

# PROJEKTAUFGABE AE

## Remove Duplicates - Spotify playlist cleaner

RAPHAEL DRECHSLER

### INHALTSVERZEICHNIS

|     |  |    |
|-----|--|----|
| 1   | Problemstellung "Remove Duplicates"                    | 2  |
| 2   | Anwendungsszenario Spotify Playlists                   | 4  |
| 3   | Funktionsweisen der Algorithmen                        | 4  |
| 3.1 | Funktionsweise Haupt-Algorithmus: dict-like            | 4  |
| 3.2 | Funktionsweise Vergleichs-Algorithmus 1: Naiver Ansatz | 5  |
| 3.3 | Funktionsweise Vergleichs-Algorithmus 2: python dict   | 6  |
| 4   | Messdaten  | 6  |
| 4.1 | Test-Daten-Generator                                   | 6  |
| 4.2 | Echt-Daten   | 7  |
| 4.3 | Skalieren der Daten-Listen                             | 8  |
| 5   | Vorbereiten der Messungen                              | 8  |
| 5.1 | Profiling und Optimierung Haupt-Algorithmus            | 9  |
| 5.2 | Profiling und Optimierung Vergleichs-Algorithmus 1     | 10 |
| 5.3 | Profiling und Optimierung Vergleichs-Algorithmus 2     | 11 |
| 5.4 | Wie wird gemessen?                                     | 12 |
| 6   | Vergleich Echt vs. Test-Daten                          | 12 |
| 6.1 | Haupt-Algorithmus                                      | 13 |
| 6.2 | Vergleichs-Algorithmus 1                               | 13 |
| 6.3 | Vergleichs-Algorithmus 2                               | 13 |
| 7   | Vergleich mit Theoretischer Laufzeit                   | 13 |
| 7.1 | Haupt-Algorithmus                                      | 14 |
| 7.2 | Vergleichs-Algorithmus 1                               | 15 |
| 7.3 | Vergleichs-Algorithmus 2                               | 15 |
| 8   | Vergleich der Algorithmen                              | 16 |
| 8.1 | Grafischer Vergleich                                   | 16 |
| 8.2 | Statistischer Hypothesen-Test                          | 17 |
| 9   | Fazit  | 18 |

# 1 PROBLEMSTELLUNG "REMOVE DUPLICATES"

## Problem

Gegeben ist eine Sequenz. Diese Sequenz enthält ggf. Duplikate. Ziel des umzusetzenden Algorithmus ist das Entfernen der Duplikate aus dieser Sequenz.

Für die Umsetzung eines entsprechenden Algorithmus sollen die zwei folgenden Ansätze betrachtet werden:

- Naiver Ansatz: Nutzung eines einfachen Arrays
- Ansatz im Fokus: Nutzung einer Hash-Table

## Naiver Ansatz

In der Implementierung nach dem naiven Ansatz würden alle Daten einer Sequenz in einem Array gespeichert werden, insofern sie nicht bereits im Array enthalten sind. Beim betrachten eines Elementes aus der Sequenz muss also in der naiven Umsetzung das komplette Array nach einem identischen Element durchsucht werden. Im Hinblick auf die Laufzeit schlimmsten Fall, muss daher jedes Element im Array mit dem aktuell untersuchten Element der Sequenz verglichen werden. Bei einer Sequenz-Liste der Größe  $n$  müssen also im worst-case

$$\sum_{k=1}^{n-1} k \quad (1)$$

Vergleichsoperationen durchgeführt werden. Betrachtet man die folgende Umformung entsprechend der Gaußschen Summenformel

$$\sum_{k=1}^{n-1} k = \frac{(n-1)^2 + (n-1)}{2} = \frac{1}{2}(n-1)^2 \quad (2)$$

ergibt sich für eine Umsetzung dieses Ansatzes eine theoretische obere Komplexitätsgrenze von

$$O(n^2) \quad (3)$$

## Ansatz Hash-Table

Der Ansatz "Hash-Table" setzt an dem Punkt der Komplexitätsbetrachtung an. Würde man für die Einordnung der Elemente in das Array eine Direktadressierung verwenden, so würde man ein Element selbst als Schlüssel interpretieren, mit dem ein Feld im Array adressiert wird.

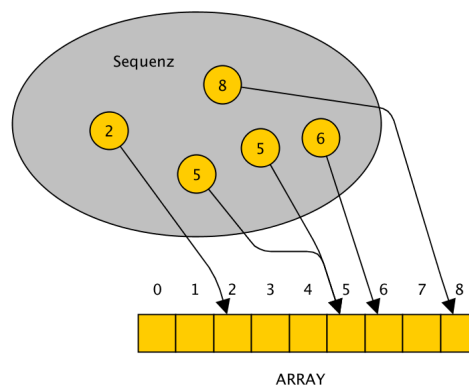


Abbildung 1: Skizze: Prinzip der direkten Adressierung nach [1]

Der Zeitaufwand für die Prüfung, ob ein Element bereits im Array gespeichert ist, wäre dabei  $O(1)$ .

Ist das Universum  $U = 0, 1, \dots, m$ , in denen sich die Schlüssel der Elemente befinden klein, lässt sich somit schnell auf ein Array der Größe  $A[0..m-1]$  zugreifen. Ab einer bestimmten Größe des Universums kann eine Umsetzung der Direkt-Adressierung Aufgrund der erforderlichen Größe des Ziel-Arrays nicht mehr sinnvoll bzw. möglich sein.

Die Hash-Table löst dieses Problem, indem sie das große Universum  $U$  einem kleineren Array  $A[0..m-1]$  gegenüberstellt. Die Adressierung der Array-Felder pro Element erfolgt weiterhin auf Grundlage des Element-Wertes. Um nur existierende Schlüssel zu erhalten wird zur Ermittlung des Schlüssels eine Funktion  $h$  (sogenannte Hash-Funktion) eingesetzt, welche die Werte der Elemente im Universum  $U$  auf existierende Schlüssel abbildet. vgl.[1]

$$h : U \rightarrow 0, \dots, m-1 \quad (4)$$

Die Skizze gestaltet sich wie folgt:

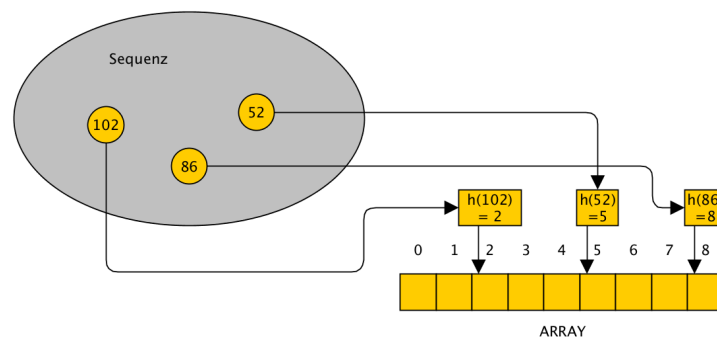


Abbildung 2: Skizze: Prinzip der Adressierung in einer Hash-Tablenach [1]

Dadurch, dass gilt  $|U| > m$  wird über Anwendung des Schubfachprinzips ersichtlich, dass die Hash-Funktion  $h$  nicht injektiv ist.

