Class 8: Static Embeddings

Theme: Text

Computational Analysis of Text, Audio, and Images, Fall 2023

Aarhus University

Mathias Rask (mathiasrask@ps.au.dk)

Aarhus University

Today's Menu

Beyond BoW

Embeddings

Word2Vec

Lab

Table of Contents

Recap on Vectorization

Recap:

1. What's the main reason we need to vectorize text when using machine learning?
2. Explain the fundamentals of BoW vectorization. How does it work, what's the assumption?

Example:

	jeg 0	elsker 1	slik 2	chokolade 3	er 4	min 5	favorit 6
$\mathcal{D}_1$	1	1	1	0	0	0	0
$\mathcal{D}_2$	0	0	0	1	1	1	1

Two drawbacks:

$\rightarrow$ Sparse and inefficient representation

$\rightarrow$ Similar words have orthogonal representations

Beyond BoW

We want a representation of words that are short and dense which capture meaning and relations

How can we obtain that?

- ⇝ From vectorization of documents to vectorization of words: *word embeddings*

Word embeddings are widely used in political science nowadays:

1. Learning representations for ‘downstream’ tasks (e.g. classification)
2. Learning word usage and meaning (semantics) directly

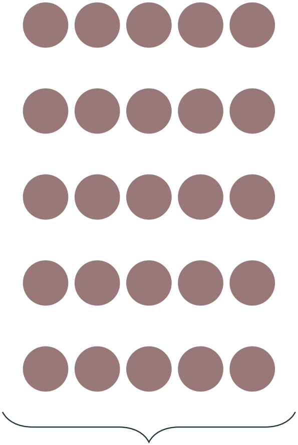
Table of Contents

Intuition

"Embeddings" are designed to represent words in a short and dense format while still maintaining meaning and relations:

- General term that refers to representing discrete features (e.g. word, document, actors) as a real-valued vector with d-dimensions: $X \in \mathbb{R}^d$
- From fixed-length vectors of length-$|\mathcal{V}|$ to fixed-length vectors of d-length

Example I



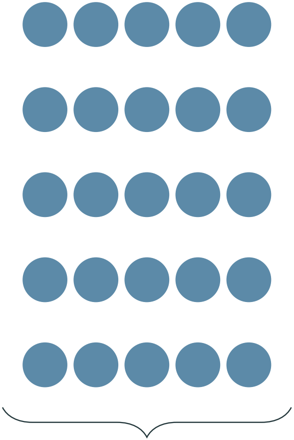
Collection of movies

Let's say we want to embed movies using $d = 5$ embeddings:

1. crime
2. comedy
3. drama
4. horror
5. romance

- ▷ The Godfather (1972): $[0.80, 0.20, 0.90, 0.30, 0.20]$
- ▷ Dumb and Dumber (1994): $[0.20, 0.90, 0.30, 0.01, 0.40]$
- ▷ *Not* a probability distribution!

Example II



Collection of people

Person characteristics

1. Age
2. Height (cm)
3. Weight (kg)
4. Skin color
5. Hair-color

- Embedding: $[28, 184, 79, 0.1, 2]$

The Distributional Hypothesis

The core idea about embeddings is that we want to represent words such that semantically related words are closer to each other

$\rightsquigarrow$ *The distributional hypothesis*:

- Words that occur in *similar contexts* tend to have *similar meaning*
  - $\rightsquigarrow$ "We know a word by the company it keeps" (Firth, 1957)
- Formalizes the very intuitive idea that contexts give meaning to words
  - $\rightsquigarrow$ Context $\neq$ co-occurrence

Embeddings in The Social Sciences

The semantic similarity conveyed by embeddings is a *powerful* and *flexible* tool:

- Semantic changes
- Semantic differences

⇝ The core idea is that the similarity between embeddings is informative about the semantic similarity of the concept we want to measure

⇝ How can we define similarity?

  ▷ Cosine similarity!

