UNIVERSITÀ DEGLI STUDI DI UDINE

Dipartimento di Scienze Matematiche, Informatiche e Fisiche



RELAZIONE PROGETTO DI ALGORITMI E STRUTTURE DATI

A cura di:

* Raffaele Marseu, matricola 133879
* Maksym Sernyak, matricola 143087
* Danjel Stefani, matricola 142863
* Marco Venir, matricola 144762

Anno Accademico 2019/2020

Indice

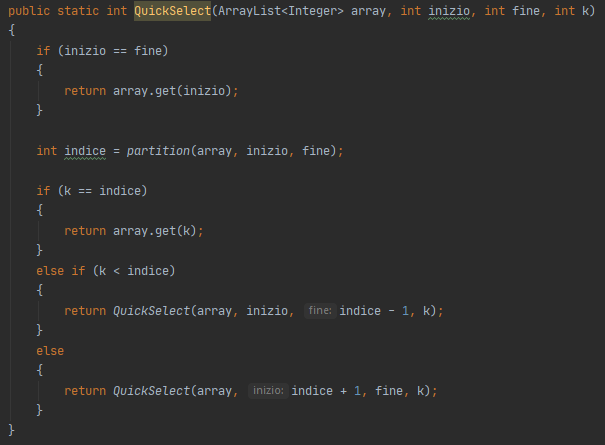
1. Introduzione Pag. 1
2. Presentazione algoritmi Pag. 1
   1. QuickSelect Pag. 1
   2. HeapSelect Pag. 2
   3. Median of Medians Select Pag. 4
3. La funzione “*Creatore*” Pag. 6
4. Il metodo per misurare il tempo Pag. 6
5. Presentazione dei risultati e considerazioni Pag. 12

1. Introduzione

Abbiamo realizzato i programmi in Java sul programma Intellij, come ambiente di sviluppo e di run dei programmi. Abbiamo usato un computer con uno Xeon W3680 (3.33 GHz di base, 3.60 GHz Intel TurboBoost 1.0) con 6 core e 12 threads, equipaggiato con 20gb di ram DDR3 1333 MHz di velocita’. Il sistema operativo usato e’ Windows 10 Pro. Come JDK abbiamo usato la versione 14.0.0.

2. Presentazione algoritmi

2.1. QuickSelect

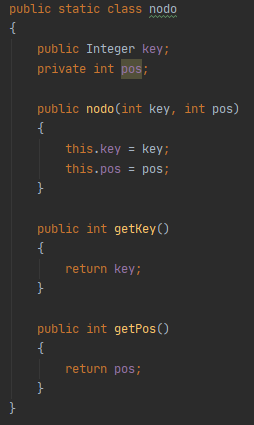
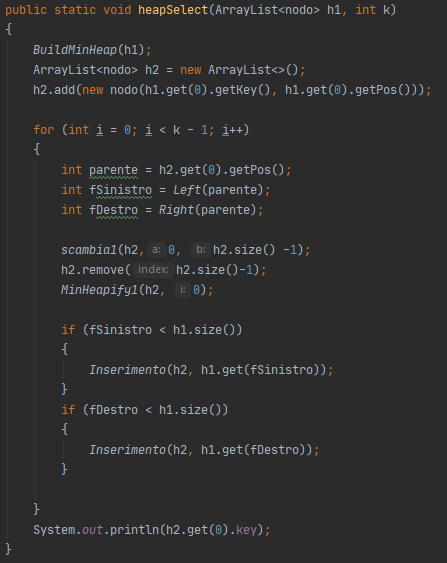


Nel codice di QuickSelect controlliamo se è presente un solo elemento e in caso affermativo lo ritorniamo. *“partition”* (codice omesso, perché da manuale) divide l’array in varie parti e nella variabile “indice” salva il valore dell’elemento di pivot, il quale viene confrontato con k. Se il valore di entrambe le variabili coincide, ritorna l’elemento che si trova in posizione di k.

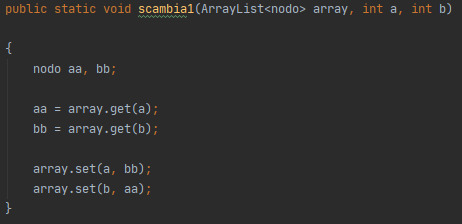
In seguito, controlliamo se l’indice trovato con partition e’ maggiore di k; se si, richiamiamo Quickselect con la variabile “fine” uguale a “indice – 1”. Altrimenti richiamiamo Quickselect con la variabile “inizio” uguale a “indice + 1”.

