University of Zurich[UZH]

Institute of Computational Linguistics

Introduction to Machine Learning

Lesson 5

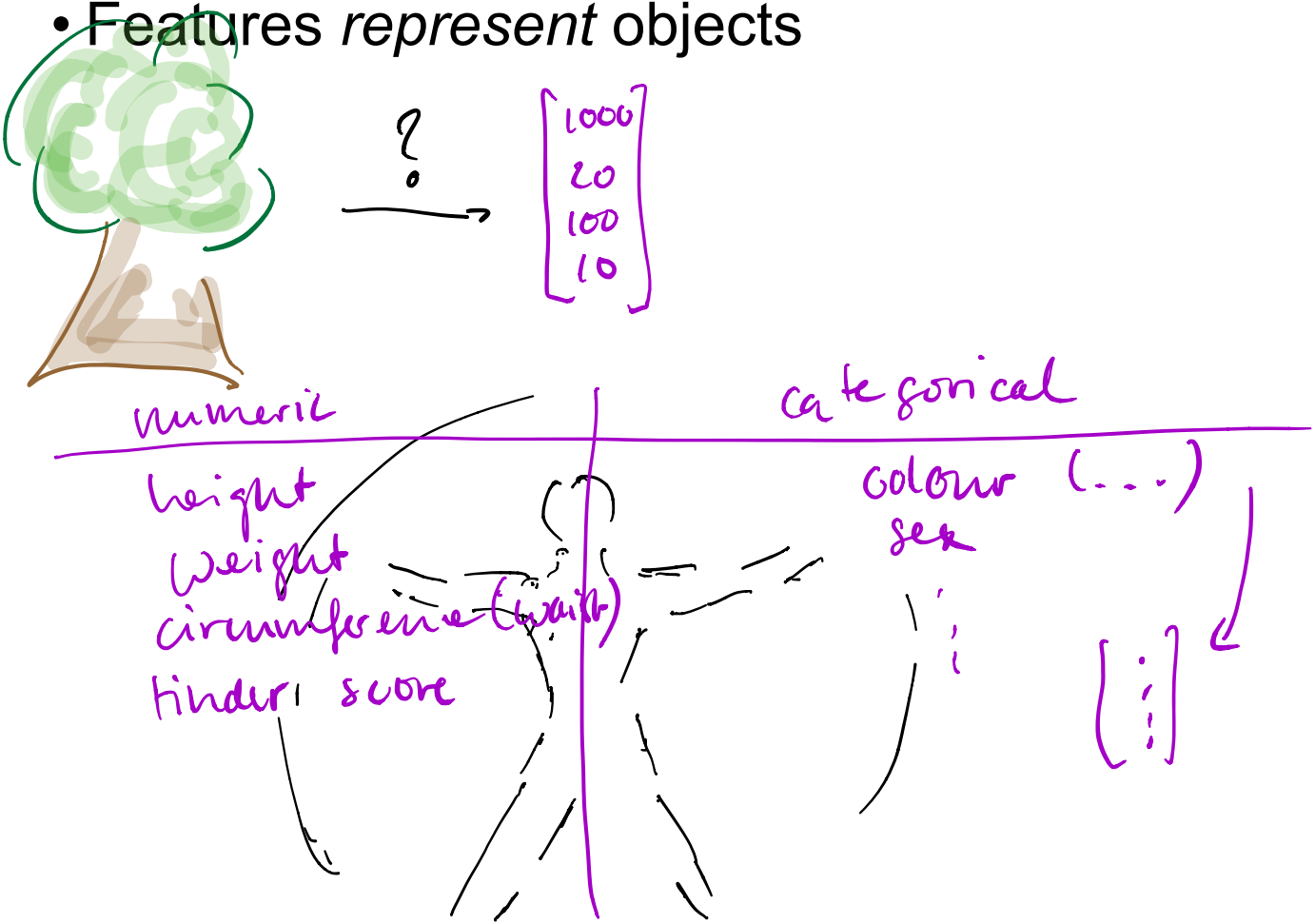
Mathias Müller, Phillip Ströbel

Now

- Feature extraction
- Feature extraction and preprocessing for text

Features?

- Features *represent* objects



Feature extraction

Feature extraction

$$\longrightarrow \begin{bmatrix} 1000 \\ 20 \\ 10 \end{bmatrix}$$

$$\longrightarrow \begin{bmatrix} 100 \\ 2 \\ 5 \end{bmatrix}$$

and text?

