| Collection | Characters | Documents | Avg. doc. len. | gzip-compr. | xz-compr. |
|---|---|---|---|---|---|
| enwiki-sml | 68,210,334 | 4,390 | 15,537.66 | 36.60 | 26.15 |
| proteins | 58,959,815 | 143,244 | 411.60 | 52.24 | 11.31 |

Table 1: Statistics of the character based collections.

| Identifier | sdsl type |
|---|---|
| GREEDY | doc_list_index_greedy<> |
| SADA | doc_list_index_sada<csa_sada<enc_vector<>, 32, 1000000, text_order_sa_sampling<sd_vector<>>>> |

Table 2: Class definition of character indexes used in the experiment.

| Collection | Index size in MiB (fraction of original collection) | |
|---|---|---|
| | GREEDY | SADA |
| enwiki-sml | 130.49 (2.01) | 203.66 (3.13) |
| proteins | 161.67 (2.87) | 136.24 (2.42) |

Table 3: Size of character indexes.

| Collection | Words | Documents | Avg. doc. len. | gzip-compr. | xz-compr. |
|---|---|---|---|---|---|
| enwiki-sml-int | 12,741,343 | 4,390 | 2,902.36 | 71.75 | 62.88 |

Table 4: Statistics of the word based collections.

| Identifier | sdsl type |
|---|---|
| GREEDY-I | doc_list_index_greedy<csa_wt<wt_int<rrr_vector<63>>, 1000000, 1000000>> |
| SADA-I | doc_list_index_sada<csa_sada_int<enc_vector<>, 32, 1000000, text_order_sa_sampling<sd_vector<>>>> |

Table 5: Class definition of word indexes used in the experiment.

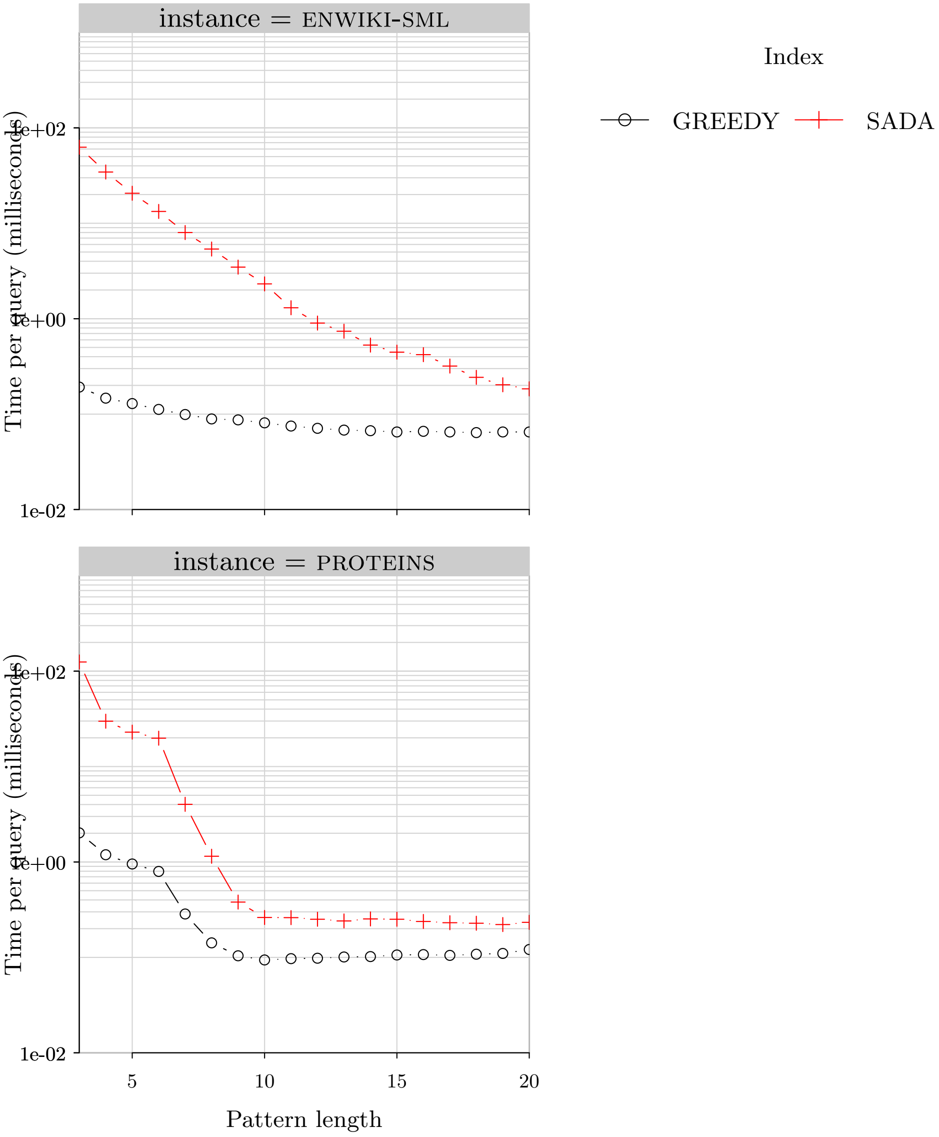

Figure 1: Average query time to find the top-10 documents (frequency measure) for different pattern length using character based indexes. For each query length, 200 pattern were queried.

