Word Embeddings

SIRCSS CTA 2024

Väinö Yrjänäinen

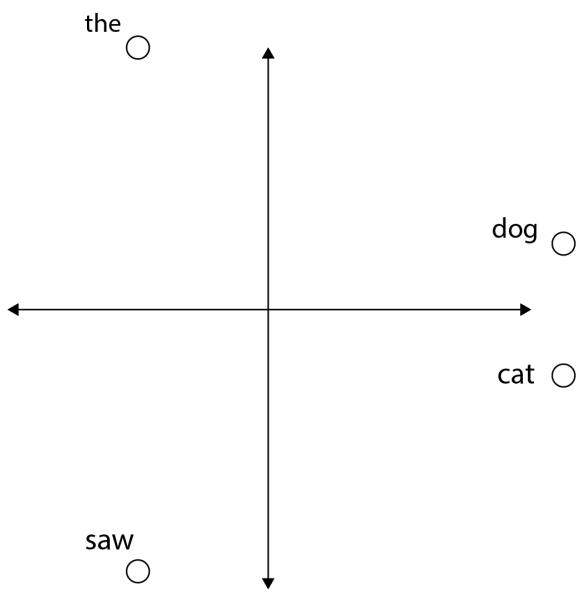
2024-06-13

omestic dog clade: familiaris, as named by Linnaeus in 1758 and, dingo named by Meyer in 1793. Wozencr aft included hallstromi (the New Guinea singing dog) as another name (junior synonym) for the dingo. W ozencraft referred to the mtDNA study as one of the guides informing his decision. Mammalogists have n oted the inclusion of familiaris and dingo together under the domestic dog clade with some debating it .In 2019, a workshop hosted by the IUCN/Species Survival Commission's Canid Specialist Group considere d the dingo and the New Guinea singing dog to be feral Canis familiaris and therefore did not assess t hem for the IUCN Red List of Threatened Species. Evolution The Cretaceous-Paleogene extinction event o ccurred 65 million years ago and brought an end to the non-avian dinosaurs and the appearance of the f irst carnivorans. The name carnivoran is given to a member of the order Carnivora. Carnivorans posses s a common arrangement of teeth called carnassials, in which the first lower molar and the last upper premolar possess blade-like enamel crowns that act similar to a pair of shears for cutting meat. This dental arrangement has been modified by adaptation over the past 60 million years for diets composed o f meat, for crushing vegetation, or for the loss of the carnassial function altogether as in seals, se a lions, and walruses. Today, not all carnivorans are carnivores, such as the insect-eating aardwolf.T he carnivoran ancestors of the dog-like caniforms and the cat-like feliforms began their separate evol utionary paths just after the end of the dinosaurs. The first members of the dog family Canidae appear ed 40 million years ago, of which only its subfamily the Caninae survives today in the form of the wol

Simple methods

- Counting words / n-grams
  - Conceptually straightforward and computationally fast
  - Not very flexible
  - All words are treated as independent categories

“the dog saw the cat” →



$$\text{dog} = \begin{bmatrix} 0.23 \\ -1.47 \\ 1.01 \\ 1.33 \\ \cdots \\ -0.56 \end{bmatrix}, \text{cat} = \begin{bmatrix} 0.42 \\ -1.07 \\ 0.31 \\ 1.54 \\ \cdots \\ 0.14 \end{bmatrix} \in \mathbb{R}^{100}$$

```
from probabilistic_word_embeddings.embeddings import Embedding

e = Embedding(saved_model_path="embedding.pkl")
dog = e["dog"]
cat = e["cat"]
```

Using word embeddings

- Distance / similarity
  - Cosine distance / similarity
  - How similar are words x, y
  - What is the most similar word to word x
- Analogies
  - $\text{king} - \text{man} + \text{woman} \approx \text{queen}$
- And more

Cosine similarity

$$\text{cossim}\ (\rho_v, \rho_w) = \frac{\rho_v \cdot \rho_w}{\|\rho_v\|\|\rho_w\|}$$

