DATASCI207-005/007 Applied Machine Learning

Vilena Livinsky, PhD(c)

School of Information, UC Berkeley

Week 9: 03/05/2025 & 03/06/2025

Today's Agenda

- Embeddings for Text
- Walkthroughs:
  - Embedding example
  - Modeling with text

Words & Meaning

What is a word?

- How to represent word meaning?

Ludwig Wittgenstein:

- “The meaning of a word is its use in the language.”

Vector Semantics

- Words that occur in similar **contexts** tend to have similar meanings
- Embeddings:
  - **learning** representations of the meaning of words directly from their distributions in texts
- There are *self-supervised* ways to learn representations of the input
  - vs. by hand via feature engineering
  - concern of NLP research

Ref.: Speech and Language Processing. Daniel Jurafsky & James H. Martin. C

Vector Semantics

- representing a word as a point in a multidimensional semantic space that is derived from the distributions of word neighbors
  - Word vectors: *embeddings*
    - a stricter application: to only dense vectors, ex.: word2vec
- vector semantic models can be learned automatically from text without supervision

For example, suppose you didn't know the meaning of the word *ongchoi* (a recent borrowing from Cantonese) but you see it in the following contexts:

(6.1) Ongchoi is delicious sauteed with garlic.

(6.2) Ongchoi is superb over rice.

(6.3) ...ongchoi leaves with salty sauces...

And suppose that you had seen many of these context words in other contexts:

(6.4) ...spinach sauteed with garlic over rice...

(6.5) ...chard stems and leaves are delicious...

(6.6) ...collard greens and other salty leafy greens

Fig. 6.1 shows a visualization of embeddings learned for sentiment analysis, showing the location of selected words projected down from 60-dimensional space into a two dimensional space. Notice the distinct regions containing positive words, negative words, and neutral function words.

Ref.: Speech and Language Processing. Daniel Jurafsky & James H. Martin. C

Distance: Tokens (One-Hot Encoding)

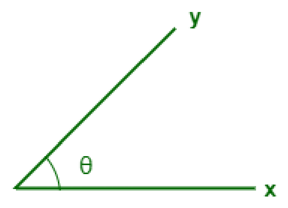
$$d(A,B) = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2}$$

A and B in the i-th dimension

Term	One-Hot Encoding
loan	[1, 0, 0]
investment	[0, 1, 0]
insurance	[0, 0, 1]

Equal Distance: The Euclidean distance between any two different vectors is the same.

- Distance between "loan" and "investment": $\sqrt{(1-0)^2+(0-1)^2+(0-0)^2} = \sqrt{1+1+0} = \sqrt{2}$
- Distance between "loan" and "insurance": $\sqrt{(1-0)^2+(0-0)^2+(0-1)^2} = \sqrt{1+0+1} = \sqrt{2}$
- Distance between "investment" and "insurance": $\sqrt{(0-0)^2+(1-0)^2+(0-1)^2} = \sqrt{0+1+1} = \sqrt{2}$



$$\cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2}\cdot\sqrt{\sum\limits_{i=1}^{n} B_i^2}},$$

cosine similarity always belongs to the interval [-1, 1]

Cosine Similarity between "loan" and "insurance":

- Dot product: $[1,0,0]\cdot[0,0,1] = 0$
- Magnitudes: $\|[1,0,0]\| = 1$ and $\|[0,0,1]\| = 1$
- Cosine similarity: $\frac{0}{1\cdot 1} = 0$

Cosine Similarity between "loan" and "investment":

- Dot product: $[1,0,0]\cdot[0,1,0] = 0$
- Magnitudes: $\|[1,0,0]\| = 1$ and $\|[0,1,0]\| = 1$
- Cosine similarity: $\frac{0}{1\cdot 1} = 0$

Cosine Similarity between "investment" and "insurance":

- Dot product: $[0,1,0]\cdot[0,0,1] = 0$
- Magnitudes: $\|[0,1,0]\| = 1$ and $\|[0,0,1]\| = 1$
- Cosine similarity: $\frac{0}{1\cdot 1} = 0$

Building Language Models: A Basic Overview

Preprocessing → Tokenization/ Vectorization → Embedding → Modeling

Preprocessing

- **Normalize** text
  - Lowercase
  - Remove punctuation
  - Etc.
- Lemmatization: sing: sing, sang, sung
- Stop word removal
- URL formatting
- Etc.