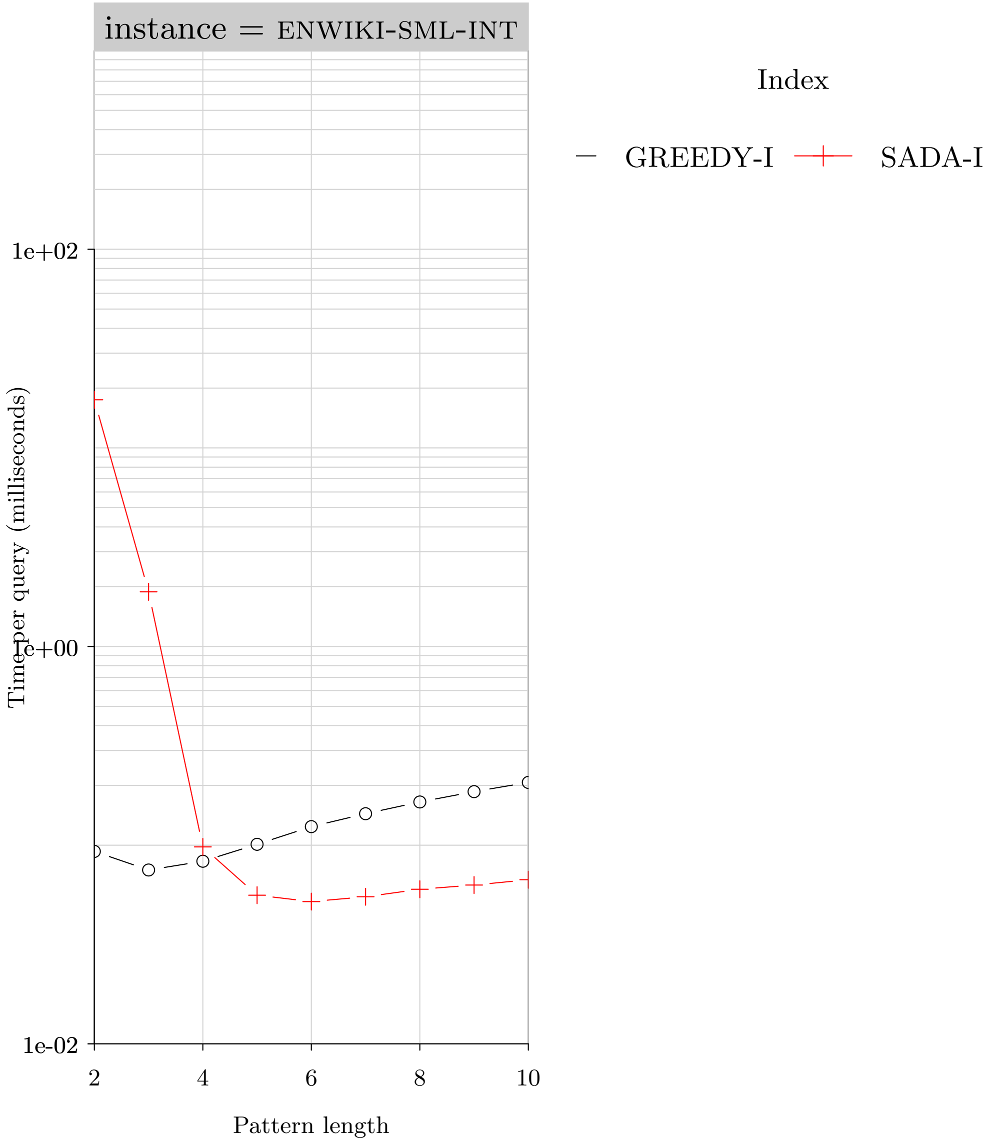

Figure 2: Average query time to find the top-10 documents (frequency measure) for different pattern length using word bases indexes. For each query length, 200 pattern were queried.

| Collection | Index size in MiB (fraction of original collection) | |
|---|---|---|
| | GREEDY-I | SADA-I |
| ENWIKI-SML-INT | 38.05 (1.32) | 49.55 (1.72) |

Table 6: Size of word indexes.