Word embeddings

TA Arun Kumar Rajasekaran

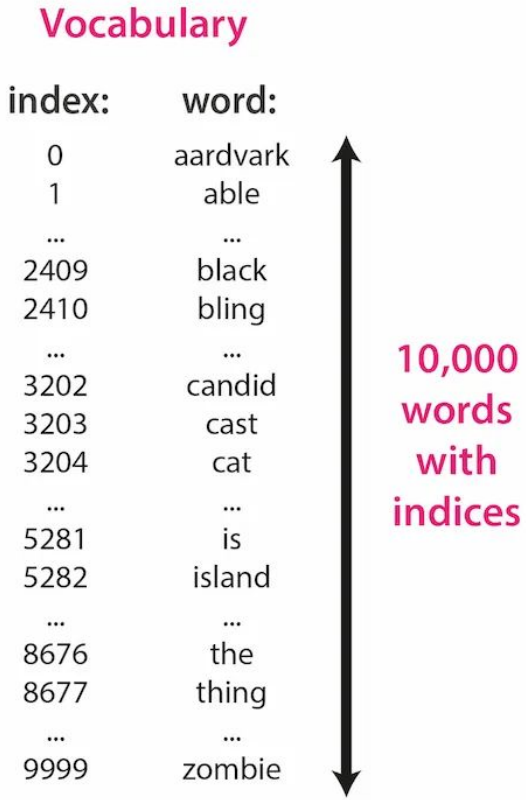
What and Why?

Word embeddings are a type of word representation that allows words with similar meaning to have a similar representation.

How can we best numerically represent textual input?

Word Embeddings in NLP is a technique where individual words are represented as real-valued vectors in a lower-dimensional space and captures inter-word semantics. Each word is represented by a real-valued vector with tens or hundreds of dimensions.

Why not index all the known words? Can we?



One-hot vector encoding

Feature (Color)
Red
Green
Yellow
Green
Red

One Hot Encoded Vector	Red	Green	Yellow
[1,00]	1	0	0
[0,1,0]	0	1	0
[0,0,1]	0	0	1
[0,1,0]	0	1	0
[1,00]	1	0	0

SCALER Topics

Issues with one-hot vector representation

The similarity issue....

(cant distinguish closely related words - example, cat/tiger, man/boy/male)

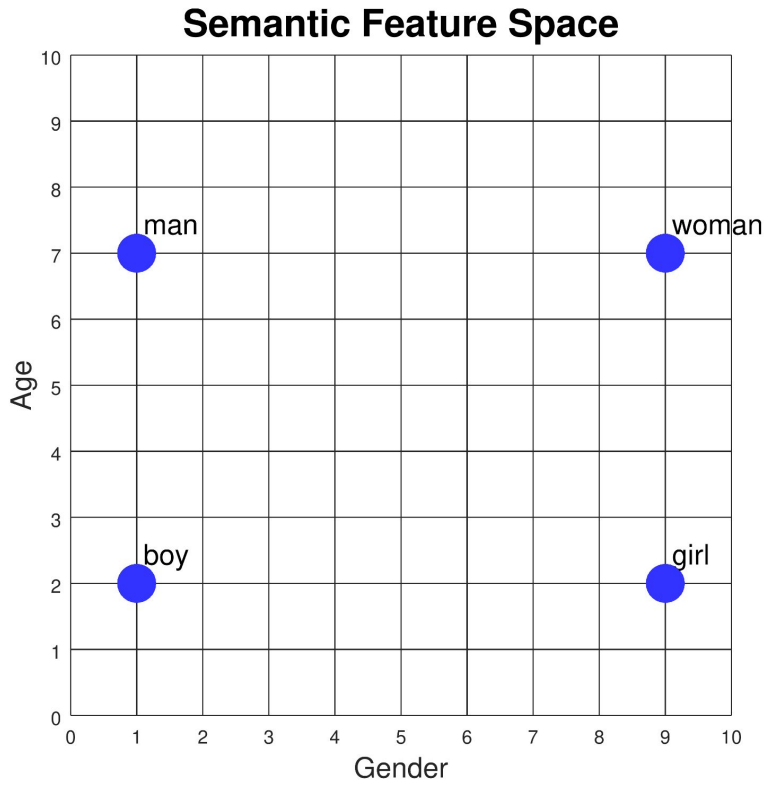
The vocabulary size issue…

(it can really explode)

The computational issue…..

