Preprocess: add {} manually

TODO: doppelte NAGs!!!!!

Concept

* general problem: structured data easier to evaluate than unstructured data
  + possibility to build statistics (how good is a product rated?, political attitude?, player/game statistics in chess -> average count of good/bad moves in a chess game, average count of good/bad moves by a specific player (-> talent scouting?))
  + 🡪 EXTRACT KNOWLEDGE OUT OF UNSTRUCTURED INFORMATION
* general [in this case]: comments [chess annotations] should be converted to symbols (=classes) [chess symbols]
  + step 1: define input and output
    - possible input types (symbol: detection of handwritten letters, word/sentence: detection of language, page/file: detection of author
    - possible output types (letter, language, author)
      * "precision": language = language family|language|dialect -> chess: good move vs. brilliant/good/slightly good move
  + step 2: find a database (or similar source) with sufficient information to extract data from
  + step 3: define conditions that filtered data sets has to fulfill
    - language restriction
    - minimal comment length
    - output known/unknown 🡪 supervised/unsupervised learning (in beginning)
  + step 4: tokenize the text
    - handling of punctuation
    - special letter or word combinations
  + step 5: token preprocessing
    - remove stopwords
    - lowercase
    - stemming
  + step 6: definition of features
    - absolute count of letters/words
    - relative count of letters/words
    - input as set or list? If list:
      * bigrams, trigrams
      * word context (e.g. via word2vec)
  + step 7: feature selection
    - or attribute selection included in classifier?
  + step 8: algorithm / classifier selection
    - text classifier (typically not rules)
  + step 9: evaluation of results of classifier(s)
    - accuracy
    - order of classes
* how to handle order of classes?
* how to handle differences in class counts?
* how to handle too many attributes? -> attribute selection

Motivation: Learning from chess annotations