Useful string things

**string.find(s, sub[, start[, end]])**

Return the lowest index in s where the substring sub is found such that sub is wholly contained in s[start:end]. Return -1 on failure. Defaults for start and end and interpretation of negative values is the same as for slices.

**string.rfind(s, sub[, start[, end]])**

Like find() but find the highest index.

**string.count(s, sub[, start[, end]])**

Return the number of (non-overlapping) occurrences of substring sub in string s[start:end]. Defaults for start and end and interpretation of negative values are the same as for slices.

**string.join(words[, sep])**

Concatenate a list or tuple of words with intervening occurrences of sep. The default value for sep is a single space character. It is always true that string.join(string.split(s, sep), sep) equals s.

**string.replace(s, old, new[, maxreplace])**

Return a copy of string s with all occurrences of substring old replaced by new. If the optional argument maxreplace is given, the first maxreplace occurrences are replaced.

Text mining ideas

* If there is a + then there are two ingredients and if not one
* Amounts and ingredients start with a number/ some end with the amount and units in brackets
* Units have punctuation
* Function have max ‘x’ characters
* Trade names start with a capital letter – apart from one with EC on different line
* Function is one of a small number of options – same for every year? Ish, create a list being careful with names that are on different lines – include space
* Split string and add to database depending on certain criteria
* How to keep entries together? Some way of bundling into sets of three? Creating an ordered array?
* Ingredients if there isn’t a bracket pair up with next element if that is bracketed
* Make an array for each set of information by using capital letter and brackets as markers?
* Write functions to determine the structure of the text
* Function which standardises the format – looks for capital letters to find tradenames, if there is a space after its name is finished – using the fact that if the line isn’t in one string it will be the following entry
* If number next to letter (not punctuation or other symbol) don’t include

Write a program that finds the pages with pesticides used:

* Starts with the page with ‘PESTICIDES’ and disregards everything before
* Ends with where ‘BROADBALK’ , does this always stop after one full page of entries?