Goethe Universität
Institut für Informatik
Texttechnologie

Text2Scene

SoSe 2021
Fingerübung

Alexander Mehler Alexander Henlein

24.04.2020

Grundlegendes

Lesen Sie bitte die gesamte Aufgabe erst vollständig durch, bevor Sie diese bearbeiten.

1 Handwerkszeug

Machen Sie sich mit dem **Handwerkszeug** des Praktikums vertraut:

- Unterschreiben und Zuschicken[1] der im **Olat** hinterlegten Datenschutz relevanten Daten.
- Richten Sie das **LaTeX-Plugin** für PowerPoint, gemäß der Anleitung im Olat, ein.
- Legen Sie ein für die Fingerübung ein entsprechendes **Projekt** in **GitLab**[2] oder **GitHub**[3] an.
- Richten Sie sich eine geeignete Pythonumgebung (z.B. **Anaconda**[4]) oder Mavenprojekt für Java ein.

2 Einarbeitung in Python oder Java

Ziel der Fingerübung ist es, sich mit Python bzw. Java und IsoSpace Daten vertraut zu machen.

2.1 Daten

Die Daten erhalten Sie von der SpaceEval Homepage. Laden Sie sich dazu bitte die **Trainingsdaten** herunter (https://alt.qcri.org/semeval2015/task8/index.php?id=data-and-tools).

[1] henlein@em.uni-frankfurt.de
[2] https://gitlab.texttechnologylab.org/
[3] https://github.com/
[4] https://www.anaconda.com/products/individual

2.2 Vorverarbeitung

Schreiben Sie ein Tool, mit dessen Hilfe Sie die Daten einlesen können und in einem geeigneten Dateiformat abspeichern können. Zusätzlich sollen Sie die Texte mit Spacy[5] (Python) oder StanfordCoreNLP[6] (Java) tokenisiert und mit Part-of-Speech Tags (PoS) versehen werden. Nutzen Sie das Tool auch, um Satzgrenzen zu erkennen. Die unter Umständen entstehenden Konflikte zwischen Tokengrenzen und IsoSpace-Entities müssen dabei nicht beachtet werden.

2.3 Auswertung

Anschließend sollen Sie folgende Punkte auswerten:

- Wie oft kommen welche PoS-Tags vor?
- Wie viele [SpatialEntities, Places, Motions, Locations, Signals, QsLinks, OLinks] gibt es?
- Wie oft kommen welche QsLink Typen vor? (DC,EC, ...)?
- Verteilung der Satzlänge graphisch darstellen (x: Satzlänge, y: Wie häufig)?
- Welche Links (QSLinks, OLinks) werden von welchen Präpositionen (markiert durch SPATIAL_SIGNAL) getriggert (z.B. wie oft werden QSLinks durch die Präposition „on“ getriggert)?
- Welches sind die fünf häufigsten „MOTION“ Verben (und wie oft kommen diese vor)?

2.4 Visualisierung

Visualisieren Sie das Dokument **Bicycles.xml** und **Highlights_of_the_Prado_Museum.xml** grafisch. Stellen Sie dazu alle räumlichen Entitäten (*PLACE*, *LOCATION*, *SPATIAL_ENTITY*, *NONMOTIONEVENT*, *PATH*) als Knoten da farblich zugeordnet nach Klasse und beschriftet mit deren entsprechenden Textbeschreibung.

Entitäten, die dabei mit einem *METALINK* verknüpft, sollen als **ein** Knoten dargestellt werden (mergen durch Koreferenz). Als Kanten sollen *OLINKS* und *QSLINKS* zwischen den Knoten eingezeichnet werden, wieder farblich erkennbar und *relType* als Label. Die Trigger müssen dabei nicht beachtet werden. Dabei ist es ihnen überlassen, welches Tool Sie zur Visualisierung verwenden (Neo4j, Graphviz, Networkx, ...).

Aufgabenumsetzung

Das Projekt soll dabei in **GitLab/GitHub** hinterlegt und aktuell gehalten werden.

Abgabe der Aufgabe

Ihre fertige Projekt schicken Sie dem Dozenten als **Git Link per Email oder Discord** zu. Bitte achten Sie darauf, dass das Projekt direkt **lauffähig** ist. Alle Ergebnisse sollen zusätzlich zusammengetragen und als .pdf abgegebenen werden.

Aufgabenzeitraum

Der Aufgabenzeitraum für diese Aufgabe ist **vom 16.04.2021 bis 28.04.2021 (Abgabe)**. Die Bearbeitungszeit beträgt also **13 Tage**.

[5] https://spacy.io/
[6] https://stanfordnlp.github.io/CoreNLP/

Nächster Praktikumstermin

Der nächste Termin ist am **30.04.2021**. An diesem Tag werden die Teilnehmer in ihre gewählten Gruppen eingeteilt und erhalten eine Einarbeitungsaufgabe.