**Some Basic pre-processing steps for Text Analysis**

In a real world scenario the text we get for analysis is not the ideal one to analyse. The text may have something we’re interested in and something we’re not at all interested. For example, If we want system to analyse the speech of a person, we may be interested in the topic(s) he is caring about or delivering in his speech and may not be interested in the words like ‘and’, ‘an’, ‘the’, ‘to’, ‘in’ etc. These kind of words generally don’t contain any meaning or at least don’t infer what the topic is about. We consider these words as “Stop Words”.

Pre-processing of text is very important and mandatory step in order to make prepare data for processing. Pre-processing is not just about taking things out and removing stop words, it may also be about putting things. I will discuss in a while about different pre-processing steps and tools used.

**Tokenisation & Normalisation:** finding boundaries between word-like entities in character string

**Fixing Misspellings:** where possible

**Stemming, lemmatisation, POS-tagging:** finding slightly deeper identities between words (fished, fishing)

**Removing Stop Words:** maximising the content-full words in the document/corpus

**Entity Extraction:** identifying conceptual entities behind words

Now let’s look at the Tokenization, what it is, how it works, why it is beneficial and it what cases.