Machine Learning Methods for Language Processing

University of Applied Sciences
Computer Science and Digital Communication

April 5, 2024

Exercise outline

Programming task - statistical language model (2-gram, 3-gram)

1. load corpus
2. tokenize loaded corpus
3. split corpus into training (80%) and testing chunks (20%)
4. obtain vocabulary of tokenized training corpus
5. write the vocabulary of tokenized training corpus alphabetically sorted to file
6. count n-grams over tokenized training corpus
7. write the n-gram counts to a file
8. calculate n-gram probabilities from n-gram counts
9. calculate perplexity over tokenized test corpus

Exercise - technical conditions

to be considered

- python 3 and pycharm
- each group uses one corpus:
  - corpus 1 Shakespeare - Hamlet
  - corpus 2 Shakespeare - Macbeth
  - corpus 3 Milton - Paradise
  - corpus 4 Blake - Poems

Exercise

to be considered

- design wrt object-oriented paradigma
- think of appropriate and efficient data structures
- think of methods to be implemented for the desired functions
- think of input and output
- verify the vocabulary in terms of quality
- verify the counts and probabilities in terms of plausibility
- analyse the perplexity