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 without losing information and applications of frequency count

So we saw how to get the maximum marks from this stack of cards, but when we did it the first time, what we had done was, we had kept aside the maximum marks as we were going along.

Professor G. Venkatesh: The card with the maximum marks.

Professor Madhavan Mukund: The card with the maximum marks. And later on we said we do not want to do that. We do not want to keep aside a card.

(Refer Slide Time: 00:35)

We just did this iteration where we kept track of the maximum marks. But now I realize that we have lost some information which we had earlier when we kept the card aside with the maximum marks. Then we actually knew at the end who had the maximum marks. This way, we just know that the maximum marks is say, 280 or something but we really have no idea at the end who got this marks.

So, so I think we need to somehow be able to do that also in our iterations. So…

Professor G. Venkatesh: So, is there a way to keep, so what you want is that we want to not only keep track of the maximum marks, but we want to keep track of that person or card which had, where we found that maximum mark, right? We want to keep track of that.

Professor Madhavan Mukund: Exactly and it is enough if we know this card number like 0 or…

Professor G. Venkatesh: It should be enough to know the card, in fact actually it is better to, because the name might be the same on many cards, it is better to keep track of the card number where we saw the maximum mark.

Professor Madhavan Mukund: So then what we will do is every time you see a new card, we will compare two things. We will compare the marks and if we update the maximum, then we will also update the …

Professor G. Venkatesh: update simultaneously the card number

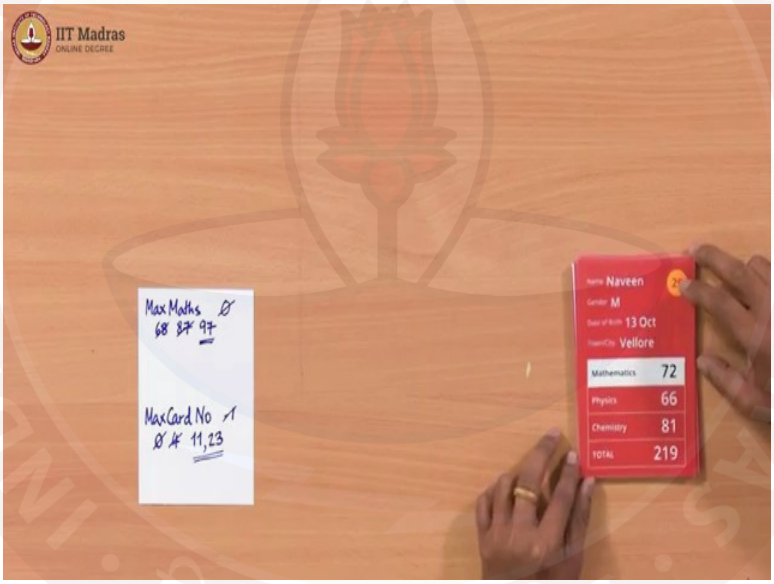
Professor Madhavan Mukund: Card number.

Professor G. Venkatesh: Should we try that?

Professor Madhavan Mukund: Yeah, let us try that.

Professor G. Venkatesh: We will do it for maths?

(Refer Slide Time: 01:47)



Professor Madhavan Mukund: Let us do it just for maths since we have already done for the total, let us do it for maths. Okay.

Professor G. Venkatesh: So here we go, so the name method. So we are iterating for all the cards. Yeah, so we will have… two variables.

Professor Madhavan Mukund: So, we will have this Maths maximum. We will have, I will call it MaxCard Number, right.

Professor G. Venkatesh: Max card number (reference) refers to the card number which has the maximum marks, alright.

Professor Madhavan Mukund: So initially of course the maximum marks is 0 and we have no idea what the card number is. So…

Professor G. Venkatesh: so should we set it at minus 1?

Professor Madhavan Mukund: Say, minus 1 because we know that all cards…

Professor G. Venkatesh: Because it starts at 0. Yeah. So minus 1 should be alright. Minus 1 is not a number is there would not be there at any card. Okay, so first card is 68.

