Entwurf
Version 0.8

CIP – Projekt FS 2020 – Zeitplan + Deliverables

Die/der Gruppenleiter*in ist die **Schnittstelle zum Dozenten**. Wie in einem Praxis-Projekt. Diskutieren Sie viel untereinander und fragen Sie bei Unklarheiten (über die/den Gruppenleiter*in) den Dozenten!

10.3.2020 (23.59 Uhr)

Der "Plan" für Ihr Gruppen-CIP-Projekt muss schriftlich formuliert und elektronisch (per Mail) an den Dozenten geschickt werden (erwin.mathis@hlsu.ch).

Überlegen Sie sich: Welche Fragestellung(en) wollen wir mit dem CIP-Projekt beantworten? (Umfang max. 2 A4 – Seiten => Word File)

13.3.2020

Besprechung der CIP-Projekt-Idee mit Dozenten.
Die "abgesegnete" Projekt-Idee gilt als eine Art "Vertrag" für Ihr CIP-Projekt.

17.10.2020 (23.59 Uhr)

=> hat am Rande auch mit dem CIP-Projekt was zu tun …!
Abgabe der persönlichen Crawler-Aufgaben auf Ilias (Optional => Erklärung in Vorlesung!)

14.3. – 27.3.2020

(Vorlesungszeit + Homeworkzeit)
Arbeiten an Ihrem Projekt z.B.

- **Tableau Prep:** Knowhow-Aufbau (kann in anderen Vorlesungen wieder verwendet werden!)
  Ein grosser Teil (d.h. 60 – 80%) des CIP-Projekt soll in Tableau-Prep 'erledigt' werden.
- **Daten crawlen** mit Beautiful Soup 4 von mindestens einer zu ihrem CIP-Projekt passenden Website (nicht von einem Rest-Service oder API etc. crawlen oder nur in Absprache mit dem Dozent (über den "Vertrag"))
  Diese 'gecrawlten' Daten müssen eventuell einem Cleaning-Prozess unterzogen werden z.B. über Pandas oder nativ Python
- **Pandas** muss irgendwo im CIP Projekt verwendet werden z.B. um die 'gecrawlten' Website-Daten zu bereinigen … oder …
  Beispiele was mit Pandas gemacht werden kann:
  - Fehlende Daten herausfiltern,
  - Fehlende Daten ev. sinnvoll ersetzten
  - (dokumentieren und begründen, warum diese Ersetzung so gewählt wurde!)
  - ev. überflüssige Duplikate ersetzen,
  - Daten transformieren
  - Ausreisser erkennen
  - Reguläre Ausdrücke einsetzen (z.B. beim Suchen und Ersetzen …)
  - ….
- **Export (Load)** der bereinigten Daten auf die bereitgestellte **SQLServer** Datenbank

27.3.2020 (8.15) Uhr eine kleine Probe-CIP-Modulendprüfung: **Obligatorisch für ALLE!**

3.4.2020 (oder schon früher) - **Deadline:** 3.4.2020 – 11.30 Uhr

Erstellen einer **Videoaufnahme mit Ton** oder eines **Screencast mit Ton** der eigenen Präsentation des CIP-Projekt (mit PowerPoint).
Dieses **Video/Screencast-Dokument** muss dem Dozenten über einen **Download-Link** oder über z.B. einen **USB-Stick** zur Verfügung gestellt werden (spätestens bis am 3.4.2020 um 11.30 Uhr)

Ebenfalls elektronisch muss dem Dozenten über einen **Download-Link** oder über z.B. einen **USB-Stick** zur Verfügung gestellt werden (wieder spätestens am 3.4.2020 um 11.30 Uhr):

Ein **Zip-File**, welches **alle notwendigen Daten**, alle **Skript's (z.B. Python, Pandas), Tableau-Prep-Files** etc. plus auch die **Projekt-Beschreibung**, die **Präsentation** und die **3 oder 4 Zeiterfassungs-Excel Files** enthält, damit das Projekt auch ausserhalb ihrer individuellen virtuellen Maschine nachvollzogen werden kann.
Es geht darum, dass Sie ihre Projektbeschreibung etc. und dazugehörenden Daten allen anderen Studierenden zur Verfügung stellen sollen – als '**Beispiel-Rucksack**' für die Praxis.
Falls gewisse Daten (von der Filegrösse her, aus rechtlichen Gründen etc.) nicht weitergegeben werden können, muss das mit dem Dozenten explizit besprochen werden.

