MALIS
Master in Library and Information Sciences

Fakultät für Informations- und Kommunikationswissenschaften

Institut für Informationswissenschaft

Prof. Dr. Konrad Förstner

Technology Arts Sciences TH Köln

foerstner@zbmed.de
+49 (0)221 478-789090

(Stand 06.03.2020)

Aufgabe MALIS 19.3 WPM_T9.1: Data Science / Data Librarianship / IT-Praxis

Beschreibung von datenintensiven und datenfokusierten Aktivitäten im eignen Arbeitsalltag / in der eigenen Institution

Aufgabenstellung:

- Tragen Sie zusammen, welche Prozesse in Ihrer eigenen Arbeitsalltag und/oder ihre Institution datenintensiv und/oder datenfokusiert ist. Dies kann zum Beispiel Forschungdatenmanagement, aber auch die Pflege von Metadaten sein.
- Beschreiben Sie an Hand eines oder mehrerer ausgewählten Beispiels wie dort mit den Daten umgegangen wird und arbeiten Sie heraus wo sie Potential sehen, durch den Einsatz von Software, diesen Prozess zu vereinfachen, nachhaltiger und/oder transparenter zu gestalten.
- Schreiben Sie den Text in Markdown und nutzen Sie für die Erarbeitung git und/oder GitHub (alternative GitLab). Die TadempartnerInnen sollen gegenseitig mindestens 3 Issue (Tickets) mit Verbesserungsvorschlägen schreiben sowie mindestens einen Pull-Request schreiben.
- Umfang: 1000 – 1500 Wörter (exklusive von Quellenverweisen)

Termin und Aufgabentyp:

10.05.2019 23:59 (TA) (Typ: EA = Einzel-Aufgabe / TA = Tandem-Aufgabe / GA = Gruppen-Aufgabe)

Abgabe durch Einsenden eines Links zu einem offen GitHub-Repositorium oder durch Einladen zu deinem privaten Repositoriums (Nutzername: „konrad“)

Bewertungskriterien:

- Inhaltliche Qualität
- Nutzung der vorgeschriebenen Werkzeug und Kommunikation mit dem Tandempartner
- Form (Anfertigung der Arbeit gemäß den Standards des wissenschaftlichen Arbeitens), allerdings keine besondere Formatierung nötig.

Arbeitsbelastung:
24 Stunden

Anteil an der Modulnote:

30 %

MALIS
Master in Library and Information Sciences

Fakultät für Informations- und Kommunikationswissenschaften

Institut für Informationswissenschaft

Prof. Dr. Konrad Förstner

Technology Arts Sciences TH Köln

foerstner@zbmed.de
+49 (0)221 478-789090

(Stand 01.06.2020)

Aufgabe MALIS 19.3 WPM_T9.2: Data Science / Data Librarianship / IT-Praxis

Implementation von Software-Lösungen datenintensiven Prozesse und Datenanalysen

Aufgabenstellung:

- Python:
  - Option 1: Erstellen Sie in einem Jupyter Notebook oder einem Python-Skript eine Datenanalyse der offenen Daten der Seattle Public Library (Checkouts[1] oder Inventory[2]). Formulieren Sie fünf unterschiedliche Fragen, die Sie mit dem Daten beantworten möchten und implementieren Sie die Lösung. Mindestens drei der Fragestellungen sollten zu einer Grafik führen, die Sie mittels Python erzeugen (weitere Details unten).
  - Option 2: Sollten Sie selber ein datenintensives Problem aus Ihrem Arbeitsumfeld haben, das mittels Python (als Script oder in einem Jupyter Notebook) gelöst werden können, können Sie dieses ebenfalls nach Rücksprache mit dem Dozenten bearbeiten. Der Rahmen sollte in etwa Option 1 entsprechen. Hierzu verfassen sie bitte ein kurzes Exposé (400 – 600 Wörter). (Weitere Details unten)
- Shell:
  - Option 1: Erstellen Sie ein Shell-Script, das die bereitgestellte (siehe Moodle-Kursraum) Text-Datei bereinigt und eine neue Datei erstellt, die nur bestimmte Spalten enthält (weitere Details unten).
  - Option 2: Sollten Sie selber ein datenintensives Problem aus Ihrem Arbeitsumfeld haben, das mittels eines Shell-Scrips gelöst werden kann, können Sie diese ebenfalls nach Rücksprache mit dem Dozenten bearbeiten. Der Rahmen sollte in etwa Option 1 entsprechen. Hierzu verfassen sie bitte ein kurzes Exposé (400 – 600 Wörter).
- Alle Lösungen sollen ausreichend im Quell-Code / in Markdownzellen Dokumentiert werden.