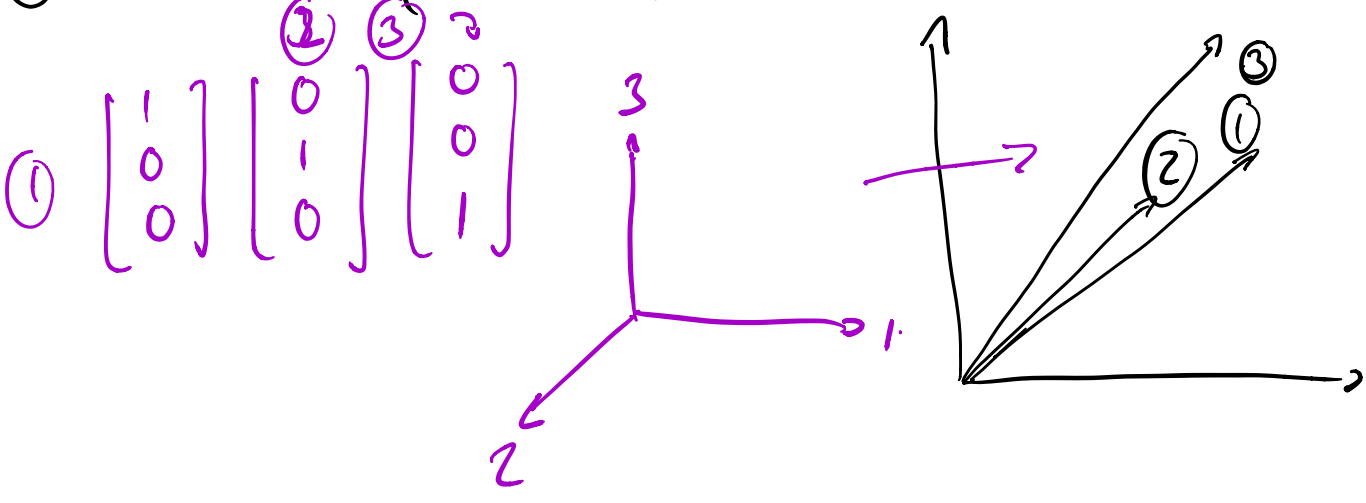
"Supercalifragilistischexpialigetisch" $\longrightarrow \begin{bmatrix} 2 \\ 8 \\ 7 \end{bmatrix}$

Why is raw text not a good representation?

corpus → corpora (adj.)

- (What are good features, then?)

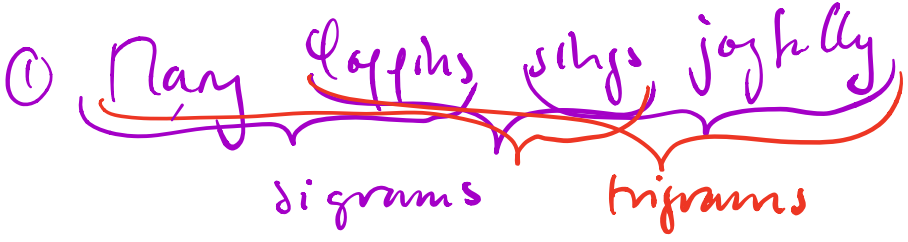
① The game of life is a game of everlasting learning.
② The unexamined life is not worth living!
③ Never stop learning.



Kinds of features from text

Based on counts:

