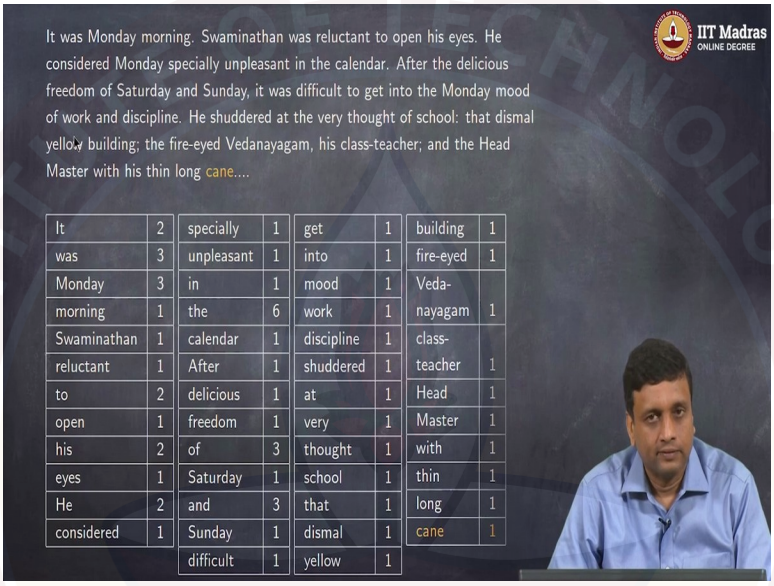
IIT Madras

ONLINE DEGREE

Computational Thinking
Professor Madhavan Mukund
Department of Computer Science
Chennai Mathematical Institute
Professor G Venkatesh
Department of Electrical Engineering
Indian Institute of Technology Madras
Max in a single iteration and max in two iterations (non-nested)

(Refer Slide Time: 00:14)



So, we have looked at how to count the cards in a particular set of cards. And now, let us look at this dataset we had, which had a paragraph of words. And supposing we want to count how many times each word occurs. So we really want to find out which word in this paragraph is most frequent. So, first of all, that means we have to count every word and one difficulty we have is that we do not really know in advance, how many words there are and which words will appear.

So, let us just read through the paragraph. So here it read; It was Monday morning. Swaminathan was reluctant to open his eyes. He considered Monday specially unpleasant in the calendar. After the delicious freedom of Saturday and Sunday, it was difficult to get into the Monday mood of work and discipline. He shuddered at the very thought of school: that dismal yellow building; the fire-eyed Vedanayagam, his class-teacher; and the Head Master with his thin long cane.

So, as far as we are concerned, a word is a single unit. So, there are some hyphenated words like fire-eyed and class-teacher. So, we will treat these as single words. Of course, we will ignore commas, full-stops and other punctuation. So, the way we go about it is to, as we do with a normal iterator, go through each word. Imagine that this is a stack of cards. So, we will start with the first card and as we see each card, we will increment the count associated with the word on that card.

So, the first word in this paragraph is 'It'. And at this point, we do not have a counter for 'It'. So we create one and set its value to 0. And since we have seen now the first occurrence of the word 'It', we increment this count and make it 1. So, at this point, we have one counter for one word 'It' and its value is 1, because we have seen one copy. In the same way, we go on to 'was'. Now, this is a new word. We do not have a counter for 'was'. So again, we have to create a counter, set it to 0 and then increment it to indicate that we have seen one word 'was'.

The same happens when we see 'Monday'. Again, 'Monday' is not a word we have seen before and notice as we go along, we have to keep looking at all the counters we have already created to see if it is one of them or not. So, at this point, we have to compare it with 'It' and 'was' and discover that it is neither of those. So, 'Monday' now gets initialized to 0 and then incremented to 1, 'morning' also gets initialized to 0 and then incremented to 1, 'Swaminathan' also gets initialized to 0 and incremented to 1.

And now we come to the first interesting case, which is the word 'was'. So, we realize that we have seen 'was' before. We already have a counter for 'was' and its value is currently 1. So we increment it and make it 2. Moving on, we come to a new word 'reluctant'. So, 'reluctant' is a new counter which is incremented to 1, 'to' is a new counter which is incremented to 1. So, we are using the word itself to name the counter. So these are, the words in these boxes at the bottom are like variables.