Professor Madhavan Mukund: Yeah and since we have no card so far, we will make this the current maximum and this is card 0. So, we just keep track of it. So what this says is that, so far among the cards we have seen, card 0 is the maximum one and it has 68 marks

Professor G. Venkatesh: 68, alright. So, second one is 62.

Professor Madhavan Mukund: okay that is less, so we do not have to update.

Professor G. Venkatesh: So we do not update the MaxMarks, nor do we…

Professor Madhavan Mukund: We do not update the marks, we do not update the card number also.

Professor G. Venkatesh: So next one is 57.

Professor Madhavan Mukund: Again, nothing to change. 42. Nothing to change. 87. So, 87, we now update 68 to 87.

Professor G. Venkatesh: Now we have to also. So this is card number 4.

Professor Madhavan Mukund: Now we have a new candidate. Okay.

Professor G. Venkatesh: 71.

Professor Madhavan Mukund: 71 is smaller.

Professor G. Venkatesh: 81.

Professor Madhavan Mukund: 81 is also still smaller. Okay.

Professor G. Venkatesh: 84.

Professor Madhavan Mukund: No, still smaller

Professor G. Venkatesh: okay alright. 74.

Professor Madhavan Mukund: No, still smaller.

Professor G. Venkatesh: 63.

Professor Madhavan Mukund: No change.

Professor G. Venkatesh: 64.

Professor Madhavan Mukund: No change.

Professor G. Venkatesh: 97, yeah…

Professor Madhavan Mukund: okay, 97, so we have to say 97 and this is card number 11. So, we have now card number 11. Okay.

Professor G. Venkatesh: 52.

Professor Madhavan Mukund: No change.

Professor G. Venkatesh: 65.

Professor Madhavan Mukund: No change.

Professor G. Venkatesh: 89.

Professor Madhavan Mukund: No.

Professor G. Venkatesh: 76.

Professor Madhavan Mukund: No.

Professor G. Venkatesh: 87.

Professor Madhavan Mukund: No.

Professor G. Venkatesh: 62.

Professor Madhavan Mukund: No.

Professor G. Venkatesh: 72.

Professor Madhavan Mukund: No.

Professor G. Venkatesh: 56.

Professor Madhavan Mukund: No.

Professor G. Venkatesh: 93. Still less than 97.

Professor Madhavan Mukund: Still less than 97.

Professor G. Venkatesh: Yeah, 78.

Professor Madhavan Mukund: No.

Professor G. Venkatesh: 62. Oh, here we have another 97.

Professor Madhavan Mukund: So, so now this is tricky because we have the same maximum but we have two different cards which have the maximum.

Professor G. Venkatesh: What do we do?

Professor Madhavan Mukund: So, I think we have to… (both) Maybe make this variable at the bottom, keep not just one card number, maybe a sequence or a set of card numbers. So, we will keep it as 97, but we will this 11, maybe comma 23 to show that there are two cards, yes okay.

Professor G. Venkatesh: 44.

Professor Madhavan Mukund: No.

Professor G. Venkatesh: 87.

Professor Madhavan Mukund: No.

Professor G. Venkatesh: 74.

Professor Madhavan Mukund: No.

Professor G. Venkatesh: 81.

Professor Madhavan Mukund: No.

Professor G. Venkatesh: 74.

Professor Madhavan Mukund: No.

Professor G. Venkatesh: 72.

Professor Madhavan Mukund: Okay.

Professor G. Venkatesh: Okay, so we are done. We do not have any more cards…

Professor Madhavan Mukund: and we have now recreated the information which we had when we were keeping the cards aside explicitly and one interesting thing we have found is that there could be a situation where the quantity we are trying to track, the maximum or the minimum is actually…

Professor G. Venkatesh: So, I presume that when you are doing this, we have comma, 11 comma 23. And then later on, if you found another card which was greater, you would strike out both.

