

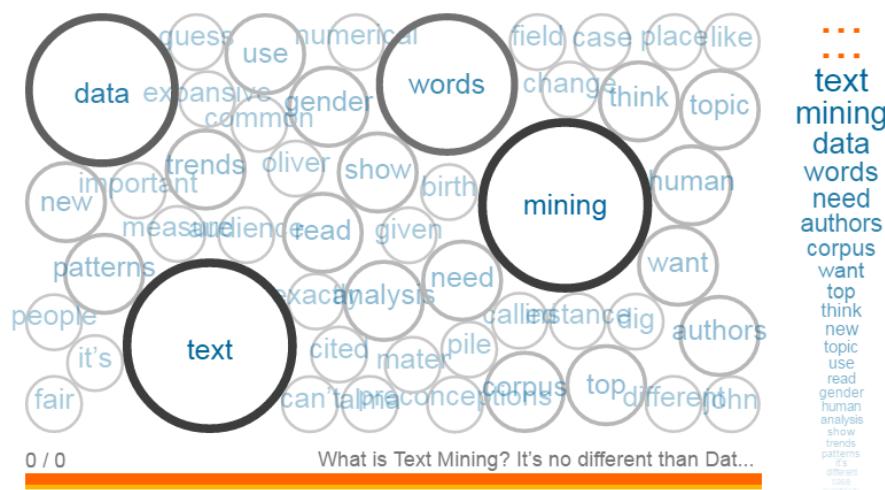
Text Mi·n·ing (noun)

What is Text Mining? It's no different than [Data Mining](#), but in this case your data is text, rather than numerical statistics. If you want to measure *words* – such as how often John Oliver urges his audience to “be fair”, or who the top cited authors are in a given field – you will need to do some text mining. People call this type of analysis “mining” because the amount of data can be expansive, and you will need to sort out the valuable data. For instance, the most common words in any set (or ‘corpus’) will be what are called “stop words”: ‘the’, ‘and’, ‘but’, ‘like’, etc. If we start by stripping these words from our data, we will have a much smaller pile to dig through.

Why is text mining important? Can't we already guess the results? Well, not exactly. First, we all have preconceptions that change our interpretation of a text or speech. If we zoom out and rearrange the words (say, by frequency, topic, or context), we might see new patterns and lines of interest that would otherwise have gone unnoticed. Second, indications can be so inconspicuous as to escape notice by even the most perceptive human minds. Think of anonymous or [pseudonymous authors](#): we can use text mining to learn their identity by comparing subtle traits in the vocabulary and grammar across a huge corpus of books. Third, researchers are ambitious. We want to analyse more data than we have time to read. With text mining, we can quickly pull out the top fifty words in a website or a book – think of the ubiquitous word cloud. Or we can see [which scientific articles present new information](#) out of the mass of literature published every year.

We all participate in text mining every day, and have for years. Ever look at the news trends that show up on your Facebook? Search through your email for a person or topic, or use your email’s spam filter? Read about studies that uncover [racial](#) and [gender bias](#), or consider the generated recommendations by Amazon and Netflix?

By no means does text mining replace human analysis, but it can show trends and patterns in data that are either too micro or too macro for us to effectively study.



Visualization of this text created using [Voyant Tools](#)