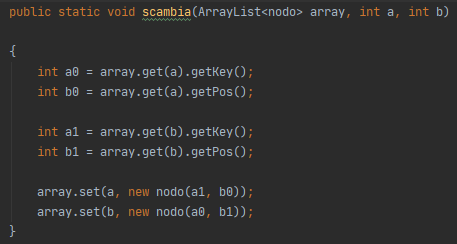
2.2. Heap Select

Per implementare questo algoritmo abbiamo utilizzato la classe “*nodo*” per memorizzare la posizione e il valore di ogni elemento nella heap in modo da rendere unico ogni nodo ed evitare problemi di ricerca (ad esempio in caso di più nodi con valore uguale).

“HeapSelect”, all’inizio costruisce la heap “h1” a partire dall’array fornito con  BuildMinHeap (codice omesso, perché da manuale). Dopodiché si crea la heap “h2” inizialmente vuota. In essa viene inserito il primo elemento della heap “*h1”*. Durante ogni ciclo *for*, si calcolano le variabili *“figlio sinistro”* e *“figlio destro”* (codici omessi, perché da manuale). In seguito, si scambiano il primo e l’ultimo nodo (con il metodo *“scambia1”*) e si rimuove l’ultimo nodo dalla heap. A questo punto, se esistono i figli, si salvano nella heap *“h2”.*

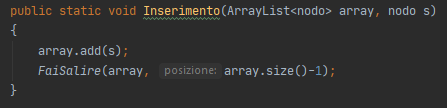




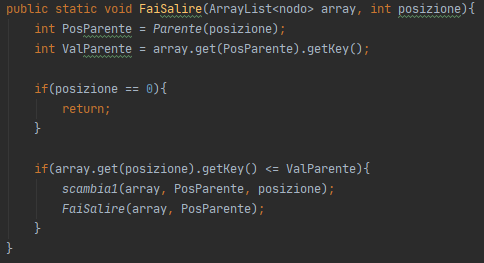
La funzione *“scambia1*” scambia semplicemente i due nodi all’interno della heap *“h2”* mentre *“scambia”*, scambia le variabili *“key”* che rappresentano i valori dei nodi. Tale funzione viene utilizzata soltanto durante la costruzione della heap *“h1”*.

*“MinHeapify1”* è identico a “*MinHeapify”* da manuale. L’unica differenza tra i due è che “*Minheapify1”* usa “*scambia1”* nel codice.

Per caricare i figli nella heap *“h2”* viene usata la funzione “*Inserimento*”.

