Filter Matching

2020-10-18

A mathematical analysis of pattern matching in digit sequences, specifically examining how many numbers contain a given pattern like '01'. This post develops a combinatorial approach to count pattern occurrences in n-digit numbers, accounting for over-counting through inclusion-exclusion principles. The analysis reveals that as the number of digits increases, most numbers will contain any given pattern, with connections to the mathematical properties of irrational numbers like .

blog: https://tetraquark.vercel.app/posts/pattern_matching/?src=pdf

email: quarktetra@gmail.com

Problem Statement

Consider the set of numbers $[0, 10^N - 1)$ where N is the number of digits. How many of these numbers will have a certain filter pattern, i.e., for a given N, how many of the numbers will contain 01?

- $N < 3$: We can immediately see that the numbers will not have any 01 in them .
- $N = 3$: We can list the numbers with the required pattern: $\{101, 201, \cdots, 901\}$.
- $N > 3$: How can we calculate the number of instances that include the pattern we are looking for?

Note that we are trying to find out whether a number contains any 01. If it does, no matter how many times 01 appears in the number, we will count it as 1. If we were simply calculating how many times the pattern appears, it would have been a much simpler problem. To illustrate, consider 50101: 01 appears twice in this number, however this is will count as one.

Total number of appearances

Consider n digit numbers. First of all, note that the numbers do not start with 0 by construction. Therefore, the leading digit, X, can only be $[1-9]$, i.e., it has 9 combinations. We can place 01 anywhere in the remaining $n-1$ digits, however, given the fact that 01 has two digits, we have $n-2$ spots available for it:

$$\underbrace{|\, X\, 0\, \overbrace{1\, \underbrace{\cdots\cdots}_{\text{n-3}}}^{\text{n-2}}\, |}_{\text{n}} \tag{1}$$

Therefore, total number of 01 appearances is:

$$\text{Total number of appearances} \simeq 9(n-2)10^{n-3}, \tag{2}$$

where the factors 9, $n-2$ and 10^{n-3} come from total possible cases for X, the possible positions of 01 and, possible combinations of $n-3$ digit numbers, respectively. We used $\simeq$ rather than $=$ since we will need to correct for the over-counting.

Multiple appearances

Eq. 2 over counts the numbers because it assumes where ever 01 lands, it creates a new number. This is not necessarily correct when we already have 01 in the $n-3$ digit number. Swapping 01's will not create a new number. We need to identify such cases and subtract them out. It will look as follows:

$$\underbrace{X\boxed{01}\boxed{01}\,\underbrace{\cdots\cdots}_{\text{n-5}}}_{\text{n}} \tag{3}$$

we have two $\boxed{01}$ objects, and $n-5$ digits, i.e., $n-3$ objects total. The total number of different combinations we can create is given by

$$\text{first correction} = 9\frac{(n-3)!}{2!(n-5)!}10^{n-5}, \tag{4}$$

where $(n-3)!$ represents different configurations of $n-3$ objects, $2!$ removes the over-counting due to the fact that swapping $\boxed{01}$'s does not give a new combination, $(n-5)!$ removes the over-counting within $n-5$ digits since that is already taken into account with the term 10^{n-5}. Subtracting this from Eq. 2, we get

$$\text{Total number of appearances} \simeq 9(n-2)10^{n-3} - 9\frac{(n-3)!}{2!(n-5)!}10^{n-5}. \tag{5}$$

This equation is almost correct. We need to check for another over-counting we did in Eq. 4, just like in the case of Eq. 2: we assumed that swapping of $n-5$ with $\boxed{01}$ creates new combinations. This is clearly not correct if $n-5$ has $\boxed{01}$'s in it. In that case, we need to consider the following:

$$\underbrace{X\boxed{01}\boxed{01}\underbrace{\boxed{01}\underbrace{\cdots\cdots}_{\text{n-7}}}_{\text{n-5}}}_{\text{n}} \tag{6}$$

The number of combinations we can create is:

$$\text{second correction} = 9\frac{(n-4)!}{3!(n-7)!}10^{n-7}, \tag{7}$$

which is the over counting we did in Eq. 4, i.e., we need to correct the correction term by subtracting this out.

$$\text{Total number of appearances} \simeq 9(n-2)10^{n-3} - 9\frac{(n-3)!}{2!(n-5)!}10^{n-5} + 9\frac{(n-4)!}{3!(n-7)!}10^{n-7}. \tag{8}$$

Note that in Eq. 7, we again assumed $n-7$ digits had no $\boxed{01}$'s in them, but this might be the case. We have to rinse and repeat: correct the correction to the correction. It might look like a never ending battle, however, we quickly notice the pattern of alternating corrections and shifted factorials. Therefore we can arrive at the final answer as a sum of alternating terms:

$$\text{Total number of appearances} = 9\sum_{j=2}(-1)^j\frac{(n-j)!}{j!(n-2j+1)!}10^{n-2j+1}, \tag{9}$$

where the summation will gracefully terminate for $j > \frac{n+1}{2}$ since factorial of negative numbers is defined as ∞. Note that this is the number of matches for the set of all n digit numbers. If we wanted to compute the total of all matches up to and including N digit numbers, we just sum over n:

$$\text{Total number of appearances up to N digits} = 9\sum_{n=3}^{N}\sum_{j=2}(-1)^j\frac{(n-j)!}{j!(n-2j+1)!}10^{n-2j+1}, \tag{10}$$

where the summation over n can be started from 3 since we know there is no match for $n < 3$.

An approximate solution

If we are just interested in an estimate of the fraction of the numbers that will match the filter pattern, we can come up with a quick estimation. We have an n- digit number, we are sliding a pattern of 01 over $n-1$ digits to check if they match. The probability that we will not observe 01 is 0.99. We need to slide the pattern and keep checking for a total of $n-2$ times. The probability that there will be no match is 0.99^{n-2}, which means the probability of matching is $1-0.99^{n-2}$. This shows that as n increases, most of the files will have 01's in it.

The story of π

π is an irrational number and its digits don't repeat themselves. Many mathematicians believe that it contains every possible finite sequences of numbers in it. How is π relevant to this problem? It just illustrates that if the number is long enough, it will contain any pattern you can imagine. Do you want to find where your name appears in π, convert your name to ASCII here, and search for it in π at this site.

Results

Table 1: The table of computed numbers.

digits	Count	Count.cumul.	PCT.cumul.	PCT.appr.	first	second	third	fourth
1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0
3	9	9	0.9	1	9	0	0	0
4	180	189	1.89	1.99	180	0	0	0
5	2691	2880	2.88	2.97	2700	-9	0	0
6	35730	38610	3.86	3.94	36000	-270	0	0
7	444609	483219	4.83	4.9	450000	-5400	9	0
8	5310360	5793579	5.79	5.85	5400000	-90000	360	0
9	61658991	67452570	6.75	6.79	6.3e+07	-1350000	9000	-9
10	701279550	768732120	7.69	7.73	7.2e+08	-18900000	180000	-450
11	7851136509	8619868629	8.62	8.65	8.1e+09	-2.52e+08	3150000	-13500
12	8.681e+10	9.543e+10	9.54	9.56	9e+10	-3.24e+09	50400000	-315000