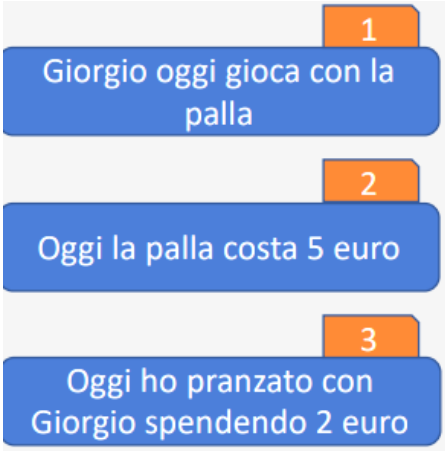
13 - Inverted Index

Un inverted index è una struttura dati in cui si collegano i contenuti alle loro posizioni in un documento o in un insieme di documenti.

Struttura dati ausiliaria che serve per effettuare operazione efficienti in un set di documenti e stringhe.

Esempio

A ogni stringa è associato un indice che verrà usata nel calcolo di occorrenze.

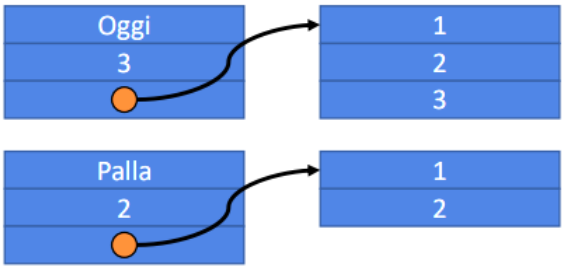


Parola	Totale	Documento
Giorgio	2	1,3
Oggi	3	1,2,3
Gioca	1	1
Palla	2	1,2
Euro	2	2,3
Pranzato	1	3
…		…

La chiave è quindi la parola a cui viene associato la posizione all'interno del documento (o il nome/numero). Viene poi mantenuta anche il numero totale di occorrenze

Gli identificatori dei documenti sono memorizzati all'interno di una lista detta posting lista che non è altro che un database di tutti i documenti che si vuole indicizzare.

Rappresentazione in memoria dell'esempio precedente



File di partenza

- **lista.h, lista.cc, tipo.h, tipo.cc**
  - Per la gestione della posting list
- **inverted**
  - Che contiene i dati relativi all'inverted index
    - Prima riga: numero di parole

        - Dalla seconda in poi: PAROLA, TOTALE, [IDENTIFICATIVI]
- **doc**
    - Che contiene i dati relativi ad un documento
        - Struttura: ID, [ELENCO PAROLE]

Obiettivi:

1) Costruire il tipo di dato parola (struct nuova da creare, uguale a quella dell'esempio sopra) da inserire nell'header file parola.h
2)