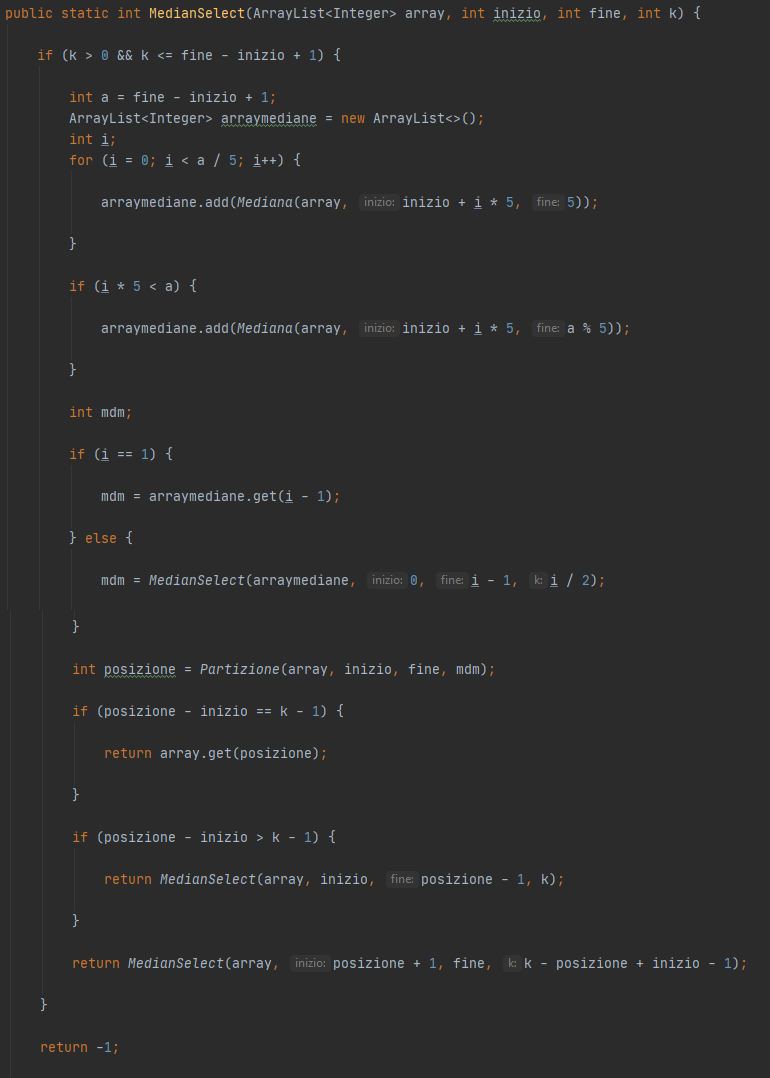


Il metodo “*Inserimento*” richiama la funzione “*FaiSalire*”. Questa funzione, controlla se il valore dell’elemento è minore del valore del genitore e in tal caso scambia i due nodi. Ripete ricorsivamente *“FaiSalire”* dalla posizione del genitore fino a quando il valore non si trova nella posizione giusta, ovvero è più grande del proprio genitore oppure è nella prima posizione.



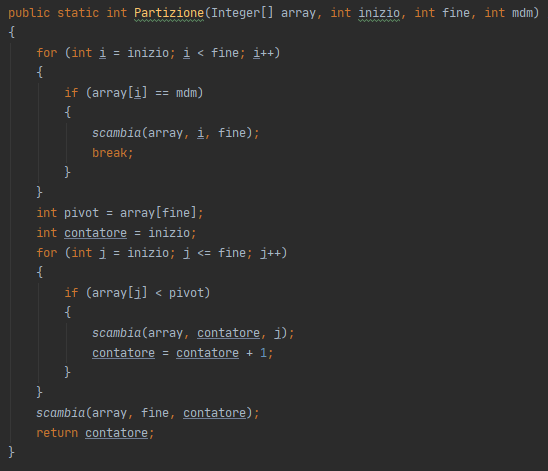
2.3. Median of Medians Select

Questo algoritmo sfrutta la mediana delle mediane per trovare l’elemento richiesto.



In primo luogo, confronta il valore di k con la dimensione dell’array che contiene le mediane. Nella condizione usiamo “fine - inizio + 1” in quanto indica il numero di elementi nell’array che parte da zero. Il ciclo *for* calcola le posizioni facendo una suddivisione nell’array originale in gruppi da 5 elementi. La suddivisione dell’array potrebbe non essere perfetta in quanto potrebbe non contenere 5 elementi e la mediana di questi viene calcolata con un if a parte. Dopodiché inizializziamo la variabile *“mdm”* che serve per memorizzare il valore della mediana delle mediane. La variabile *“mdm”* è inizializzata separatamente, come la variabile *“i”*, in quanto le utilizzeremo in seguito.

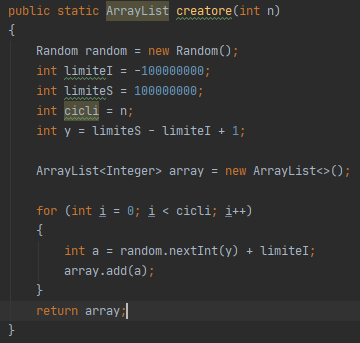
La variabile posizione viene calcolata dalla funzione “*Partizione*” (partition modificata). Se posizione - inizio corrisponde a k-1, viene ritornato il valore che si trova nella posizione *“posizione”* dell’array. Se invece è maggiore, facciamo una ricorsione della funzione con la variabile *“fine”* modificata con il valore “posizione - 1”. Nel caso nessuna delle due condizioni venga soddisfatta, la funzione esegue una ricorsione con la variabile “inizio” modificata con il valore “posizione + 1” e la variabile *“k”* modificata con il valore “k - posizione + inizio - 1”.



Nella funzione *“Partizione*” il for iniziale sposta la variabile *“mdm”* in posizione *“fine”*. Il pivot corrisponde alla nostra variabile *“mdm”* e viene eseguita una normale funzione di partizione (identica a quella di QuickSelect). Il codice della funzione “*scambia*” è omesso, data la semplicità del codice in quanto scambia gli elementi, date le posizioni nell’array.

3. La funzione “*Creatore*”

In questa funzione utilizziamo il “*Random*” predisposto dalla libreria di Java. Impostiamo i limiti inferiore e superiore che rappresentano il minimo e massimo numero che il *“Random”* può selezionare. Per fare in modo che possano essere presenti anche numeri negativi, utilizziamo la variabile *“a*” in modo che l’intervallo di scelta sia *“y”*. Nella variabile *“y”*, il “+1” viene aggiunto, perché la funzione “nextInt()” non considera l’ultimo elemento del range. Sottraendo il limite inferiore, sarà possibile avere anche numeri negativi nell’array.