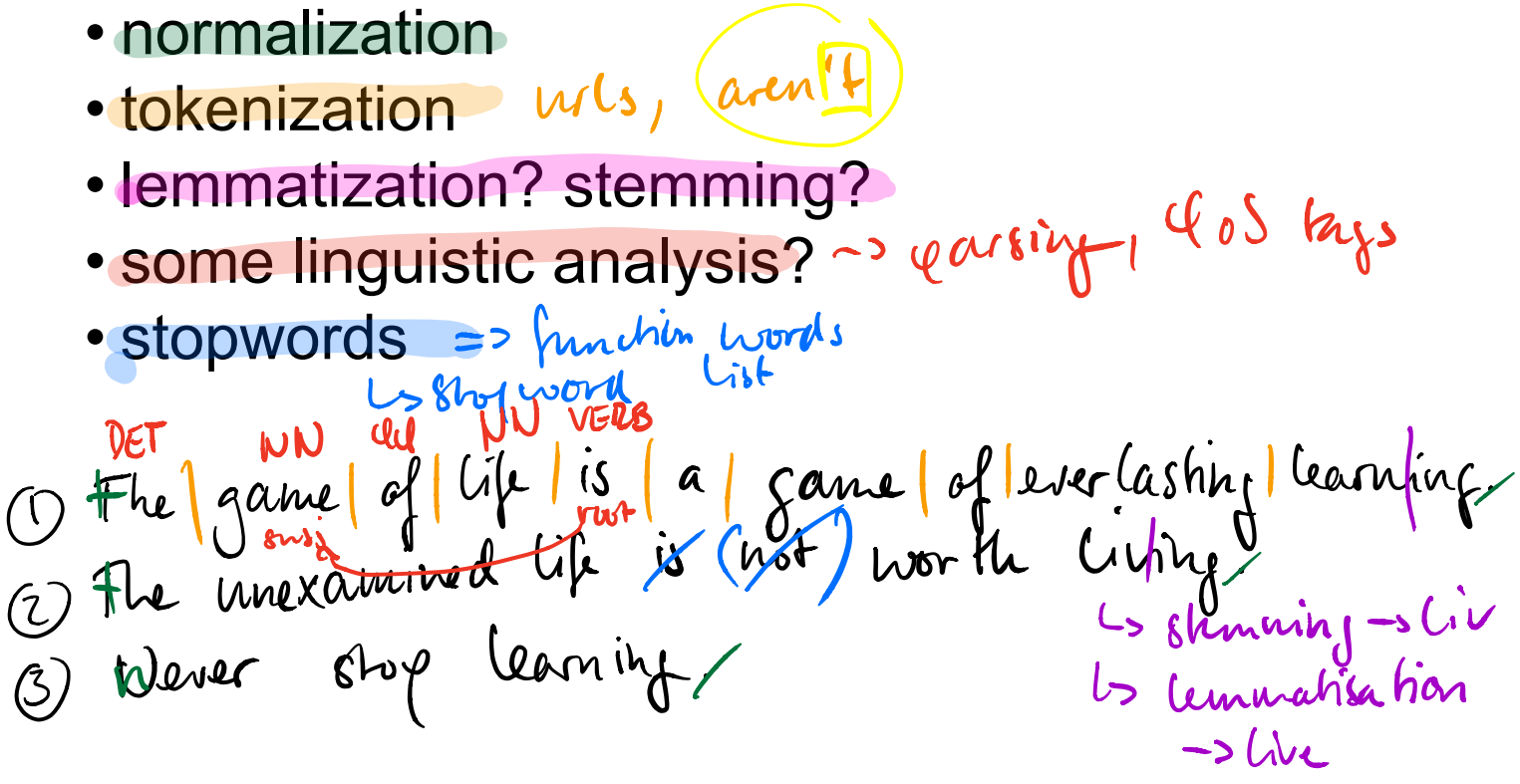
- word counts
- ngram counts
- TF/IDF counts



after preprocessing! ?

Kinds of preprocessing for text features

- normalization
- tokenization
- lemmatization? stemming?
- some linguistic analysis?
- stopwords



Frequency counts ("vectorization")

- count how often each word appears in a text

① The game of life is a game of everlasting learning.
② The unexamined life is not worth living.
③ Never stop learning.

tokens = 20

index: The game of life is a everlasting learning unexamined not worth living never stop

① [1 2 2 1 1 1 1 1 0 0 0 0 0 0]
② [1 0 0 1 1 0 0 0 1 1 1 1 0 0]
③ [0 ——— 0 1 0 ——— 0 1 1]

|V| = 14

TF/IDF vectorization

- importance weighting for frequency counts

① The game of life is a game of everlasting learning.
② The unexamined life is not worth living!
③ Never stop learning.

TF

$$\frac{\#\text{ of occurrences}}{\text{length of document}}$$

- how often a term occurs in a document

① The game of Life is a everlasting learning.
[0.1 0.2 0.2 0.1 0.1 0.1 0.1 0.1 ... 0]

② The unexamined life is not worth living.
[0.14 ——————————————— 1 ——— 0]

③ Never stop learning.
[$0.\overline{3}$ $0.\overline{3}$ $0.\overline{3}$]

IDF

$$\log\left(\frac{\#\text{ of documents}}{\#\text{ of documents where } t \text{ occurs}}\right)$$

- inverse of how many documents contain the term

$\Rightarrow$ learning

$$IDF(\text{learning}) = \log\left(\frac{3}{2}\right) = 0.4$$

TF/IDF scores

- multiply TF and IDF

① TF(learning) = 0.1 · 0.4 = 0.04

③ TF(learning) = 0.3 · 0.4 = 0.12

IDF(learning) = 0.4

in doc ③ learning is more important!

because 0.12 >> 0.04

① The game life is a everlasting learning. 0.1 ... 0]

[TF/IDF

0.04

TF/IDF (cont.)

Sequential nature / order preserving

- count features discard ordering

(A) "Mary loves songs joyfully"
= "loves joyfully Mary songs"

n-grams

① The game of life is a game of everlasting learning.
② The unexamined life is not worth living!
③ Never stop learning.

Summary

- Feature extraction is a necessary step for text data
- text is frequently **normalized** as a preprocessing step before extracting features
- **vectorization** is a common feature extraction technique

Practical

- Notebook 5.ipynb