Working with Text Data | Basic definitions

February 22, 2021

> "The goal is to amplify reading"

Possible types of strings

- Categorical Data
- Free strings that can be (semantically) mapped to categories
- Structured string data
- Text data

Typology

Corpus A dataset of *documents*

Document A single data point in corpus, a single text.

Definitions

bag-of-words A text that reduced only to the count of how often the words appear in it after discarding the structure (chapters, paragraphs, sentences, formatting etc.) in it.

tokenization Split each document into the words that appear in it (tokens).

vocabulary building create a vocabulary of all unique words appear in any of the documents and order them (for example alphabetically).

encoding For each document, count how often each of the words in vocabulary appear in that document.

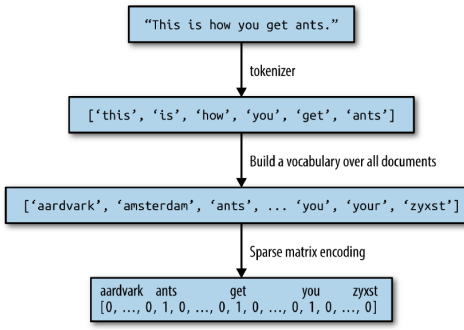

Figure 7-1. Bag-of-words processing

Rescale with tf-idf

$$tfidf(w,d) = tf \log\left(\frac{N+1}{N_w+1}\right) + 1 \tag{1}$$

N Number of documents

N_w Number of documents word w is appearing in.

tf Number of w occurences in the specific document we are transforming.

n-Grams and the shortcomings of bag-of-words

The main disadvantage of the bag-of-words approach is that it is completely overlooking the order of words, the structure of the text. [1]. n-Grams (bigrams, trigrams ...) approach is trying to solve this problem.

[1] *it's bad, not good at all == it's good, not bad at all*

This feature can be applied by defining `ngram_range` in `CountVectorizer` or `TfidfVectorizer`.