Hochschule RheinMain
Prof. Dr. Adrian Ulges

04.11.2020

Anwendungen der KI (WS 21/22)

Aufgabenblatt 3

zu bearbeiten bis: 11.11.2020

Aufgabe 3.1 (Regular Expressions)
Schreiben Sie reguläre Ausdrücke, mit denen Sie innerhalb der Wikibase möglichst gut folgende Entitäten finden:

a) Namen von Personen (Z.B. Angela Merkel; John W. Booth; ...)

b) Daten (10.09.2020; Sep 14th 1999; ...)

Testen Sie ihre regulären Ausdrücke und bringen Sie Beispiele ins Praktikum mit.
Hinweis: Versuchen Sie durch die Verwendung mehrerer verschiedener regulärer Ausdrücke ihren Recall zu erhöhen.

Aufgabe 3.2 (Positive Pointwise Mutual Information)
Finden Sie die 100 Wortpaare mit der höchsten **PPMI**-Ähnlichkeit innerhalb der Wikibase.

- Tokenizen Sie dafür die Dokumente in einzelne Wörter (nutzen Sie hierfür `nltk` mit der Funktion `word_tokenize()`). Lowercasen Sie außerdem Ihre Tokens, und entfernen Sie Zahlen und Sonderzeichen.
- Erstellen Sie ein **Vokabular** der $n = 2000$ häufigsten Worte.
- Erstellen Sie sich eine Matrix $\in \mathbf{R}^{n \times n}$ (nutzen Sie hierfür `numpy`) und füllen sie mit Ihren berechneten PPMI-Werten. Wählen Sie dabei bei der Berechnung der PPMI als Kontext jeweils ein Fenster von ± 5 Worten.
- Geben Sie die 100 Wortpaare mit der höchsten PPMI-Ähnlichkeit aus.

Hinweis: Testen Sie ihr Programm zunächst auf einem Ausschnitt der Wikibase, um schnell zu iterieren. Starten Sie dann einen Durchlauf für die komplette Wikibase.
Hier ist ein netter Tipp, um die Positionen der größten Werte in einer Matrix zu finden.