Week 7 Introduction

Topic Introduction

Week 7 Lesson Video

(21:52)

Watch this video for a comprehensive introduction to this week's learning content. You can follow along in Blackboard as you watch, with the video organized into the following chapters:

- 7.1 Lesson: Vector Models
- 7.2 Lesson: Word Embeddings (05:28)
- 7.3 Lesson: Sentence and Document Embeddings (17:58)

Required Resources

The following resources are required for your learning this week. Make sure you review everything linked below, as you may be tested on the concepts in an upcoming Knowledge Check or Homework.

Reading | Brownlee, J. (2019, August 7). A gentle introduction to the bag of words model (https://machinelearningmastery.com/gentle-introduction-bag-words-model/). *Machine Learning Mastery*.

- A blog post covering the main points, with Python, by *the* Jason Brownlee (1 of my favorite machine learning bloggers).

Reading | Tam, A. (2025, September 12). Word embeddings in language models (https://machinelearningmastery.com/word-embeddings-in-language-models/). *Machine Learning Mastery*.

- Another useful post on *Machine Learning Mastery.*

Video | Taher Pilehvar, M. (2020, December 25). Word embeddings (embeddings in NLP) (https://www.youtube.com/watch?v=UELW9EnibYk) [Video]. YouTube.

- A comprehensive tutorial including many details not covered here, well worth watching the whole thing!

Video | Taher Pilehvar, M. (2020, December 26). Sentence and document embeddings (embeddings in NLP) [Video]. (https://www.youtube.com/watch?v=oaX_th7mVIl) YouTube.

- A shorter tutorial by the same author, also excellent. More advanced tutorials on embeddings are presented in the author's YouTube channel.

Key Terms

- **Token**: A basic unit of text, typically a word or subword, used in NLP processing.
- **Integer Encoding**: A simple token representation using an index from a vocabulary list.
- **One-Hot Encoding**: A sparse vector where only 1 position is 1 (hot), and all others are 0, representing a token.
- **Bag-of-Words (BOW)**: A vector that counts how often each word appears in a document, ignoring word order.
- **Term Frequency Vector**: Another name for a BOW vector, focused on raw word counts.
- **Cosine Similarity**: A metric that compares two vectors by the angle between them, commonly used to measure text similarity.

- **Embedding**: A dense, low-dimensional vector that captures semantic relationships between words or texts.
- **Word2Vec**: A method that learns word embeddings by predicting surrounding words (skip-gram or CBOW).
- **Skip-gram**: A Word2Vec approach where a word is used to predict its surrounding context.
- **Sentence Embedding**: A single vector that represents an entire sentence, capturing its overall meaning.
- **Document Embedding**: A dense vector representing a whole document for tasks like classification or clustering.
- **Doc2Vec**: An extension of Word2Vec that learns embeddings for documents as well as words.