Weiter muss dem Dozenten je **1x ausgedruckt** (ja echt auf 'richtigem' Papier 😉!) spätestens am 3.4.2020 um 11.30 Uhr zur Verfügung gestellt werden:

- **Projekt-Beschreibung** (Inhalts-Details siehe unten). Die Projektbeschreibung muss so erstellt werden, dass Sie für alle IDS-Studiengänger FS2020 verständlich ist und problemlos nachvollzogen werden kann. Screenshots von Tools / Daten sind erlaubt und erwünscht, wenn Sie das Verständnis der Arbeit fördern. Fliesstext > 10 Zeilen "stinkt" 😉!
- **Individuelle Zeiterfassung jedes Gruppenmitglieds** im File 'Z_Erfasstung_GruppeXX_Name_Vorname.xlsx (XX, Name und Vorname sind angepasst!) Dieses Excel-File muss 1x pro Gruppenmitglied individuell ausgedruckt abgegeben werden.
- (ca.) **1 A4 Blatt** mit den genauen Angaben, wo der Dozent zur Korrektur das CIP-Projekt auf den virtuellen Maschinen genau findet (Gruppen Nummer + Projekttitel gehören auch auf das Blatt). Das heisst: Wo genau (d.h. in welchen Verzeichnissen auf den VM's) sind z.B. die Tableau Prep-Dateien, wo sind die Python Skripts, wo sind die Jupyter-Notebooks etc.

  Der Dozent geht davon aus, dass der **Schluss-Projekt-Zustand** auf ALLEN (!) virtuellen Maschinen der Gruppe **genau gleich** sein wird. Das heisst: **Ihr ganzes gemeinsames CIP-Projekt muss am Schluss auf ALLEN VM's der Gruppe problemlos laufen.**
  Empfehlung: Verwenden Sie möglichst von Beginn weg auf allen VM's die gleichen Verzeichnisstrukturen. Die Erfahrung aus den letzten Semestern zeigt: Wenn man erst am Schluss alle Projekt-Daten auf allen Virtuellen Maschinen in eine gleiche Verzeichnis-Struktur übertragen will, wird es unnötig kompliziert und aufwendig!
  Glauben Sie mir: Diesen Hinweis haben in der Vergangenheit jeweils eine oder zwei Gruppen nicht beachtet – Sie hatten am Schluss leider immer unnötigen Stress.

Inhalt der CIP-Projektbeschreibung (… das ist ein **Vorschlag** – Änderungen erlaubt)

- Titelblatt (mit Gruppenmitglieder und Projekttitel)
- Inhaltsverzeichnis
- Aufgabenstellung (Vertrag, Ziel …), eigene Fragestellung(en)
- Beschreibung des eigenen CIP – Prozess (als eine Art Management-Summery)
- Datenquellen und Daten (genau mit URL z.B. beim Crawlen)
- Verwendete Tools
- Lösungsschritte:
    - Collection / Extract
    - Integration / Preprocessing / Transformation
    - Load der Resultate auf DB (SQL Server)
- Anworten auf die Fragestellung(en)
- Goody: Graphische Interpretation / Analyseversuch (optional!)
- Schlusswort
- Lesson Learned / Reflexion

Power-Point Präsentation

- lassen Sie ihre Kreativität walten!
- … seien Sie aber auch "seriös"!
- Sie wollen Ihr CIP-Projekt "gut" erscheinen lassen - aber bitte keine Trump-Show!
- d.h. seien Sie ehrlich …! (… uuups ist diese Aussage politisch korrekt? 😉)
- "data Science – technische" Fakten sind interessant …
- PPT-Präsentation sollte die gleichen Aussagen/Folgerungen wie die CIP-Projektbeschreibung haben
- Zeitumfang ca. 10 Min. (ausser Sie haben etwas anderes mit dem Dozenten abgemacht)

Ich wünsche Ihne viel Erfolg und vor allem viel Spass – auch wenn das Projekt mit Aufwand verbunden ist!!

Erwin Mathis