Validierung und statistische Auswertung von dblp.xml

Softwaredokumentation

Informatik Praktikum (Gr. Studienprojekt) 2014/2015

Betreuer: Dr. Michael Ley

Praktikant: Serge Oliver, Djomo Moungoue

Matrikelnummer: 1050291

Meilensteine: Erster Termin 🡪 Softwarespezifikation 🡪 Softwaredokumentation

Daten: 30.10.14 27.11.14 17.03.15

Inhaltsverzeichnis

**I - Einführung3**

**II - Softwarespezifikation3**

II.1 – Ausgewertete Statistiken (Darstellung als Säulendiagramm)5

II.2 – Häufigkeitsverteilung von Daten (Darstellung als Box-Plot-Diagramm)5

**II – Inhalt des Softwarepakets3**

II.1 – Quellcode ( 136 KB) und Output-Dateien (3.44 MB)5

**IV – Entwicklungs- und Testumgebung3**

**V – Veranschaulichung von Endergebnissen3**

V.1 – Startseite mit der Liste von Statistiken5

V.2 – Darstellung von Säulendiagramm und Box-Plot-Diagramm5

V.3 – UML-Paketdiagramm5

V.4 – Liste von Aureißer5

V.5 – Baumstruktur von dblp.xml5

**VI - Schluss3**

# **I – Einführung**

Die dblp computer science bibliography ist ein online Verweis für bibliographische Informationen über meistens Publikationen im Bereich Informatik. Binnen einem halbjährlichen Studienprojekt wurde ich aufgefordert eine Software zu entwickeln, welche die zugrunde liegende Datei der dblp *(dblp.xml)* validiert und statistisch auswertet. Die dblp.xml wurde am Mittwoch, dem 12. November 2014 von der dblp Webseite auf meinem Laptop heruntergeladen. Die war im Zeitpunkt des Downloads circa 1.4 GB groß. Die entwickelte Software speichert die generierten Daten je nach Art und Zweck in html, csv und log Dateien. Die html Dateien enthalten die Daten zur Darstellung von Säulendiagrammen und Box-Plot-Diagrammen auf die Website. Die csv und log Dateien bestehen jeweils aus den über die Häufigkeitsverteilung von Daten ermittelten Werten und einigen Ausreißern. Die Software lief circa zwei Minuten in der Entwicklungsumgebung.

# **II – Sofwarespezifikation**

## **II.1 – Ausgewertete Statistiken (Darstellung als Säulendiagramm)**

Zahl von

Autoren/Editoren Namen, die gleich viel Zeichen umfassen

### Titeln, die gleich viel Zeichen umfassen

### Titeln, die gleich viel Wörtern umfassen

### Querverweisen mit gleichem Umfang von Seiten

### Büchern mit gleich viel Querverweisen

### bearbeiteten Publikationen pro Monat

### bearbeiteten Publikationen pro Jahr

### Autoren/Editoren, die in demselben Jahr für das erste Mal publizierten

### generierten elektronischen Versionen von Publikationen pro Jahr

### Feldern pro Publikation

## **II.2 – Häufigkeitsverteilung von Daten (Darstellung als Box-Plot-Diagramm)**

Häufigkeitsverteilung von

### Zeichen in Autoren/Editoren Namen

### Zeichen in Titeln

### Wörtern in Titeln

### Seiten in Querverweisen

### Querverweisen in Büchern

### bearbeiteten Publikationen pro Monat

### bearbeiteten Publikationen pro Jahr

### ersten Publikationsjahr von Autoren/Editoren

### generierten elektronischen Versionen von Publikationen pro Jahr

### Feldern pro Publikation

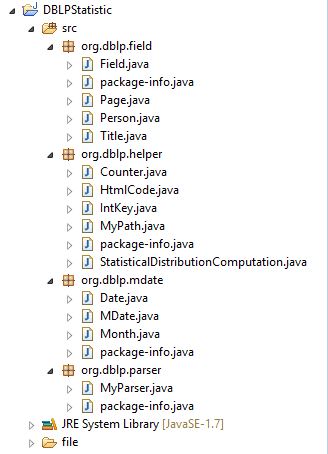
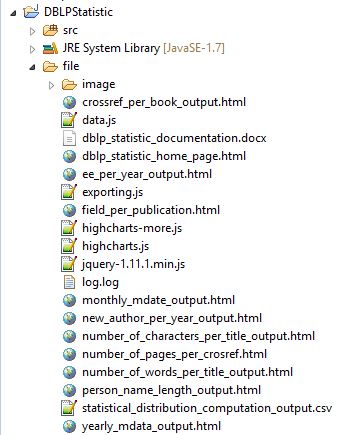
Liste von berechneten Maßen