---

1 https://dev.socrata.com/foundry/data.seattle.gov/tmmm-ytt6

2 https://data.seattle.gov/Community/Library-Collection-Inventory/6vkj-f5xf

- Versionieren Sie dabei den Quellcode regelmäßig mittels git und GitHub (mindestens 10 Commits) in Ihrem Git-Repositorium des Moduls.
- Das Ergebnis soll in einer kurzen (voraussichtlich 8 min Vortrag + 2 min Fragen) Präsentation den Plenum vorgestellt werden.

Termin und Aufgabentyp:

21.06.2019 23:59 (EA) (Typ: EA = Einzel-Aufgabe / TA = Tandem-Aufgabe / GA = Gruppen-Aufgabe)

Abgabe durch Einsenden eines Links zu einem offenen GitHub-Repositorium oder durch Einladen zu deinem privaten Repositoriums. Der letzte Commit der vor Ablauf der Abgabefrist getätigt wurde wird gezählt.

Bewertungskriterien:

- Inhaltliche Qualität und Funktionstüchtigkeit der Lösung;
- Der Code ist gut leserlich und verständlich - Variablen (und falls Funktionen selber erstellte wurden auch diese) sind sinnvoll benannt.
- Es wurde kurze Dokumentation zum Code bereitgestellt (als Kommentare im Code oder als Markdown-Zellen in Jupyter Notebook), der das Problem sowie den Lösungsansatz und Nutzung der Lösung kurz beschreibt (mindestens 400 Wörter)
- Es gab mindestens 10 git-Commits im Laufe der Entwicklung

Arbeitsbelastung:
56 Stunden

Anteil an der Modulnote:

70 %

Detail zu den Aufgaben:

Für Python nutzen Sie bitte die folgenden zwei Datenset der Seattle Library:

- Checkouts
- Inventory

Die Daten könne über diese beiden URLs bezogen werden.

```
"https://data.seattle.gov/resource/tmmm-ytt6.csv?$where=checkoutyear=2018&$limit=10000"

"https://data.seattle.gov/resource/6vkj-f5xf.csv?$limit=5000"
```

Beachten Sie, dass standardmäßig ein Limit von 1000 Zeilen gesetzt ist, das Sie in der URL anpassen können (im Beispiel 5000). Für die Checkout-Daten können Sie auch das Jahr

angegeben (im Beispiel 2018). Die Datensets sind relativ groß - so gab es im Jahr 2018 2665099 Ausleihungen. Die CSV-Datei dazu hat eine Größe von etwa 650 Mb.

Für die Shell-Aufgabe nehmen Sie Datei „2020-05-23-Article_list_dirty.tsv“, die in Moodle abgelegt ist. Erzeugen Sie mittels eines Shell-Scripts eine neue Datei „2020-05-23-Dates_and_ISSNs.tsv“, die nur die Spalten der ISSNs und Veröffentlichungsjahren enthält. Die Zeilen sollen nicht redundant sein. Sie werden sehen, dass die Datei nicht sauber strukturiert ist. Das heißt Ihr Shell-Script benötigt Befehle, die unnötige Zeilen entfernen und bestimmte Zeilen verändern. Die Lösung kann mit den Unix-Shell-Befehle grep, sed, cut, sort, uniq erreicht werden. Meine Musterlösung hat insgesamt 6 Befehl (wahrscheinlich werden Sie mehr benötigen). Es muss also kein großes Shell-Skript werden. Lesen Sie als Einführung die Library Carpentry Lesson „Working with free text“. Lesen Sie sich selbstständig nach Bedarf in die nötigen Befehle ein. Die Ergebnis-Datei „2020-05-23-Dates_and_ISSNs.tsv“ ist auch im Moodle zu finden, damit Sie die Ergebnisse Ihres Shell-Scriptes damit abgleichen können.