Task Area Infrastructure/Operations

Jupyter Notebooks: Ein Angebot für die Text+ Community

George Dogaru, Gesellschaft für wiss. Datenverarbeitung Göttingen (GWDG),
Melina Jander, Niedersächsische Staats- u. Universitätsbibliothek (SUB)

Was sind Jupyter Notebooks?

Jupyter Notebooks sind Dokumente und Code. Sie sind über eine Web-Oberfläche bedienbar und bestehen aus anzeigbaren Markdown- und ausführbaren Code-Zellen (Python/R/JavaScript, etc.). Code-Zellen erzeugen Output-Zellen, welche normalen Text enthalten können, oder auch Bilder, Videos, Visualisierungen, (dynamisches) HTML (s. Abbildungen 2 und 3). Es ist möglich, in derselben Umgebung Code zu schreiben und auszuführen, Daten zur Anzeige zu bringen, Arbeitsschritte zu erklären. Die Jupyter-Anwendungen (Jupyter Notebook, Jupyterlab) können auf einer lokalen Maschine betrieben werden oder in einer Cloud-Infrastruktur.

Jupyter Notebooks in Text+

Infrastruktur-Aufgaben

Zugang zu verschiedenen Jupyterhubs (Multi-User Lösungen für Jupyter Notebooks) ermöglichen

Einbindung von Storage

Möglichkeiten anbieten, um besondere Software-Anforderungen zu erfüllen (z.B.: Notebook x setzt voraus, dass auf dem System Java in der Version "openjdk 17.0.3" vorhanden ist)

Ggf. spezielle Hardware verfügbar machen (GPU, HPC)

Inhalte (Notebooks, Code, Dokumentation)

Demos: Einsatzmöglichkeiten durch Beispiele verdeutlichen

Zwei Demos unter gitlab-ce.gwdg.de/textplus/code/jupyter-demos:

- XML-Datenverarbeitung mit XSLT, Erstellung von METS-Daten
- Verarbeitung der Text+ User Stories
- Weitere Demos folgen

Produktiv einsetzbare Notebooks für bestimmte Aufgaben erstellen

Community Activities

Präsentieren der Einsatzmöglichkeiten

Bereits verfügbare Beispiele:

- Download von Daten aus dem Internet,
- Datengewinnung aus HTML mit beautifulsoup,
- XML-Datenverarbeitung mit XSLT 3.0
- Erstellung einer METS-Datei, die im DFG-Viewer angezeigbar ist
- Daten-Upload in ein Nextcloud-Verzeichnis in der Academiccloud
- Verarbeitung von tabellarischen Daten mit pandas, Konversion nach JSON, HTML

Ermittlung von Bedarfen und Anforderungen

Entwicklung von Ideen zusammen mit der Community

Demo: Verarbeitung der Text+ User Stories

Input: User Stories (TSV-Tabellen u. HTML-Dateien)