* Der obere Whisker
* Das obere Quartil
* Die Median
* Das untere Quartil
* Der untere Whisker
* Die Varianz
* Die Standard Abweichung
* Die Gesamtsumme

Ausführliche Liste in der Datei statistical\_distribution\_computatio\_output.csv

# **III – Inhalt des Softwarepakets**

## **III.1- Quellcode ( 136 KB) und Output-Dateien (3.44 MB)**



**IV – Entwicklungs- und Testumgebung**

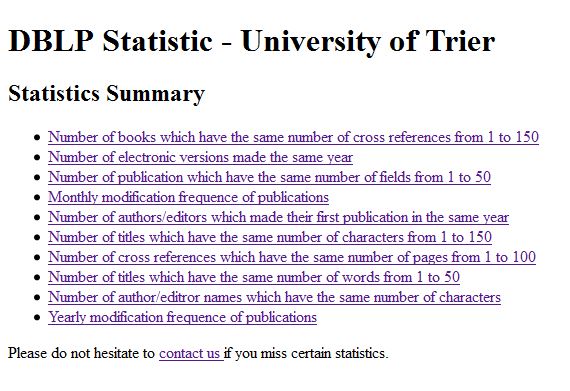
|  |  |
| --- | --- |
| IDE | Eclipse SDK Version 3.8.0 |
| VM Arguments | * ${-DensityExpensionLimit=2500000} * ${-Xms1024M} * ${-Xms1024M} |
| Externe Bibliotheken | * xerces.jar * jquery-1.11.1.min.js * highcharts.js * highcharts-more.js * data.js * exporting.js |
| Browser | Mozilla Firefox 36.0.1 |
| Betriebssystem | Microsoft Windows 8 |
|  |  |

# 

# **V – Veranschaulichung von Endergebnissen**

## 

# **V.1 – Startseite mit der Liste von Statistiken**



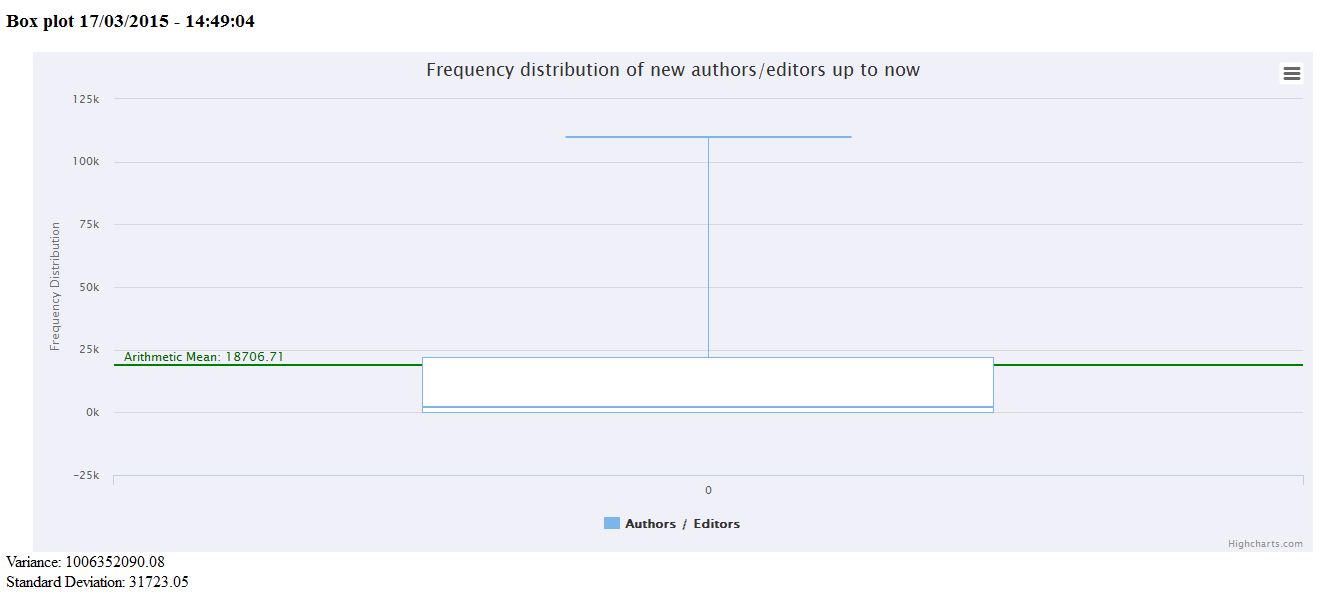
**Abbildung 1** Startsete, die die Liste von Statistiken enthält

