Statistical Language Modeling with NLTK

Natural Language Understanding Lab

Evgeny A. Stepanov,
Mahed Mousavi, Gabriel Roccabruna

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

Objectives

- Understanding:
  - relation between lexicon and language model
  - smoothing techniques
  - OOV effects and remedies
- Learning how to:
  - prepare data for ngram modeling
  - count ngrams in a corpus
  - train an ngram model with NLTK
  - use ngram model to
    - compute probability of a sequence
    - generate a sequence
  - evaluate ngram model

Outline

1. Ngram and Ngram Counting
    - Counting Ngrams
        - Preparing Data
        - Ngram Extraction
    - `NgramCounter` class
    - *Exercise*: 10 min
2. Vocabulary and Basic Usage
    - Counts, Vocabulary Membership, and Lookup
    - Cut-Off and Vocabulary Size
    - *Exercise*: 10 min
3. Training Language Model
    - Computing Ngram Probabilities
    - Training Ngram Models with NLTK

Outline

4. Using Ngram Language Models
    - Computing Probability of a Sequence (Scoring)
        - Scoring with NLTK LM
        - *Exercise*: 15 min
        - OOVs
        - Missing Ngram (Data Sparseness)
    - Generating Sequences
    - LM Evaluation
    - Handling Data Sparseness: Smoothing Techniques
    - NLTK Language Models
5. ***Lab Exercise***: Language Model Evaluation: 30 min

Recommended Reading

- Dan Jurafsky and James H. Martin. Speech and Language Processing (3rd ed. draft)
  - Chapter 3: N-gram Language Models