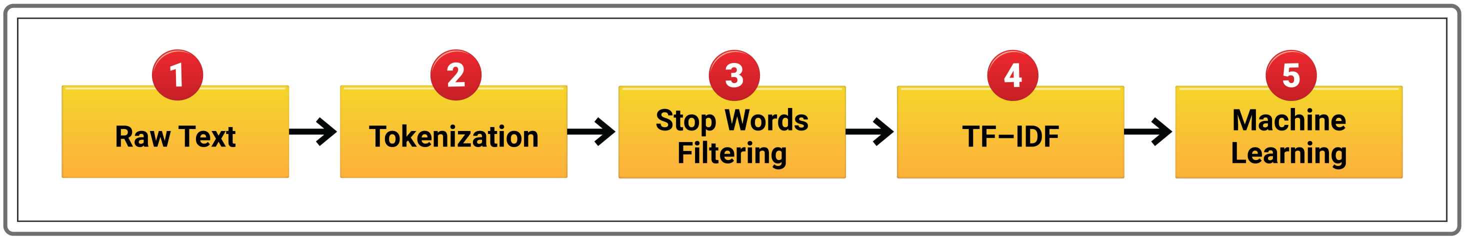
16.5.5

NLP Pipeline

Now that you understand the many steps needed to get regular text into a machine-readable state, it's time to get the process going. This process is called building an NLP pipeline.

NLP is complicated. To manage it, you must **build an NLP pipeline**, a process breaking NLP down into a series of smaller, less complex tasks. Below we'll provide a high overview of this process, and in the next section, we'll dive deeper with the code.

Each step of the NLP pipeline involves a separate task. The output data from one step, in turn, becomes the input data for the next step, with an opportunity to evaluate and refine each task, if needed. A basic NLP pipeline follows:



Here's a breakdown of each step:

1. **Raw Text:** Start with the raw data.
2. **Tokenization:** Separate the words from paragraphs or sentences, into individual words.
3. **Stop Words Filtering:** Remove common words like "a" and "the" that add no real value to what we are looking to analyze.
4. **Term Frequency-Inverse Document Frequency (TF-IDF):** Statistically rank the words by importance compared to the rest of the words in the text. This is also when the words are converted from text to numbers.
5. **Machine Learning:** Put everything together and run through the machine learning model to produce an output.

We'll cover each step of the pipeline in more depth in the next section.