Vector Semantics

Natural Language Understanding Lab

Evgeny A. Stepanov,
Mahed Mousavi, Gabriel Roccabruna

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

Objectives

- Understanding:
  - different methods of representing words as vectors
  - vectors and similarity between vectors
  - evaluation of word embeddings
- Learning how to:
  - train word embeddings with `gensim`
  - use pre-trained word embeddings for similarity computation

Outline

1. Words as Vectors (Embeddings)
    - One-Hot Encoding
2. Co-Occurence Matrices and Word as Vectors
    - Term-Document Matrix
    - TF-IDF
    - Term-Term Matrix
    - Pointwise Mutual Information (PMI) and Positive Pointwise Mutual Information (PPMI)
    - Building Co-Occurence Matrix
    - *Exercises*: 20 min
3. Vector Similarity
    - Dot Product
    - Cosine Similarity
    - *Exercises*: 20 min

Outline

4. Training Word Embeddings with `gensim`
   - Word2Vec
5. Pre-Trained Embeddings
   - Word Embeddings in `spacy`
   - Evaluation: Analogy Task
6. ***Lab Exercise***: Analogy Task: 30 min

Recommended Reading

- Dan Jurafsky and James H. Martin. Speech and Language Processing (3rd ed. draft)
  - Chapter 6: Vector Semantics and Embeddings