Fälle in denen gilt

$$h(a) = h(b) | a, b \in U, a \neq b \quad (5)$$

werden als Kollision bezeichnet. vgl.[1]

Bei der Umsetzung des Algorithmus muss also eine entsprechende Strategie zur Auflösung solcher Kollisionen mit betrachtet werden.

### Kollisions-Auflösung im python dict-Objekt

Als Kollisions-Auflösungs-Strategie soll im Rahmen der Umsetzung das Verfahren implementiert werden, welches im python-Standard beim Zugreifen auf *dict*-Objekte genutzt wird. Durch die ggf. notwendige Behandlung von Kollisionen ergibt sich für einen Zugriff auf ein Feld des Arrays bei einer entsprechenden Implementierung der folgende Zeitbedarf.

| Operation             | average-case | worst-case |
|-----------------------|--------------|------------|
| Element einfügen      | $O(1)$       | $O(n)$     |
| Auf Element zugreifen | $O(1)$       | $O(n)$     |

Tabelle 1: Zugriffs-Komplexität im python dict [2]

## 2 ANWENDUNGSSZENARIO SPOTIFY PLAYLISTS

Als Anwendungsfall soll das disjunkte Vereinigen von Titeln mehrerer Spotify-Playlists betrachtet werden.

In Spotify lassen sich für eine geöffnete Playlist per Tastatur-Kurzbefehl "Strg + A" und "Strg + C" alle Titel der Playlist in die Zwischenablage kopieren. Die Titel werden dabei als URI repräsentiert.

---

```
https://open.spotify.com/track/757530vPBymdi31CtXstxP
https://open.spotify.com/track/1Qi256uJuMihknGuuFcQoC
https://open.spotify.com/track/0JFBf2PloRfMkPg5DjXhDx
https://open.spotify.com/track/7iXF2W9vKmDoGAhIHdpyIa
```

---

Abbildung 3: Beispiel: Spotify-Titel URIs

Die URIs werden durch den Anwender in einem separatem *.txt*-File gesammelt. Somit werden mehrere Playlists in diesem *.txt*-File vereinigt.

Die Bereinigung der Duplikate soll nun der zu implementierende Algorithmus übernehmen.

Nach Abschluss der Bereinigung lassen sich alle URIs des *.txt*-Files in die Zwischenablage kopieren und per Tastatur-Kurzbefehl "Strg + V" in Spotify in eine (sinnvollerweise neue) Playlist einfügen.

## 3 FUNKTIONSWEISEN DER ALGORITHMEN

Im Rahmen der Lösungs-Umsetzung des "Remove Duplicates"-Problems sollen drei verschiedene Implementierungen in der Sprache python betrachtet werden.

Dabei soll das Hauptaugenmerk auf der ersten Implementierung liegen. Diese soll die Funktionalität der *dict*-Objekte in python nachempfinden, indem Sie die entsprechende hash-Funktion und die entsprechende Kollisions-Auflösungs-Strategie implementiert.

Zum Vergleich sollen zwei weitere Implementierungen betrachtet werden:

- der in Kapitel 1 beschriebene naive Ansatz
- eine Lösung, welche die tatsächlichen python *dict*-Objekte nutzt

Im Folgenden werden die Implementierungen mittels Pseudo-Code beschrieben.

### 3.1 Funktionsweise Haupt-Algorithmus: dict-like

Zunächst wird eine hash-table in Form eines Arrays initialisiert.

#### Haupt-Algorithmus: Hash-Tabelle initialisieren

---

**Ergebnis :** table, tableSize

```
1 tableSize ← 8;
2 table ← Array der Größe tableSize;
```

---

Anschließend werden die Elemente (URIs) aus der Sequenz-Datei des Nutzers in die angelegte hash-table geschrieben. Tritt dabei ein Element doppelt auf, wird dieses als Duplikat erkannt und ignoriert.

Als hash-Funktion wird dabei die python-Standardfunktion *hash()* eingesetzt. Die Strategie zur Kollisionsauflösung entspricht derjenigen, die bei python *dict*-Objekten Anwendung findet und lässt sich dem folgenden Pseudocode entnehmen.

## Haupt-Algorithmus: Elemente in Array einfügen

---

```

Daten : Sequenz – Datei
Ergebnis : table
1 für Jede URI in Sequenz – Datei tue
2   index ← hash(URI);
3   uriHandled ← false;
4   pertub ← None;
5   tue
6     wenn bereits Element in table[index] enthalten dann
7       wenn table[index] == URI dann
8         //Duplikat liegt vor - URI wird ignoriert
9         uriHandled ← true;
10      sonst
11        wenn pertub besitzt Wert dann
12          pertub = pertub >> 5;
13          index ← ((5 * slotindex) + 1 + pertub)%tableSize;
14        sonst
15          pertub = index;
16        Ende
17      Ende
18    sonst
19      table[index] ← URI;
20      uriHandled ← true;
21    Ende
22  solange !uriHandled;
23  wenn URIs in table >  $\frac{2}{3}$  * tableSize dann
24    tue
25      | tableSize ← tableSize * 2
26    solange 4 * |URIs in table| >= tableSize;
27    newTable ← table;
28    Kopiere alle URIs in table unter Errechnung neues Schlüssels in newTable;
29    table ← newTable;
30 Ende

```

---

Anschließend werden die in der hash-table enthaltenen URIs in eine Output-*.txt*-Datei geschrieben. Diese enthält damit die Duplikat-freie Liste der URIs und kann vom Anwender genutzt werden.

### 3.2 Funktionsweise Vergleichs-Algorithmus 1: Naiver Ansatz

Der in Kapitel 1 skizzierte naive Ansatz gestaltet sich grundlegend wie folgt.

#### Algorithmus Naiver Ansatz

---

```

Daten : Sequenz – Datei
Ergebnis : table
1 table ← [];
2 für Jede URI in Sequenz – Datei tue
3   uriInTable ← false;
4   für Jede processedUri in table tue
5     | wenn processedUri == URI dann
6       | uriInTable ← true;
7   Ende
8   wenn uriInTable dann
9     | //Duplikat erkannt, ignoriere es
10  sonst
11    | table.append(URI);
12  Ende
13 Ende

```

---

Anschließend erfolgt die Ausgabe der Duplikat-freien Liste per Erzeugen einer Output-Datei.

