Predicting the next letter

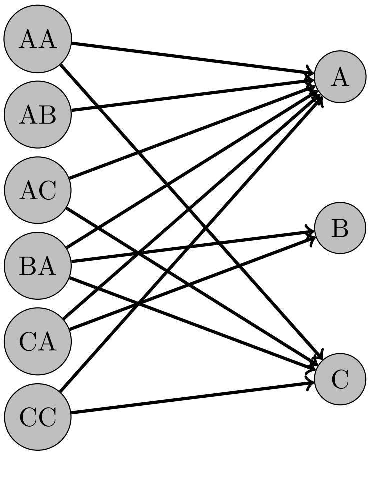
Peter Rowlett

Consider these 12 strings (a string is a sequential set of characters like a word or sentence, and can contain letters, numbers and punctuation).

AAB
AAC
ABA
ACA
ACC
BAA
BAB
BAC
CAA
CAB
CCA
CCC

What we would like to do here is to generate strings of letters that look like they belong in this set. Clearly they should only use the characters 'A', 'B' and 'C', but there is more to it than this. For example, notice that 'AB' is never followed by anything other than 'A'? So we aren't just choosing letters at random.

Each of these strings is three letters long. We are going to think about which two letters at the start are followed by which letter next. We can represent this information in a graph, where an arrow from node 1 to node 2 means the pair of letters in node 1 is followed by the letter in node 2.



We can also write this information as a transition table – the pairs of letters across the top are followed by the letters down the left hand side with the proportions listed.

	AA	AB	AC	BA	CA	CC
A	0	1	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{2}$
B	$\frac{1}{2}$	0	0	$\frac{1}{3}$	$\frac{1}{2}$	0
C	$\frac{1}{2}$	0	$\frac{1}{2}$	$\frac{1}{3}$	0	$\frac{1}{2}$

Note that we could write this as a square transition matrix, where the entries down the left hand side are pairs of letters rather than individual letters. For example, when ‘AA’ is followed by ‘C’ it produces the sequence ‘AAC’ which begins ‘AA’ and ends ‘AC’. When ‘BA’ is followed by ‘C’ it also ends ‘AC’, but when ‘AC’ is followed by ‘C’ it ends with ‘CC’. Doing this, we could consider it an eigenvector problem and find the ranking as we did with PageRank or consider the long-term probabilities for pairs of letters. However, this would just tell us what pairs of letters are in what proportion in the original set of 12 strings, and we can count this from the 12 strings: for example, ‘AB’ and ‘BA’ occur three times in the 12 strings and are thus the most common letter pairs. Doing this via matrices wouldn’t tell us anything new, and really we aren’t interested in the long-term next letter.

What we can do with our table is use it to look up the probabilities that a pair of letters is followed by each of the three letters. For example, if we see ‘AB’ it is certain this will be followed by ‘A’; if we see ‘CA’ there is a probability of 0.5 that it is followed by ‘A’ and a probability of 0.5 that it is followed by ‘B’; and so on.