Differences

$$\rho_v - \rho_w$$

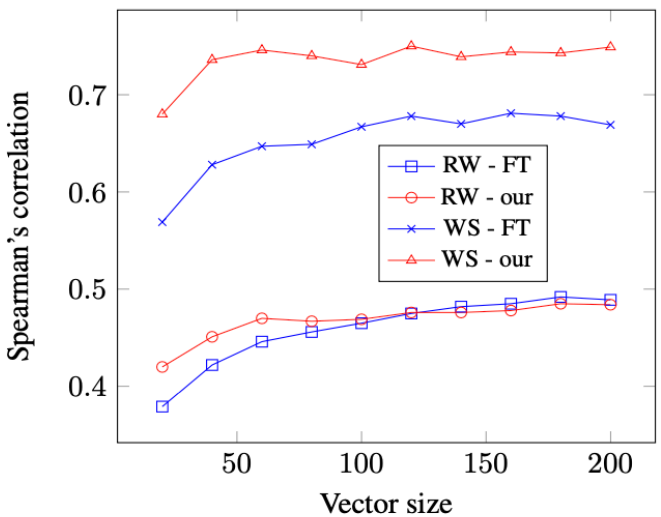


Figure 1: Spearman's rank correlation coefficient for RW-STANFORD (RW) and WS-353-ALL (WS) on the fastText model (FT) and our, with different vector size. Training is done on the corpus A of 50M tokens.

```
e1 = Embedding(saved_model_path="dict2vec_embedding.pkl")
e2 = Embedding(saved_model_path="fasttext_embedding.pkl")

similarity1 = cossim(e1["dog"], e1["cat"])
similarity2 = cossim(e2["dog"], e2["cat"])
```

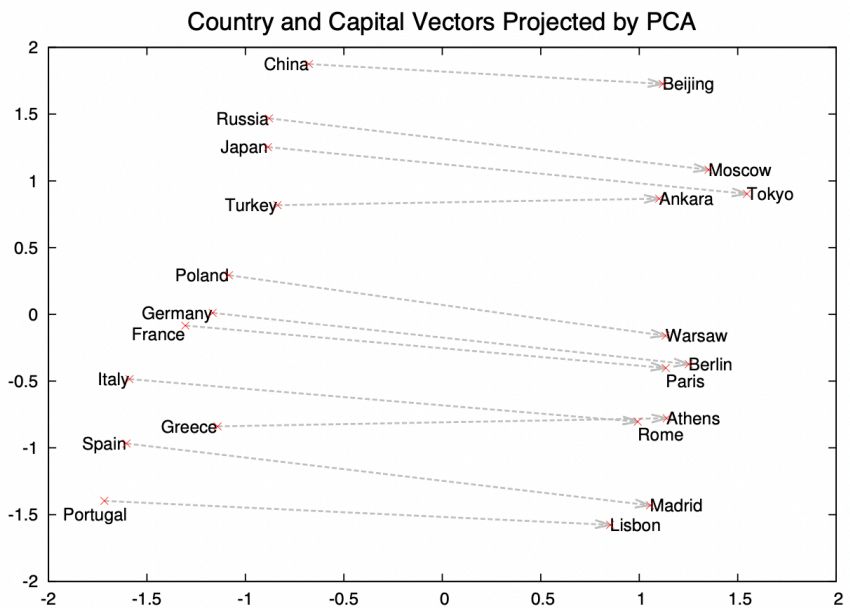


Figure 2: Two-dimensional PCA projection of the 1000-dimensional Skip-gram vectors of countries and their capital cities. The figure illustrates ability of the model to automatically organize concepts and learn implicitly the relationships between them, as during the training we did not provide any supervised information about what a capital city means.

```
diff1 = e["china"] - e["beijing"]
diff2 = e["germany"] - e["berlin"]
```

Train yourself or use pretrained embeddings?

- For research, you can do two things
  - Use a pre-trained embedding
  - Gather data and train your own embedding
- It is crucial what data has been used for training
- Own embeddings can be more specific, but require work and good data

Training word embeddings

“you shall know a word by the company it keeps”

— Firth (1957)

omestic dog clade: familiaris, as named by Linnaeus in 1758 and, dingo named by Meyer in 1793. Wozenc raft included hallstromi (the New Guinea singing dog) as another name (junior synonym) for the dingo. Wozencraft referred to the mtDNA study as one of the guides informing his decision. Mammalogists have noted the inclusion of familiaris and dingo together under the domestic dog clade with some debating it.In 2019, a workshop hosted by the IUCN/Species Survival Commission's Canid Specialist Group conside red the dingo and the New Guinea singing dog to be feral Canis familiaris and therefore did not asses s them for the IUCN Red List of Threatened Species. Evolution The Cretaceous-Paleogene extinction even t occurred 65 million years ago and brought an end to the non-avian dinosaurs and the appearance of th e first carnivorans. The name carnivoran is given to a member of the order Carnivora. Carnivorans poss ess a common arrangement of teeth called carnassials, in which the first lower molar and the last uppe r premolar possess blade-like enamel crowns that act similar to a pair of shears for cutting meat. Thi s dental arrangement has been modified by adaptation over the past 60 million years for diets compose d of meat, for crushing vegetation, or for the loss of the carnassial function altogether as in seals , sea lions, and walruses. Today, not all carnivorans are carnivores, such as the insect-eating aardwo lf.The carnivoran ancestors of the dog-like caniforms and the cat-like feliforms began their separate evolutionary paths just after the end of the dinosaurs. The first members of the dog family Canidae a ppeared 40 million years ago, of which only its subfamily

