Universität Trier
Professur für Softwaretechnik
Prof. Dr. Stephan Diehl
Lucas Kreber, M.Sc.

Fortgeschrittene Softwaretechnik
(WS 2024/25)
Fünftes Übungsblatt

Mining von Softwarearchiven / Quantitative Analyse

Aufgabe 1: Relative Häufigkeit mit BOA (4 Punkte)

a) Implementieren Sie ein BOA-Skript, das für jedes Projekt in temporären Variablen (vom Typ float) die Anzahl der Statements und Try-Statements für den aktuellen Stand jedes Projektes (also alle letzten Dateirevisionen) berechnet und das Verhältnis der beiden Zahlen (also wieviel Prozent der Statements sind Try-Statements) für jedes Projekt ausgibt. Anschließend führen Sie Ihr Skript auf den Datensätzen 2022 Feb/Python und 2022 Jan/Java aus und laden die Ausgabe herunter.

b) Erweitern Sie Ihr Skript so, dass es das Verhältnis der beiden Zahlen für den Stand des Projektes am Ende jeden Jahres (2000 bis 2022) berechnet und ausgibt.

Hinweis: In Python gibt es die Anweisung `open ... as ...`, die syntaktischer Zucker für ein komplexes TRY-Statement ist. Für den Python-Datensatz gibt es daher das zusätzlich Attribut `StatementKind.WITH`.

Aufgabe 2: Statistische Auswertung mit Python (3 Punkte)

Für diese Aufgabe erweitern Sie das Python-Skript `uebung5.py`, das Sie in StudIP finden. Da wir zwei sehr große Datensätze (aus Augabe 1a) haben, gehen wir von der schwächeren **Annahme aus, dass diese nicht normalverteilt** sind und verzichten darauf auf Normalverteilung zu testen. Als Signifikanzniveau verwenden Sie $\alpha = 1\%$.

a) Wählen Sie einen geeigneten statistischen Test mit Hilfe des Entscheidungsbaums aus der Vorlesung zur quantitativen Analyse aus und berechnen Sie mit der in den Modulen `scipy.stats` oder `researchpy` verfügbaren Implementierung dieses Tests, ob der Unterschied der Mittelwerte für Java- und Python-Projekte statistisch signifikant ist. Der Test liefert Ihnen auch die Effektstärke (Cohen's d).

b) Bestimmen Sie experimentell eine Schranke für die Sample-Größe (Parameter `sampleSize` von `readSample()`), ab der der statistische Test nicht mehr signifikant ist.

Aufgabe 3: Interpretation (1 Punkte)

Interpretieren Sie die Effektstärke für die gesamte Stichprobe (Aufgabe 2a). Welchen Schluss kann man aus Aufgabenteil 2b für Studien zu ähnlichen Fragestellungen ziehen.

Abgabe: Dienstag, der 14.01.2025 zum Vorlesungsbeginn via StudIP

Abgabeformat: Die Abgabe der Übungsblätter erfolgt über Stud.IP. Bitte beachten Sie die folgenden Hinweise. Abgaben, die nicht dem beschriebenen Format entsprechen, werden ignoriert und mit 0 Punkten bewertet:

- Bei **normalen Übungsblättern**:
  Die Abgabe erfolgt als **einzelne PDF-Datei** (keine Ordner, kein ZIP-Archiv, nicht in mehreren Teilen, keine anderen Dateiformate). **Bitte vermerken Sie in dieser PDF-Datei gut lesbar Ihren Namen und Ihre Matrikelnummer.** Verwenden Sie folgendes Schema für den Dateinamen der PDF-Datei: `fst24-uebxx-123456` (`xx` durch Nummer des entsprechenden Übungsblattes ersetzen, z.B. ueb01 für das erste Blatt, und `123456` durch Ihre Matrikelnummer ersetzen).

- Bei **Übungsblättern mit Programmieraufgaben**:
  Neben der evtl. vorhandenen PDF-Datei soll nur der Quellcode abgeben werden (also keine `class`-Dateien o.Ä.). **Bitte vermerken Sie in jeder Quellcodedatei Ihren Namen und Ihre Matrikelnummer als Kommentar.** Packen Sie in diesem Fall alle benötigten Dateien (Quellcode und ggf. PDF-Datei) als **ZIP-Datei** mit dem Namen `fst24-uebxx-123456.zip` (`xx` durch Nummer des entsprechenden Übungsblattes ersetzen, z.B. ueb01 für das erste Blatt, und `123456` durch Ihre Matrikelnummer ersetzen).