

**Dokumentation Belegaufgabe**

**Häufigkeitsberechnung mit Spark**

Belegarbeit im Modul

**Programmierkonzepte und Algorithmen**

Hochschule für Technik und Wirtschaft (HTW) Berlin

Fachbereich 4: Informatik, Kommunikation und Wirtschaft

Studiengang Master *Angewandte Informatik*

Felix Riehm (xxxxxx)

Nicolas Tschirr (561206)

15.06.2020

Inhaltsverzeichnis

[1 Beschreibung und Erklärung der Aufgabe 1](#_Toc42424943)

[2 Lösungsbeschreibung 1](#_Toc42424944)

[3 Code-Fragmente 1](#_Toc42424945)

[4 Screenshots der Ergebnisse 1](#_Toc42424946)

[5 Tests der Anwendung 1](#_Toc42424947)

[6 Graphische Visualisierung 1](#_Toc42424948)

[7 Fazit 1](#_Toc42424949)

# 1 Beschreibung und Erklärung der Aufgabe

Bei der bearbeiteten Aufgabe sollte eine Häufigkeitsberechnung von Wörtern aus verschiedenen Texten und Sprachen vorgenommen werden. Es gab als Grunddatensatz eine Vielzahl von verschiedenen Texten in acht verschiedenen Sprachen. Um die Textdateien, welche als .txt Format zur Verfügung standen, zu analysieren, wurde das Framework für verteiltes Rechnen bzw. Clustercomputing Apache Spark verwendet.

Die Häufigkeitsberechnung wurde dabei in zwei Unteraufgaben aufgeteilt. Im ersten Schritt sollte für jeden vorhandenen Text eine eigene Häufigkeitsberechnung durchgeführt werden, sodass die vorkommende Anzahl der Wörter in jedem vorliegenden Text einsehbar ist. Im zweiten Schritt wurden die Häufigkeiten dann nach Sprache gruppiert und es sollte eine Liste der zehn am häufigsten vorkommenden Wörter pro Sprache erzeugt werden. Bei dieser Liste sollten die Stoppwörter, also für den Text keine Relevanz besitzende Wörter, ausgeschlossen werden.

# 2 Lösungsbeschreibung

# 3 Code-Fragmente

# 4 Screenshots der Ergebnisse

# 5 Tests der Anwendung

# 6 Graphische Visualisierung

Durchlaufen mit 8 virtuellen Kernen: 4:49

Durchlaufen mit 4 virtuellen Kernen: 6:08

doppelt durchlaufen: 09:50

dreifach: 14:38

Sprachen einzeln durchlaufen:

Dutch: 10s

English: 1:08 (1:18)

French: 1:07 (2:25)

German: 1:09 (3:34)

Italian: 1:08 (4:42)

Russian: 1:25 (6:07)

Spanish: 0:34 (6:41)

Ukrainian: 0:53 (7:34)

# 7 Fazit