Semantic Changes (Hamilton *et al.*, 2016)

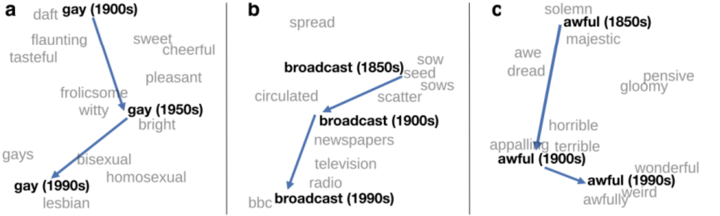
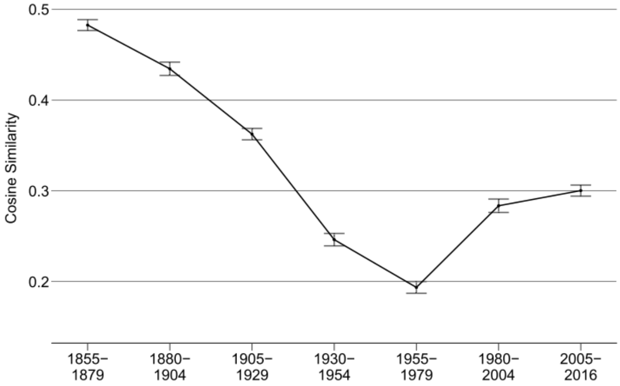
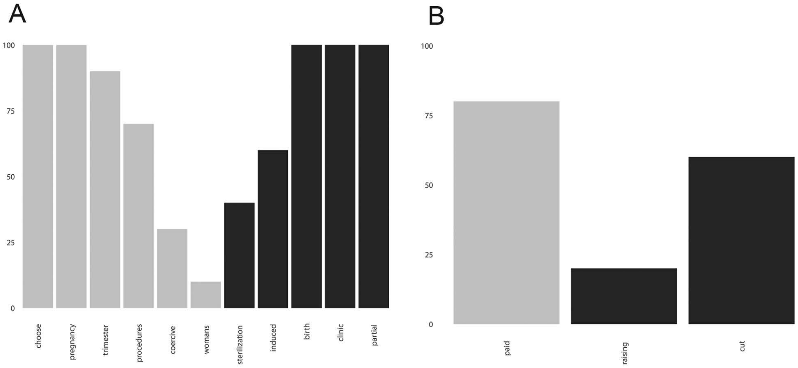


Figure 1: Two-dimensional visualization of semantic change in English using SGNS vectors.[2] **a**, The word *gay* shifted from meaning "cheerful" or "frolicsome" to referring to homosexuality. **b**, In the early 20th century *broadcast* referred to "casting out seeds"; with the rise of television and radio its meaning shifted to "transmitting signals". **c**, *Awful* underwent a process of pejoration, as it shifted from meaning "full of awe" to meaning "terrible or appalling" (Simpson et al., 1989).

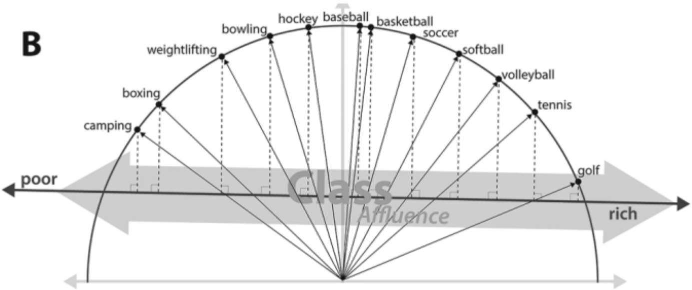
“Equality” - “Social” Cosine Similarity (Rodman, 2020)



Partisan Differences in Word Choice (Rodriguez and Spirling, 2022)

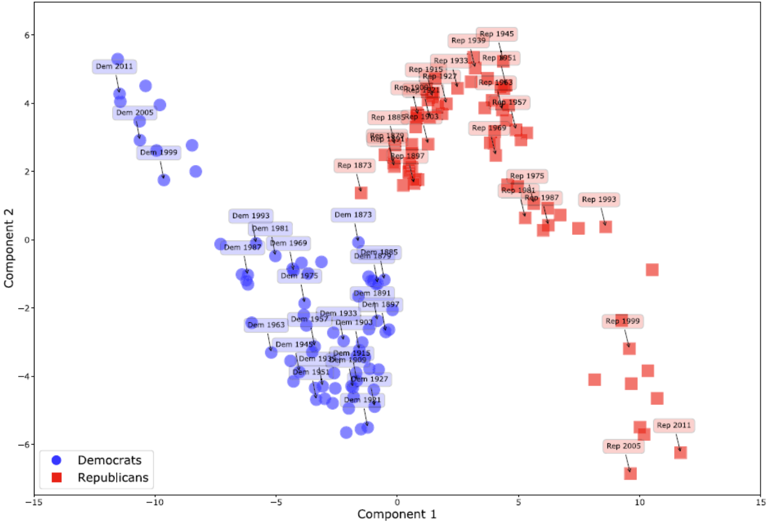


Cultural Dimensions of Word Embeddings (Kozlowski *et al.*, 2019)



Ideology and Word Embeddings (Rheault and Cochrane, 2020)

(a) Two-Dimensional Projection



Exercise

Discuss with your neighbors how word embeddings can be combined with dictionaries.

