UNIVERSITY OF
LUCERNE

The ABC of Computational Text Analysis

#10 NLP with Python

Alex Flückiger

Faculty of Humanities and Social Sciences
University of Lucerne

02 May 2024

Recap last lecture

- from unique words to contextual embeddings 🧮

  more granular representations are more effective

- modern, data-driven NLP is both powerful and biased 🚨

  there is nothing like raw data

  reflect the representation and decisions behind it

Outline

- get some organizational stuff done
- 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 29 April 2024

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

Your mini-projects

- Your project idea is recorded here
- You are ready to work on it (self-paced)
- Reach out if you are stuck! 🤯

A primer on classic 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

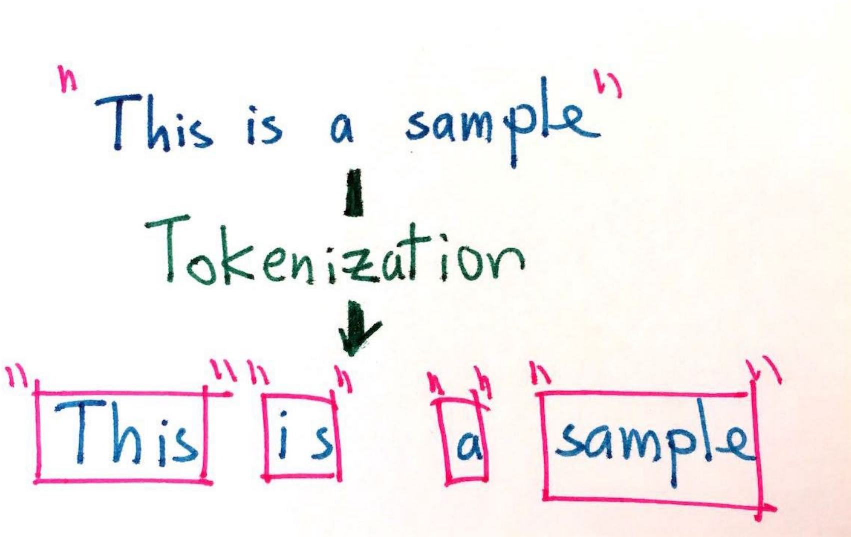
  `texts` → `text`

  `goes` → `go`

- stop words ~ functional words

  lacking deeper meaning

  `the`, `a`, `on`, `and` …



Segmenting a text into tokens

`Let's tokenize this sentence! Isn't is easy?` 🤓

Classic processing steps in NLP

1. Tokenizing

    segmenting text into words, punctuation etc.

2. Tagging part-of-speech (POS)

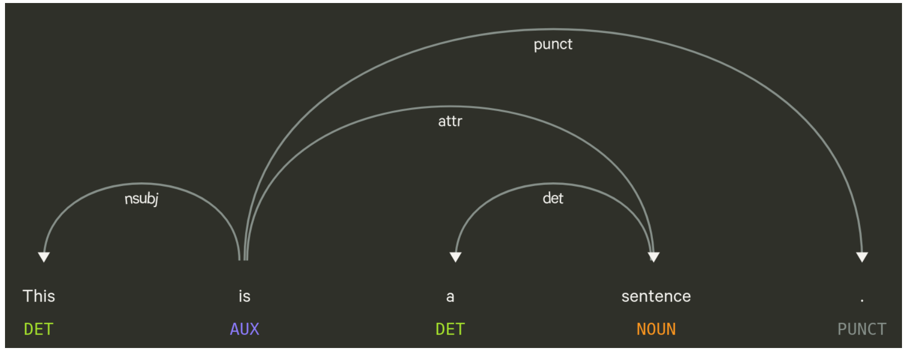
    assigning word types (e.g. verb, noun)

3. Parsing

    describing syntactic relations

4. Named Entity Recognition (NER)

    organizations, persons, locations, time etc.



Automatically inferred information of a sentence

👉 Catch up on NLP with

Jurafsky and Martin (forthcoming)

Let’s apply this in practice ✨

Questions?

References

Jurafsky, Dan, and James H. Martin. forthcoming. *Speech and Language Processing*. 3rd (Feb 3, 2024 draft). London: Prentice Hall. https://web.stanford.edu/~jurafsky/slp3/.