So we are giving, creating a variable for every word in our text. Similarly, 'open' is for the first time. So, 'open' gets incremented to 1, 'his' is appearing for the first time, incremented to 1, 'eyes' is for the first time, incremented to 1. Now, we come to an end of a sentence but it does not bother us. We keep going. So, 'He' is coming for the first time. Similarly, 'considered' is coming for the first time. Now, we see 'Monday'. So 'Monday' has appeared again.

Now, you might realize that it is a bit tedious as we go along even manually to go back and check. Every time, we have to go through the entire sequence of words. So, if we are actually going to do this in some automated way, we need an efficient way to check which counters have been created and which counters need to be created. Otherwise, you might accidentally create a new counter for ‘Monday’ and have two different counts for Monday rather than a single count.

But, anyway, assuming we got it right, we see ‘Monday’ has already appeared. So now, ‘Monday’ comes for the second time. So ‘Monday’ gets value 2. ‘specially’ is a new word, that is 1, ‘unpleasant’ is a new word, so that is 1, ‘in’ is a new word, so that is 1, ‘the’ is a new word, so that is 1, ‘calendar’ is a new word, so that is 1. ‘After’ is again a new word, so that is 1. Now, we come to ‘the’ and ‘the’ we have seen before. So, ‘the’ was previously 1, becomes 2. ‘delicious’ is a new word, becomes 1, ‘freedom’ is a new word, becomes 1, ‘of’ is a new word, becomes 1.

‘Saturday’ is a new word, becomes 1, ‘and’ is a new word, becomes 1, ‘Sunday’ is a new word, becomes 1. So what you can see is that in this particular paragraph, as we go along, we are creating a counter each time we see a word for the first time and the number of counters is getting rather large. So, we have to keep track of a large number of variables. And not only that, each time we have to decide whether to create a new variable or not. So, it is a little bit tricky, as I said, to implement this and not make a mistake.

So, now let us go to ‘it’. Now, here we have to make a decision. So, notice that we have a counter called ‘It’ but to the very first word in the paragraph which is a capital ‘It’ and this is a small letter ‘it’. But, let us assume that a capital ‘It’ and a small letter ‘it’ are actually the same word. So, we will just treat this as a repeat of the word ‘It’ and make the counter for ‘It’ 2, even though it is spelt with a capital letter once and with a small letter again. ‘was’ again is a word we have seen before. So, ‘was’ goes from 2 to 3. ‘difficult’ is new.

‘to’ we have seen before, so it goes from 1 to 2. ‘get’ is new, ‘into’ is new. ‘the’ we have seen before, so it goes from 2 to 3. ‘Monday’ comes now for the third time, so ‘Monday’ goes from 2 to 3. ‘mood’ is new. ‘of’ we have seen before, so ‘of’ goes from 1 to 2. ‘work’ is new. ‘and’ has been seen before. So it goes from 1 to 2. ‘discipline’ is new. ‘He’ again, we have seen before.

This time, we have actually seen it twice with a capital H. As we said, it does not matter whether it is capital or small. But ‘He’ goes from 1 to 2.

‘shuddered’ is new. ‘at’ is new. ‘the’ is old, so it goes from 3 to 4. ‘very’ is new. ‘thought’ is new. ‘of’ is old, it goes from 2 to 3. ‘school’ is new. ‘that’ is new. ‘dismal’ is new. ‘yellow’ is new. ‘building’ is new. ‘the’ is old, it goes from 4 to 5. ‘fire-eyed’, remember we said that ‘fire-eyed’ we will treat as a single word for our purposes. So, ‘fire-eyed’ becomes 1. ‘Vedanayagam’ is a new word, becomes 1. ‘his’ we have seen before, it goes from 1 to 2. ‘class-teacher’, again we are treating as a single word, though it is hyphenated. So it goes from 0 to 1.