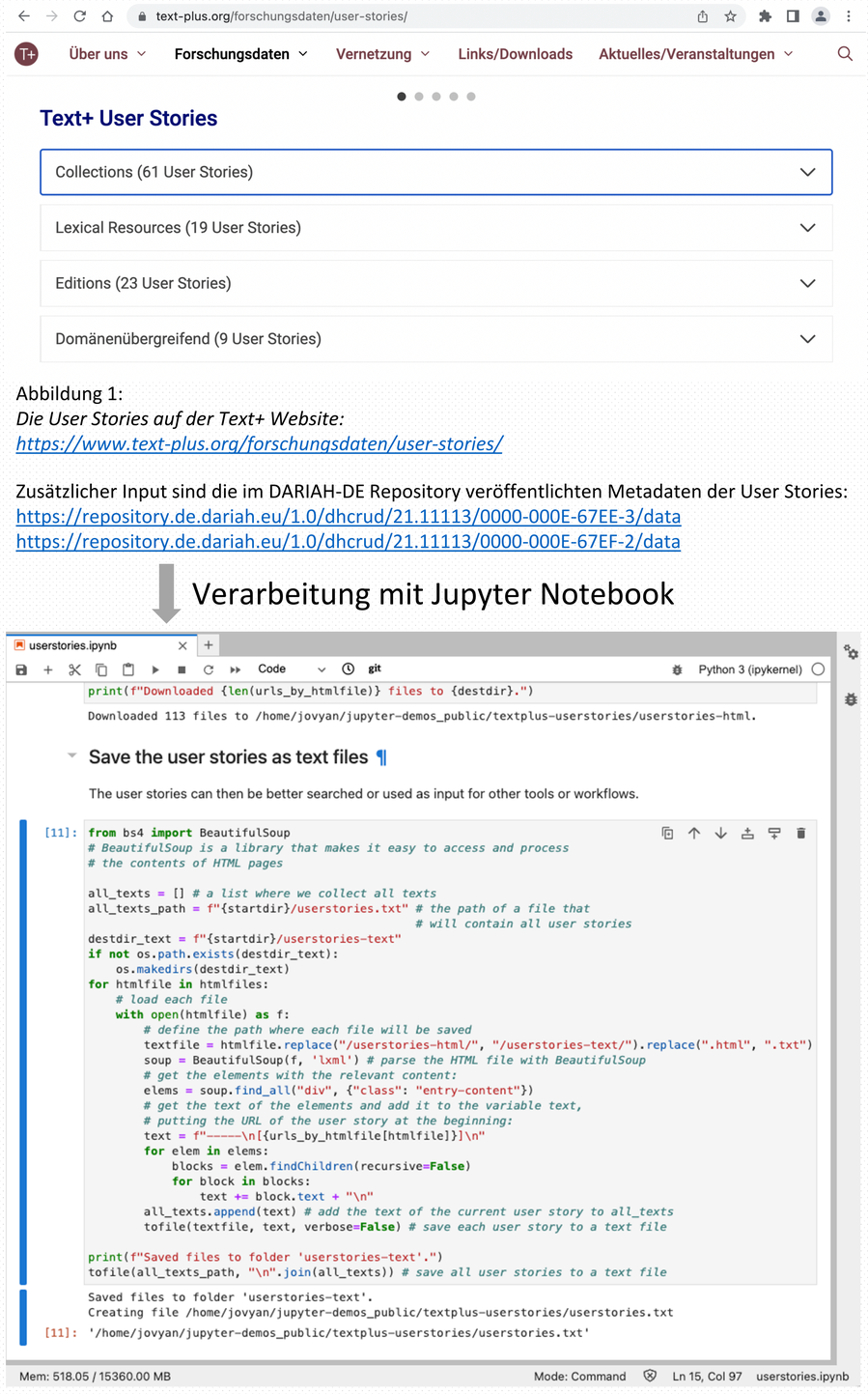

Abbildung 1:
Die User Stories auf der Text+ Website:
https://www.text-plus.org/forschungsdaten/user-stories/

Zusätzlicher Input sind die im DARIAH-DE Repository veröffentlichten Metadaten der User Stories:
https://repository.de.dariah.eu/1.0/dhcrud/21.11113/0000-000E-67EE-3/data
https://repository.de.dariah.eu/1.0/dhcrud/21.11113/0000-000E-67EF-2/data

Verarbeitung mit Jupyter Notebook

Abbildung 2:
Auszug aus dem Notebook "textplus-userstories/userstories.ipynb" im Repository https://gitlab-ce.gwdg.de/textplus/code/jupyter-demos.