Professor Madhavan Mukund: Yes. So hypothetically if we had got a 99, which was say card number 40, then we would replace 97 by 99 and we would remove both these numbers and put 40 over there. Okay.

Okay, so now we know that we can actually refer back to the, the information which is the maximum, not just, you know, because this is usually how we will require it, right. We will need to go back and say, oh this person supposing we were going to nominate them for a prize or something, we would need to go back and find out who got 97, right.

So this way and we are using the card number so that we do not have to worry about all the other data. Once we get to the right card, we can pick up all the other information, their name, their gender, date of birth.

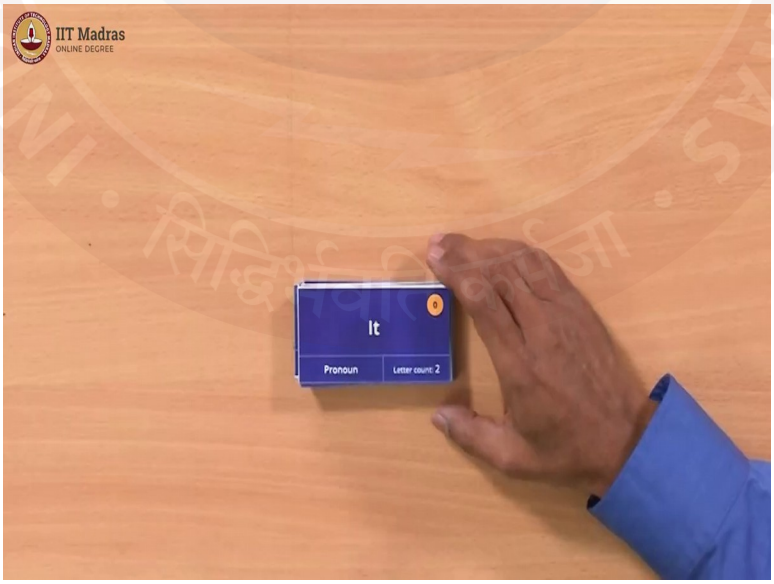
Professor G. Venkatesh: So, this is, this would be very similar to what we had done earlier where we took the card and kept it aside. Yes. Because the equivalent act of keeping the card aside or having one card which is basically is carrying the information the maximum is keeping the index of that card.

Professor Madhavan Mukund: Yes exactly.

Professor G. Venkatesh: Because you can always, when that card you can always take out if you know the index. You can go searching for the index and you can take it out.

Professor Madhavan Mukund: Yeah, but then if we keep it aside, then we have this awkward thing that we have three piles. We have one pile to be seen, one pile which is visited and then we have a third one which is the current maximum. So now, it is a little neater to keep it this way, because then depending on what information we want to keep, we can keep it here. Okay. And not get cluttered with data.

(Refer Slide Time: 06:55)



Professor G. Venkatesh: So, we went through this, this problem that you did with the slides, right? Which is set these cards, these words, the paragraph, you counted how many times each word occurred in the paragraph, kept track of that and also you found the one that occurred the maximum number of times.

But what is the use case for something like that? I mean, why do we need to do something like that? Why do we need to keep track of how many times a word occurs in a paragraph?

Professor Madhavan Mukund: So actually, this turns out to be quite useful when you are searching for, say a document. So we all use search engines when we look for information on the internet, for instance. So, we type out some words about what we want to find. So, supposing we want to find out capital of Italy.

Then there are some words in that question that we are asking are important, like 'capital' and 'Italy'. But the word 'of' is less important. It occurs in many phrases. So, very often, what the search engine is trying to do is match the words that you have asked for with words that appear in the document. But words which appear very frequently will occur in all documents. So they do not serve to separate out documents saying this is different and that is different.

Professor G. Venkatesh: So frequently occurring words should be…

Professor Madhavan Mukund: So, ideally I mean if you are doing it, one simple way to do it is to remove all the frequently, okay. So, you keep only the in, relatively rare words and then the rare words are better ways of separating out documents which are different from those which are similar.

