



Next >

< Previous

∷ Hide menu

Lecture: Word Embeddings

- Video: Week Introduction
 1 min
- Video: Overview 2 min
- Reading: Overview
- Video: Basic Word Representations 3 min
- Reading: Basic Word Representations
 5 min
- Video: Word Embeddings
 3 min
- Reading: Word Embeddings
- Video: How to Create
 Word Embeddings
 3 min
- Reading: How to Create Word Embeddings?
 4 min
- Video: Word Embedding
 Methods
 3 min
- Reading: Word
 Embedding Methods
 4 min
- Video: Continuous Bagof-Words Model
- Reading: Continuous Bag of Words Model
 3 min
- Video: Cleaning and Tokenization
 4 min

Cleaning and Tokenization

Before implementing any natural language processing algorithm, you might want to clean the data and tokenize it. Here are a few things to keep track of when handling your data.

- Letter case "The" == "the" == "THE" → lowercase / upper case
- Punctuation ,!.? →. "'«»'" → ∅ ...!!???? →.
- Numbers 1 2 3 5 8 → Ø 3.14159 90210 → as is/<NUMBER>
- Special characters ∇ \$ € § ¶ ** → Ø
- Special words
 ⊕ #nlp → :happy: #nlp

You can clean data using python as follows:

You can add as many conditions as you want in the lines corresponding to the green rectangle above.