Output A: HTML-Tabelle mit User Stories gruppiert nach Thema

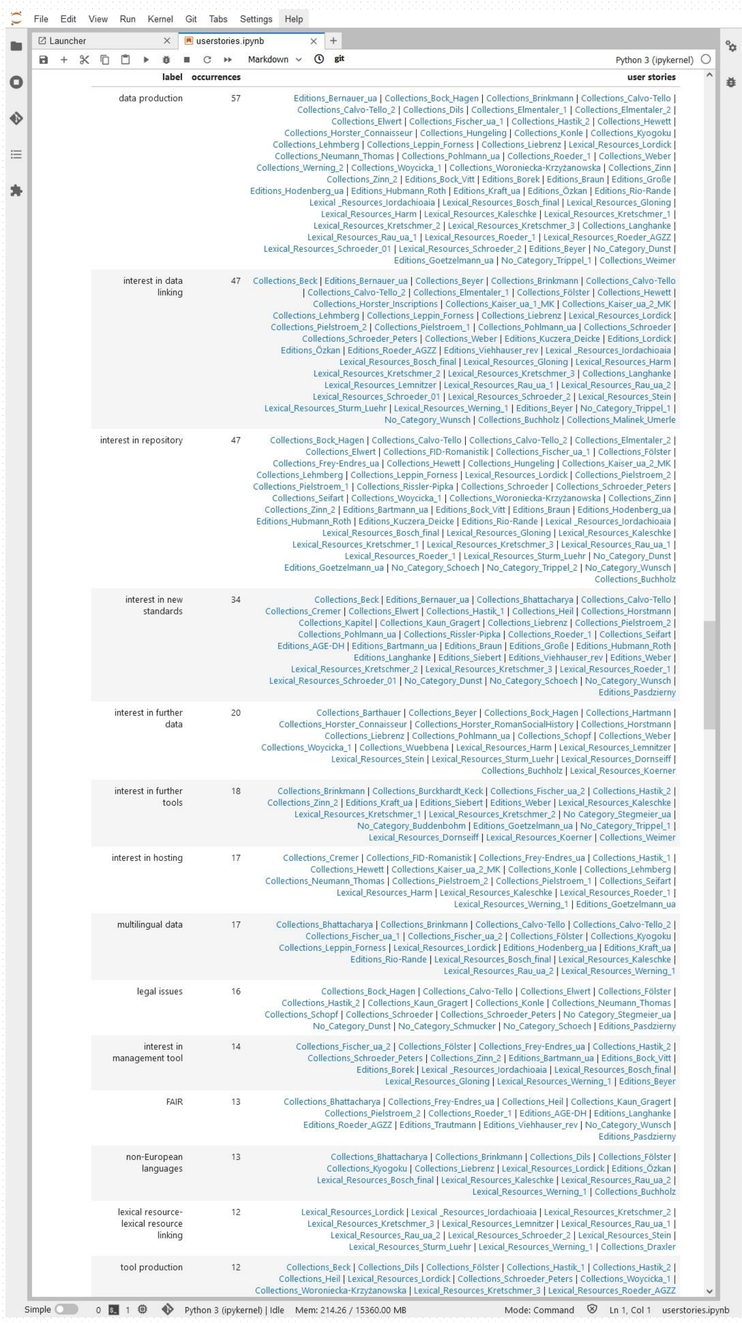

Abbildung 3:
Die User Stories wurden im Zuge der Veröffentlichung analysiert und einer oder mehreren Kategorien zugeordnet (z.B. „interest in repository"). Die User Stories und die Daten aus der Analyse wurden mit dem Notebook verarbeitet, um eine Übersicht zu erstellen (eine HTML-Tabelle), die einen komfortablen Zugang von jeder Kategorie zu den dazugehörigen User Stories ermöglicht.

Output B: Textdateien mit den einzelnen User Stories (hier Suchergebnisse)

Abbildung 4:
Die User Stories liegen als einzelne Textdateien vor und sind somit gut durchsuchbar.
Hier so sehen sind Suchergebnisse und das Suchprotokoll in Notepad++. Die Suchbegriffe sind in den blau hinterlegten Zeilen abgebildet.

Deutsche Forschungsgemeinschaft

Das NFDI-Konsortium Text+ wird gefördert durch die Deutsche Forschungsgemeinschaft (DFG) unter der Projektnummer 460033370.

Text+ ist Teil der Nationale Forschungsdaten Infrastruktur

Kontakt: textplus-operations-office@lists.gwdg.de
Poster: Jupyter Notebooks: Ein Angebot für die Text+ Community
Inhalt: Jupyter Notebooks als anwendungsbezogenes Angebot für die Text+ Community
Verantwortlich: George Dogaru, Melina Jander (0000-0003-1646-6836)
Event: Text+ Plenary 2022, 12./13. September 2022, Mannheim