1. Referenz für aktuellste Recherchen/Ansätze: DeepAPI

Basis für DeepAPI (Natural Language – Java API Path):

Zitate aus zugehörigem Paper:

„Finally, we obtain a database consisting of 7,519,907 <API-sequence, annotation> pairs.“

„We train our models in a server with one Nvidia K20 GPU. The training lasts ∼240 hours with 1 million iterations.“

<https://docs.oracle.com/javase/7/docs/api/>

Immer erster Satz der Doku benutzt:

Beispielpaar:

API sequence: InputStream.read🡪OutputStream.write

Annotation: copies bytes from a large inputstream to an outputstream.

Benutzes framework: GroundHog (development is over) – nachfolger Blocks schlecht dokumentiert

2. Referenz für aktuellste Recherchen/Ansätze: Neural Text Generation

- Deep learning für Translation tasks (English-French)

- siehe unten

Probleme:

StackOverflow-Daten:

* Stark unterschiedliche Formulierungen mit sehr variierendem Vokabular

🡪 semantisch nicht gruppierbar

* Leute die auf StackOverflow Fragen stellen haben nicht viel Ahnung vom Thema – Qualität gering
* Länge der Fragen extrem unterschiedlich (1-wort <> Roman)
* Im Vergleich zu DeepAPI-Datensatz (40k <–> 7,5 Mio.) winzig

Difflib-Ansatz:

* Vgl. StackOverflow-Daten-Probleme: Daten stark variierendes Vokabular, „willkürliche“ Fragestellungen etc.
* Close\_matches sind Wort/Buchstaben-closematches, nicht semantische closematches
* Selbst wenn vorheriges nicht immer der Fall ist: Laufzeit sehr hoch

Classification-Ansatz:

* Sehr viele Klassen 🡪 ungeeignet für classifier
* StackOverflow-Daten Inkonsistenz, mangelnde Integrität
* Numpy Reference nicht umfangreich genug

Schwierigkeit bei Seq2Seq im Vergleich zu DeepAPI:

* StackOverflow Daten bedingt brauchbaur (s.o. Inkonsistenz)
* Python Doku (v.a. numpy Doku) nicht so brauchbar wie JavaDocs (umfamg & form)
* Idee: Aus NumPy source informationen gewinnen (Beachte benötigter Umfang, vgl 7,5Mio bei DeepAPI)
* Beachte: Aus Referenz 2: 5.2 Common Issues/Rare and out of vocabulary Words:

„For languages with very large vocabularies, especially languages with rich morphologies, rare words become problematic when choosing a tokenization scheme that results in more token labels than it is feasible to model in the output softmax.“