Ad esempio:

Limite Inferiore = -100;

Limite superiore = 100;

Range Y = Limite Superiore - Limite Inferiore + 1 = 100 + (-100) + 1 = 201;

Random Y = 10;

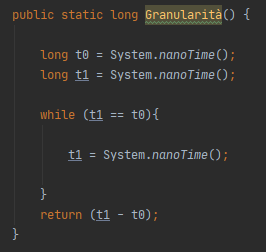
Numero finale = Random Y + Limite inferiore = 10 + (-100) = -90;

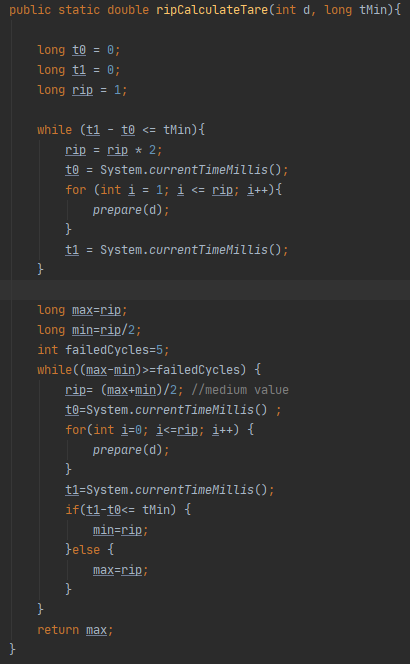
Nell’implementazione di tutti gli algoritmi, abbiamo utilizzato un range che va da

-100.000.000 a 100.000.000.

4. Il metodo per misurare il tempo

Per granularità (o risoluzione) si intende l’intervallo di misurazione necessario al sistema per caricare una singola istruzione, in questo caso noi carichiamo e tentiamo di ottenere il tempo minimo registrabile e ritorniamo la differenza tra i due risultati.



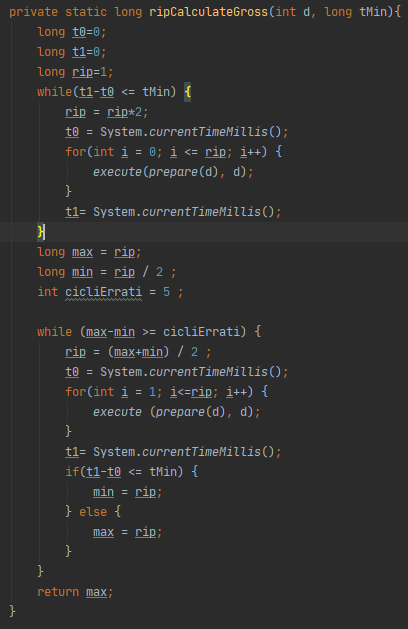


In “*ripCalculateTare*” calcoliamo il numero di ripetizioni minimo per far si che l’esecuzione di *“prepare”* sia registrato come tempo di preparazione e venga registrato come dato per conoscere il tempo per il caricamento dati tenendo conto anche dei cicli errati. Il tempo di *“prepare”* viene registrato per poterlo sottrarre al tempo totale ed avere il tempo di esecuzione vero e proprio dell’algoritmo.

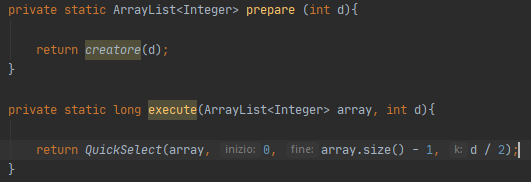
Nel primo while verifichiamo ogni quante ripetizioni la differenza dei due tempi (t0 e t1) continua ad essere minore del tMin (ovvero la granularità fratto la percentuale di errore dei dati, nel nostro caso 1%). La crescita di *“rip”* è esponenziale. Nel secondo while invece, considerando le ripetizioni, troviamo il tempo minore senza cicli falliti tramite il metodo numerico di bisezione.

*“ripCalculateGross”* calcola il tempo di esecuzione del programma in totale, compreso del tempo di esecuzione dell’algoritmo, il tempo per il caricamento dei dati e come per *“ripCalculateTare”*, vengono tenuti in considerazione anche i cicli errati che possono

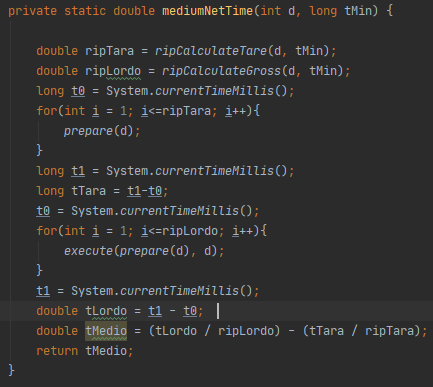
essere presenti. In questo caso abbiamo inserito un numero di al più 5 cicli errati.