‘and’ now goes to 3. ‘the’ goes to 6. ‘Head’ is 1. ‘Master’ is 1. ‘with’ is 1. ‘his’ was already 2, so it goes to 3. ‘thin’ is 0, goes to 1. ‘long’ is new, goes to 1 and finally, ‘cane’ is new and goes to 1. So, we have something like almost fifty distinct words in this paragraph. So, overall this paragraph has something like sixty-five or sixty-six words. And of those, some are repeated. So, we actually have fifty distinct words and of course, if you go to a longer piece of text, the number of words in the text could be even more.

So, we have to figure out in the long run, how to keep these counts in an efficient way, but right now, the important thing to notice is that you actually need to keep a variable number of count variables. And we have to name each one so that we recognize which one we have to increment. And as we go along, we create them as and when we need them. So, we cannot say in advance that we will need twelve or fifteen or fifty or sixty. We have to create them as we go along and attach them in some way.

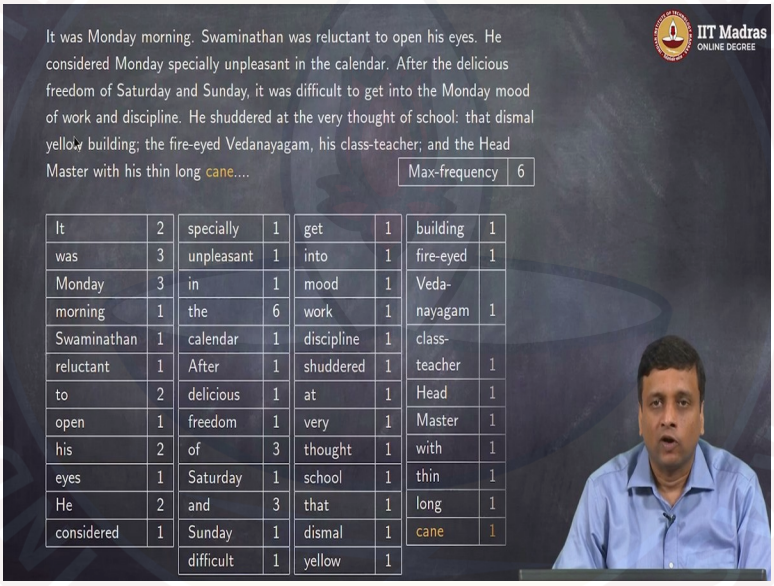
Here we have attached them by just using the same name for the variable as the word itself, but we need to find out later on, how to do this a little more effectively. So, one of the things we would like to do with this word frequency count is to understand which word appears the most frequently amongst these. So, the way we would have to do it now, given that we have about forty-eight words and forty-eight counts, is to go through each of these counts and keep track of the maximum.

So, we have to a second iteration which is almost as long as the first iteration. Remember there were around sixty-five, sixty-six words. And then we would discover as we go along that ‘the’ with 6, of course here we can see it, but remember that a computational process cannot see the

input the way we think we can see it when we visualize it. And imagine there were not sixty-six words but six thousand six hundred words or six million words; even we would not be able to glance at it and decide.

So at, at this point, the only thing that we could do to find out the maximum frequency word is to go through all these counts again starting from 'It', 'was', 'Monday', 'morning' and so on and keep track of the maximum number that appears among the counts. But we saw that we could try to combine these two things. We could try to combine keeping track of the frequency as it is coming up and the currently available maximum count as it is coming up. So, let us try and do that with this dataset.

(Refer Slide Time: 09:38)



So, we start again with our paragraph, the same paragraph as before. So, now we are going to keep, in addition to the count for every word and its frequency, we are also going to keep track of the maximum frequency we have seen so far. So initially the maximum frequency that we have seen is 0 and the first word we see is 'It' and its count is 0. But when 'It' becomes 1 after seeing 'It', we compare it to the maximum frequency and we increment the (max) maximum frequency also to 1.

The next word is 'was', which is again a new counter as far as words go. But since its frequency is 1 and the maximum frequency is already 1, we do not have to make any update. So, we go

along as before. So, we create a new counter for ‘Monday’. Again, its frequency is the same as the maximum, so there is no change. We create a new one for ‘morning’, again no change in the maximum, ‘Swaminathan’, no change in the maximum.

