CPSC 4300/6300: Applied Data Science

2024

Preparing text data

Dr. Aaron J. Masino
Associate Professor, School of Computing

College of
ENGINEERING, COMPUTING
AND APPLIED SCIENCES

Text representation

- How can we represent text for computation?
  - Tokenization
  - Token embeddings

Patient Visit Summary Note
Date: [Date of Visit]
Patient: Mr. [Patient's Last Name], 63-year-old male
Medical History: Known history of Type 2 Diabetes Mellitus, Hyperlipidemia
Medications: Metformin 500mg BID, Atorvastatin 20mg daily, Lisinopril 10mg daily
Allergies: No known drug allergies
Social History: Non-smoker, consumes alcohol socially, retired engineer, lives with spouse
Chief Complaint:
Patient presents for routine follow-up of diabetes and high cholesterol.
Assessment/Plan:
1.Type 2 Diabetes Mellitus: Stable. Continue current regimen of Metformin. Consider HbA1c check if not performed in the last 3 months.
2.Hyperlipidemia: Continue Atorvastatin. Reinforce dietary advice regarding low cholesterol diet. Consider lipid panel if not done in the past year.
3.Hypertension: Blood pressure slightly elevated. Continue current dose of Lisinopril and reinforce lifestyle modifications for blood pressure control (DASH diet, weight loss, regular exercise).
Laboratory Tests Ordered:
•Hemoglobin A1c (HbA1c)
•Fasting lipid panel
•Comprehensive metabolic panel (CMP)
•Thyroid-stimulating hormone (TSH)
•Urine Albumin-to-Creatinine Ratio (UACR)

Tokenization

Continue Atorvastatin. Reinforce dietary advice regarding low cholesterol diet.

Input text

Continue Atorvastatin.

Reinforce dietary advice regarding low cholesterol diet.

Sentence tokenization

Continue | Atorvastatin | . | Reinforce

diet | ary | advice | regard | ing

low | cholesterol | diet | .

Word / subword tokenization

Process splits sentences into their constituent linguistic tokens (word stems, suffix, etc)

4 2 10 9 5 1 8

0 6 7 3 5 10

Map to vocabulary token indices

Vocabulary

Token	Index
advice	0
ary	1
Atorvastatin	2
cholesterol	3
continue	4
diet	5
ing	6
low	7
regard	8
reinforce	9
.	10
<eoi>	11

One-hot embeddings

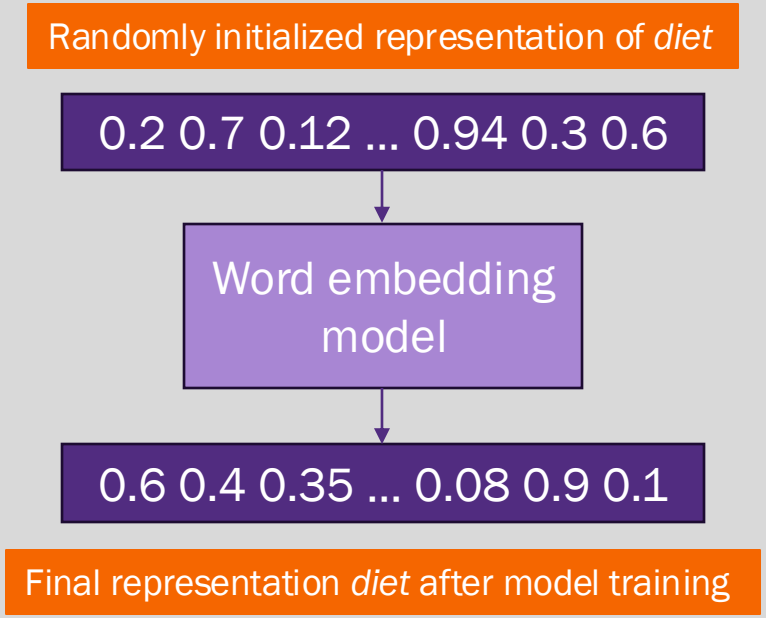
Reinforce → 9	0	0	0	0	0	0	0	0	0	1	0	0
diet → 5	0	0	0	0	0	1	0	0	0	0	0	0
ary → 1	0	1	0	0	0	0	0	0	0	0	0	0
advice → 0	0	0	1	0	0	0	0	0	0	0	0	0
regard → 8	0	0	0	0	0	0	0	0	1	0	0	0
ing → 6	0	0	0	0	0	0	1	0	0	0	0	0
low → 7	0	0	0	0	0	0	0	1	0	0	0	0
cholesterol → 3	0	0	0	1	0	0	0	0	0	0	0	0
diet → 5	0	0	0	0	0	1	0	0	0	0	0	0
. → 10	0	0	0	0	0	0	0	0	0	0	0	1

Every word is assigned to a *V-dimensional* vector where *V* is the size of the vocabulary

Vocabulary	
Token	**Index**
advice	0
ary	1
Atorvastatin	2
cholesterol	3
continue	4
diet	5
ing	6
low	7
regard	8
reinforce	9
.	10
<eoi>	11

Token embeddings

- In practice, our vocabulary will contain *O(10K)* to *O(100K)* tokens
- We could encode the tokens with *V* dimensional one-hot vectors (*V* is length of vocabulary)
  - Computationally inefficient
  - Similar tokens lack similar vectors
- Instead, tokens are represented with *N<<V* dimensional vectors called embeddings
  - N is typically *O(100)*
  - Embeddings are randomly initialized or initialized from other methods (e.g., word2vec)
  - Embeddings may be updated during model training



Token embedding properties

- Most token embedding models yield tokens that encode semantic information
- Tokens with similar semantics are closer to each other in their latent space than tokens with different semantics
- Semantic “operations” are possible to obtain tokens with “root” meanings