### 3.3 Funktionsweise Vergleichs-Algorithmus 2: python dict

Die Funktionsweise des zweiten Vergleichs-Algorithmus ist diejenige, deren Verhalten in der Umsetzung des Haupt-Algorithmus angestrebt wird.

Folglich zeichnet sich der Quellcode der Implementierung durch die Nutzung von python-Standard-Operatoren aus. Der wesentliche Teil der Implementierung umfasst die folgenden Zeilen Code:

---

```
uriDict = {}
for uri in sequenzFile:
    if uri not in uriDict:
        uriDict[uri] = 1
    else:
        #ignore duplicate
        pass
```

---

Analog zu den anderen zwei Implementierungen erfolgt im Anschluss eine Ausgabe aller Elemente des dicts per Output-File.

Alle drei initialen Implementierungen finden sich im Abgabeordner unter */Code/02\_ImplementierungOhneVerbesserung*. Zum Aufruf der Algorithmen über die Kommandozeile muss die zu untersuchende *.txt*-Datei als Parameter angegeben werden. Beispielhaft sei hier ein Aufruf der ersten Implementierung gezeigt.

---

```
python 1_dictLike.py AimeeMann_980DS.txt
```

---

## 4 MESSDATEN

Zur Beschaffung von möglichst großen und verschiedenen Playlists als Messdaten wurde zunächst ein Test-Daten-Generator implementiert. Dieser wird in 4.1 näher beschrieben.

Im Folgenden wurde ein kurzes Programm implementiert, welches die Spotify Web-API nach Titeln anfragt und somit echte Datensätze erhält. Dieser Echt-Daten-Abgreifer wird in 4.2 näher beschrieben.

Zudem wurde für das Durchführen von Messreihen mit wachsender Playlist-Größe ein Vorgehen etabliert, um die benötigten *.txt*-Files zu erzeugen. Dieses wird in 4.3 erläutert.

### 4.1 Test-Daten-Generator

Der Aufruf des Generators erfolgt mit zwei optionalen Parametern und einem obligatorischem Parameter.

---

```
testDataGenerator.py <n> [dupFactor] [fileSuffix]
```

---

Der Parameter *n* entspricht der Anzahl der Elemente, welche die zu generierende Liste enthalten soll. Der *dupFactor* entspricht der Wahrscheinlichkeit mit der ein Element in der Liste ein Duplikat ist. Der *fileSuffix* kann bei mehreren ausgegebenen Dateien mit gleicher Parametrisierung für eine unterschiedliche Benennung genutzt werden.

Ein Beispiel für einen Aufruf könnte wie folgt aussehen.

---

```
python testDataGenerator.py 100000 .4 Run1
```

---

Im Wesentlichen erzeugt der Test-Daten-Generator eine *.txt*-Datei im sequence-File-Format. Die dynamischen Teile der URIs, (welche einen base62-Code [3] der Länge 22 darstellen) werden dabei zufällig erzeugt und ergeben im Regelfall keine tatsächlich existierende URI.

Die Funktionsweise des Generators ist dabei wie folgt zu beschreiben.

### Algorithmus Test-Daten-Generator

---

```

Daten : n, [dupFactor], [fileSuffix]
Ergebnis : .txt – File
1 wenn dupFactor nicht gegeben dann
2   | dupFactor ← zufälliger Faktor aus {0.0, 0.1, ..., 0.9};
3 linkList ← [];
4 für i in {0, ..., n – 1} tue
5   | wenn i > 0 dann
6     | wenn Zufallsentscheidung mit Wahrscheinlichkeit dupFactor = positiv dann
7       | linkList.append(linkList[i – 1]);
8     | sonst
9       | linkList.append(neue zufällige URI);
10    | Ende
11   | sonst
12     | linkList.append(neue zufällige URI);
13   | Ende
14 Ende
15 erzeuge neues .txt – File ggf mit Suffix fileSuffix;
16 Schreibe in 1. Zeile von .txt – File "sequenceString";
17 Schreibe Elemente aus linkList in Zufalls-Reihenfolge in .txt – File;
```

---

Die resultierende Datei (Beispielsweise "*generatedTestData100000\_0.4dupFac\_Run1.txt*") ist dann bereit zur Verarbeitung durch einen implementierten Algorithmus oder lässt sich zum Erstellen mehrerer Playlist für eine Messreihe einsetzen. (Siehe 4.3)

Die Implementierung des Algorithmus findet sich im Abgabeordner unter *Code/01\_Daten\_Generieren/01\_TestDaten/testDataGenerator.py*.

## 4.2 Echt-Daten

Für das Beschaffen von großen Playlists mit echten Track-URIs wurde eine Java-Anwendung umgesetzt, welche mithilfe eines Wrappers [4] die Spotify API abfragt. Um eine beliebig große Menge an URIs abzufragen, wurden das unten dargestellte Vorgehen für das Abfragen der API implementiert. Dabei müssen eine Mindest-Anzahl an URIs, die erhalten werden sollen und die URI eines Interpreten als Start-Interpreten angegeben werden.

Der entsprechende Java-Projektordner findet sich im Abgabeordner unter *Code/01\_Daten\_Generieren/02\_EchtDaten/realDataGetter*.

### Algorithmus Echt-Daten-Abgreifer

---

```

Daten : minimalNoOfUris, startArtist
Ergebnis : outputFile.txt
1  Autorisierungsprozess zwischen Anwendung und Spotify API;
2  outputTracklist  $\leftarrow$  [];
3  searchProcessedArtists  $\leftarrow$  [];
4  searchUnprocessedArtists  $\leftarrow$  [];
5  processedArtists  $\leftarrow$  [];
6  searchUnprocessedArtists.add(startArtistUri);
7  tue
8      für artist in api.getRelatedArtists(searchUnprocessedArtists.get(0)) tue
9          wenn artist nicht in processedArtists dann
10             searchUnprocessedArtists.add(artist); für album in
11                 api.getAlbumsForArtist(artist) tue
12                     für track in album tue
13                         outputTracklist.add(track);
14                     Ende
15                 Ende
16             processedArtists.add(artist);
17         Ende
18     searchProcessedArtists.add(searchUnprocessedArtists.get(0));
19     searchUnprocessedArtists.remove(0);
20 solange |outputTracklist|  $\leq$  minimalNoOfUris UND
21     searchUnprocessedArtists ist nicht leer;
22     Erzeuge outputFile.txt;
23     Schreibe in Zeile 1 von outputFile.txt "sequenceString";
24     für URI in outputTracklist tue
25         Schreibe URI in outputFile.txt;
26     Ende

```

---

Analog zu 4.1 kann die generierte Datei im Anschluss verarbeitet werden.

