The myth of unstructured text

Data needs to be characterized, i.e. labeled within a category. Once assigned, data is easy to retrieve. People who create tables, spreadsheets, databases abide by this vision. Once everything is properly labeled, computers make it easy to retrieve, filter, sort, calculate, visually present this data in synthetic form, that can be presented to show trends, previsional results, project management, results, returns on investments. Managers tend to rely on this paradigm of data analysis for most of their activities.

In this perspective, text appears as unstructured data. However, nothing can be further from the reality. The oxymoron “Artificial Intelligence” exploits that ignorance, while being based on a long tradition based on the structure of text.

Artificial intelligence is the latest attempt to digitize textual data, i.e. to analyze it into chunks that can be described with numbers. The concept of vector databases, at the heart of the “large language model” paradigm, consists in recording each word, associating it with its immediate neighbors to create expressions, expand it to phrases and sentences, and spit out the phrases based on a large number of occurrences. This calculation, based on a huge number of occurrences, is what gives the illusion that computers “understand” what we are asking them. What they are doing is simply retrieve patterns that resemble the question in a prompt. The first search engines only retrieved words that were spelled right. Then the engines evolved to accept some variations, such as capitalization or accented letters, then synonyms. The current large language models are based on algorithms that expand these associations to return whole sentences that they are extracting from the accumulate knowledge corpora, and their results are fine-tuned by an army of humans who curate the results, in order to eliminate the most flagrant errors returned by automatic processes. But they can’t track everything that is spit out by the AI algorithms.

In *Runaway Technology*, Joshua A.T. Fairfield shows that law is a technology. Similary, music is a technology. Text is also a technology, and this is what I am going to talk about here.

However, there is much more out there that is not reachable by this dominant paradigm. And we are so blinded by the fact that computers just crunch numbers that we are failing to see things that are in front of us and that are technologies, heavily structured but in a different way. Text is a technology. Law is a technology. Music is a technology. Etc.

[^ See also the book on law as technology: *Runaway Technology* Can Law Keep Up, by Joshua A. T. Fairfield].

But structure existed long before computers.

Nothing can be further from the reality. Text is a technology, made of several layers, and well-structured. To understand it, let’s go back for a while to elementary school, and remember what we learned them about the basics of language, as well as arithmetic fundamentals.

To write, we use a limited set of well-defined characters or ideograms, which vary according to the language we are using. A set of symbols complement the alphabet, that serve as punctuation, mathematical operators, other symbols. Emojis are used as symbols to express emotions.

Written text is made of words, which are a sequence of letters, separated by white space. Sentences are made of words. The initial word of a sentence starts with a capitalized letter. A sentence ends with a dot. Sentences can be divided into kinds: questions, answers, assertions. Words play various roles within a sentence. They can be articles, nouns, verbs, adjectives. Grammar rules indicate how sentences can be constructed, how names can be changed to indicate a plural form. Verbs are conjugated according to certain rules to express the time in which their action occurs. The rules that govern written text are well established and quite constraining, they vary by language.

Numbers also follow well-established rules. The most common numbering system taught at school and used is the decimal system, made of 10 digits, from 0 to 9, that, when combined, allowed the creation of numbers. Arithmetics defines a set of rules that define how to operate on numbers. Numbers can appear also in text. Furthermore, every number can be represented in an alternate numbering system. The binary system uses only 0 and 1 to represent numbers. For example, 0 is 0, 1 is 1, 2 is 10, 3 is 11, 4 is 100, etc.

Text is therefore pretty well-structured. Considering text as unstructured is therefore misleading, as it ignores all the implicit rules which are at play. Worse, it is like pretending that they don't exist.