La stessa metodologia viene usata sia in *“ripCalculateTare”* sia in *“ripCalculateGross”*. Nella prima viene considerato il tempo di esecuzione della preparazione dei dati che andranno inseriti nell’algoritmo, mentre in *“ripCalculateGross”* viene considerato tutto il processo, ovvero dalla creazione dei dati allo svolgimento dell’algoritmo.

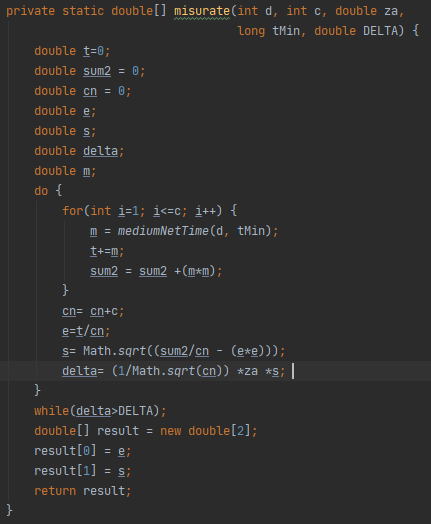


La funzione *“prepare”* inizializza l’algoritmo che successivamente andrà’ ad inserire i dati nell’array per lo svolgimento dell’algoritmo. *“execute*” invece esegue effettivamente l’algoritmo, senza contare come siano stati prodotti i dati in input.



In questa parte implementiamo il calcolo del tempo caricando il valore minimo (esecuzione della parte di caricamento dati) e il valore massimo, ovvero sfruttiamo il tempo impiegato dall’algoritmo per calcolare i risultati finali, prendendo in input i dati creati randomicamente.

In “mediumNetTime”, usiamo i vari metodi già realizzati (*“RipCalculateGross”*, *“ripCalculateTare”*, *“prepare”* ed *“execute”)*. Con il primo for controlliamo nuovamente il tempo di preparazione dei dati mentre nel secondo for, calcoliamo i tempi dell’esecuzione dell’algoritmo. Tutto ciò serve per calcolare il tempo medio sottraendo i tempi messi a rapporto con i valori precedentemente calcolati.

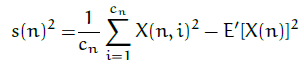


*“Misurate”* è la parte più importante dell’intero nucleo dell’algoritmo per la misurazione dei tempi, poiché vengono caricate tutte le variabili necessarie per inserire i valori delle ripetizioni, del *“delta”* e dell’errore della ripartizione della gaussiana (za). Tutto ciò serve proprio per collocare in un unico punto tutti gli input necessari alla ricerca dei tempi.

Con il do-while verifichiamo quando l’errore, ovvero la variabile *“DELTA”*, è maggiore del valore calcolato (variabile *“delta”*). “*delta*” è il parametro che serve a calcolare l’intervallo di confidenza. Se i valori superano l’1%, viene caricato il valore successivo. Nel for invece, viene tenuto il conto dei cicli di misurazioni da effettuare, ovvero tutti i risultati del tempo medio, tramite la funzione *“mediumNetTime”*.

In output sono riportate le variabili *“e”* ed *”s”* che sono rispettivamente la media dei tempi e la deviazione standard.

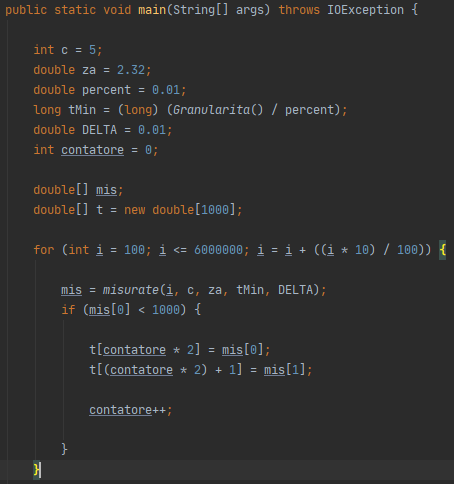
Per calcolare la varianza campionaria abbiamo seguito questa formula. Mettendola sotto radice, troviamo la deviazione standard:



Nella formula, la prima parte (X(n,i)²) della sommatoria è rappresentata dalla variabile *“sum2”,* la seconda parte (dopo il segno meno), è rappresentata dalla variabile *“e”* al quadrato.

Nel main abbiamo inizializzato le variabili principali, tra cui *“c”* che rappresenta il numero di misurazioni dell’algoritmo. Aumentando il valore, il numero di ripetizioni aumenta e di conseguenza migliora la precisione e aumenta il tempo totale di calcolo dei tempi.