#### 4.3 Skalieren der Daten-Listen

Um später ganze Messreihen auszuführen, bei denen die Anzahl der Elemente, die in den Playlists enthalten sind, im Verlaufe der Messreihe zunimmt, wurde ein python-Skript erstellt, welches ein Sequenz-File einliest und nach eine Neue Datei erstellt, welche nur die Anzahl an Zeilen (zuzüglich der Zeile "sequenceString") beinhaltet, welche als zweiter Parameter übergeben wird.

Der Aufruf erfolgt beispielsweise wie folgt:

---

```
python cutToLen.py generatedTestData100_0.5dupFac.txt 40
```

---

Um viele Dateien für eine Messreihe zu generieren, wird ein kurzes shell-Skript je Messreihe angepasst und anschließend eingesetzt.

Das python- und das shell-Skript finden sich im Abgabeordner unter *Code/01\_Daten\_Generieren/03\_DatenSkalieren*.

## 5 VORBEREITEN DER MESSUNGEN

Bevor der die Algorithmen-Laufzeiten untereinander und mit den jeweiligen theoretischen Laufzeiten verglichen werden, soll zunächst das Profiling und das Optimieren der einzelnen Implementierungen erfolgen. (Kapitel 5.1 - 5.3)  
Die Vorbereitung der Messungen wird mit Kapitel 5.4 abgeschlossen, in dem beleuchtet wird, welche Messungen auf welche Weise erfolgen.



## 5.1 Profiling und Optimierung Haupt-Algorithmus

Generell soll für die Laufzeit-Messungen nur das Abarbeiten der Track-URIs, also das schreiben dieser in die hash-table bzw. das dict im Fokus stehen. Daher werden für alles Weitere die Funktionalitäten für das Schreiben des Ergebnis-Files auskommentiert.

Für den Hauptalgorithmus liegt die entsprechende *.py*-Datei im Abgabeordner unter *Code/03\_Messen\_und\_Verbessern/01\_Algo\_1/1\_dictLike\_improvement0.py*.

Für das Profiling wird ein mittels *testDataGenerator.py* erstelltes *.txt*-File mit Größe  $n = 100.000$  und der Duplikat-Wahrscheinlichkeit von  $\text{dupFac} = 2\%$  verwendet. Das Profiling erfolgt über die Nutzung des python-Profilers *cProfile* [5].

Die erste Visualisierung des Profils mittels *Snakeviz* zeigt:

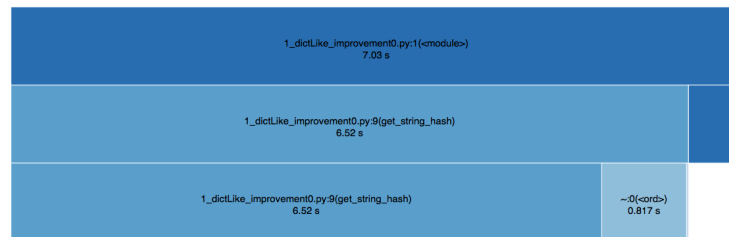


Abbildung 4: Darstellung initiales Profiling von Algorithmus 1

Die Funktion *get\_string\_hash* macht einen Großteil der Arbeitszeit aus. Diese Funktion ist der python-Standard-Funktion *hash()* nachempfunden. Das Implementieren dieser Funktion sollte dem tieferen Verständnis dienen.

Die wesentliche Verbesserung besteht also im Ersetzen der *get\_string\_hash*-Funktion durch den python-Standard.

Der angepasste Algorithmus ist im Abgabeordner unter *Code/03\_Messen\_und\_Verbessern/01\_Algo\_1/1\_dictLike\_improvement1.py* enthalten.

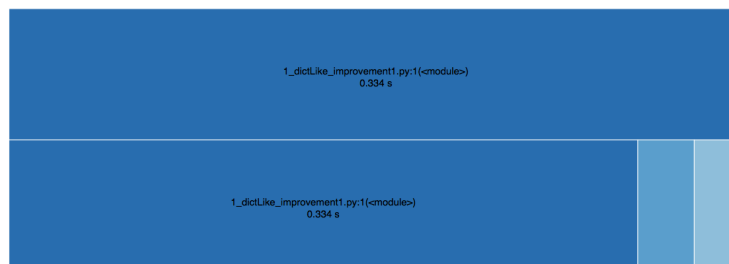


Abbildung 5: Darstellung Profiling von Algorithmus 1 nach 1. Verbesserung

Würde man diesem Prinzip der Optimierung folgen, also alles, wofür es python-Standard-Funktionalitäten gibt mit python-Standard-Funktionalitäten lösen, so würde der auf diesem Wege final optimierte Algorithmus zu dem zweiten Vergleichs-Algorithmus (Kapitel 3.3.) identisch sein.

Daher soll im nächsten Schritt einer potentiellen Verbesserung das Tailoring angesetzt werden.

Betrachtet man die URIs, wird klar, dass die hash-Funktion nur auf den 22-stelligen Base62-Code angewendet werden muss, da der restliche Teil für alle URIs identisch ist.

<https://open.spotify.com/track/757530vPBymdi31CtXstxP>  
<https://open.spotify.com/track/1Qi256uJuMihknGuuFcQoC>

Der entsprechend angepasste Algorithmus ist im Abgabeordner unter *Code/03\_Messen\_und\_Verbessern/01\_Alg\_1/1\_dictLike\_improvement2.py* enthalten. Die Ausgabe-Funktion muss für eine Anwendung um den statischen Teil der URIs erweitert werden. Dies wurde in der Datei, obgleich die Ausgabefunktion für die Messungen auskommentiert bleibt, getan.

Die Visualisierung des Profilings zeigt:

*Code/03\_Messen\_und\_Verbessern/01\_Alg\_1/1\_dictLike\_improvement1.py* enthalten.

Da die Profile bezüglich ihrer gesamt-benötigten Laufzeit zu nahe beieinander lagen, wurde das Profiling der vorherigen Version *1\_dictLike\_improvement1.py* und das Profiling der Verbesserten Variante mit einem *.txt*-File der Größe  $n = 1.000.000$  wiederholt.