So the words which are similar will have more rare words in common. The ones which are different will have fewer rare words in common and that is one way to match up also queries with responses. So when Google or any other search engine puts up a list of responses, it is actually giving you in some order.

It is not just putting up a list of internet documents which match your query but it is trying to suggest that this is the most relevant, this is the next relevant.

Professor G. Venkatesh: Most relevant is in terms of most number of words.

Professor Madhavan Mukund: Somehow matching…

Professor G. Venkatesh: Matching the number of words. Correct. And in that, it is removing all these words which are most frequently occurring in.

Professor Madhavan Mukund: Correct. So that means that we have to go back to the, say the words in the paragraph and we have to, in some sense, extract or remove. Remove the words which are too frequent to get a reduced set of words and use that as our, even though we might lose the meaning of the (para), may not make sense anymore as a language, just in terms of this word by word comparison, it is helpful to do that.

Professor G. Venkatesh: Okay. So, in this particular thing, you did the count, right, which is, so should we say that, we found that most of the words occur only once.

Professor Madhavan Mukund: Exactly.

Professor G. Venkatesh: And some words are occurring more than once. So, should we say that things which are occurring more than once are…

Professor Madhavan Mukund: Yeah, so, let us assume that we want to keep track of only those words which occur once, say because that is…

Professor G. Venkatesh: Unique kind of thing.

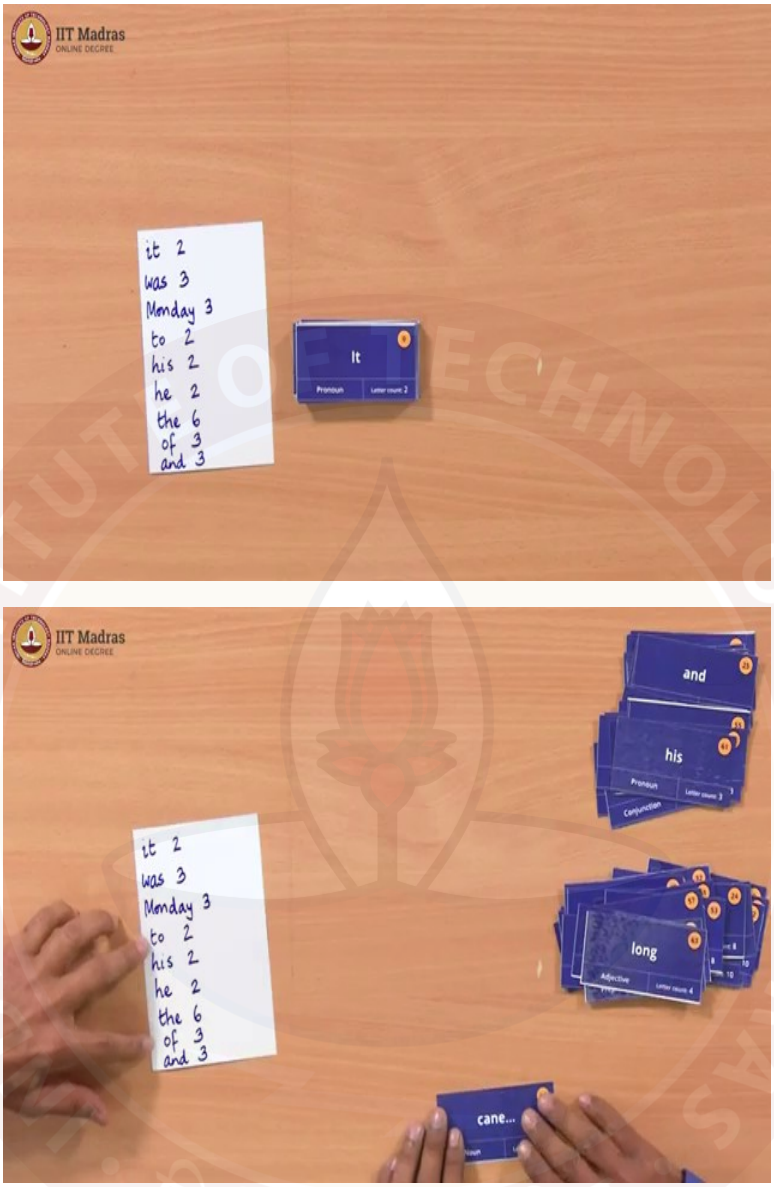
Professor Madhavan Mukund: Unique words and everything which occurs multiple times let us believe that they…

Professor G. Venkatesh: They are frequently occurring.

