The ABC of Computational Text Analysis

#10 NLP with Python

Alex Flückiger

Faculty of Humanities and Social Sciences
University of Lucerne

12 May 2022

Recap last Lecture

introduce Python 🐍

- working with VS Code Editor
- learning programming concepts & syntax

data types, loops, indexing…

Outline

- get the organizational stuff done

  evaluation, mini-project, assignment #3

- let’s do serious NLP! ✨
- code interactively

  interrupt, ask, and complement

Organizational

Course Evaluation

Tell me... 📣

Please follow the link in the email

- received on 9 May 2022 (or similar)
- by the University of Lucerne, Faculty of Humanities and Social Sciences

Thanks for any constructive feedback,
be it sweet or sour! 🙏

Assignment #3 ✍️

- get/submit via OLAT

  starting tomorrow

  deadline 20 May 2022, 23:59

- use the OLAT forum

  subscribe to get notifications

- ask friends for support, not solutions

Requirements of Mini-Project

present project on 2 June 2022

- analyze any collection of documents

  compare historically

  compare between actors

- apply quantitative measures + interpretation

  executable script
  multiple documents

- form groups of 2-4 people

❗ share your project idea here by 19 May 2022

Optional Seminar Paper

- writing a seminar paper (6 ECTS)
- get in touch to discuss your idea

A Primer on Old School NLP

What is a Word?

- words ~ segments between whitespace
- yet, there are …
  - contractions: `U.S.`, `don't`
  - collocations: `New York`

Token

- token ~ computational unit

  representation of words

- lemma ~ base form of a word

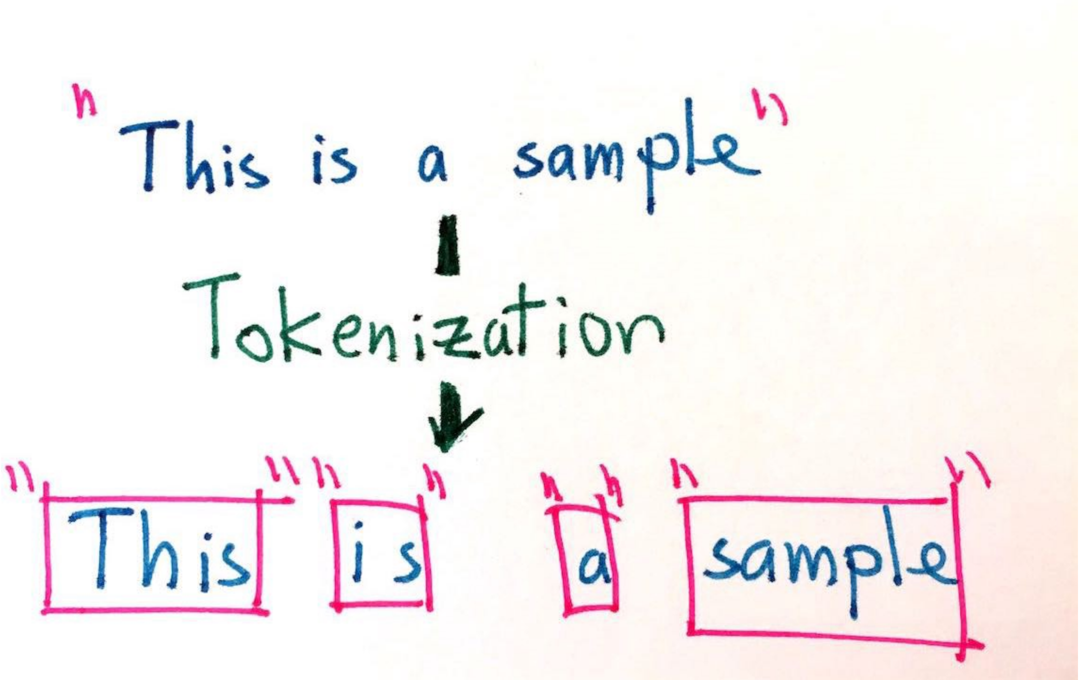
  `texts` → `text`

  `goes` → `go`

- stop words ~ functional words

  lacking deeper meaning

  `the`, `a`, `on`, `and` …



Tokenizing a sentence (Medium)

Common Processing Steps in NLP

1. Tokenizing

   segmenting text into words, punctuations etc.

