Word Vectors (Part 1)

CS114B Lab 6

Kenneth Lai

March 10, 2022

The Distributional Hypothesis

- "You shall know a word by the company it keeps." (Firth 1957)
- "It may be presumed that any two morphemes A and B having different meanings, also differ somewhere in distribution: there are some environments in which one occurs and the other does not." (Harris 1951)
- "The similarity of the contextual representations of two words contributes to the semantic similarity of those words." (Miller and Charles 1991) (emphasis mine)

The Distributional Hypothesis

- "You shall know a word by the company it keeps." (Firth 1957)
- "It may be presumed that any two morphemes A and B having different meanings, also differ somewhere in distribution: there are some environments in which one occurs and the other does not." (Harris 1951)
- "The similarity of the contextual representations of two words contributes to the semantic similarity of those words." (Miller and Charles 1991) (emphasis mine)

- Words can be represented by (abstractions over) their contexts

Distributed Representations of Words

- Representations of (contexts of) words as embeddings in some vector space

Distributed Representations of Words

- Representations of (contexts of) words as embeddings in some vector space
- Two approaches to distributed, distributional representations (Baroni et al. 2014):

Distributed Representations of Words

- Representations of (contexts of) words as embeddings in some vector space
- Two approaches to distributed, distributional representations (Baroni et al. 2014):
  - Count-based

Distributed Representations of Words

- Representations of (contexts of) words as embeddings in some vector space
- Two approaches to distributed, distributional representations (Baroni et al. 2014):
  - Count-based
    - Count occurrences of words in contexts, optionally followed by some mathematical transformation (e.g. tf-idf, PPMI, SVD)

Distributed Representations of Words

- Representations of (contexts of) words as embeddings in some vector space
- Two approaches to distributed, distributional representations (Baroni et al. 2014):
  - Count-based
    - Count occurrences of words in contexts, optionally followed by some mathematical transformation (e.g. tf-idf, PPMI, SVD)
  - Prediction-based

Distributed Representations of Words

- Representations of (contexts of) words as embeddings in some vector space
- Two approaches to distributed, distributional representations (Baroni et al. 2014):
  - Count-based
    - Count occurrences of words in contexts, optionally followed by some mathematical transformation (e.g. tf-idf, PPMI, SVD)
  - Prediction-based
    - Given some context vector(s) **c**, predict some word **x** (or vice versa)
    - a.k.a. language modeling-based

Count-Based Word Vectors

- What are contexts?

Count-Based Word Vectors

- What are contexts?
    - Contexts are documents
        - Term-document matrices

Count-Based Word Vectors

- What are contexts?
  - Contexts are documents
    - Term-document matrices
  - Contexts are words (within some window)
    - Term-term matrices

Term-Document Matrices

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Figure 6.3 The term-document matrix for four words in four Shakespeare plays. The red boxes show that each document is represented as a column vector of length four.

- Document vectors: coordinates are counts of each word in the document

Term-Document Matrices

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Figure 6.5 The term-document matrix for four words in four Shakespeare plays. The red boxes show that each word is represented as a row vector of length four.

- Word vectors: coordinates are counts of the word in each document

Term-Term Matrices

	aardvark	...	computer	data	result	pie	sugar	...
cherry	0	...	2	8	9	442	25	...
strawberry	0	...	0	0	1	60	19	...
digital	0	...	1670	1683	85	5	4	...
information	0	...	3325	3982	378	5	13	...

Figure 6.6 Co-occurrence vectors for four words in the Wikipedia corpus, showing six of the dimensions (hand-picked for pedagogical purposes). The vector for *digital* is outlined in red. Note that a real vector would have vastly more dimensions and thus be much sparser.

- Word vectors: coordinates are counts of times the row (target) word and the column (context) word co-occur in some context in some training corpus

Term-Term Matrices

	aardvark	...	computer	data	result	pie	sugar	...
cherry	0	...	2	8	9	442	25	...
strawberry	0	...	0	0	1	60	19	...
digital	0	...	1670	1683	85	5	4	...
information	0	...	3325	3982	378	5	13	...

Figure 6.6 Co-occurrence vectors for four words in the Wikipedia corpus, showing six of the dimensions (hand-picked for pedagogical purposes). The vector for *digital* is outlined in red. Note that a real vector would have vastly more dimensions and thus be much sparser.

- Word vectors: coordinates are counts of times the row (target) word and the column (context) word co-occur in some context in some training corpus
    - e.g., in a 4 word window (4 words to the left and 4 words to the right)

Weighting Terms in Vectors

- Words that occur frequently in some contexts are important

Weighting Terms in Vectors

- Words that occur frequently in some contexts are important
  - But words that occur frequently in every context are not!

Weighting Terms in Vectors

- Words that occur frequently in some contexts are important
  - But words that occur frequently in every context are not!
- Term frequency-inverse document frequency (tf-idf)
  - Words that occur in most/all documents are less important
  - More useful for document vectors

Weighting Terms in Vectors

- Words that occur frequently in some contexts are important
  - But words that occur frequently in every context are not!
- Term frequency-inverse document frequency (tf-idf)
  - Words that occur in most/all documents are less important
  - More useful for document vectors
- Positive pointwise mutual information (PPMI)
  - How often do two words co-occur in some context, compared with what we would expect by chance?