(lots of zeros.. ML doesn’t like many zeros)

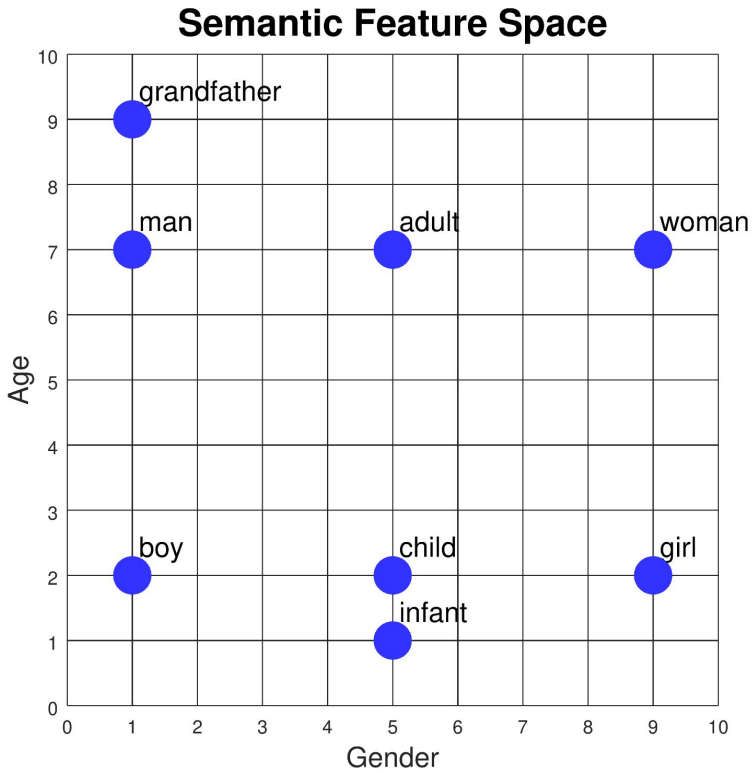
Semantic feature space



Word Coordinates

	Gender	Age
man	[1,	7]
woman	[9,	7]
boy	[1,	2]
girl	[9,	2]

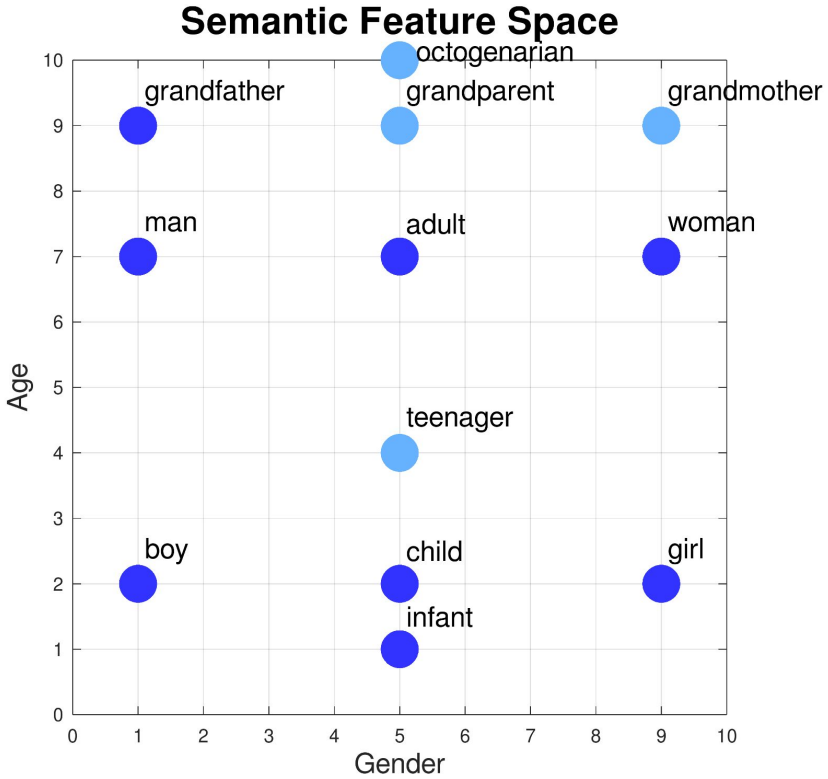
Semantic feature space



Word Coordinates

	Gender	Age
grandfather	1	9
man	1	7
adult	5	7
woman	9	7
boy	1	2
child	5	2
girl	9	2
infant	5	1