Die Verbesserung ist marginal. Bei mehrfachem Messen zeigte sich, dass durch Schwankungen anhand Gesamt-Laufzeit nicht erkannt werden konnte, welche Implementierung schneller ist. Durch das Profiling jedoch stellte sich heraus, dass die benötigte Zeit für das Anwenden der Hash-Funktion bei in jedem Profil für den zweiten Algorithmus geringer war.

| Durchlauf | Vor Tailoring gesamt | Nach T. gesamt | Vor T. hash() | Nach T. hash() |
|-----------|----------------------|----------------|---------------|----------------|
| 1         | 2,65s                | 2,54s          | 0,185s        | 0,143s         |
| 2         | 2,61s                | 2,62s          | 0,183s        | 0,146s         |
| 3         | 2,74s                | 2,43s          | 0,188s        | 0,138s         |

Tabelle 2: Ergebnisse des Tailorings: Benötigte Zeit gesamt und für hash-Funktion

Abschließend seien hier die Zeitbedarfe für je einen Durchlauf einer Verbesserungs-Stufe des ersten Algorithmus für die oben verwendete Playlist mit  $n = 100.000$  Elementen dargestellt:

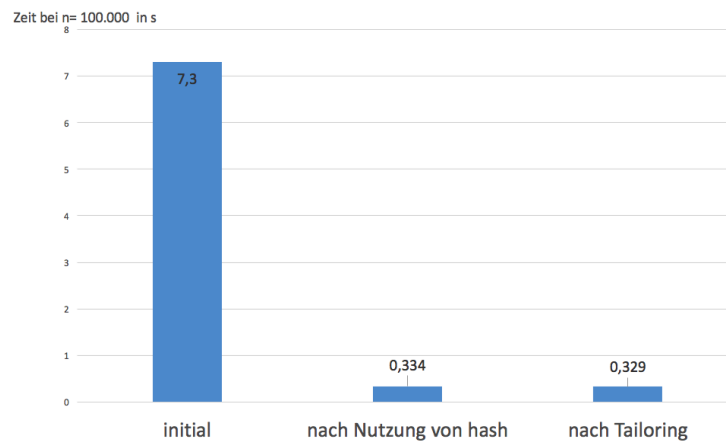


Abbildung 6: Zusammenfassung der Optimierung von Algorithmus 1

## 5.2 Profiling und Optimierung Vergleichs-Algorithmus 1

Analog zu der in Kapitel 5.1 gewonnenen Erkenntnis, soll die Verbesserung des Algorithmus in zwei Schritten ablaufen:

1. Ersetzen von implementierten Funktionalitäten, die mit python-Standards gelöst werden können (Iterieren durch Array ablösen durch Verwendung des *in*-Operators)
2. Anpassen des Algorithmus analog zum in 5.1 durchgeführten Tailoring

Die entsprechenden Implementierungen sind im Abgabeordner unter *Code/03\_Messen\_und\_Verbessern/02\_Alg\_2/* enthalten.

Die zusammengefasste Verbesserung gestaltet sich für ein Beispiel der Laufzeit mit  $n = 10.000$  für Algorithmus 2 wie folgt:

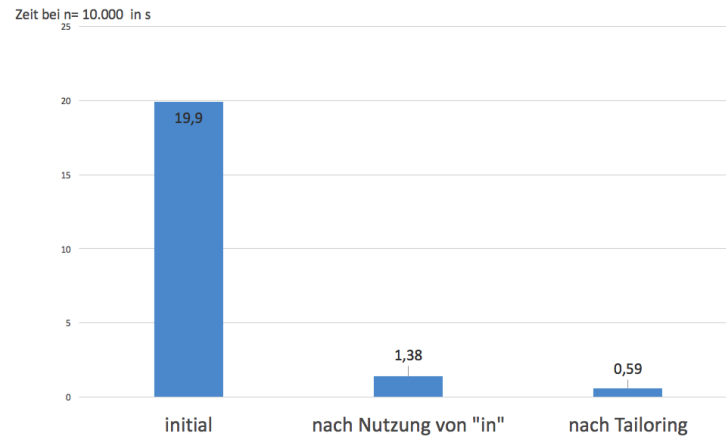


Abbildung 7: Zusammenfassung der Optimierung von Algorithmus 2

Die Verbesserung ist ähnlich zu der in Kapitel 5.1 erzielten, wobei jedoch das Tailoring für Algorithmus 2 eine größere Wirkung zeigt.

### 5.3 Profiling und Optimierung Vergleichs-Algorithmus 2

Da in dieser Umsetzung bereits die python-Standardfunktionen genutzt werden, soll hier nur das Tailoring betrachtet werden. Die entsprechenden Implementierungen sind im Abgabeordner unter *Code/03\_Messen\_und\_Verbessern/03\_Alg\_3/* enthalten.

Die Verbesserung wird erst ersichtlich, wenn man die Anzahl der URIs in der Liste deutlich erhöht. Ab  $n = 10.000.000$  zeigt sich der Effekt:

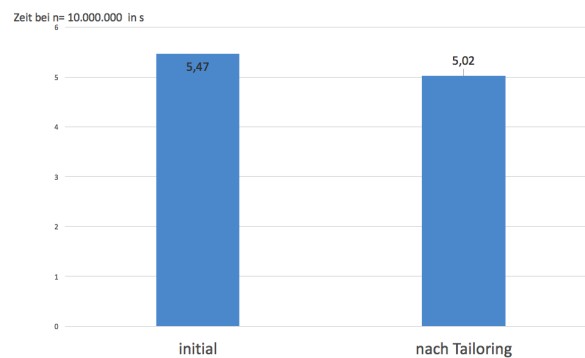


Abbildung 8: Zusammenfassung der Optimierung von Algorithmus 3

## 5.4 Wie wird gemessen?

DRT Damit soll Optimierung in diesem Projekt abgeschlossen betrachtet ... und gehen ins Rennen für Zeitmessungen.  
Fertig zu Nutzung unter 05...

Die Laufzeit wird mithilfe der Funktion *time()* des python-Moduls *time* gemessen. Die Funktion *time()* liefert die seit dem 1.1.1970 vergangene Zeit in Sekunden als Gleitkommazahl. [6] Nach dem folgenden Schema wird somit die Laufzeit errechnet und in der Konsole ausgegeben.

---

```
import time

#read file...

#start time measuring
start = time.time()

#actual algorithm...

#stop time measuring
end = time.time()

#print it to console
elapsedTime = str(end-start)
print elapsedTime.replace('.', ',')

#close file
#write output (commented out)
```

---

DRT Für messen ganzer Reihe so:

1. Erzeugen Ausgangsdaten per *testDataGenerator.py* oder Java Anwendung *real-DataGetter*
2. Erstellen mehrerer Playlists mit steigender Titelzahl mittels *cutToLen.py* und *shCutToLen.sh*
3. Ausführen der zu untersuchenden Implementierung für alle generierten *.txt*-Files mittels shell-Skript *shRunScript.sh*

DRT Das sh skript und die zur messung bla sind in 04

DRT Was fangen wir damit an?