Nel for memorizziamo tutti i tempi nell’array “t”. Il “contatore \* 2”, serve per inserire i dati che verranno trascritti sul foglio Excel. I dati verranno caricati con un ordine tale che il primo dei due risultati sia sempre il tempo e il secondo la deviazione standard. Abbiamo quindi inserito questo contatore solo per ordinare gli elementi dell'array “t”.



Una volta che l’algoritmo ha finito, importiamo i dati in output direttamente su Excel. La libreria che abbiamo usato per farlo è “hu.blackbelt.bundles.poi:org.apache.poi:4.0.1\_1”.

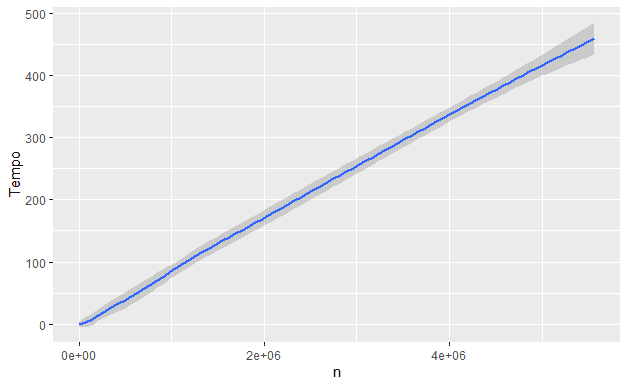


5. Presentazione dei risultati e considerazioni

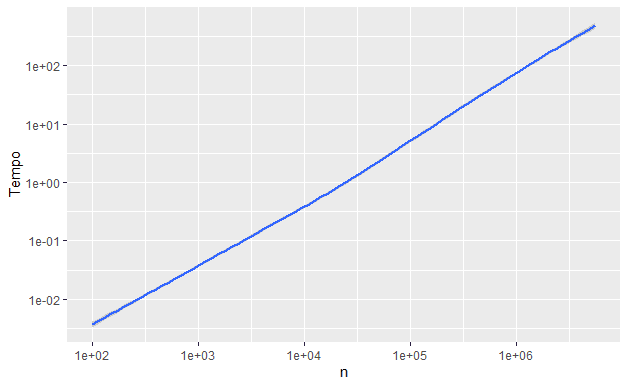
Per fare i grafici abbiamo usato R. Nei grafici seguenti il Tempo è sempre rappresentato in millisecondi e collocato sull'asse y mentre *“n”* rappresenta la grandezza del vettore ed è collocato sull’asse x. I tempi sono stati presi con un aumento del 10% della lunghezza dell’array (il valore successivo è maggiore del 10% rispetto al precedente). Il limite superiore della grandezza del vettore è stato fissato a 6 milioni di elementi.

Nei grafici la linea in azzurro rappresenta regressione lineare dei dati mentre la parte in grigio esprime l’errore standard.