# **V.2 –Darstellung von Säulendiagramm und Box-Plot-Diagramm**

# **Säulendiagram**

**Abbildung** **2** Anzahl von Autoren/Editoren die publizierten für das erste Mal demselben Jahr

# **Box Plot-Diagramm**

****

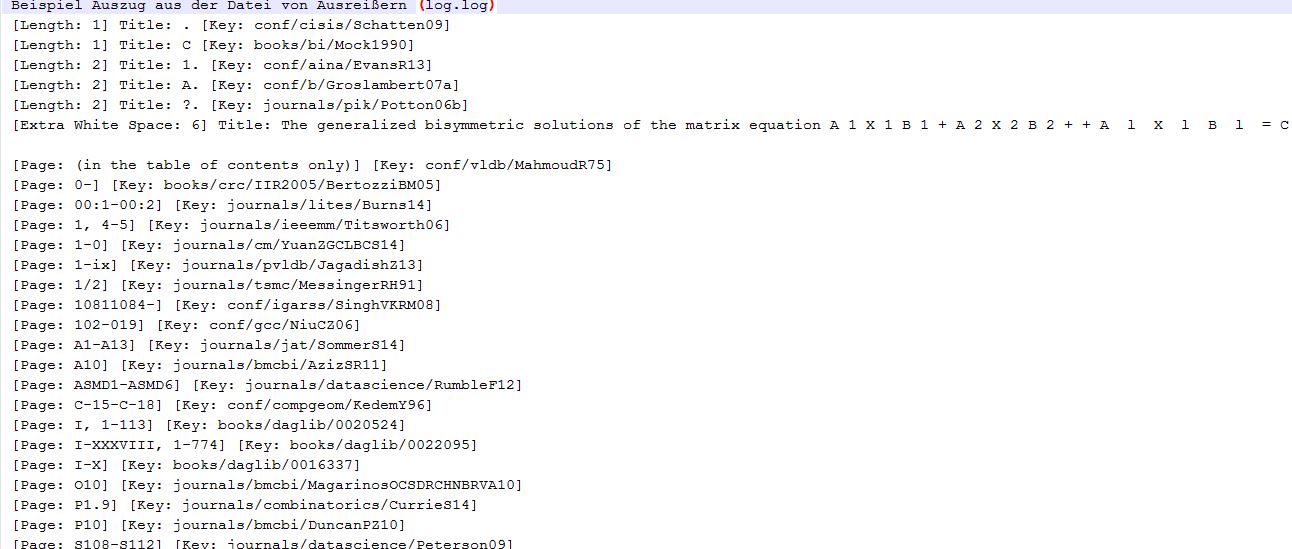
**Abbildung 3** Fünf-Nummer Verteile über die neuen Autoren/Editoren pro Jahr

# **V.3 – UML- Paketdiagramm**

## 

**Abbildung 4** Grobe Darstellung der Inhalt des Projektpaket

## **V.4 – Liste von Ausreißern**



**Abbildung 6** Auszug aus der log Datei

Manche Publikationen haben eine Groß Zahl von Felder wegen zu viele *cite* Felder

**Zum Beispiel:** Book key: books/aw/AbiteboulHV95 enthält 751 Felder (circa 740 *cite* Felder)

Ausführliche Liste in der Datei *log.log*

# **V.5 – Baumstruktur von dblp.xml**

# 

**Abbildung 5** Baumstruktur der dblp.xml

# **VI – Schluss**

Die Software ist laut die durchgeführten Tests zuverlässig und hat eine Laufzeit von circa zwei Minuten auf meinem Laptop. Alle Diagramme bis auf das Box-Plot-Diagramm der Statistik über Felder in Publikationen lassen sich gut darstellen. Ich habe die Ursache bislang nicht herausfinden können.