The Number of Nodes in a Trie of English Words

DONG Yuxuan <https://www.dyx.name>

06 Jan 2022 (+0800)

This note documented an experiment on how the nodes in a trie increases with inserted English words.

The Implementation of the Trie

The trie used in this experiment is a simple C implementation. A node is represented by a `trienode` struct. The `next` field of `trienode` is an array of pointers to children. The `stop` field denotes where there is a string stops at the node. The `ALPSIZE` macro represents the size of the alphabet. I defined `ALPSIZE` to be `0x100` (256). This is too large for English words, but it won't affect the number of nodes, and brings brevity to the code. The global variable `root` points to the root of the trie. The global variable `nnodes` stores the number of the nodes.

```
#define ALPSIZE 0x100

struct trienode {
        struct trienode *next[ALPSIZE];
        int stop;
} *root;

int nnodes = 1;
```

The `triepush` function inserts a string into the trie, and returns the new number of the nodes.

```
int triepush(unsigned char *key)
{
        struct trienode *p;

        for (p = root; *key != '\0'; p = p->next[*key++])
                if (NULL == p->next[*key]) {
                        p->next[*key] = malloc(sizeof *p);
                        memset(p->next[*key], 0, sizeof *p);
                        ++nnodes;
                }
        p->stop = 1;
        return nnodes;
}
```

The Data Set of Words

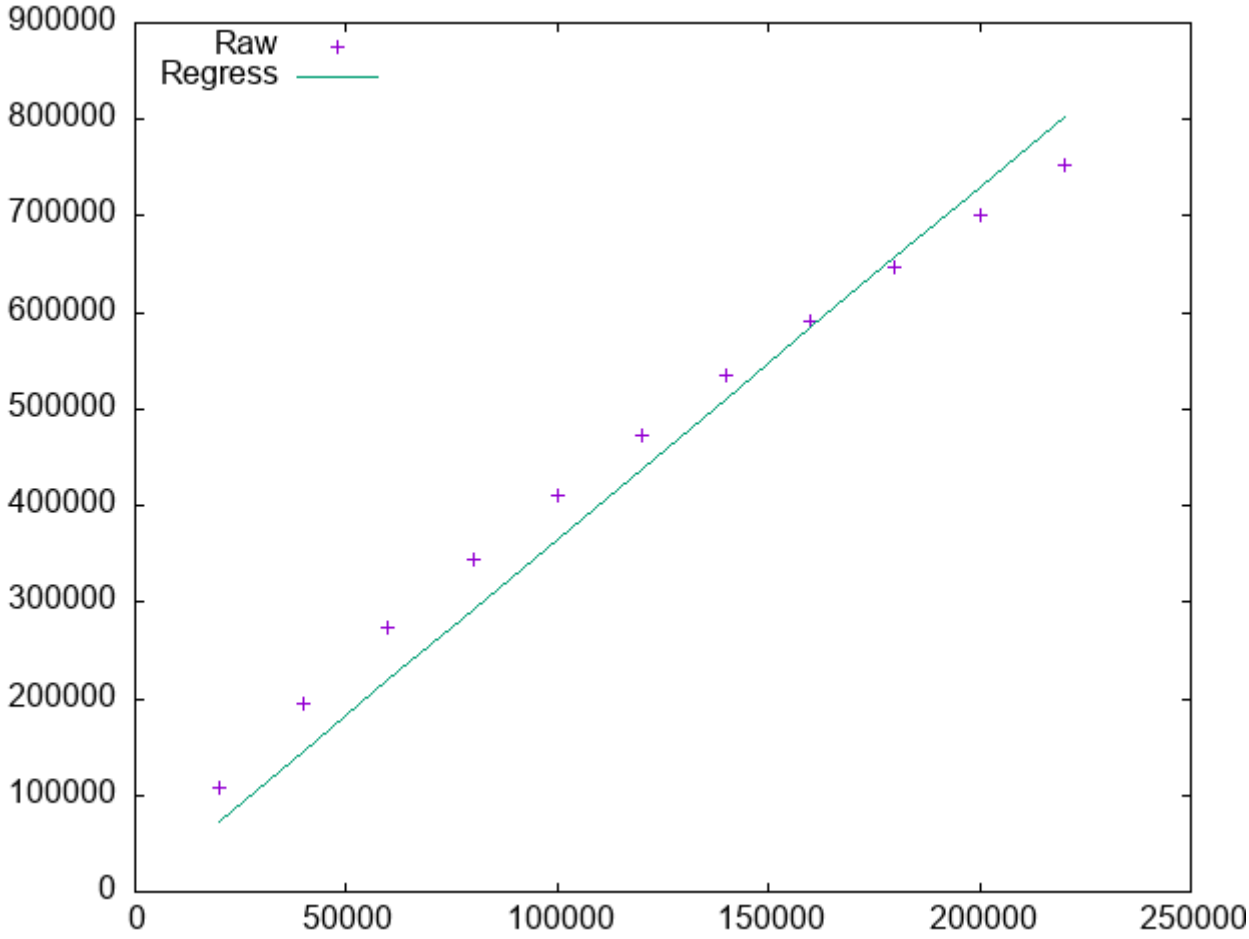
The file `/usr/share/dict/words` shipped with most Unix-compatible systems contains English words one per line. In macOS 12.1, it's a symbolic link to `/usr/share/dict/web2`. According to the `README` file, it contains 234,936 words[1] from Webster's Second International. This file is the dataset I used in this experiment.

The `/usr/share/dict/web2` file is nearly lexicographically sorted[2], case insensibly. Although the order doesn't affect the final number of the nodes, it does affect the growth curve. Thus the file was shuffled in the experiment. I named the shuffled file `web2`, without the leading path.

```
$ head -n3 web2          $ tail -n3 web2
Teleut                   bass
biliously                unsatirized
palaeofauna              sulphobismuthite
```

Conclusions

The result is shown in the following graph. The X-axis is the number of words. The Y-axis is the number of non-root nodes.



Denoting the number of words as x, the number of non-root nodes as y, the linear regression shows that:

```
y ~= 3.64 * x
```

Appendix: Resources

All the data and source code are archived in a gzipped tarball[3].

The tarball contains a makefile to generate:

- `plot.png`: the above graph
- `k`: the slope of the regressed line

Clang/GCC, Awk, and Gnuplot[4] are required.

1. There are actually 235886 lines of the file, according to the output of the `wc` command.↩

2. The only exception is that `pliers` is put after `plies`, according to the report of `diff` and `sort`.↩

3. <https://www.dyx.name/notes/ntrienodes/ntrienodes.tar.gz>↩

4. <http://www.gnuplot.info>↩