**Using Markov Chains for Spam detection**

* Using Markov Chains, we can tell the probability of transition between two states.

*Assumption of a Markov is chain is that the future state depends only on the current state*

* Here each state is the word in a tweet ( we can also use the same logic with character in a word)
* *Eg: Possibility of 'f' coming after 'z' is almost zero but possibility of 'u' coming after 'q' is almost 100%*
  + We'll be able to find the probability of transition from one word/character to another word/character.
  + If the probability is very low (less than threshold), then we label is as spam

* Given we train the model on sufficiently large 'not spam' data, we'll be able to label spam (sentences that make no sense and words that are gibberish) with good accuracy

Eg:  The columns here are future states, rows are current states.

This example here shows probability of next day being 'Rainy' when today is 'Sunny' as 0.19

Machine generated alternative text:
States 
Sunny 
Rainy 
Snowy 
Sunny 
0.8 
0.2 
0.1+ 
Rainy 
0.19 
0.7 
0.2 
Snowy 
0.01 
0.1 
0.7 

With good training data, we'll create a large multi-dimensional transition matrix (similar to above matrix) where rows and columns are the corpus of words in the tweets

We'll get probabilities of next word (in columns) when the current word is 'x' ( in rows)

Then label spam based on threshold