1. Vglf Test real
2. theo lauf
3. grafischer vgl
4. statistischer hypotest

## 6 VERGLEICH ECHT VS. TEST-DATEN

DRT Genommen wird: Echt-Daten: 3 versch Ausg-künstlert Liste entspr gekürzt uaf n 3x Test Daten der länng original erfragte list, dup nachempfunden auch auf n Messungen 3x Mittelwert:  
Ergebnis so:

## 6.1 Haupt-Algorithmus

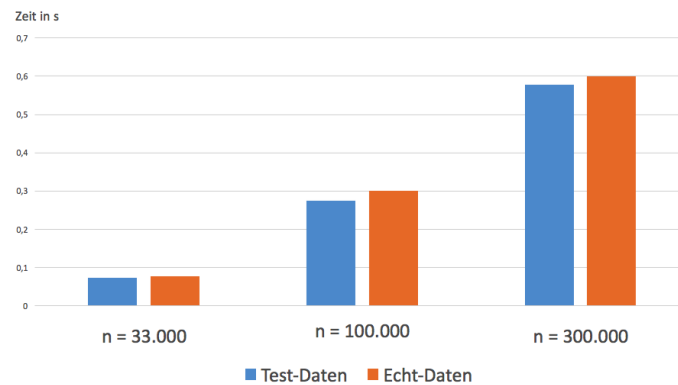


Abbildung 9: DRT

DRT

Kann man jetzt nicht so genau sagen. Nehmen aber im Folgenden REAL

## 6.2 Vergleichs-Algorithmus 1

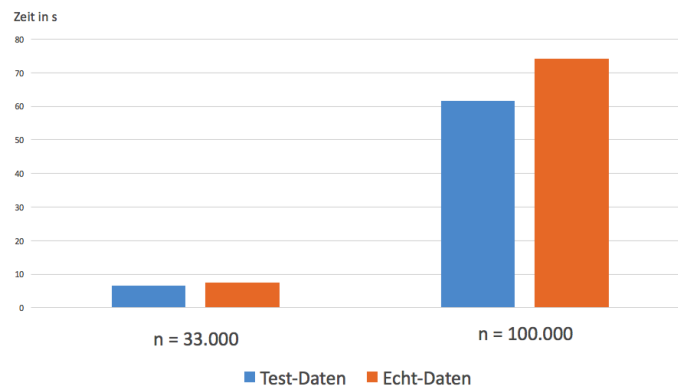


Abbildung 10: DRT

DRT

Analog 1 Nach dem Vergleich: Auf Echt.

## 6.3 Vergleichs-Algorithmus 2

DRT

Hier halt nicht so stark. Nach dem Vergleich: Auf Echt.

# 7 VERGLEICH MIT THEORETISCHER LAUFZEIT

betrachtet: Aufscheteigendes n

Dazu: Boxplot.

Als versch Werte pro n:

- Drei Versch Ausgangskünstler

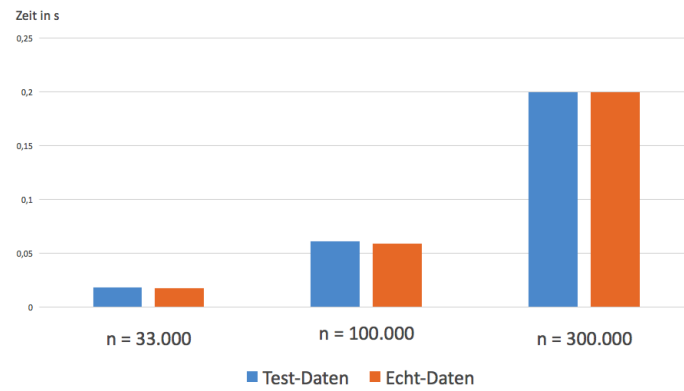


Abbildung 11: DRT

- Drei mal

Einmal so, zum besseren Betrachten Skalieren mit theo laufz.

