Datastructuren: Opdracht 4
Hashing

Deadline: Zondag 24 November, 23:59

Voor deze opdracht wordt gevraagd een snellere toegang (constante tijd) te realiseren voor genealogische records. De totale data set bestaat uit 16 miljoen records, er zijn twee subsets beschikbaar gesteld met de meest voorkomende voornamen[1] en achternamen[2].

Het is de bedoeling om de toegang O(1) te maken door middels de voor- of achternaam key een adres in een hash tabel te berekenen dat zoveel mogelijk perfect en minimaal is.

In deze opdracht is het de bedoeling dat er drie hashtabellen worden geïmplementeerd, allen met een eigen hashfunctie:

- twee van de hashfuncties besproken in het hoorcollege. (Bijvoorbeeld digit selection, folding, zie sheets)
- een zelf ontworpen hashfunctie.

De hashtabellen moeten naast deze hashfunctie ten minste een `insert`, `remove` en `contains` methode bevatten. Maak gebruik van overerving, zodat je dit drietal functies slechts een maal hoeft te implementeren. Zorg dat collisions op een juiste manier worden afgehandeld, bijvoorbeeld door middel van Chained Hashing (zie college sheets).

Om te bepalen of de ontworpen hashfunctie perfect en minimaal genoeg is het nuttig om per adres in de hash tabel ook een integer bij te houden (initieel 0) die bijhoudt hoeveel keys op de berekende index terechtkomen. Door statistieken (loadfactor, gemiddelde multipliciteit, geheugenfactor) op te maken van hoe vaak elke hash-tabel entry aangeslagen is kun je aantonen dat de ontworpen hash-functie als resultaat een betere hash invulling opleveren.

Probeer eerst met behulp van bijvoorbeeld in het college voorgestelde hash aanpakken voor de voor- en achternamen set apart een goed scorende hash functie te maken en laat zien hoe goed die op de andere set werkt; kun je een hash functie bedenken die voor beide sets goed presteert? Laat dat dan middels statistieken zien. Maak ook een kort LaTeXverslag waarin je de hash functies uitlegt en de statistieken presenteert.

[1] `http://www.liacs.nl/home/jvrijn/ds2013/assets/voornamen2013.txt`
[2] `http://www.liacs.nl/home/jvrijn/ds2013/assets/achternamen2013.txt`

Voor het implementeren van de hashtabel en het inladen van de namen mag gebruik gemaakt worden van de C++ Standard Library. De uitwerking dient uiterlijk zondag 24 November, om 23:59, via email opgestuurd te worden naar Jan van Rijn (email: j.n.van.rijn@liacs.leidenuniv.nl).