Analogies

Unlike other text representations, word embeddings are capable of solving analogies:

- Son is to father as daughter is to X
- Copenhagen is to Denmark what London is to X
- Denmark is to Copenhagen what England is to X

Textbook example:

$$\text{king} + \text{woman} - \text{man} = \text{queen}$$

What's the intuition behind this logic?

- The operation (woman − man) captures a gender dimension
- Starting at king means we are "walking" one step in the vector space along the gender dimension
- This means we can consider *directions* and not only *distances*

Exercise

Discuss with your neighbor how we can construct neural networks that use the distributional hypothesis to generate embeddings:

1. What's the input?
2. What's the output?
3. How do we specify d when we implement the net? (recall that d is the dimension of the embeddings)
4. How do we get annotated data? I.e. how can we train a network in a supervised manner?

See tutorial for a hands-on example using PyTorch

Table of Contents

Word2Vec: Overview

Word2Vec is one of the possible embedding algorithms that exist: learns *dense* representations that capture word *relations and meaning*

- Revolutionized NLP – $40,736$ citations – when introduced 10 years ago (Mikolov *et al.*, 2013)
- Learned word vectors/embeddings are typically around $50 - 1000$ with $d \in \mathbb{Z}$ with values $X \in \mathbb{R}^d$
- Individual values can not be interpreted $\rightsquigarrow$ but related words should have vectors closer to each other in the d-dimensional space

Word2Vec: Algorithms

CBOW

- Objective:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c,\, j\neq 0}\log(p(w_t \mid w_{t+j}))$$

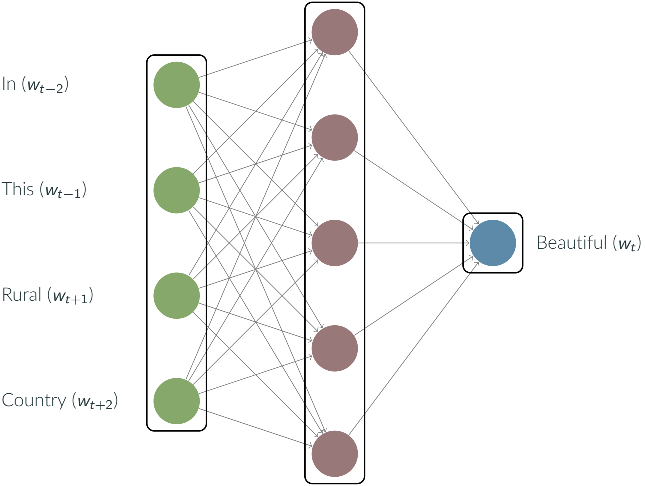
Skip-gram

- Objective:

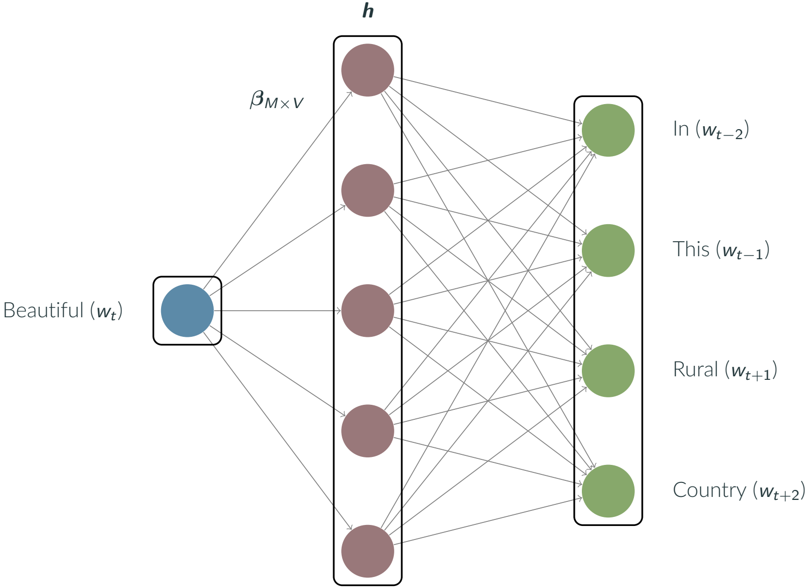
$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c,\, j\neq 0}\log(p(w_{t+j} \mid w_t))$$