2. Tagging part-of-speech (POS)

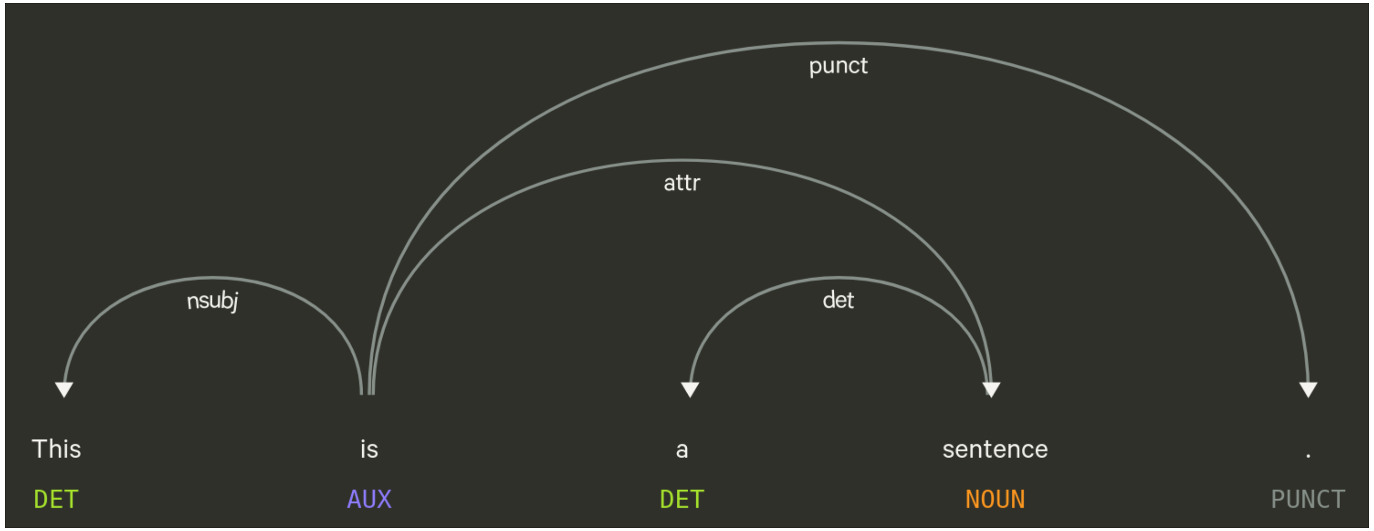
   assigning word types (e.g. verb, noun)

3. Parsing

   describing syntactic relations

4. Named Entity Recognition (NER)

   organizations, persons, locations, time etc.



🤓 Catch up on NLP with Jurafsky and Martin (forthcoming)

Modules/Packages

No programming from scratch 🎉

- packages provide specific functionalities
- packages need to be installed first

NLP Packages

- spaCy

  industrial-strength Natural Language Processing (NLP)

- textaCy

  NLP, before and after spaCy

- scattertext

  beautiful visualizations of how language differs across corpora

Deep Dive into NLP for Social Science

- check code on GitHub
- run code on Binder launch binder

Resources

tutorials on spaCy

- official spaCy 101
- official online course spaCy
- Hitchhiker’s Guide to NLP in spaCy

Questions?

References

Jurafsky, Dan, and James H. Martin. forthcoming. *Speech and Language Processing*. 3rd (Draft of December 30, 2020). London: Prentice Hall. https://web.stanford.edu/~jurafsky/slp3/.