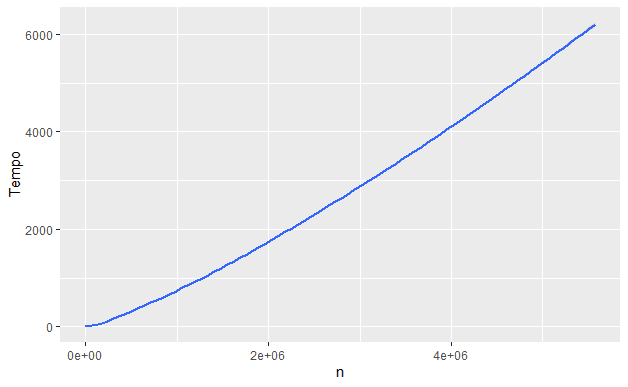
QuickSelect con scala lineare



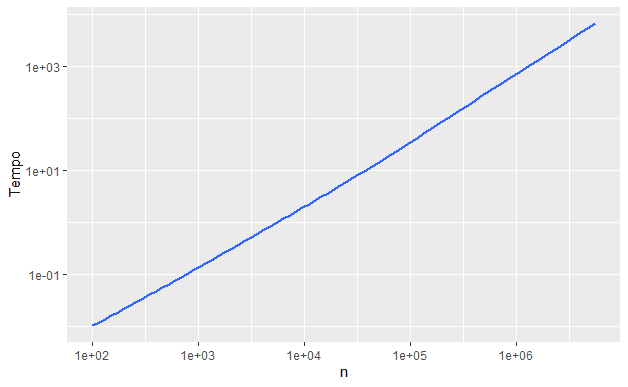
QuickSelect con scala logaritmica



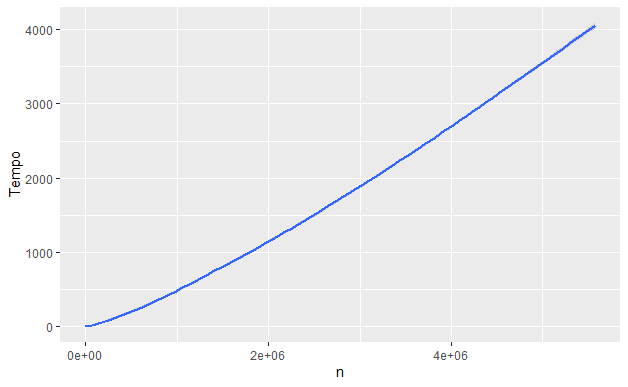
HeapSelect con k =  n / 2 e scala lineare



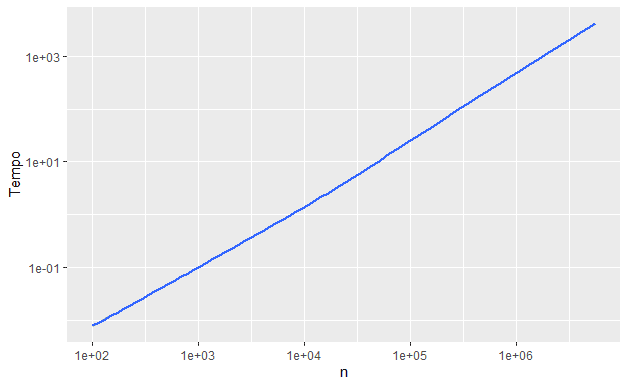
HeapSelect con k =  n / 2 e scala logaritmica



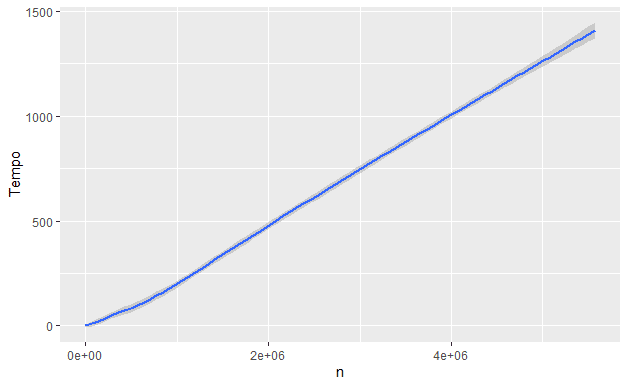
HeapSelect con k = n / 3 e scala lineare



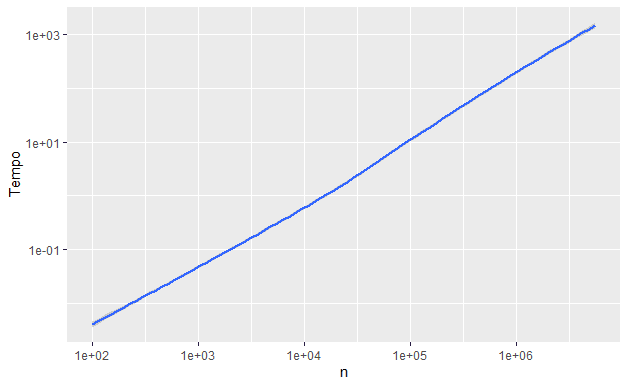
HeapSelect con k = n / 3 con scala logaritmica



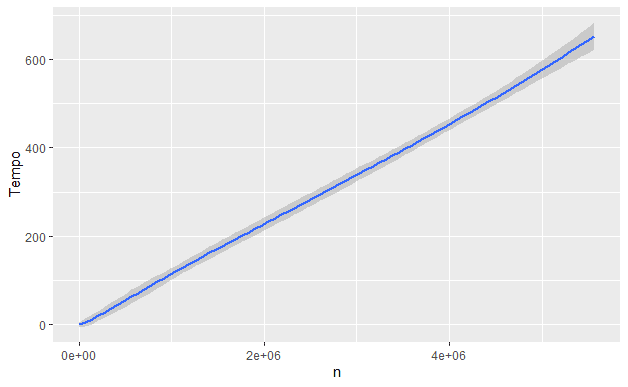
HeapSelect con k = n / 10 e scala lineare