Professor Madhavan Mukund: They are something which is not very useful for us for calculation.

Professor G. Venkatesh: Okay. So, do we have…

(Refer Slide Time: 09:44)



Professor Madhavan Mukund: Yeah. If we go back to the counts that we had, it turns out that out of the 48 words which were there in this paragraph, there are only some 9 words or so which occurred more than once. Okay. And these are these 9 words, ‘it’, ‘was’, ‘Monday’, ‘to’, ‘his’, ‘he’ ‘the’, ‘of’, ‘and’.

Professor G. Venkatesh: So, we should remove these words?

Professor Madhavan Mukund: Yeah, so what we need to do is go through this list, now that we have, so we went through this list once to create these counts. And now we have to go through

the list a second time, keeping these counts in mind to remove the words that appear in the smaller list of 9. So, we do not want these words. We only want words which are not in this.

Professor G. Venkatesh: So, it is there are 2 iterations. One iteration basically the first pass was to create the counts. Yes. And then the second iteration we are doing…

Professor Madhavan Mukund: is a kind of filtering…

Professor G. Venkatesh: is a filtering…

Professor Madhavan Mukund: where our filtering condition is that the word should, the word will be retained if it is not in this list. It will be moved into a discarded pile if it is in this list. But we are not actually counting or accumulating. We are just separating. We are segregating the words into two groups.

Professor G. Venkatesh: So, should we, you want to try that, I mean take this set of words. (Yeah) Ofcourse, there is this problem, right, which is that, 'it' what you have written here is small letters whereas…

Professor Madhavan Mukund: Correct, correct, correct. We did… So that we had decided that we will ignore, I mean so it is a decision that we have to make whether we are. So, if we had counted 'it' with a small letter and 'It' with a capital separately, they would have each had a count of 1. And then we would have actually kept it. Yeah. But since we decided when counting that we will ignore the capitals, then here also we should assume that capital 'It' is the same as.

Professor G. Venkatesh: So the filtering is a little bit more complex, not just filtering for 'it', all small letters, you are filtering for 'it', whether it is small letters or capital letters.

Professor Madhavan Mukund: So we have to take the word on the card and in some sense, remove all the capital letters or make all the capitals to small. So, make it into lower case. Yeah, right. So, let us…

Professor G. Venkatesh: and you might find some cards with some full-stop and all.

Professor Madhavan Mukund: Exactly. We also have to put the punctuation marks which is not relevant to the word itself. So that also we have to ignore it. Okay.

Professor G. Venkatesh: So, let us go through that. Let us, so we are going to basically take card by card, move it to another file. Same…

Professor Madhavan Mukund: So we have two piles. So, one pile of cards we keep and one pile of cards we discard (okay) and ofcourse the pile which have we have not seen so…

Professor G. Venkatesh: Alright. So, the first card is 'It'. So, 'it' is on this list, discard. So we discard it. So I am keeping it separately. So this is the discarded pile, alright. Next is 'was' is also discarded pile. Okay. 'Monday'…

Professor Madhavan Mukund: also to be discarded.

Professor G. Venkatesh: alright so we are discarding the thing. 'Morning'

Professor Madhavan Mukund: So morning has a full stop which we ignored and we keep morning because it is not in the list so therefore we are not counting it.

Professor G. Venkatesh: So here is the list of all the things we are keeping, right? Yeah. This is the list of discarded and this is the list we are keeping. Yeah. 'Swaminathan'?

Professor Madhavan Mukund: Keep it.