[...] as the oldest domesticated species, dogs' minds inevitably have been shaped by millennia of contact with humans [...]

[...] as the oldest domesticated species, dogs' minds inevitably have been shaped by millennia of contact with humans [...]

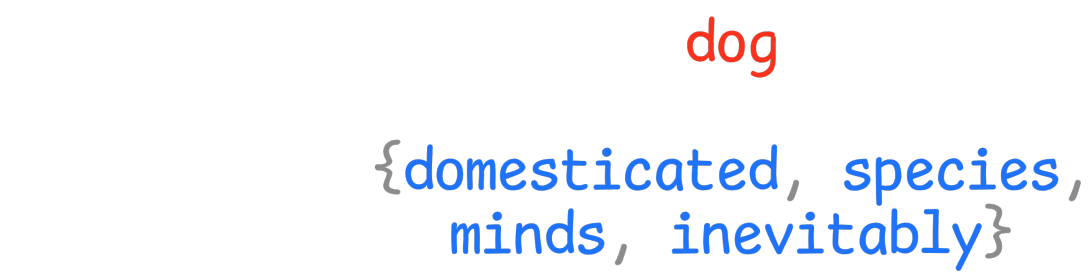


Figure 5: Context window, symmetric and size 2

Word2Vec – concept

- Continuous bag-of-words CBOW

Does the word x appear, given that its context window is $y_1, y_2, \ldots, y_{ws}$

- Skip-gram

Do words x and y appear in the same word window?

Continuous bag-of-words

[...] as the oldest domesticated species,
[???] minds inevitably have been shaped
by millenia of contact with humans

Whole context window simulatenously!

Skip-gram

[...] as the oldest domesticated species
[???] minds inevitably have been shaped
by millenia of contact with humans

Skip-gram

[...] as the oldest domesticated species
[???] minds inevitably have been shaped
by millenia of contact with humans

Skip-gram

[...] as the oldest domesticated species
[???] minds inevitably have been shaped
by millenia of contact with humans

[...] as the oldest domesticated species
[???] minds inevitably have been shaped
by millenia of contact with humans

One context word at the time!

Word2Vec - mathematical definition

- Each word is assigned a word vector ρ and a context vector α
  - $\rho, \alpha \in \mathbb{R}^D$, where D is the *dimensionality* of the embedding
- Conditional probabilities are functions of these vectors
  - $P(x \mid y_1, ..., y_w) = f\left(\rho_x, \alpha_{y_1}, ..., \alpha_{y_w}\right)$

Skip-gram

Do words x and y appear in the same word window?

$$p(\text{x and y co-occur}) = \sigma\left(\rho_x^T \alpha_y\right)$$

where $\sigma(\cdot)$ is the logistic function.

Skip-gram, numerical example

$$\rho_{\text{dog}} = \begin{bmatrix} 0.23 \\ -1.47 \\ 1.01 \\ 1.33 \\ \cdots \\ -0.56 \end{bmatrix}, \alpha_{\text{cat}} = \begin{bmatrix} 0.42 \\ -1.07 \\ 0.31 \\ 1.54 \\ \cdots \\ 0.14 \end{bmatrix}$$

Do the words *dog* and *cat* appear in the same word window?

$$p(\text{dog and cat co-occur}) = \sigma\left(\begin{bmatrix} 0.23 & -1.47 & 1.01 & 1.33 & \dots & -0.56 \end{bmatrix} \begin{bmatrix} 0.42 \\ -1.07 \\ 0.31 \\ 1.54 \\ \dots \\ 0.14 \end{bmatrix} \right)$$

$$= \sigma(6.2325) = 0.998$$

$$\rho_{\text{dog}} = \begin{bmatrix} 0.23 \\ -1.47 \\ 1.01 \\ 1.33 \\ \cdots \\ -0.56 \end{bmatrix}, \alpha_{\text{university}} = \begin{bmatrix} -2.03 \\ 0.07 \\ 1.01 \\ -0.34 \\ \cdots \\ 0.89 \end{bmatrix}$$

