



Style markers based on stop-word list

Jan Rygl, Marek Medved'

{rygl, xmedved1}@fi.muni.cz

NLP Centre, Faculty of Informatics, Masaryk University

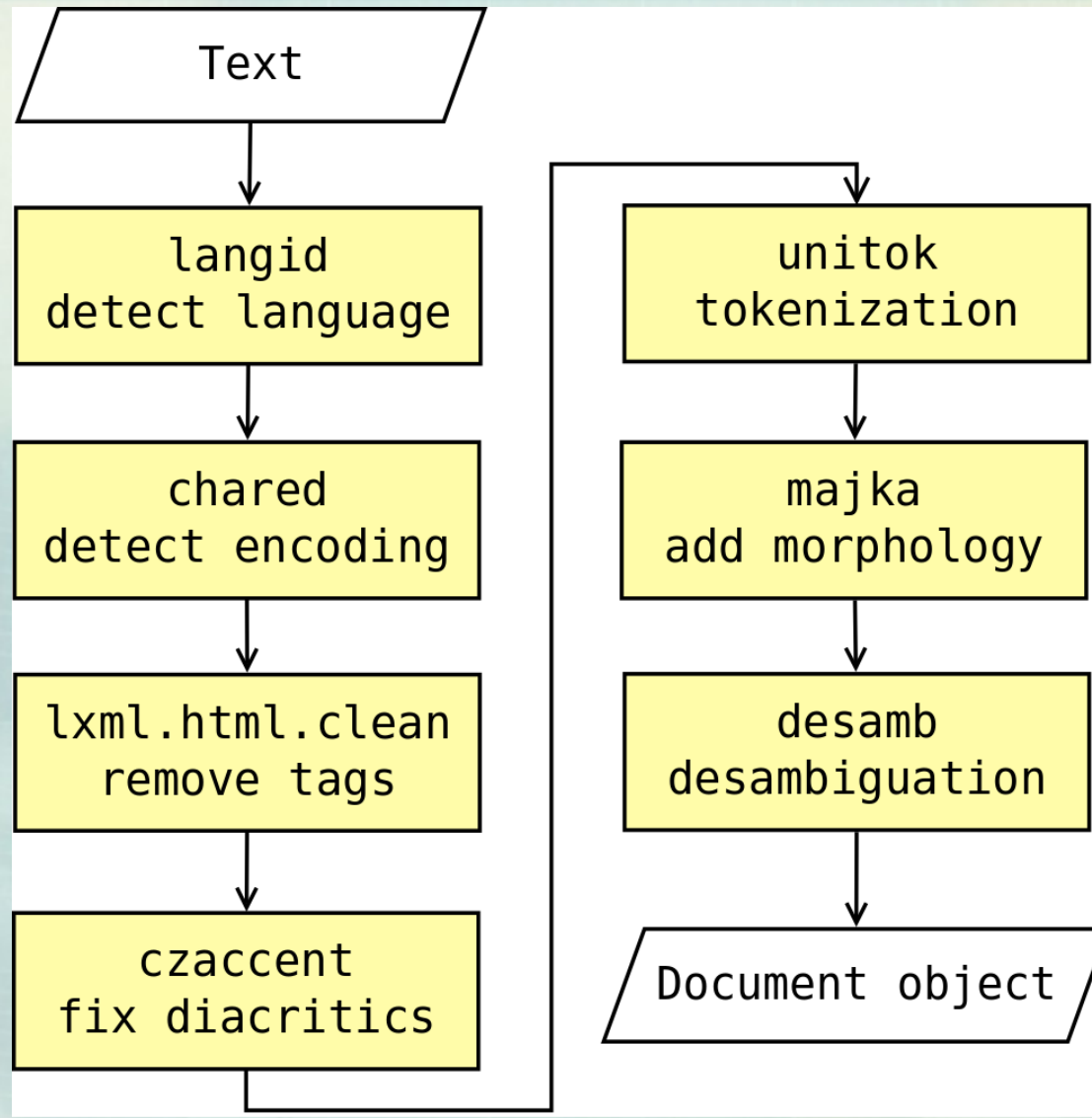
Stop-word style markers

- Efficient to solve many stylometric tasks
- Easy to implement
- Difficult to collect for analysed domain

Style marker usage

- Solve problems depending on style of author (gender, age, authorship detection)
- Preprocess document
- Analyse text using style markers
- Convert output to features
- Process by machine learning / statistically

Document preprocessing



Style marker analysis

- Presence of stop word in text
 - $\{0, 1\}$
- Relative frequency of stop word in text
 - $<0, 1>$
- Normalized difference of relative frequencies of stop word in text and in corpus

Stylometry analysis

- Find style markers which are:
 - Unique (feature values are typical only for one label)
 - Consistent (typical values are seen in almost all documents with that label)

Generation of stop words

- Use existing corpora (NLP Centre)
- Use existing tools:
 - Generate token frequencies with Sketch engine:
freqs corpora '[]' 'word 0 tag 0' > word_freq
freqs corpora '[]' 'lemma 0 tag 0' > lemma_freq
 - Filter out noise
 - Select lower limit for frequencies (more stop words, slower machine learning)

Thank you for attention

