Machine Translation
WiSe 2015/2016

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Words, Sentences, Corpora - Exercises

Dr. Mariana Neves

October 19th, 2015

Goal

- Get familiar with the dataset and HANA
- Get interesting (creative) insights from the text

Exercises

- Check tokenization and word segmentation
  - Compound words
  - Contractions
  - Casing

Exercises

- Analyze the vocabulary
  - Size
  - Word tokens, word types
  - Functional and non-functional words
  - Stem and lemma
  - Suffixes, Prefixes
  - Distribution of words, Zipf's law

Exercise

- Deadline
  - Sunday, Nov 1st, 23:59
- Hand-in
  - SQL file with queries
- Presentation
  - Monday, Nov 2nd
  - More details next week
- Not graded, but all teams need to present at some point