GloVe: Global Vectors for Word Representation

PRESENTED BY:

FANTASTIC FOUR

Introduction

- GloVe is an unsupervised learning algorithm for obtaining vector representations for words
- Training is performed on aggregated global word-word co-occurrence statistics from a corpus
- Resulting representations showcase interesting linear substructures of the word vector space.

Model Overview

- GloVe is essentially a log-bilinear model with a weighted least-squares objective.
- Intuition – Ratios of word-word co-occurrence probabilities have the potential for encoding some form of meaning

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

Cost function

$$J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij}\right)^2$$

$$f(x) = \begin{cases} (x/x_{\max})^{\alpha} & \text{if } x < x_{\max} \\ 1 & \text{otherwise} \end{cases}$$

Properties of GloVe

- Nearest neighbours
- Linear substructures

Nearest Neighbours

- The Euclidean distance between two word vectors is a measure of semantic similarity of the corresponding words.

3. litoria

4. leptodactylidae

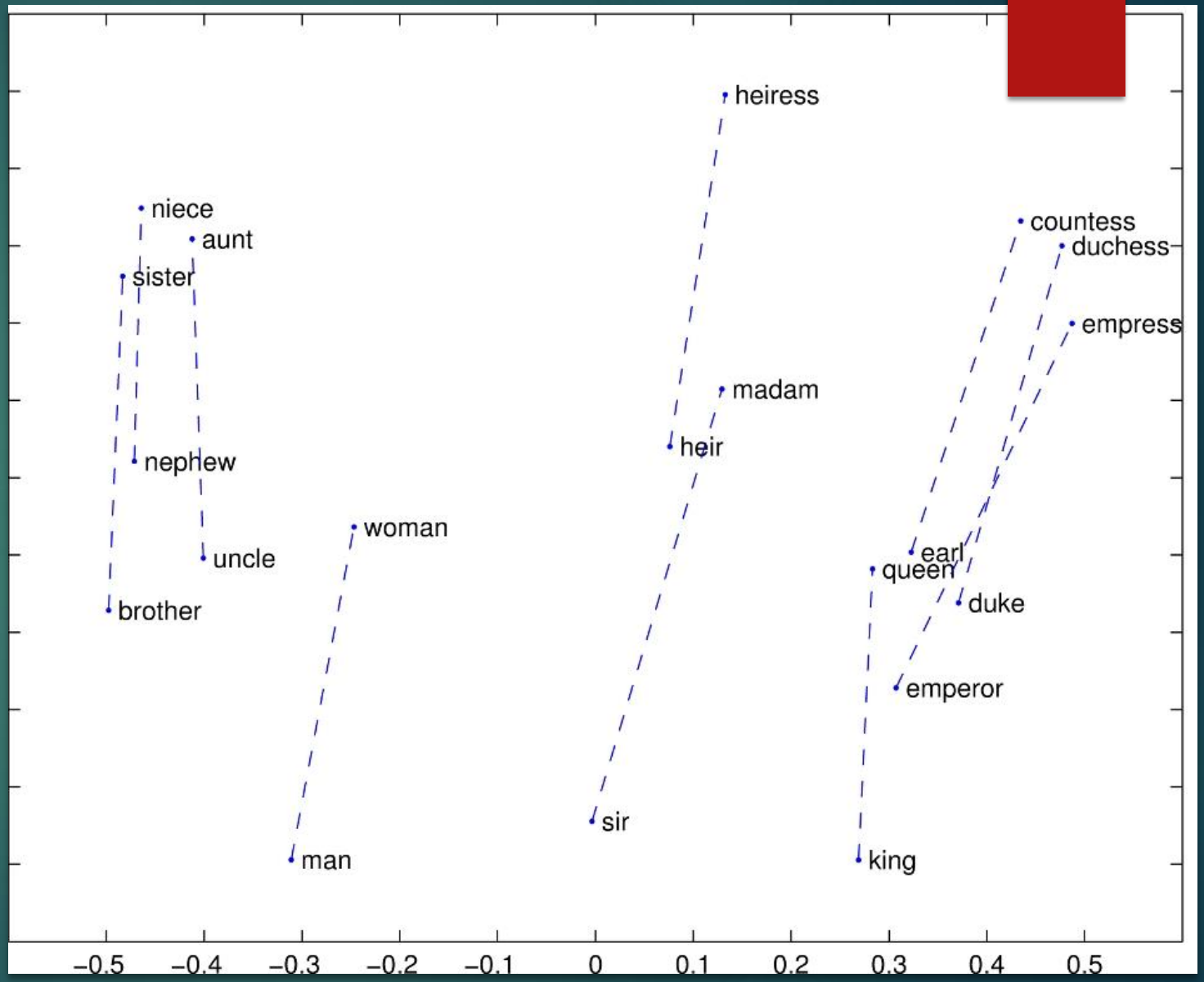
5. rana

7. eleutherodactylus

Linear Substructures

- Nearest neighbor evaluations produce single scalar that quantifies the relatedness of two words.
- Problematic since two given words almost always exhibit more intricate relationships.
- A natural and simple candidate for an enlarged set of discriminative numbers is the vector difference between the two word vectors.
- GloVe is designed in order that such vector differences capture as much as possible the meaning specified by the juxtaposition of two words

- The underlying concept that distinguishes man from woman, i.e. sex or gender, may be equivalently specified by various other word pairs, such as king and queen or brother and sister.
- The vector differences man - woman, king - queen, and brother - sister might all be roughly equal.



Results

Advantages

- Fast training
- Good performance even with small corpus, and small vectors
- Early stopping. We can stop training when improvements become small.

Disadvantages

- Uses a lot of memory: the fastest way to construct a term-co-occurrence matrix is to keep it in RAM as a hash map and perform co-occurrence increments in a global manner
- Sometimes quite sensitive to initial learning rate