Corpus and Lexicon

Natural Language Understanding Lab

Evgeny A. Stepanov,
Mahed Mousavi, Gabriel Roccabruna

SISL, DISI, UniTN & VUI, Inc.
evgeny.stepanov@unitn.it

Objectives

- Understanding:
  - relation between corpus and lexicon
  - effects of pre-processing (tokenization) on lexicon
- Learning how to:
  - load basic corpora for processing
  - compute basic descriptive statistic of a corpus
  - building lexicon and frequency lists from a corpus
  - perform basic lexicon operations
  - perform basic text pre-processing (tokenization and sentence segmentation) using python libraries

Outline

1. Corpora and Counting
    - Loading corpus in NLTK
    - Computing Corpus Descriptive Statistics
        - *Exercise*: 15 min
2. Lexicon
    - Computing Lexicon and Lexicon Size
    - Computing Frequency List
        - *Exercises*: 15 min
    - Lexicon Operations: Cut-Off and Stopwords
        - *Exercises*: 15 min
3. Basic Text Pre-processing
    - Tokenization and Sentence Segmentation
4. ***Lab Exercise***: 30 min

Recommended Reading

- Dan Jurafsky and James H. Martin. Speech and Language Processing (3rd ed. draft)
  - Chapter 2: Regular Expressions, Text Normalization, Edit Distance
- Steven Bird, Ewan Klein, and Edward Loper. Natural Language Processing with Python
  - Chapter 2: Accessing Text Corpora and Lexical Resources
  - Chapter 3: Processing Raw Text