Do the words *dog* and *university* appear in the same word window?

$$p(\text{dog and university co-occur}) = \sigma\left(\begin{bmatrix} 0.23 & -1.47 & 1.01 & \dots & -0.56 \end{bmatrix} \begin{bmatrix} -2.03 \\ 0.07 \\ 1.01 \\ -0.34 \\ \dots \\ 0.89 \end{bmatrix} \right)$$

$$= \sigma(-0.13) = 0.4675$$

Continuous bag-of-words

Do words x and y appear in the same word window?

$$p(\text{x occurs given } C = y_1, ..., y_{ws}) = \sigma\left(\rho_x^T \sum_{y \in C} \alpha_y\right)$$

where C is the context window, and $\sigma(\cdot)$ is the logistic function.

Negative sampling

- In the data, we can easily find instances where words do co-occur
  - “the dog saw the cat”
  - Eg. (dog,saw), (dog,cat) and (the, cat) co-occur
- We do not have direct examples of words not co-occuring
- We need negative examples, otherwise we will just have all vectors be the same
- Solution: pick at random

Negative sampling

- For each word x in the data
  - Sample $ns \in \mathbb{N}$ negative samples y'_1, y'_{ns} from the empirical distribution of word types
  - Add a term $1 - p(\text{x and } y' \quad \text{co-occur})$ in the likelihood

Negative sampling

Say we only have the data point *dog co-occurs with cat*. We draw a negative sample *university* from the empirical distribution of words in the data. The likelihood is

$$p(\text{data}) = \underbrace{p(\text{dog and cat co-occur})}_{\text{positive sample}}\underbrace{(1 - p(\text{dog and university co-occur}))}_{\text{negative sample}}$$

$$= 0.998 \cdot (1 - 0.4675) = 0.531435$$

For a real dataset, the log likelihood is

$$\log p(\mathcal{D} \mid w, c) = \sum_{i=1}^{N} \left(\underbrace{\sum_{y \in C_i} \log \sigma\left(\rho_{x_i}^T \alpha_y\right)}_{\text{positive samples}} + \underbrace{\sum_{y \in C_{ns}} \log\left(1 - \sigma\left(\rho_{x_i}^T \alpha_y\right)\right)}_{\text{negative samples}} \right)$$

where $\mathcal{D} = (x_1, ..., x_N)$ is the dataset indexed by $i \in \{1, ..., N\}$, C_i is the context window at i and C_{ns} is a randomized context window generated from the empirical distribution of words.

Different word embeddings

Probabilistic word embeddings

- Word2Vec can be formulated as a probabilistic language model
  - Same likelihood
  - Adds a prior on the parameters, such as $\rho_x \sim \mathcal{N}(0, \lambda_0 I)$ for all word types $x \in W$
- Estimated via maximum a posteriori estimation or variational inference

Bernoulli embeddings (Rudolph et al 2016)

$$\log p(w, c \mid \mathcal{D}) = \log p(\mathcal{D} \mid w, c) + \sum_{x \in W} \lambda_1 \left\| \rho_{x,t} \right\|^2 + \sum_{x \in W} \lambda_0 \| \alpha_x \|^2$$

Probabilistic word embeddings – advantages

- Straightforward way of including prior knowledge
  - Regularize models for data with natural divisions to subsets, for example regularize models that model different time periods

Probabilistic word embeddings – advantages

- Straightforward way of including prior knowledge
  - Regularize models for data with natural divisions to subsets, for example regularize models that model different time periods
  - Incorporate additional information about the words, such as dictionary information

Probabilistic word embeddings – advantages

- Straightforward way of including prior knowledge
  - Regularize models for data with natural divisions to subsets, for example regularize models that model different time periods
  - Incorporate additional information about the words, such as dictionary information
  - Enforce certain properties of the embedding, such as sentiment aspects

Dynamic probabilistic word embeddings

- Different set of word vectors for each time period $t \in \{1, ..., T\}$
- For example, there is a separate vector for "dog" for each year
  - $\rho_{\text{dog},2012}, \cdots, \rho_{\text{dog},2024}$
- You can calculate all the measurements for each year, or between years
  - Distance between $\rho_{\text{dog},2012}$ and $\rho_{\text{dog},2013}$
  - How does the similarity of $\rho_{\text{dog,t}}$ and $\rho_{\text{cat,t}}$ develop over time t?