Now, we come to ‘was’. So, ‘was’ is a count that is already created. So we increment the count from 1 to 2. And now when we compare to the maximum frequency, we find that ‘was’ has a higher value than the current maximum frequency. So, we also update the maximum frequency to 2 to indicate that overall, the maximum frequency word we have seen has occurred two times. Notice that we do not know which word it is. Okay? We only know that some word has appeared two times.

So, we keep going. So, ‘reluctant’ is 0 to 1, ‘to’ is 1, ‘open’ is 1, ‘his’ is 1, ‘eyes’ is 1, ‘He’ is 1, ‘considered’ is 1. ‘Monday’ goes from 1 to 2. But since we already have a word of maximum frequency 2, we do not have to update the maximum frequency. We just leave it as 2. ‘specially’ is 1, ‘unpleasant’ is 1, ‘in’ is 1. So, for all of these, the count is well below the maximum frequency. So, we do not make any update. ‘the’ is 1, ‘calendar’ is 1, ‘After’ is 1. ‘the’ becomes 2, but again 2 is the same as the (max) maximum frequency currently, so no update. ‘delicious’ is 1, ‘freedom’ is 1, ‘of’ is 1, ‘Saturday’ is 1.

‘and is 1, ‘Sunday’ is 1. ‘It’ is 2, but again 2 is the same as the maximum frequency, so no change. Now, we come to ‘was’ again. So, ‘was’ is currently 2. Now we have seen it for the third time and since it is 3, we now see that we have exceeded the maximum frequency seen so far. So, we make the maximum frequency also 3. So, now we have seen some word three times. We do not know which one. Of course, we look at it as we go along, we know it is ‘was’, but if we needed to know which word it was, we should have recorded that separately.

We are not doing that. We are just keeping track of the number of times we saw the most frequent word. So continuing, we see ‘difficult’ 1, ‘to’ becomes 2, but that is less than 3, ‘get’ is 1, ‘into’ is 1. ‘the’ is now 3 which is the same as the maximum frequency, so no change. ‘Monday’ is 3, again the same as the maximum frequency, so no change. ‘mood’ is 1. ‘of’ is now 2, ‘work’ is 1, ‘and’ is 2, ‘discipline’ is 1, ‘He’ is 2, ‘shuddered’ is 1, ‘at’ is 1.

Now, ‘the’ appears for the fourth time. So, we increment the count of ‘the’ to 4. And this now is bigger than the maximum frequency. So, the maximum frequency goes from 3 to 4. ‘very’ is 1,

‘thought’ is 1. ‘of’ is 3, but 3 is less than the maximum frequency 4, so no change. ‘school’ is 1, ‘that’ is 1, ‘dismal’ is 1, ‘yellow’ is 1, ‘building’ is 1. ‘the’ comes again. Now, ‘the’ goes from 4 to 5 and so the maximum frequency is 4 and since ‘the’ has appeared 5 times, we have to update 4 to 5. So, now we know that some word has appeared 5 times.

‘fire-eyed’ is 1, ‘Vedanayagam’ is 1, ‘his’ is 2, ‘class-teacher’ is 1, ‘and’ is 3. Now, ‘the’ appears again. So, ‘the’ now moves from 5 to 6. 6 is higher than the maximum frequency. So, we update the maximum frequency to 6. ‘Head’ is 1, ‘Master’ is 1, ‘with’ is 1, ‘his’ is now 3, ‘thin’ is 1, ‘long’ is 1 and ‘cane’ is 1. So, we have essentially recreated the same thing we did earlier, which is to keep track of the frequency of every individual word in this paragraph, with the additional information that some word in this list appeared 6 times.

So, we have simultaneously computed the maximum frequency of the maximum, most frequent occurring word, of course, importantly, without knowing which word it is, although we could have kept track, so we could have also kept a second variable which said which is the most frequent word and initially it was ‘was’. If you remember, up to step 3, it was ‘was’ and then it changed to ‘the’. But right now, we have been able to keep track of this and the important thing is, in a single iteration, without going back and iterating through all the counts again, we have not only kept the frequency of every word, we have also kept the maximum frequency.

And likewise, you can imagine, you can keep the minimum frequency and so on. And of course, if you keep minimum frequency, in this list you can see there are several words with frequency 1, so if you want to know which word is of minimum frequency, you might have to keep track of a lot of information. So, these are all different things that we will tackle as we go along.