Word Embeddings and Vector Space Models

Summer

Introduction

This document provides an overview of word embeddings and vector space models, focusing on their role in NLP and distributional semantics. Key methods, applications, and case studies are discussed.

Distributional Semantics

- **Definition**: "Words with similar distributional properties have similar meanings" (Sahlgren, 2006).
- **Key Ideas**:
  - Words co-occurring frequently share similar meanings.
  - Contextual relationships define semantic meaning.
- **Applications**:
  - Topic modeling
  - Conceptual mapping
  - Word embeddings

Vector Space Models (VSMs)

- **Document-Term Matrices**:
  - Represent documents as vectors.
  - Similarity measured using cosine similarity.
- **Term-Term Matrices**:
  - Context window determines associations.
  - Useful for semantic similarity tasks.
- **Latent Semantic Analysis (LSA)**:
  - Reduces dimensionality.
  - Captures latent relationships.

Word Embeddings

- **Word2Vec**:
  - Continuous Bag of Words (CBOW) and Skip-Gram models.
  - Efficient representation of word relationships.
- **GloVe**:
  - Combines global and local co-occurrence information.
- **Applications**:
  - Semantic analysis
  - Document classification
  - Sentiment analysis

Collocations and Associations

- **Collocations**:
  - Frequently co-occurring word pairs.
  - Statistical measures: Mutual Information (MI), log-likelihood, chi-square.
- **Associations**:
  - Looser semantic relationships.
  - Captured via large context windows.

Case Studies

- **Migration and Sentiment**:
  - Kernel Density Estimation to analyze temporal sentiment changes.
- **Associations to Cider, Wine, and Beer**:
  - Analyzed using word embeddings and conceptual maps.

Conceptual Maps

- Visual representations of semantic relationships.
- Derived from Kernel Density Estimation and large-scale embeddings.
- Applications include sentiment analysis and trend detection.

References

- Sahlgren, M. *The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations Between Words.*
- Deerwester, S. *Latent Semantic Analysis for Document Classification.*
- Schneider, G. *Text Analytics in the Digital Humanities.*