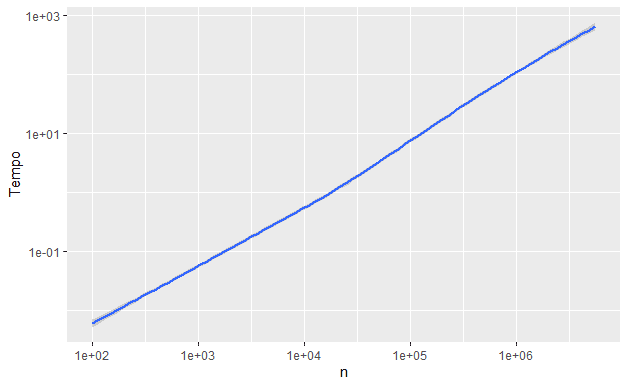
HeapSelect con k = n / 10 con scala logaritmica



Median Of Medians Select con scala lineare



Median Of Medians Select con scala logaritmica



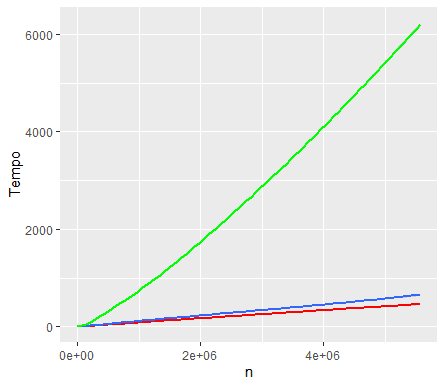
Sostanzialmente i grafici risultano coerenti con le nostre aspettative:

a) QuickSelect ha una complessità temporale di O(n) nel caso medio e di Θ(n²) nel caso peggiore. Dal grafico si evince che abbiamo un andamento lineare, il che soddisfa le nostre aspettative sui tempi.

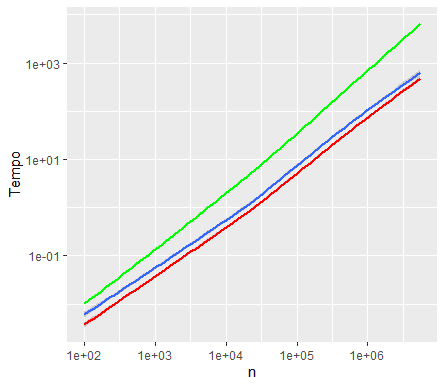
b) HeapSelect ha una complessità temporale di O(n + klogk). I tempi di questo algoritmo variano in base al k. Dai grafici è evidente che all’aumentare di k, aumentano i tempi.

c) Median Of Medians Select ha una complessità temporale di Θ(n). Dal grafico si può osservare che abbiamo un andamento lineare, il che rispetta le nostre aspettative.

Tutti i tre algoritmi in scala lineare



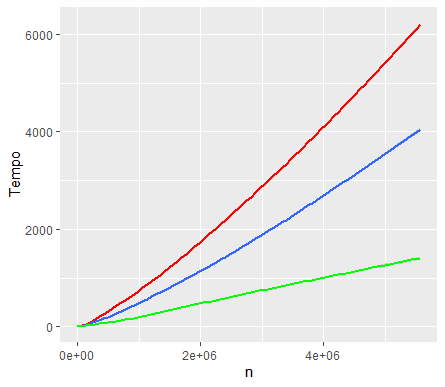
Tre algoritmi su grafico logaritmico



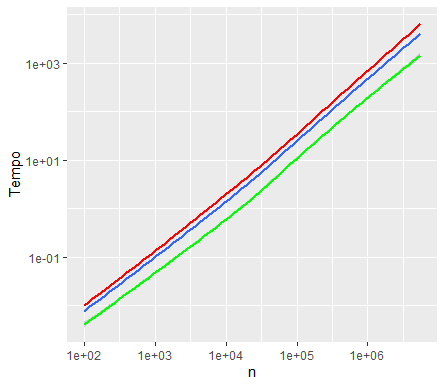
(BLU Median of Medians Select, VERDE HeapSelect, ROSSO QuickSelect)

Come si può vedere l’algoritmo più veloce è QuickSelect mentre quello più lento è HeapSelect. QuickSelect e Median of Medians Select hanno prestazioni simili fino ad n uguale a un milione circa. HeapSelect cresce molto più velocemente degli altri due, perché la creazione della heap impiega un certo tempo per essere costruita. In questo grafico, tutti e tre gli algoritmi sono stati eseguiti con lo stesso k che equivale a n/2. Per questo, dato che il k è in posizione centrale nell’array, i tempi di HeapSelect crescono molto in fretta. Se paragonato ad altri HeapSelect con k più vicini all’inizio dell’array, si nota che questi ultimi sono più veloci.

Tutti gli HeapSelect in scala lineare



Tutti gli HeapSelect su grafico logaritmico



(ROSSO k = n/2, i BLU k = n/3, VERDE k = n/10)