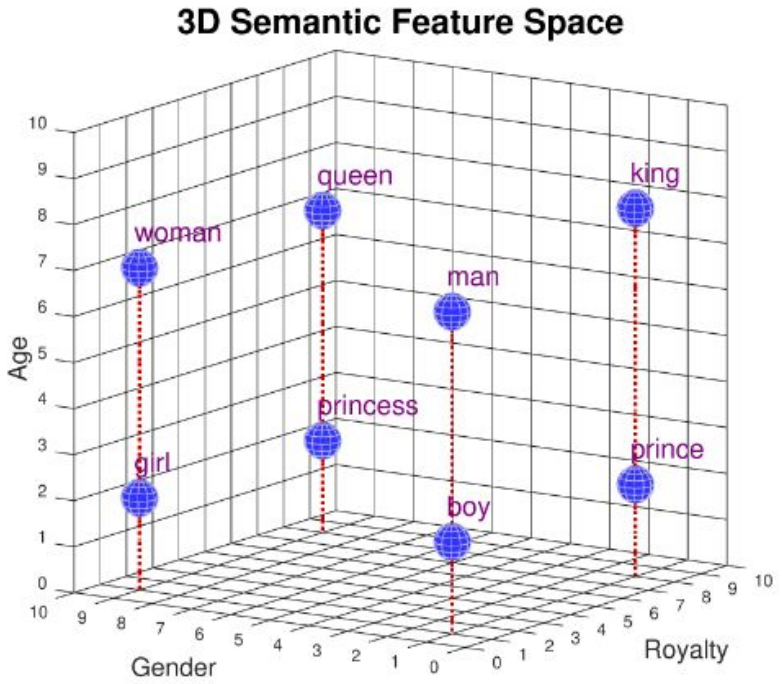
Semantic feature space



Word Coordinates

	Gender	Age
grandmother	[9,	9]
grandparent	[5,	9]
octogenarian	[5,	10]
teenager	[5,	4]

Increasing dimensionality …



Word Coordinates

	Gender	Age	Royalty
man	[1,	7,	1]
woman	[9,	7,	1]
boy	[1,	2,	1]
girl	[9,	2,	1]
king	[1,	8,	8]
queen	[9,	7,	8]
prince	[1,	2,	8]
princess	[9,	2,	8]

Word Embeddings

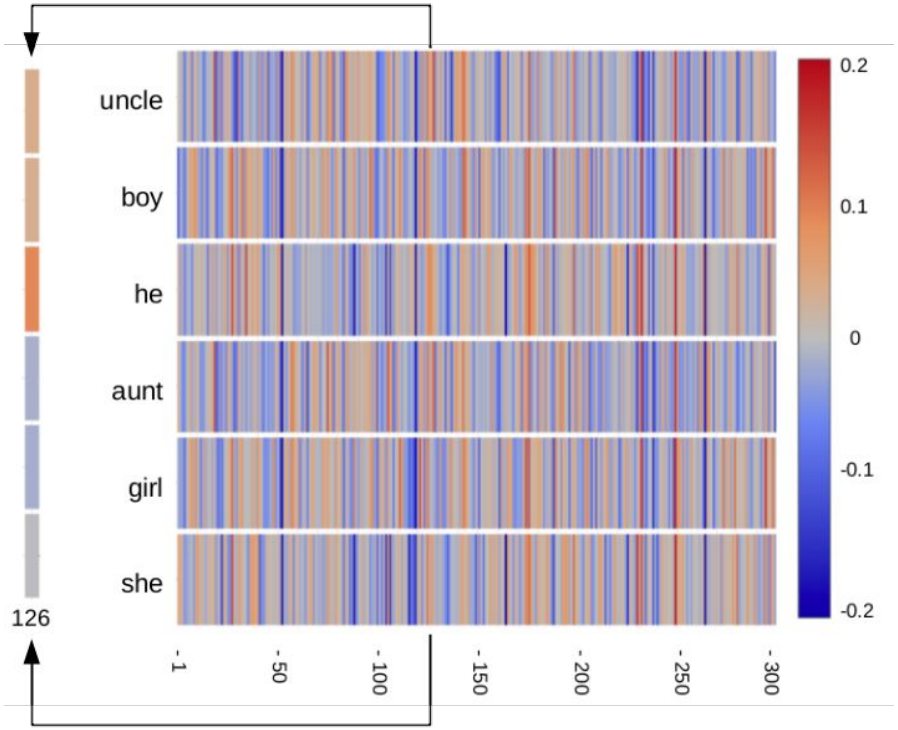
To represent the complexity of a typical 50,000 word English vocabulary requires hundreds of features

Instead we can let the computer create the feature space for us by supplying a machine learning algorithm with a large amount of text, such as all of Wikipedia, or a huge collection of news articles.

The algorithm discovers statistical relationships between words by looking at what other words they co-occur with. It uses this information to create word representations in a semantic feature space of its own design.

These representations are called word embeddings……..

A typical embedding might use a 300 dimensional space, so each word would be represented by 300 numbers. "Uncle", "boy", and "he" are male words, while "aunt", "girl", and "she" are female words. Each word is represented by 300 numbers with values between -0.2 and +0.2.



Word2Vec

Word2vec, one of the most popular techniques to create Word Embeddings, was created in 2013 by a team of Google researchers.

GloVe

GloVe creates the embeddings by generating a matrix with the number of occurrences of the surrounding words and performing statistics on that matrix.

FastText

FastText splits the word into smaller parts (eg. parts= <pa, ar, rt, ts>) and tries to learn embeddings based on that. It has the advantage of creating embeddings for words that it has never seen during training.

FlMo

FLMo creates the Word Embeddings based on the context, which means that the same word can have different embeddings according to the words nearby.