Neural networks

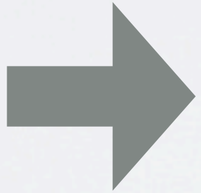
Natural language processing - preprocessing

NATURAL LANGUAGE PROCESSING

Topics: tokenization

- Typical preprocessing steps of text data
  - tokenize text (from a long string to a list of token strings)

"He's spending 7 days in San Francisco."



"He"
"'s"
"spending"
"7"
"days"
"in"
"San Francisco"
"."

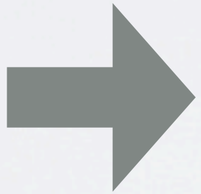
  - for many datasets, this has already been done for you
  - splitting into tokens based on spaces and separating punctuation is good enough in English or French

NATURAL LANGUAGE PROCESSING

Topics: lemmatization

- Typical preprocessing steps of text data
  - lemmatize tokens (put into standard form)

"He"	"he"
"'s"	"be"
"spending"	"spend"
"7"	"NUMBER"
"days"	"day"
"in"	"in"
"San Francisco"	"San Francisco"
"."	"."



  - the specific lemmatization will depend on the problem we want to solve
    - we can remove variations of words that are not relevant to the task at hand

NATURAL LANGUAGE PROCESSING

Topics: vocabulary

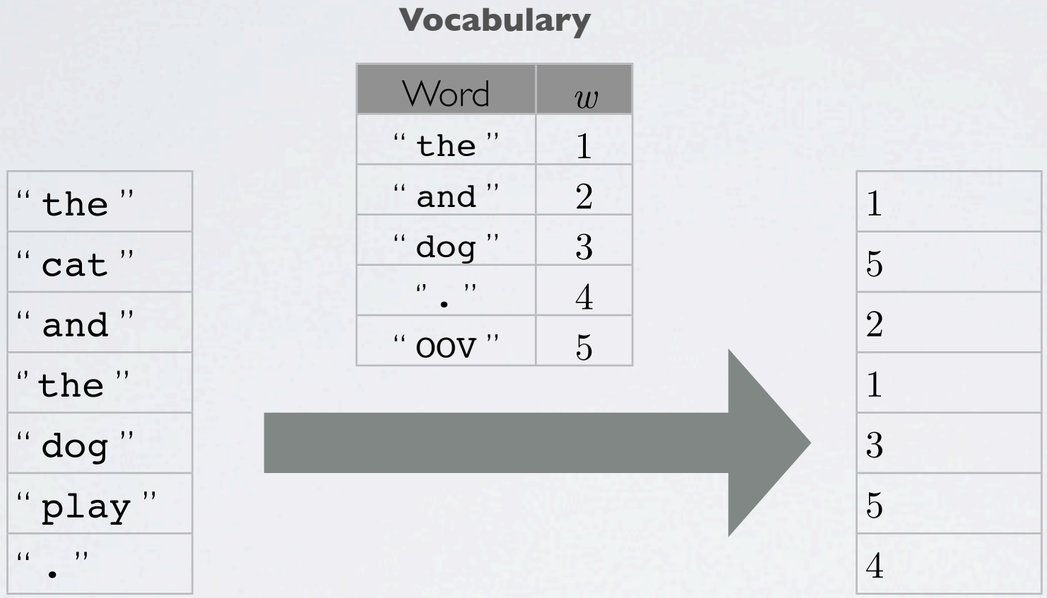
- Typical preprocessing steps of text data
  - form vocabulary of words that maps lemmatized words to a unique ID (position of word in vocabulary)
  - different criteria can be used to select which words are part of the vocabulary
    - pick most frequent words
    - ignore uninformative words from a user-defined short list (ex.: " `the` ", " `a` ", etc.)
  - all words not in the vocabulary will be mapped to a special "out-of-vocabulary" ID

- Typical vocabulary sizes will vary between 10 000 and 250 000

NATURAL LANGUAGE PROCESSING

Topics: vocabulary

- Example:



- We will note word IDs with the symbol w
  - can think of w as a categorical feature for the original word
  - we will sometimes refer to w as a word, for simplicity