NLP: Reguläre Ausdrücke

Vitor Fontanella

06.09.2022

1 Aufgaben

In dem Ordner *infekte* liegen einige Texte aus Wikipedia zum Thema Infekt. In diesen Texten werden viel verschiedene Viren erwähnt. Wir brauchen eine Liste von allen Viren, die in diesen Texten vorkommen und ihre Häufigkeiten.

1. Unbedingt gefunden werden sollen alle Viren, die mit einem Kompositum (mit oder ohne Bindestrich), das auf *virus* oder *viren* endet benannt werden. Beispiele sind: Herpes-simplex-Virus, Parainfluenzavirus, Herpes-Simplex-Typ-2-Viren, Influenzaviren, usw.
2. Ersetzen Sie in der Ergebnisliste alle vorkommen von *viren* durch *virus*, damit die Vorkommen von Namen wie *Rhinoviren* und *Rhinovirus* zusammengezählt werden.
3. Manche Viren haben Namen, die aus zwei Wörtern bestehen wie *Humanes Herpesvirus* oder *Canines Adenovirus*. Versuchen Sie auch diese zu erkennen.
4. Manche Viren haben Namen, die aus einem Virusart und den eigentlichen Namen bestehen wie *Influenza-A-Virus H1N1* oder *Hantaan-Virus HTNV*. Versuchen Sie auch diese zu erkennen.
5. In den Texten gibt es einige Stellen, wo mehrere Virennamen wie folgt koordiniert vorkommen: `Vogel- und Schweinegrippeviren` oder `Rhino-, Entero- und Mastadenoviren`. Überlegen Sie, ob und wie Sie diese Viren auch zählen können.

2 Lernziele

Am Ende dieser Lerneinheit sollen Sie in der Lage sein:

1. Mit Hilfe der Tabellen mit den Abkürzungen für reguläre Ausdrücke in Perl-Syntax komplexe reguläre Ausdrücke zu formulieren;
2. Reguläre Ausdrücke in Pythonprogramme einzubinden;
3. Mit Python und regulären Ausdrücken Muster in Texten zu finden und zu zählen.

3 Materialien

Um die Aufgabe zu lösen, müssen Sie sich vermutlich zuerst theoretisch mit dem Thema beschäftigen und einige Vorübungen dazu machen. Auf Moodle stehen hierfür folgende Materialien zur Verfügung:

- Ein Foliensatz;
- Ein Screencast zu diesem Foliensatz;
- Ein Jupyter-Notebook mit allen Beispielen aus den Folien.
- Ein Jupyter-Notebook mit einer Funktion zum Lesen aller Dateien aus dem Ordner `infekte`.