Tokenization/ Vectorization

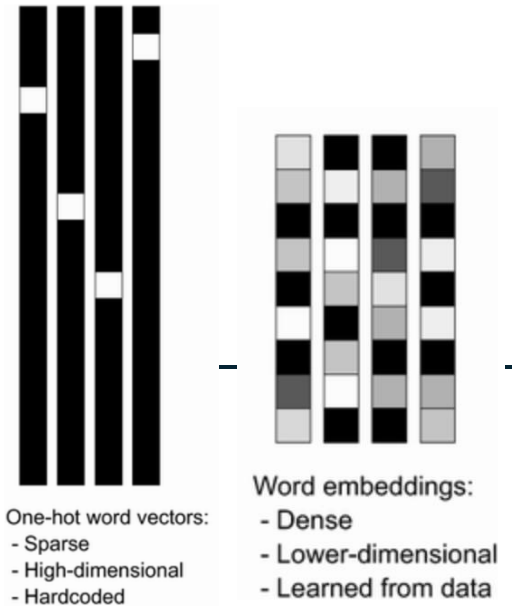
- Segment text (‘token’)
  - *Word-level tokenization*
  - *N-gram tokenization*
  - Ex.: *“the cat sat on the mat.”*
  - *Character-level*
- Map to an identifier
  - index of all terms
  - convert integer into a vector

Embedding

- **One-hot** encoding
- Learn an **embedding**
- Use an existing embedding
- Etc.

Modeling

- Train a model



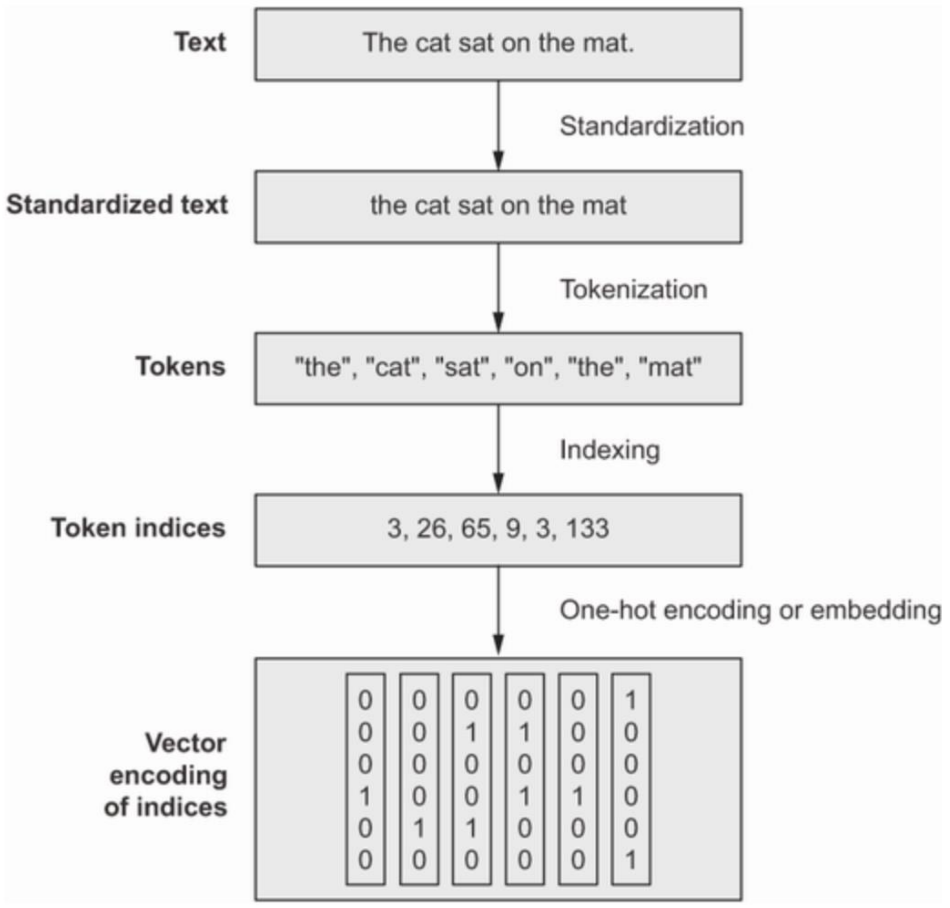


Image Ref.: Chollet, F., & Chollet, F. (2021). Deep learning with Python. Simon and Schuster.

Perfect Embedding?

- Is there some ideal word-embedding space?
  - Can we map human language perfectly so that such an embedding space could be used for any natural language processing task?