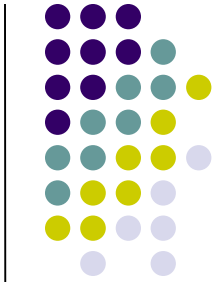
Big Data Analytics

Lab 1

UNIVERSITÄT LUZERN

Quanteda approach



`Quanteda` package has three basic types of objects:

- *`Corpus`*: it saves character strings and variables in a data frame, by also combining texts with document-level variables (where available)
- *`Tokens`*: it stores tokens in a list of vectors (in a more efficient way than character strings), while still preserving positions of words. At this stage, you can apply pre-processing
- *Document-feature matrix (`dfm`)*: it represents frequencies of features in documents in a matrix. By doing it, it does not preserve information on positions of words within each text

Text analysis with `Quanteda` via bag-of-words goes always through all those three types of objects

Quanteda approach