are tagged in this manner, it becomes much easier to perform efficient searches (or queries) of the corpus of text to extract, for example, all of the documents that mention the organization by the name of IBM and the person by the name of Jim, and so on. Thus, the corpus of text has been turned into a structured database that can be queried using the values for the entities, making it much easier to identify relevant documents, compute indices of relevance, and display (to the user) the specific places in the document where the entities of interest are found.

In a book entitled *Text Mining: A Guidebook for the Social Sciences* (2016), sociologist Gabe Ignatow and computer scientist Rada Mihalcea aim to make text mining accessible to a wider group of researchers than before, particularly in the humanities and the social sciences. As Ignatow and Mihalcea explain, text analysis has existed in various forms since the 1200s, but text mining is a fairly new set of methods, which is interdisciplinary but has its basis in computer science. Today, text mining draws on approaches such as data mining, information retrieval, computational linguistics, machine learning, and statistics. When researching digital society, one has access to very large amounts of text-based data as well as to advanced software and inexpensive but powerful programming languages such as Python and R. Taken together, these things hold the potential, Ignatow

and Mihalcea argue, to completely revolutionise text analysis in the social sciences. However, we must be aware that, even though we have a lot of text and good tools for mining it, we also have to actually interpret the patterns we can map. This is why SNA and text mining work best for social research when they are incorporated into a broader interpretative, ethnographic framework.

Some text analysis methods to start with

Corpus analysis is useful for 'distant reading' — seeing patterns in text from a holistic or large-scale perspective, which would be hard or impossible to see through close reading. Corpus analysis makes it possible to see how language is used more generally across a large number of documents (blog posts, tweets, comments, and the like). The method can respond to questions about which phrases are frequently occurring, about what types of expression would be more or less likely for a particular kind of document or author, and so on. You can start testing the method by playing around with voyant-tools.org, and dig in further by learning about the Antconc software,[1] for example, using the tutorial at The Programming Historian website.[2]

Sentiment analysis — sometimes called 'opinion mining' — is a method for determining the attitude of a speaker or writer. This can be with respect to a particular topic or with the aim of assessing the overall tone of a larger or smaller chunk of text. The method can be applied using a variety of different tools, but you can try out 30db.com or streamcrab.com to get an initial feel for the method.

Topic modelling is a form of text mining that aims to identify 'topics' in a corpus. The method processes large bodies of text to find recurring patterns of co-occurring words (topics). An open source tool for doing topic modelling is MALLET,[3] for which an accessible tutorial is also available at The Programming Historian.[4]

---

[1] www.laurenceanthony.net/software/antconc.

[2] http://programminghistorian.org/lessons/corpus-analysis-with-antconc.

[3] http://mallet.cs.umass.edu.

[4] http://programminghistorian.org/lessons/topic-modeling-and-mallet.

FURTHER READING

Rogers, Richard (2013). *Digital Methods*. Cambridge, MA: MIT Press.
In this relatively new but already influential book, Rogers – as discussed in this chapter – argues for a repurposing of digitally native tools and techniques for researching society and culture. This is about reapplying things such as search engines, crowdsourcing, tags, and likes, for use in research.

Robins, Gerry (2015). *Doing Social Network Research*. London: Sage.
This book offers hands-on guidance in how to design and carry out social network analysis research. Robins discusses topics ranging from data structures, data collection methods, and ethical issues, to techniques for analysis, visualisation, and interpretation.

Ignatow, Gabe, & Mihalcea, Rada (2016). *Text Mining: A Guidebook for the Social Sciences*. London: Sage.
Ignatow and Mihalcea's book aims to make text mining accessible to a wider group of researchers than before, particularly in the humanities and the social sciences. It addresses issues of how to deal with natural language data from the perspective of both sociology and computer science. The book covers areas such as web crawling and scraping, lexical resources, text processing, and text mining techniques from a variety of areas.