- T total number of words
- w_t target word
- c window size
- j is an index within the context window, ranging from $-c$ to c, excluding $j = 0$
- $p(a \mid b)$ is the conditional probability of observing a given b
  - $p(w_t \mid w_{t+j})$: conditional probability of target word given context words
  - $p(w_{t+j} \mid w_t)$: conditional probability context words given target word

CBOW



Skip-Gram



$\boldsymbol{h}$
$\beta_{M \times V}$
Beautiful (w_t)
In (w_{t-2})
This (w_{t-1})
Rural (w_{t+1})
Country (w_{t+2})

Skip-Gram Mechanics

Sentence: "I A mener vi altså ikke at skattelettelser og velfærd er modsætninger"

Window size: 2

$\underbrace{\boxed{\text{I}}}_{t-1}\ \underbrace{\boxed{\text{A}}}_{\textit{target}}\ \underbrace{\boxed{\text{mener vi}}}_{t+1,t+2}$ altså ikke skattelettelser og velfærd er modsætninger.

I A mener vi altså $\underbrace{\boxed{\text{ikke at}}}_{t-2,t-1}\ \underbrace{\boxed{\text{skattelettelser}}}_{\textit{target}}\ \underbrace{\boxed{\text{og velfærd}}}_{t+1,t+2}$ er modsætninger.

I A mener vi altså ikke at $\underbrace{\boxed{\text{skattelettelser og}}}_{t-2,t-1}\ \underbrace{\boxed{\text{velfærd}}}_{\textit{target}}\ \underbrace{\boxed{\text{er modsætninger}}}_{t+1,t+2}$

I A mener vi altså ikke at skattelettelser og $\underbrace{\boxed{\text{velfærd er}}}_{t-2,t-1}\ \underbrace{\boxed{\text{modsætninger}}}_{\textit{target}}$

Negative Sampling

I A mener vi altså $\underbrace{\boxed{\text{ikke at}}}_{t-2,t-1}$ $\underbrace{\boxed{\text{skattelettelser}}}_{\textit{target}}$ $\underbrace{\boxed{\text{og velfærd}}}_{t+1,t+2}$ er modsætninger.

Positive Samples

- (ikke, skattelettelser)
- (at, skattelettelser)
- (og, skattelettelser)
- (velfærd, skattelettelser)

Negative Samples

- (???, skattelettelser)
- (kørekort, skattelettelser)
- (fodbold, skattelettelser)
- (zoo, skattelettelser)

The positive and negative samples constitute the training set – no labeling required! ⇝ *self-supervision*

Practical Issues

Working with embeddings in practice involves choosing between four "hyperparameters" (Rodriguez and Spirling, 2022):

1. Window size (depends on the length of input text)
2. Dimensionality size (d)
3. Locally vs. pretrained (fixed or fine-tuned) embeddings
4. Preprocessing (huge debate!)

Table of Contents

See you next week!

Theme: Text

Computational Analysis of Text, Audio, and Images, Fall 2023

Aarhus University

References i

[1] W. L. Hamilton, J. Leskovec, and D. Jurafsky, "Diachronic word embeddings reveal statistical laws of semantic change," *arXiv preprint arXiv:1605.09096*, 2016.

[2] E. Rodman, "A timely intervention: Tracking the changing meanings of political concepts with word vectors," *Political Analysis*, vol. 28, no. 1, pp. 87–111, 2020.

[3] P. L. Rodriguez and A. Spirling, "Word embeddings: What works, what doesn't, and how to tell the difference for applied research," *The Journal of Politics*, vol. 84, no. 1, pp. 101–115, 2022.

[4] A. C. Kozlowski, M. Taddy, and J. A. Evans, "The geometry of culture: Analyzing the meanings of class through word embeddings," *American Sociological Review*, vol. 84, no. 5, pp. 905–949, 2019.

[5] L. Rheault and C. Cochrane, "Word embeddings for the analysis of ideological placement in parliamentary corpora," *Political Analysis*, vol. 28, no. 1, pp. 112–133, 2020.

[6] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013.