NLP Assignment 1

Dieses Assignment wird bewertet !

Die Zusammenarbeit unter Studierenden ist erwünscht, so lange sich diese auf die Diskussion von Konzepten oder Problemen mit python konzentrieren. Das Kopieren von Code ist nicht erlaubt, in diesem Falle werden alle Lösungen der beteiligten Parteien mit 0 Punkten bewertet; dazu werden alle Lösungen (manuell und automatisiert mit Plagiat-Checkern) auf Gruppenarbeit und Kopien untersucht.

Entwickeln Sie einen *auto-correct* Assistenten. Dieser prüft, ob ein eingegebenes Wort in Ihrem Vokabular vorhanden ist oder nicht. Ist das eingegebene Wort nicht im Vokabular vorhanden, dann schlägt der *auto-correct* Assistent eine Liste möglicher Worte vor. Diese Liste soll nach zwei Kriterien geordnet werden: nach der Edit-Distanz und der Häufigkeit der Vokabular-Worte im Korpus (d.h. in Ihrem Text).

- Wählen Sie für Ihr Vokabular einen genügend langen Text (Korpus) in Deutsch oder Englisch aus. Zum Beispiel aus https://www.gutenberg.org/.
- Wählen Sie *Damerau–Levenshtein* als Edit-Distanz.

Bewertungskriterien und Bemerkungen:

- Die Minimalanforderung ist, dass der *auto-correct* Assistent Worte mit
  - (*i*) einem fehlenden Buchstaben (*insertion*),
  - (*ii*) einem überflüssigen Buchstaben (*deletion*),
  - (*iii*) einem falschen Buchstaben (*substitution*) und
  - (*iv*) zwei benachbarten, vertauschten Buchstaben (*transposition*)

  richtig erkennen kann. Wenn der *auto-correct* Assistent auch grösserer Edit-Distanz korrekt bewältigt, umso besser.
- Der *auto-correct* Assistent soll auf eingegebene Worte ohne spürbare Verzögerung reagieren. Bevor Worte eingegeben werden, darf er aber Zeit für's *Pre-Processing* (Korpus einlesen, etc.) und andere vorbereitende Aktivitäten aufwenden.
- Aus obigem Grund muss gut geprüft werden, ob die Berechnung der Edit-Distanzen nach jeder Worteingabe genügend performant ist.

Abgabe : In Form eines `Jupyter Notebooks`
per Email an claudio.paonessa@fhnw.ch bis 16. März 2020