**Identifying Features from Text:**

**Types of Textual Features:**

Words: We could break down all the words in a document into features, but that would result in massive data frames. How do we handle **Stop Words** which are words like “the”? This word is unlikely to be very important to a classification task.

Another decision is how to normalize your corpus, should you leave all the words as they are, or should you make all the words lowercase? Sometimes uppercase words will be more important for your classification.

Other things that could be used for classification:

* Stemming/lemmatization: used to find root words,
* POS: used to find if the word is a noun or adjective etc. This can be used to identify specific spelling of things. E.g. “weather” or “whether”, if the word before was “the” then it is most likely to be “weather”.
* Grouping words of similar meaning to be one feature. E.g. big, huge, enormous.
* Features might be from a sequence of words e.g. “White House”, this would use a bigrams approach.