The Distributional Hypothesis

- "You shall know a word by the company it keeps." (Firth 1957)
- "It may be presumed that any two morphemes A and B having different meanings, also differ somewhere in distribution: there are some environments in which one occurs and the other does not." (Harris 1951)
- "The similarity of the contextual representations of two words contributes to the semantic similarity of those words." (Miller and Charles 1991) (emphasis mine)

- Similar words should have similar contexts

The Distributional Hypothesis

- "You shall know a word by the company it keeps." (Firth 1957)
- "It may be presumed that any two morphemes A and B having different meanings, also differ somewhere in distribution: there are some environments in which one occurs and the other does not." (Harris 1951)
- "The similarity of the contextual representations of two words contributes to the semantic similarity of those words." (Miller and Charles 1991) (emphasis mine)

- Similar words should have similar vectors

The Distributional Hypothesis

- "You shall know a word by the company it keeps." (Firth 1957)
- "It may be presumed that any two morphemes A and B having different meanings, also differ somewhere in distribution: there are some environments in which one occurs and the other does not." (Harris 1951)
- "The similarity of the contextual representations of two words contributes to the semantic similarity of those words." (Miller and Charles 1991) (emphasis mine)

- Similar words should have similar vectors
  - How to measure similarity (or distance) between vectors?

Euclidean Distance

- A normed vector space is a vector space with a norm

Euclidean Distance

- A normed vector space is a vector space with a norm
  - A norm generalizes the notion of "length"

Euclidean Distance

- A normed vector space is a vector space with a norm
  - A norm generalizes the notion of "length"
- On a real coordinate space, we can define the Euclidean norm

$$||\mathbf{x}|| = \sqrt{\sum_{j=1}^{n} x_j^2} = \sqrt{\mathbf{x} \cdot \mathbf{x}}$$

Euclidean Distance

- A normed vector space is a vector space with a norm
  - A norm generalizes the notion of "length"
- On a real coordinate space, we can define the Euclidean norm

$$||\mathbf{x}|| = \sqrt{\sum_{j=1}^{n} x_j^2} = \sqrt{\mathbf{x} \cdot \mathbf{x}}$$

- A norm induces a distance (induced metric)

Euclidean Distance

- A normed vector space is a vector space with a norm
  - A norm generalizes the notion of "length"
- On a real coordinate space, we can define the Euclidean norm

$$||\mathbf{x}|| = \sqrt{\sum_{j=1}^{n} x_j^2} = \sqrt{\mathbf{x} \cdot \mathbf{x}}$$

- A norm induces a distance (induced metric)
- Euclidean distance

$$d(\mathbf{x}, \mathbf{y}) = ||\mathbf{y} - \mathbf{x}|| = \sqrt{\sum_{j=1}^{n} (y_j - x_j)^2}$$

Cosine Similarity

- Euclidean distance favors long vectors

Cosine Similarity

- Euclidean distance favors frequent words

Cosine Similarity

- Euclidean distance favors frequent words
- Consider the “angle” between two vectors, rather than distance

Cosine Similarity

- Euclidean distance favors frequent words
- Consider the "angle" between two vectors, rather than distance
- Cosine similarity

$$\cos(\theta) = \frac{\mathbf{x} \cdot \mathbf{y}}{||\mathbf{x}||\ ||\mathbf{y}||}$$

Sparse and Dense Vectors

- tf-idf and PPMI vectors are long and sparse

Sparse and Dense Vectors

- tf-idf and PPMI vectors are long and sparse
- We want to learn vectors that are short and dense

Principal Component Analysis

- Idea: Given a matrix $\mathbf{X}$ (where rows are vectors), apply a linear map $\mathbf{V}$

$$\mathbf{T} = \mathbf{X} \cdot \mathbf{V}$$

Principal Component Analysis

- Idea: Given a matrix $\mathbf{X}$ (where rows are vectors), apply a linear map $\mathbf{V}$

$$\mathbf{T} = \mathbf{X} \cdot \mathbf{V}$$

- If we choose $\mathbf{V}$ correctly, then $\mathbf{T}$ is a projection of $\mathbf{X}$ onto a new coordinate space, where the first coordinate (called the first principal component) has the greatest variance, the second coordinate has the second greatest variance, etc.

Principal Component Analysis

- Idea: Given a matrix $\mathbf{X}$ (where rows are vectors), apply a linear map $\mathbf{V}$
$$\mathbf{T} = \mathbf{X} \cdot \mathbf{V}$$
- If we choose $\mathbf{V}$ correctly, then $\mathbf{T}$ is a projection of $\mathbf{X}$ onto a new coordinate space, where the first coordinate (called the first principal component) has the greatest variance, the second coordinate has the second greatest variance, etc.
  - (Specifically, the columns of $\mathbf{V}$ are the eigenvectors of $\mathbf{X}^T \cdot \mathbf{X}$)

Principal Component Analysis

- Then we can truncate our new vectors (rows of $\mathbf{T}$) to dimension L by keeping only the first L principal components

$$\mathbf{T}_L = \mathbf{X} \cdot \mathbf{V}_L$$

Principal Component Analysis

- Then we can truncate our new vectors (rows of $\mathbf{T}$) to dimension L by keeping only the first L principal components

$$\mathbf{T}_L = \mathbf{X} \cdot \mathbf{V}_L$$

- Our new vectors (rows of $\mathbf{T}_L$) are short and dense, but retain much of the variance in the original vectors

Singular Value Decomposition

- As it turns out, we can get $\mathbf{V}$ (technically $\mathbf{V}^T$) from the singular value decomposition (SVD) of $\mathbf{X}$

$$\mathbf{X} = \mathbf{U} \cdot \Sigma \cdot \mathbf{V}^T$$

References

- Baroni, Marco, Georgiana Dinu, and Germán Kruszewski. "Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors." Proceedings of ACL. 2014.
- Firth, John R. "A synopsis of linguistic theory, 1930-1955." Studies in linguistic analysis (1957).
- Harris, Zellig S. Methods in structural linguistics. 1951.
- Miller, George A., and Walter G. Charles. "Contextual correlates of semantic similarity." Language and cognitive processes 6.1 (1991): 1-28.