Professor G. Venkatesh: Okay, because it is not in the list.

Professor Madhavan Mukund: Not in this, in this frequent list.

Professor G. Venkatesh: Swaminathan starts with a capital S so presume you will first make it all small letters then see whether it is in the list.

Professor Madhavan Mukund: Yeah the same with 'Monday' also, although it turns out that all Mondays …

Professor G. Venkatesh: will always start with a capital. 'Was'?

Professor Madhavan Mukund: To be discarded.

Professor G. Venkatesh: Okay, so it goes here. 'Reluctant'?

Professor Madhavan Mukund: Keep it.

Professor G. Venkatesh: Okay. ‘Two’?

Professor Madhavan Mukund: Discard it. Okay. Because it appears in this list.

Professor G. Venkatesh: ‘Open’?

Professor Madhavan Mukund: Keep it.

Professor G. Venkatesh: Oh, okay. ‘His’?

Professor Madhavan Mukund: ‘His’ appears twice, so we discard it.

Professor G. Venkatesh: Alright. ‘Eyes’?

Professor Madhavan Mukund: Keep it.

Professor G. Venkatesh: So, here again eyes has full stop in it which we should ignore. And then we keep ‘eyes’. Yeah. ‘He’?

Professor Madhavan Mukund: Again it is a capital ‘He’, but ‘he’ with or without capital…

Professor G. Venkatesh: So here we are matching it, even though it is starting with a capital H because we first convert it into small letters (So discard it). Discard it.

Professor Madhavan Mukund: ‘Considered’ is not here on this list. Keep it.

Professor G. Venkatesh: ‘Monday’?

Professor Madhavan Mukund: ‘Monday’ is a frequent word, 3 times.

Professor G. Venkatesh: Discard. ‘Specially’?

Professor Madhavan Mukund: Yes, keep it.

Professor G. Venkatesh: ‘Unpleasant’?

Professor Madhavan Mukund: Keep it. ‘in’?

Professor G. Venkatesh: ‘in’

Professor Madhavan Mukund: surprisingly appears only once even though it is a common word, in this paragraph.

Professor G. Venkatesh: ‘the’ should definitely, definitely we discarded. ‘calendar’?

Professor Madhavan Mukund: ‘calendar’ keep it.

Professor G. Venkatesh: ‘After’?

Professor Madhavan Mukund: Keep it.

Professor G. Venkatesh: ‘the’? Gone, right.

Professor Madhavan Mukund: Discard it.

Professor G. Venkatesh: ‘delicious’?

Professor Madhavan Mukund: Keep.

Professor G. Venkatesh: ‘freedom’ keep. ‘of’?

Professor Madhavan Mukund: ‘of’ discard.

Professor G. Venkatesh: Discard, right, yeah. ‘Saturday’?

Professor Madhavan Mukund: Yes, we keep Saturday.

Professor G. Venkatesh: okay ‘and’?

Professor Madhavan Mukund: ‘and’ is too frequent.

Professor G. Venkatesh: Okay. ‘Sunday’?

Professor Madhavan Mukund: Keep it.

Professor G. Venkatesh: ‘It’?

Professor Madhavan Mukund: ‘it’ goes away.

Professor G. Venkatesh: ‘was’?

Professor Madhavan Mukund: also goes away.

Professor G. Venkatesh: ‘difficult’?

Professor Madhavan Mukund: Yes we keep difficult.

Professor G. Venkatesh: ‘to’?

Professor Madhavan Mukund: Goes away.

Professor G. Venkatesh: Goes away. ‘get’?

Professor Madhavan Mukund: ‘get’ we keep.

Professor G. Venkatesh: Okay ‘into’?

Professor Madhavan Mukund: ‘into’ we keep.

Professor G. Venkatesh: ‘the’ is gone ‘Monday’?

Professor Madhavan Mukund: ‘Monday’ is gone.

Professor G. Venkatesh: ‘mood’?

Professor Madhavan Mukund; ‘mood’ we keep.

