Zusammenfassung

Es besteht die Notwendigkeit, um die große Menge der Daten von CareLink-Website für Forschungszwecke in TK-Labors herunterzuladen. CareLink, das webbasiertes System ist entwickelt, um die Informationen von allen Diabetes-Management-Tools wie Insulin-Pumpe, kontinuierliche Glukose-Monitore, Blutzuckermessgeräte und Logbuch zu übernehmen und sie in leicht zu lesende Diagramme, Graphen und Tabellen zu organisieren. Diese Daten werden für die Erzeugung von Muster mithilfe des maschinellen Lernalgorithmus verwendet. Die Anforderung ist, diese Daten für Forscher leicht zugänglich zu machen, ohne tatsächlich die Website zu besuchen und dieser gesamter Prozess wird mit der Crawling-Methodik durchgeführt.

Crawling ist der Prozess, durch den wir verschiedene Sorten von Webseiten sammeln, um die Informationen von ihnen zu sammeln. Es ist das System für das Massen-Herunterladen von Webseiten und kann auch als ein Programm oder ein automatisiertes Skript aufgerufen werden, dass das Web automatisiert scannt. In dieser wissenschaftlichen Arbeit wird eine native Java-Anwendung für das Herunterladen der großen Mengen der Daten vom Server aufgebaut.

Gleichzeitig ist es auch notwendig, Daten vom lokalen System/USB zum Server hochzuladen. In der CareLink Website wird der Prozess des Hochladens mit Applet behandelt. Es gibt technische Schwierigkeiten, Applet aus dem Browser heraus zu nehmen und diesen Prozess automatisiert zu machen. Wir haben versucht, diesen langwierigen Prozess zu vermeiden, in dem die Webbrowser-Steuerung automatisiert wird. Nativ Java-Kompatibilität mit Selen als externe Open-Source-Bibliothek wird für den Automatisierungsbrowser verwendet, der wiederum den Prozess des Hochladens einfacher und nicht wiederholend für Endbenutzer macht.

Diese wissenschaftliche Arbeit versucht auch das Problem des Daten-Slicers zu lösen, wobei ein Algorithmus erstellt wird, der eine Reihe von Ereignissen in Abhängigkeit von einigen Kriterien erkennt und die geschnittenen Daten in den lokalen DB speichert. Die Daten, die aus verschiedenen Quellen für Analysemuster gesammelt wurden, sind meist wiederholt, schmutzig und verpassen zahlreiche Spalten von Daten. Datenanalyse/Maschinelles Lernen brauchen qualitativ hochwertige Daten zu haben, so dass die Schlussfolgerungen gezogen werden können, die für die Analyse sinnvoll sind. Für das Daten-Slicing wird der Algorithmus in Java aufgebaut und verwendet JDBC und sql mit einem Pub-Sub-Komplex-Ereignissystem, um Ereignisse zu erkennen. Für das komplexes Ereignissystem in Java wird JMS (Java Messaging Service) verwendet. JMS ist eine Java Message Oriented Middleware (MOM) API zum Senden von Nachrichten zwischen zwei oder mehrere Kunden. Es ist eine Implementierung, um das Problem des Produzenten-Verbrauchers zu behandeln. Vor der Implementierung des Code in der Produktion, ist es notwendig, Unit-Tests für jede Module zu implementieren und damit Junit wird verwendet, um Unit-Testing zu implementieren. Unit-Testing ist ein Prozess der Softwareentwicklung, bei der die kleinsten testbaren Teile einer Anwendung, sogenannten Einheiten, unabhängig für ordnungsgemäße Operationen überprüft werden. Die gesamte Projektarbeit muss in zwei Versionen, GUI-basiertes Projekt und Command Line Interface aufgebaut werden. Das Command Line-Projekt verwendet verschiedene Flaggen, um verschiedene Aktivitäten innerhalb des Projekts zu beschreiben und wird mit Apache Commons CLI erstellt, das eine externe Bibliothek für Java ist. JavaFx wird als Swing-Framework für die Erstellung eines GUI-basierten Projekts verwendet. Außerdem wird JSoup als Java-Framework verwendet, um die Website zu durchsuchen und Daten von der Website zu extrahieren oder herunterzuladen. Die langwierige Programmierungsautomatisierung erfolgt so, dass das Hochladen und das Herunterladen benutzerfreundlich, einfach und ein Klick ist. Sobald die Daten heruntergeladen und hochgeladen sind, erfolgt die Datenvalidierung und -überprüfung mithilfe von Forschern des TK Labs, die an Data Mining Technologies arbeiten.