Text processing use cases

- Extract statistical information from text data
- Search/index large text documents
- Custom parsers to extract limited information

> For instance, extract one-two fields from the json

For example

- Count number of words
- Count number of times word is present in the text
- Extract few fields values from json

Count number of words in the text

Assume that the number of words is close to the number of spaces. We will be only counting `' '` and `'\n'` characters.

For simplicity we assume that custom utf-8 characters are not used.

For huge texts small error in count is allowable.

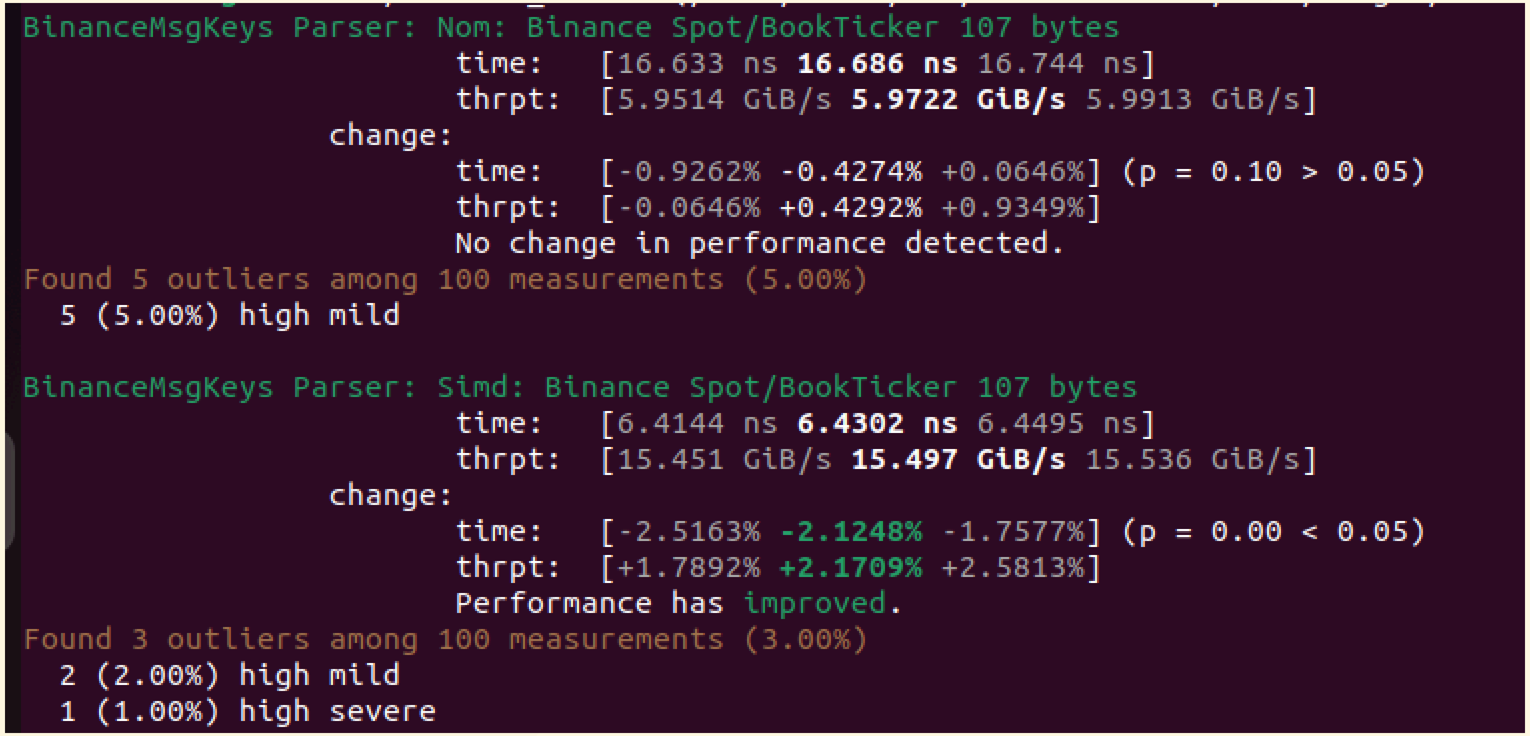
```
...
.LBB0_6:
        vmovd   xmm5, dword ptr [rdi + rax]
        vmovd   xmm6, dword ptr [rdi + rax + 4]
        vmovd   xmm7, dword ptr [rdi + rax + 8]
        vmovd   xmm8, dword ptr [rdi + rax + 12]
        vpcmpeqb        k0, xmm5, xmm1
        vpcmpeqb        k1, xmm6, xmm1
        vpcmpeqb        k2, xmm7, xmm1
        vpcmpeqb        k3, xmm8, xmm1
        vpmovm2q        ymm5, k0
        vpsubq  ymm0, ymm0, ymm5
        vpmovm2q        ymm5, k1
        vpsubq  ymm2, ymm2, ymm5
        vpmovm2q        ymm5, k2
        vpsubq  ymm3, ymm3, ymm5
        vpmovm2q        ymm5, k3
        vpsubq  ymm4, ymm4, ymm5
...
```

Count number of times word is found in the text

We will find matching substring with exact match to input query.

Extract 2 fields from small json (107b) message from Binance exchange

(no implementation provided)



```
BinanceMsgKeys Parser: Nom: Binance Spot/BookTicker 107 bytes
                        time:   [16.633 ns 16.686 ns 16.744 ns]
                        thrpt:  [5.9514 GiB/s 5.9722 GiB/s 5.9913 GiB/s]
                 change:
                        time:   [-0.9262% -0.4274% +0.0646%] (p = 0.10 > 0.05)
                        thrpt:  [-0.0646% +0.4292% +0.9349%]
                        No change in performance detected.
Found 5 outliers among 100 measurements (5.00%)
  5 (5.00%) high mild

BinanceMsgKeys Parser: Simd: Binance Spot/BookTicker 107 bytes
                        time:   [6.4144 ns 6.4302 ns 6.4495 ns]
                        thrpt:  [15.451 GiB/s 15.497 GiB/s 15.536 GiB/s]
                 change:
                        time:   [-2.5163% -2.1248% -1.7577%] (p = 0.00 < 0.05)
                        thrpt:  [+1.7892% +2.1709% +2.5813%]
                        Performance has improved.
Found 3 outliers among 100 measurements (3.00%)
  2 (2.00%) high mild
  1 (1.00%) high severe
```