Dynamic probabilistic word embeddings

- Random walk prior over the word vectors

$$\rho_{x,1} \sim \mathcal{N}(0, \lambda_0 I)$$

$$\rho_{x,t} \sim \mathcal{N}\left(\rho_{x,t-1}, \lambda_1 I\right) \quad \text{for } t \in \{2, ..., T\}$$

- Spherical prior on context vectors

$$\alpha_x \sim \mathcal{N}(0, \lambda_0 I)$$

Dynamic probabilistic word embeddings

$$\log p(w, c \mid \mathcal{D}) = \log p(\mathcal{D} \mid w, c) + \sum_{x \in W} \sum_{t=1}^{T} \lambda_1 \left\| \rho_{x,t+1} - \rho_{x,t} \right\|^2 + \sum_{x \in W} \sum_{t=1}^{T} \lambda_0 \left\| \rho_{x,t} \right\|^2 + \sum_{x \in W} \lambda_0 \left\| \alpha_x \right\|^2$$

Informative priors

- Little data but also prior knowledge about the words
  - Still good embedding quality
- If we are interested in countries, we can add a prior between the different word forms
  - finland $\sim$ finnish $\sim$ finns $\sim$ finlands'
- Also other prior sources
  - Wordnet, synonyms

Graph priors

- Connections between words can be done via graph priors
  - “These words are similar”
- Implemented in the pwe module
- Use for the dynamic prior in the Lab

```
from probabilistic_word_embeddings.embeddings import
LaplacianEmbedding
import networkx

g = networkx.Graph()
g.add_edge("finland", "finns")

e = LaplacianEmbedding(vocabulary, graph=g)
```

```
from probabilistic_word_embeddings.embeddings import
LaplacianEmbedding
import networkx

g = networkx.Graph()
g.add_edge("dog_2012", "dog_2013")

e = LaplacianEmbedding(vocabulary, graph=g)
```

Uncertainty estimation

- Many applications want to draw conclusions about the data
- Random noise can be mistaken for results
- Methods to estimate embedding uncertainty
  - Probabilistic word embeddings
  - **Bootstrap**

Bootstrapping word embeddings

- Divide corpus into documents that can be resampled
- Generate N boostrap resamples
- Train the word embedding separately on all N resamples
- Calculate quantity of interest for all N word embeddings