## 7.1 Haupt-Algorithmus

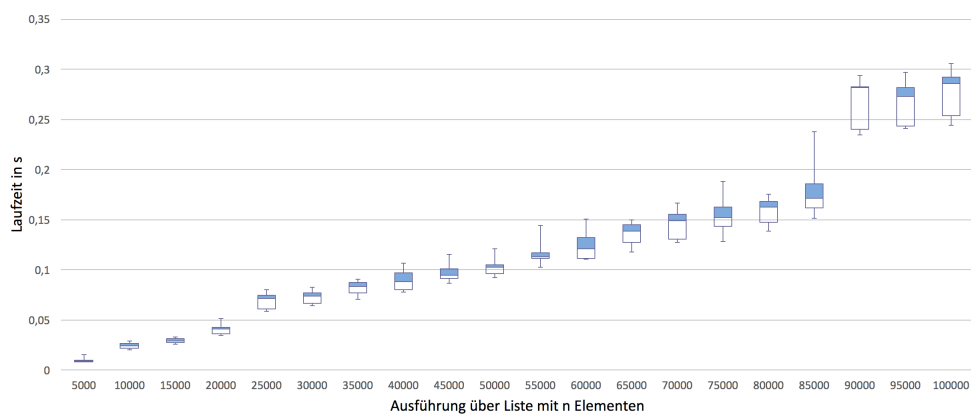


Abbildung 12: DRT

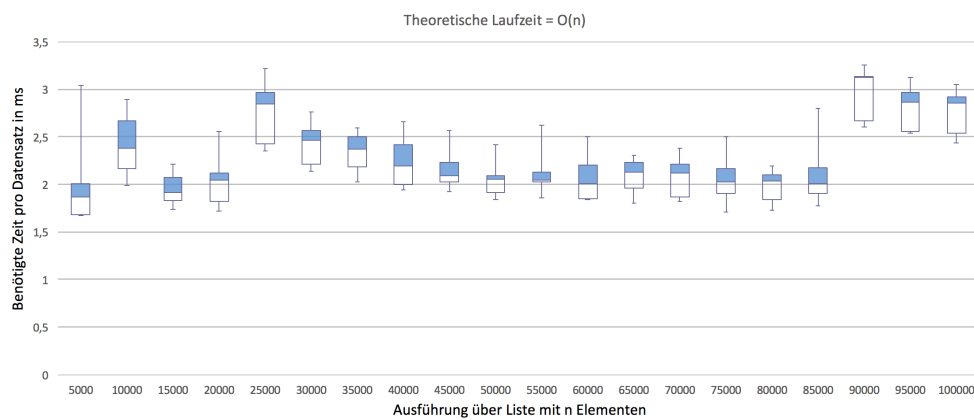


Abbildung 13: DRT

## 7.2 Vergleichs-Algorithmus 1

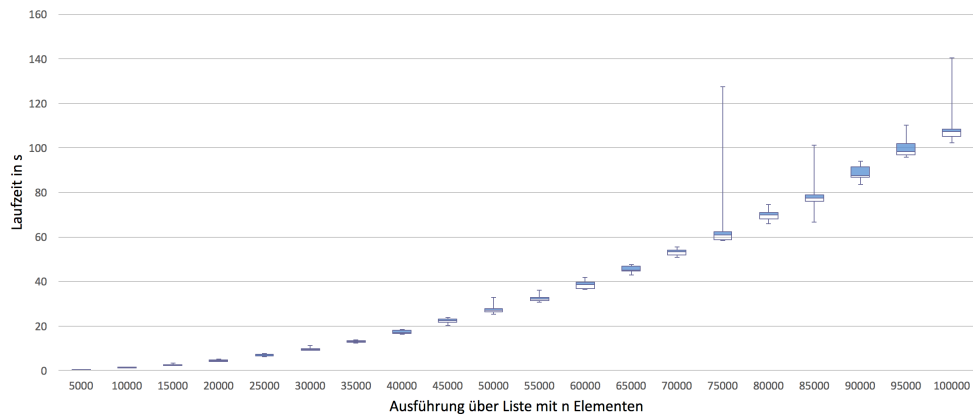


Abbildung 14: DRT

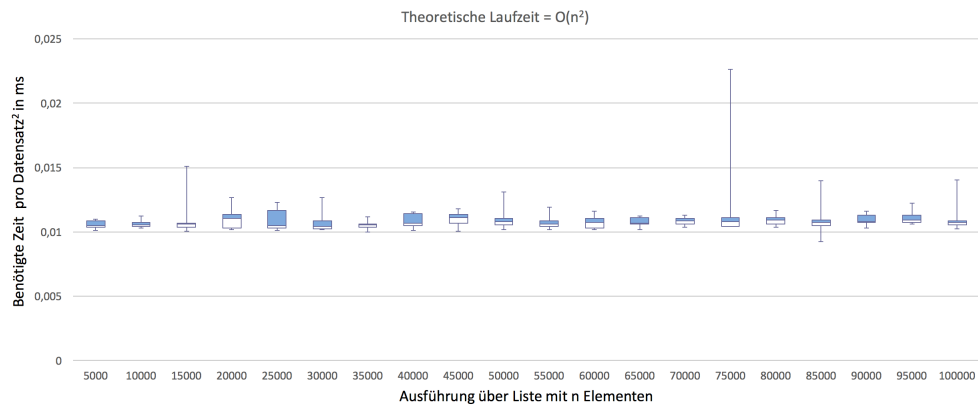


Abbildung 15: DRT

## 7.3 Vergleichs-Algorithmus 2

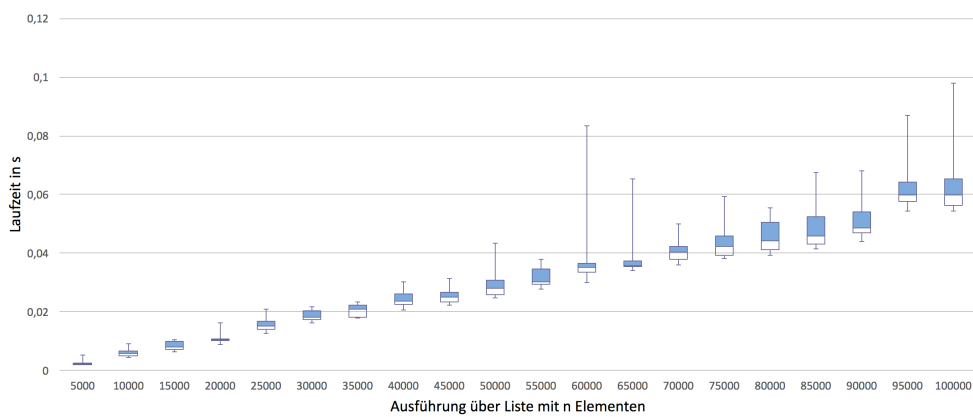


Abbildung 16: DRT

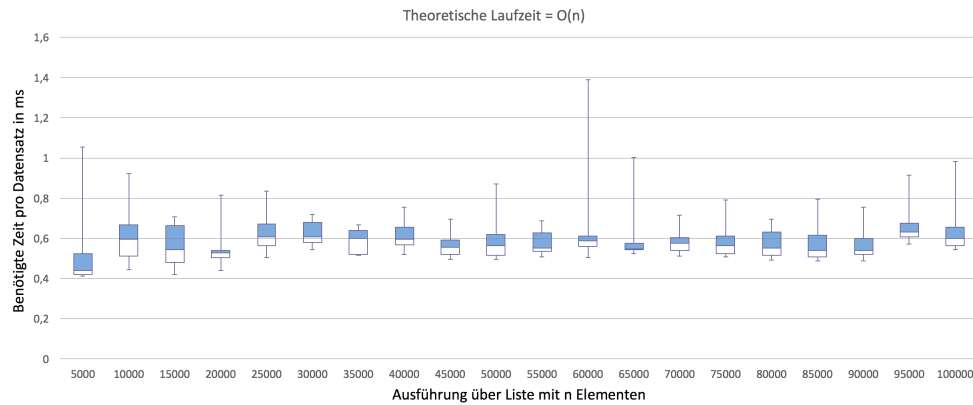


Abbildung 17: DRT

## 8 VERGLEICH DER ALGORITHMEN

### 8.1 Grafischer Vergleich

Laufzeitmessung für folgende Problem instanzen durchgeführt:

- Für wachsendes  $n = \{20, 40, \dots, 1000\}$   
 Pro  $n$ : 10 verschiedene Echtdaten-Tracklisten  
 Pro Liste: 3 Messdurchläufe

DRT Die Messreihen wurden über Ablehnung von Messungen außerhalb des Bereiches  $[\mu - \sigma, \mu + \sigma]$  bereinigt, wobei  $\mu$  der Erwartungswert und  $\sigma$  die Standardabweichung der Messwerte ist.

Hier mal Laufzeiten aller drei in einem Graph.

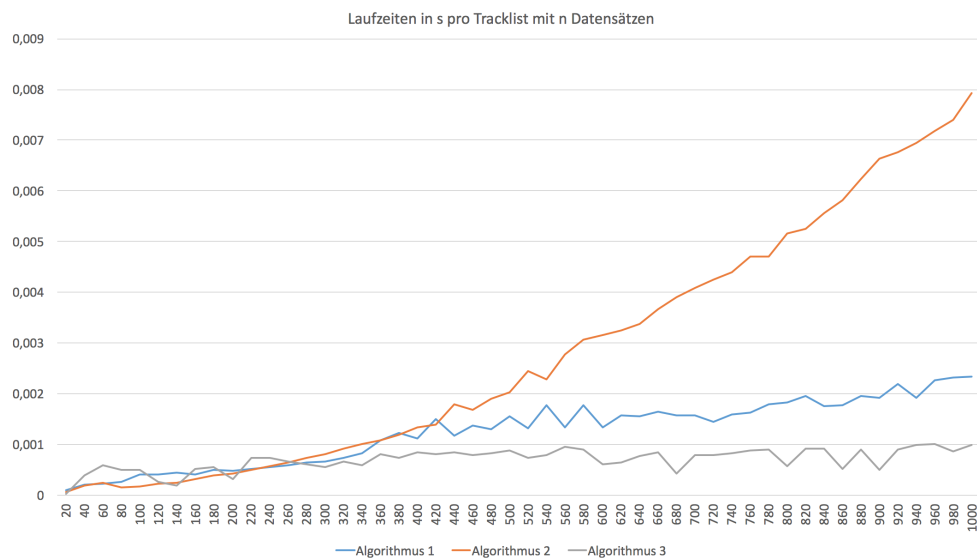


Abbildung 18: DRT

Um noch bisschen besseren Eindruck zu gewinnen:  
 Für 1 vs. 2 mal höhere laufzeiten geprüft

- Für wachsendes  $n = \{2000, 4000, \dots, 100.000\}$   
 Pro  $n$ : 5 verschiedene Echtdaten-Tracklisten  
 Pro Liste: 3 Messdurchläufe



wurde aber nur für 5 Interpreten gemacht, weil so harter Ansatz zu hart und abfrage API doch zeitintensiv.

