Data-processing:

- Removing dominant authors regarding total number of posts
- Removing bulk uploads

(1)

Uetliberg Dolder

Instagram Media Object

image → [photo] ← location

text

← likes

Image contains:

RGB or HUE values for each pixel

Text contains:

- Plain text
- Hashtags (#)
- Mentions (@)
- Hyperlinks
- HTML
→ Extremly heterogenious

Image Content Recognition:

Information extraction via *Google Cloud Vision API*

Output: Labels with certainty score: E.g (Landscape, 0.89)
→ Text

(2)

+

Text-Processing: Tokenization

Goal: harmonize text & create word tokens
Steps:
1. Remove special text patterns
2. Remove anything but letters
3. Whitespace splitting
4. Basic spell-check
5. Stemming to linguistical root
6. Convert to lowercase
7. Check minimum token length
8. Remove stop words
9. Remove topographic words

(4)

Language detection

Identifying if a media object is written in either:
- English
- German
- Swiss-german
- French
- Italian

(3)

Media Object → Final text string: „today nice walk…" (5)

Media Object Nr. *n*   Media Object Nr. 3
Media Object Nr. 2   Media Object Nr. 1
„ na…
„today nice walk"

Bag of Words: Vocabulary building

- Learn the vocabulary of all media objects
- Create an array where each word token is a unique feature in a fixed order.

Feature	1	2	...	246	...	381	...	563	...	m
Token	aal	abba		nice		today		walk		zoo

(6)

Fingerprint of each media object: Encoding

- Compute the occurrence frequency of all word-tokens of a media object compared to the entire ‚bag-of-words' vocabulary.
- Create a media object specific array with the manually assigned classification
- Count-row is the *numerical* representation of the text and image labels

Media Object Nr.1 with classification: ‚*Walking*'

Feature	1	2	...	246	...	381	...	563	...	m
Token	aal	abba	...	nice	...	today	...	walk	...	zoo
Count	0	0	...	1	...	1	...	1	...	0

(7)

Machine Learning Model

$$F(Count) = Classification$$

Parameter tuning

(8)

---

(9)