Professor G. Venkatesh: ‘of’?

Professor Madhavan Mukund: No, gone.

Professor G. Venkatesh: ‘work’?

Professor Madhavan Mukund: We keep.

Professor G. Venkatesh: ‘and’ is gone. ‘discipline’?

Professor Madhavan Mukund: Yes.

Professor G. Venkatesh: ‘He’?

Professor Madhavan Mukund: ‘He’ is gone.

Professor G. Venkatesh: ‘shuddered’?

Professor Madhavan Mukund: ‘shuddered’ we keep.

Professor G. Venkatesh: ‘at’?

Professor Madhavan Mukund: Yes.

Professor G. Venkatesh: oh surprising. ‘the’?

Professor Madhavan Mukund: ‘the’ is gone.

Professor G. Venkatesh: ‘very’?

Professor Madhavan Mukund: ‘very’, yes.

Professor G. Venkatesh: ‘thought’?

Professor Madhavan Mukund: yes.

Professor G. Venkatesh: ‘of’ is gone. ‘school’?

Professor Madhavan Mukund: Again here we dropped the punctuation mark. Yes. ‘school’, we keep it.

Professor G. Venkatesh: ‘the’?

Professor Madhavan Mukund: Gone.

Professor G. Venkatesh: ‘dismal’?

Professor Madhavan Mukund: ‘dismal’, yes.

Professor G. Venkatesh: ‘yellow’?

Professor Madhavan Mukund: Keep it.

Professor G. Venkatesh: ‘building’ with a punctuation mark is gone. No, no. The punctuation mark is removed but the word ‘building’ is kept. ‘the’ is gone. ‘fire-eyed’ is…

Professor Madhavan Mukund: kept as a single word.

Professor G. Venkatesh: ‘Vedanayagam’?

Professor Madhavan Mukund: again ignoring the comma, we keep it.

Professor G. Venkatesh: And the capital letter (yeah) ‘his’?

Professor Madhavan Mukund: ‘his’ is out.

Professor G. Venkatesh: ‘class’.

Professor Madhavan Mukund: Keep it.

Professor G. Venkatesh: ‘teacher’.

Professor Madhavan Mukund: Keep it.

Professor G. Venkatesh: ‘and’?

Professor Madhavan Mukund: ‘and’ we remove.

Professor G. Venkatesh: ‘headmaster’?

Professor Madhavan Mukund: Keep it.

Professor G. Venkatesh: ‘with’?

Professor Madhavan Mukund: ‘with’ we keep.

Professor G. Venkatesh: Surprising, ‘his’?

Professor Madhavan Mukund: ‘his’ we drop.

Professor G. Venkatesh: ‘thin’?

Professor Madhavan Mukund: ‘thin’ we keep.

Professor G. Venkatesh: ‘long’?

Professor Madhavan Mukund: ‘long’ we keep.

Professor G. Venkatesh: And ‘cane’ we keep it, alright. So these are the words. Correct. So now we have a subset of words…

Professor Madhavan Mukund: subset of words which occur, in our case exactly once. So, depending on the degree of the length of the thing, you might set the threshold not at 1 or 2, you might keep words which occur once or twice but not three times or more. So whatever that cut off is, we can go through but the thing is we need to go through the words a second time...

Professor G. Venkatesh: Second round, yeah.

Professor Madhavan Mukund: So that is something which we have not seen so far.

Professor G. Venkatesh: So this is two iterations. First iteration basically is to find the frequency count, second iteration to use that frequency count, right.

Professor Madhavan Mukund: In order to do some kind of filtering, in this case, separation into two groups right. So that is important. So sometimes we need to go through the data more than once because we need to take the result we have created out of the first pass and use it as an input to the second pass. So, we cannot do that until we have finished.

Because we thought that maximum we first started, we had to do that, we realized that we could keep a track maximum while we were counting. But here, until we have the final counts, we cannot do this filtering. So, we do not know whether to discard a word until we have seen how many times it occurs across the entire paragraph.