Sieht so aus:

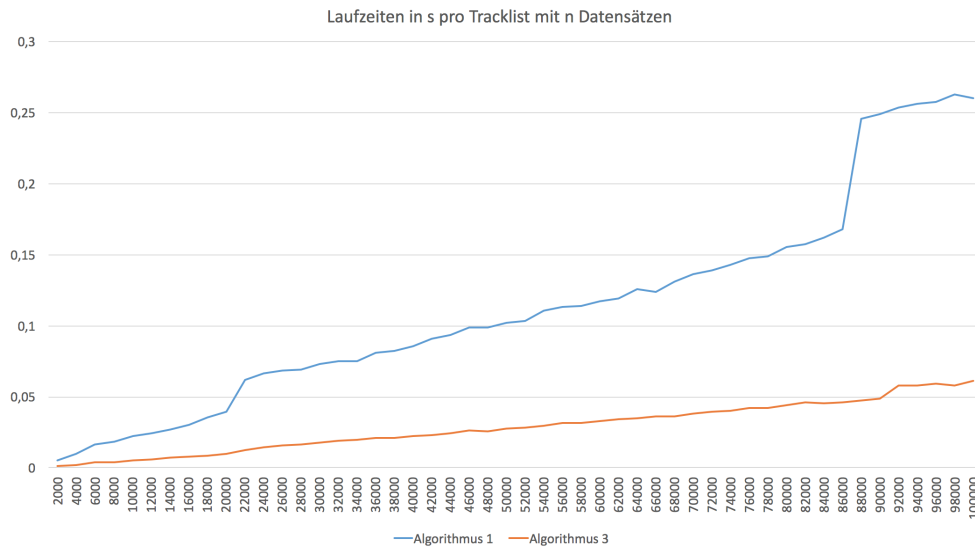


Abbildung 19: DRT

Man erkennt bei erstem deutliche Sprünge

Bei betrachtung von Algo siehe kapitel wird erkenntlich, dass das die übertragung in neue tabelle im Pseudocode in zielel ist

Das ist im original dict praktisch auch zu erhahnen

## 8.2 Statistischer Hypothesen-Test

Wenn man sich die Kurve anschaut, Grafisch anzunehmen:

$$\text{Alg3}_{\text{Laufzeit}} < \text{Alg1}_{\text{Laufzeit}} < \text{Alg2}_{\text{Laufzeit}}$$

Da können wir eigentlich nur mal schaeun ab welcher größe playlist das so stimmt. Per Hypothesentest also drei Tests:

1. Algorithmus 1 schneller als Algorithmus 2
2. Algorithmus 3 schneller als Algorithmus 2
3. Algorithmus 3 schneller als Algorithmus 1

Messungen finden dabei im Bereich  $n = \{20, 40, \dots, 1000\}$  statt.

### Algorithmus 1 schneller als Algorithmus 2

Vermutung ist alsi Algorithmus 1 ist schneller als Algorithmus 2

Null-Hypothese daher:  $H_0 : \mu_1 \geq \mu_2$

Entsprechend ergibt sich Alternativ-Hypothese:  $H_0 : \mu_1 < \mu_2$

versuch Nullhypothese zu widerlegen. Signifikanzniveau  $\alpha = 0,02\%$  und  $n = 30$  Messungen, Festlegung Fehler-Wahrscheinlichkeit: 2% ????

Bilde Differenz der Laufzeiten, welche auf selber Problemistanz gemessen wurden:

$$d_i = \text{MesswertAlg2}_i - \text{MesswertAlg1}_i$$

Errechne t-Wert  $t = \sqrt{n} \frac{\bar{d}}{\sigma_d}$  und vergleiche mit  $t(0,98;29) \approx 2,15$

Nullhypothese ablehnen, wenn  $t < 2,15$

| n    | $t = \sqrt{n} \frac{\mu_d}{\sigma_d}$ | $t(0,98;29)$ | Ergebnis                       |
|------|---------------------------------------|--------------|--------------------------------|
| 20   | -4,675596311                          | 2,15         | Null-Hypothese nicht abgelenht |
| 40   | -2,267865616                          | 2,15         | Null-Hypothese nicht abgelenht |
| ...  | ...                                   | 2,15         | ...                            |
| 400  | 1,551055352                           | 2,15         | Null-Hypothese nicht abgelenht |
| 420  | 0,540433623                           | 2,15         | Null-Hypothese nicht abgelenht |
| 440  | 4,598993754                           | 2,15         | Null-Hypothese abgehelhnt      |
| 460  | 3,17978148                            | 2,15         | Null-Hypothese abgehelhnt      |
| ...  | ...                                   | 2,15         | ...                            |
| 1000 | 35,12686104                           | 2,15         | Null-Hypothese abgehelhnt      |

Tabelle 3: Ergebnisse des Tailorings: Benötigte Zeit gesamt und für hash-Funktion

### Algorithmus 3 schneller als Algorithmus 2

Algorithmus 3 schneller als Algorithmus 1  
Königsdisziplin.

## 9 FAZIT

wen schon py dann standard funktionen

## LITERATUR/QUELLEN

- [1] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms, Third Edition*. The MIT Press, 3rd edition, 2009.
- [2] Jonathan Hartley. Time complexity - python wiki. <https://wiki.python.org/moin/TimeComplexity>. Zuletzt bearbeitet: 05.06.2017, Zugriff: 01.02.2018.
- [3] Spotify uris and ids - spotify web api user guide. <https://developer.spotify.com/web-api/user-guide/#spotify-uris-and-ids>. Zugriff: 01.02.2018.
- [4] A java wrapper/client for the spotify web api. <https://github.com/thelinmichael/spotify-web-api-java>.
- [5] The python standard library documentation - 26.4.3. profile and cprofile module reference. <https://docs.python.org/2/library/profile.html#module-cProfile>. Zugriff: 01.02.2018.
- [6] The python standard library documentation - 15.3. time — time access and conversions. <https://docs.python.org/2/library/time.html#time.time>. Zugriff: 02.02.2018.