Mikko Hyyryläinen

Type: *Taking a paper of your choice which is not using data science methods and introduce how you would use data science approaches to redo that paper 10 points per paper*

Link to the article: https://journals.sagepub.com/doi/abs/10.1111/j.1467-9558.2008.00324.x

The paper I am discussing is Gabriel Abend's (2008) *The Meaning of Theory* which was published in *Sociological Theory* (26:2). In this paper Gabriel Abend describes seven different meanings in which the word *theory* is used in sociology. The methodology – if a theoretical article can be said to have one – seems to be that Abend has been participating in the discussion on sociological theory so long that he has formed a classification scheme for the way the word theory is used in sociology. What I wondered could there be some more systematic way of answering the question of what is meant with theory in sociology and perhaps machine learning methods could provide tools for this task. I think that with machine learning methods it would be possible, at least in theory, to find a more empirical answer to the question asked by Abend. What is needed to be able to answer this question are two things: 1) we need to be able to isolate when theory is discussed in a sociological paper and 2) we need to be able to find the different meanings that theory is given from a large body of texts that discuss theory.

The two data science methods that would seem to be useful for this task are dictionary methods and topic modeling. The data set would consist of a collection of articles from sociological journals (not just leading but all kinds of sociological journals). Apparently, it would be allowed to download papers for research purposes through publishers APIs. The data set should also be quite large. In addition, we should have some kind of understanding of what is the structure of a sociological paper.

The first step would be to somehow isolate the sections of the articles that explicitly discusses theory. If we consider that the data is large enough, the first step would be to use dictionary methods to find the articles that have an explicit theory section – i.e. for example search words "theory" (or "theor", "theory", "theoret" and so on if we stemmed the texts). After that we should somehow isolate the theory section from the article. If we could download our articles in .html form we could use Beautiful Soup to find the tags that define where the theory section begins and ends. The end result would be a large body of documents that are the theory sections of each article.

This large body of texts would be then analyzed with topic modeling. In this part we would equate topic with what meaning the word theory is given in the articles. The way the data should be pre-processed would probably be to remove all general words that are used frequently when discussing theory. For example (and obviously), the word theory would be one of them. Also the names of social theorists/sociologists could perhaps be reasonable to remove since they are not necessarily relevant for the task we have. Since we know that each of them are discussed in relation to sociological theory, they are not relevant for what is meant with theory. After all these steps (isolating the theory section and pre-processing the data) we could be able to be able to use topic modeling and find the different meanings that the word theory has in sociology.

If we would be successful in finding the meanings of theory from the data, we would perhaps be able to add another aspect to the question of "What is meant with theory?" by adding the

question of “How has the way of how theory is used in sociology changed over time?”. This would probably mean that we would be looking at how the fractions of the topics in a document that represents the average document changes over some time. In addition we would perhaps be able to find if some use of the word of theory has disappeared or emerged during sociology’s history. Hence, we could be able to trace sociological trends. In addition, we would be able to test if Abend’s classification scheme was able to represent the different meanings that theory has in sociology in an adequate way.

However, this approach would probably be quite problematic. There are at least two problems that are connected to each other. If we would only use the “theory section” of the paper, I am not sure if it provides data that can be used to map the *meaning* of theory. For example, for Abend, one meaning that theory has in sociology is that it suggests that there is a causal relationship between two variables. Perhaps we would be able to find a topic in which the defining words would be concepts such as “influences” or “causes”. However I am not 100% sure how another meaning of theory, an explanation of a particular case, would be defined if we only looked at the theory sections. Especially since articles that use theory in this sense does not necessarily have an explicit theory section. Second problem is that can topic modeling find differences between topics that are about the same thing, theory, but there are some subtle differences in how the thing is discussed. Of course, topic modeling itself does not find the differences by itself, the researcher interprets the topic models and finds the differences. However, I wonder if topic modeling can provide raw material that is sufficiently nuanced to the researcher. I think that this redo would be interesting but at the same time it raises many questions (e.g. how to find “theory” from a scientific article or how computational methods can find subtle differences from texts) that are probably quite difficult to solve.