ulm university universität
uulm

Institut für Numerische Mathematik

Dr. Andreas F. Borchert
Moritz Carmesin

17. Januar 2020
Blatt 10

Systemnahe Software I (WS 2019/2020)

Abgabe bis zum 24. Januar 2020, 14:00 Uhr

Lernziele:

- Sicherer Umgang mit Zeichenketten dynamischer Länge
- Mengen auf Basis von Hash-Tabellen

Aufgabe 12: Wörter extrahieren

Schreiben Sie ein Werkzeug *extract_words*, das in Verbindung mit einem Wörterbuch arbeitet und aus der Standard-Eingabe alle Wörter extrahiert, die entweder im Wörterbuch vorkommen oder, bei Verwendung der Option „-v“, dort nicht vorkommen. Bei der Ausgabe darf jedes Wort nur einmal genannt werden. Die Reihenfolge der Wörter bei der Ausgabe darf beliebig sein.

Wörter sind hierbei alle Folgen von Klein- und Großbuchstaben, also „a“ bis „z“ und „A“ bis „Z“ – ohne Umlaute, Akzente etc. Alle andere Zeichen werden als Trenner interpretiert. So enthält die Zeichenkette „Hallo zusammen!“ die beiden Wörter „Hallo“ und „zusammen“.

Neben den Optionen ist auf der Kommandozeile die Datei mit dem Wörterbuch anzugeben. Dies muss kein besonderes Format aufweisen, d.h. Wörter können aus der Standard-Eingabe und dem Wörterbuch nach dem gleichen Verfahren extrahiert werden. Auf der Theon findet sich ein Wörterbuch mit englischen Wörtern in */usr/dict/words* und bei Debian unter */usr/share/dict/words*.

Folgende Optionen sind ebenfalls zu unterstützen:

- „-i“: Unterschiede in Bezug auf Groß- und Kleinschreibung sind zu ignorieren.
- „-v“: Es sind die Wörter auszugeben, die im Wörterbuch *nicht* vorkommen.

Bei fehlerhaften Angaben auf der Kommandozeile ist eine „Usage“-Meldung auszugeben.

Hier ist ein beispielhafter Aufruf unter Debian:

```
heim$ ./extract_words
Usage: ./extract_words [-i] [-v] dict
heim$ ./extract_words -i -v /usr/share/dict/words <extract_words.c | column
getopt          isupper         printf          tolower         stdlib
stdio           unistd          perror          stderr          bool
stdin           optind          stralloc        fprintf         fclose
argv            cmdname         ctype           fopen           const
argc            fp              int             dict
heim$
```

Verwenden Sie nur *stralloc*-Objekte und -Funktionen, um Zeichenketten zu verarbeiten. Arrays von *char* sind nicht erlaubt; Zeiger auf *char* aber schon. Um *stralloc*-Objekte verwenden zu können, ist es notwendig, die passende Header-Datei *stralloc.h* per **#include** einzubinden. Beim Zusammenbauen ist die Bibliothek „-lowfat“ anzugeben.

Das Werkzeug benötigt zwei Wortmengen. Die eine repräsentiert das Wörterbuch, die andere die Menge der auszugebenden Wörter. Die Wortmengen sollten auf Basis von Hash-Tabellen umgesetzt werden, da diese eine effiziente Suche ermöglichen. Beim Einfügen ist aber darauf zu achten, dass die Mengeneigenschaft erhalten bleibt, d.h. wenn ein Wort ein weiteres Mal in eine Menge hinzugefügt wird, sollte dies keinen Effekt haben. Sie benötigen hierfür eine Hash-Funktion für Zeichenketten. Empfehlenswert ist die Hash-Funktion von Dan J. Bernstein (siehe ganz am Ende).

Hinweise

Wie üblich ist dies wieder eine Aufgabe, die sich hervorragend dafür eignet, innerhalb einer Gruppe aufgeteilt zu werden:

- Die auf einer Hash-Tabelle basierende Datenstruktur für eine Wortmenge zusammen mit den zugehörigen Operationen ist als separates Modul zu implementieren.
- In einem weiteren Modul sollte das Extrahieren eines Worts aus einer Ein- und Ausgabe-Verbindung in ein *stralloc*-Objekt unterstützt werden. So könne eine solche Funktion aussehen:

  bool scan_word(*FILE* fp*, *stralloc* word*);

- Im Hauptprogramm sollten Sie die Argumente aus der Kommandozeile verarbeiten, die Wortmengen anlegen, befüllen und die eine davon am Ende ausgeben.

Reichen Sie bitte Ihre Lösung mit folgendem Kommando ein:

```
theon$ submit ss1 12 extract_words.c \
   [scan_word.c scan_word.c] \
   [set.c set.h]
```

Viel Erfolg!