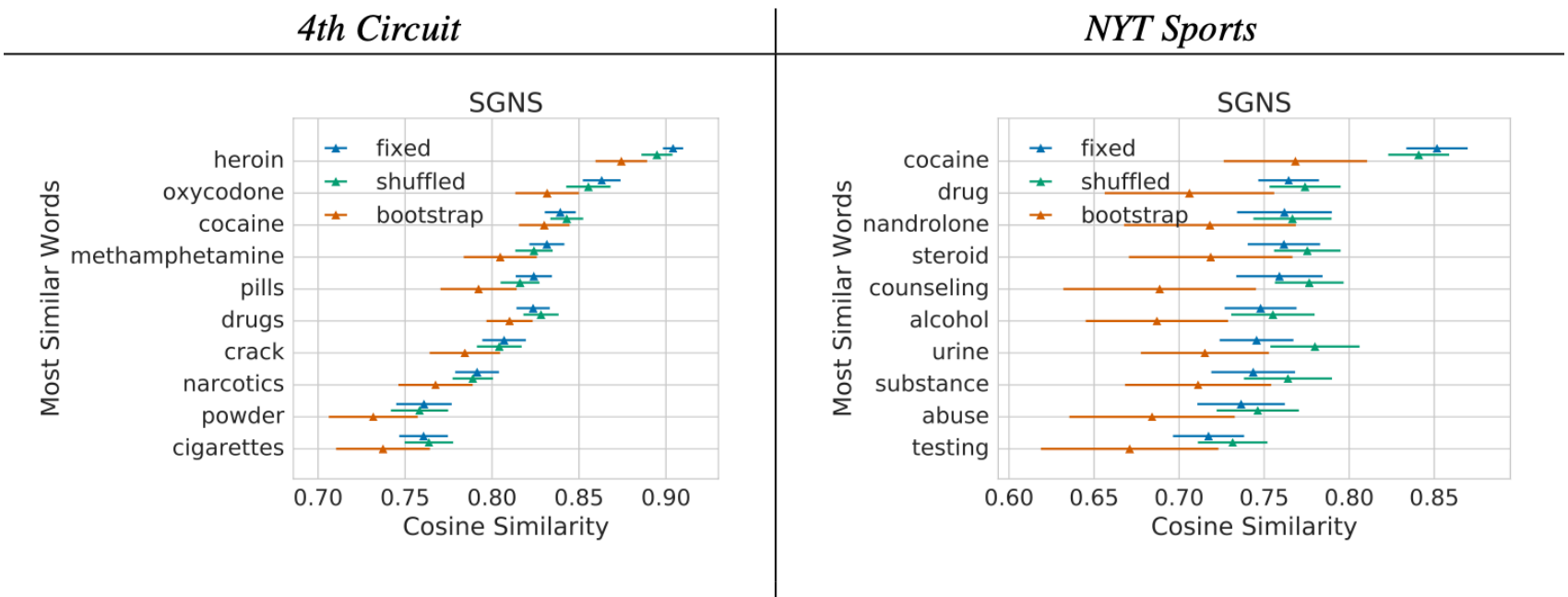
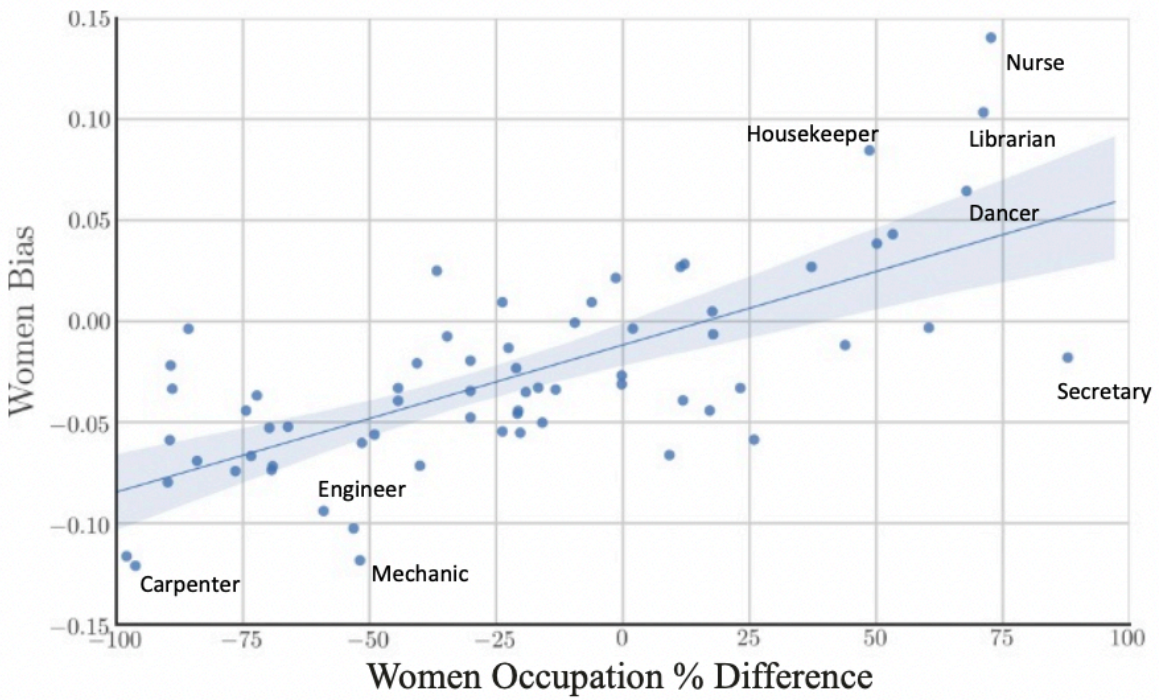


Figure 6: Most similar words to ‘marijuana’ in two corpora


Women Bias
Women Occupation % Difference
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−100
−75
−50
−25
0
25
50
75
100
Nurse
Housekeeper
Librarian
Dancer
Secretary
Engineer
Mechanic
Carpenter

Summary

- Word2vec yields vector representations of words that can be used to analyse words
  - Similrarity, aspects, analogies
- Probabilsitic embeddings enable the incorporation of prior knowledge and uncertainty estimation
  - Dynamic or grouped embeddings
  - Make estimation possible on smaller data via priors