Word Embeddings

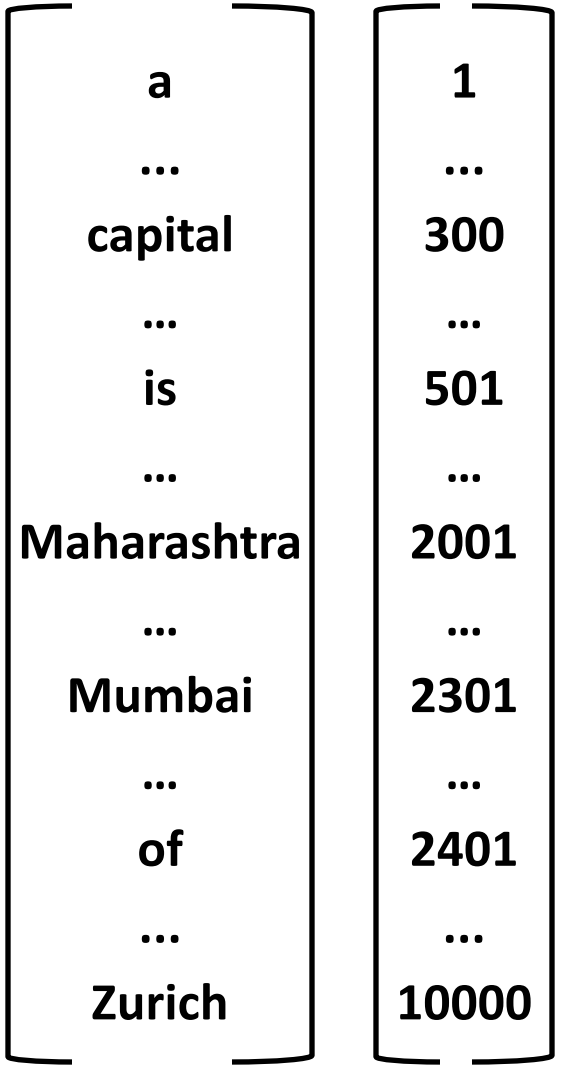
Pritam Prakash Shete
Computer Division, BARC
Centre for Excellence in Basic Sciences

Topics

- Word representation
- Featurized representation
- Visualizing word embeddings
- Using word embeddings
- Embedding matrix
- Learning word embeddings

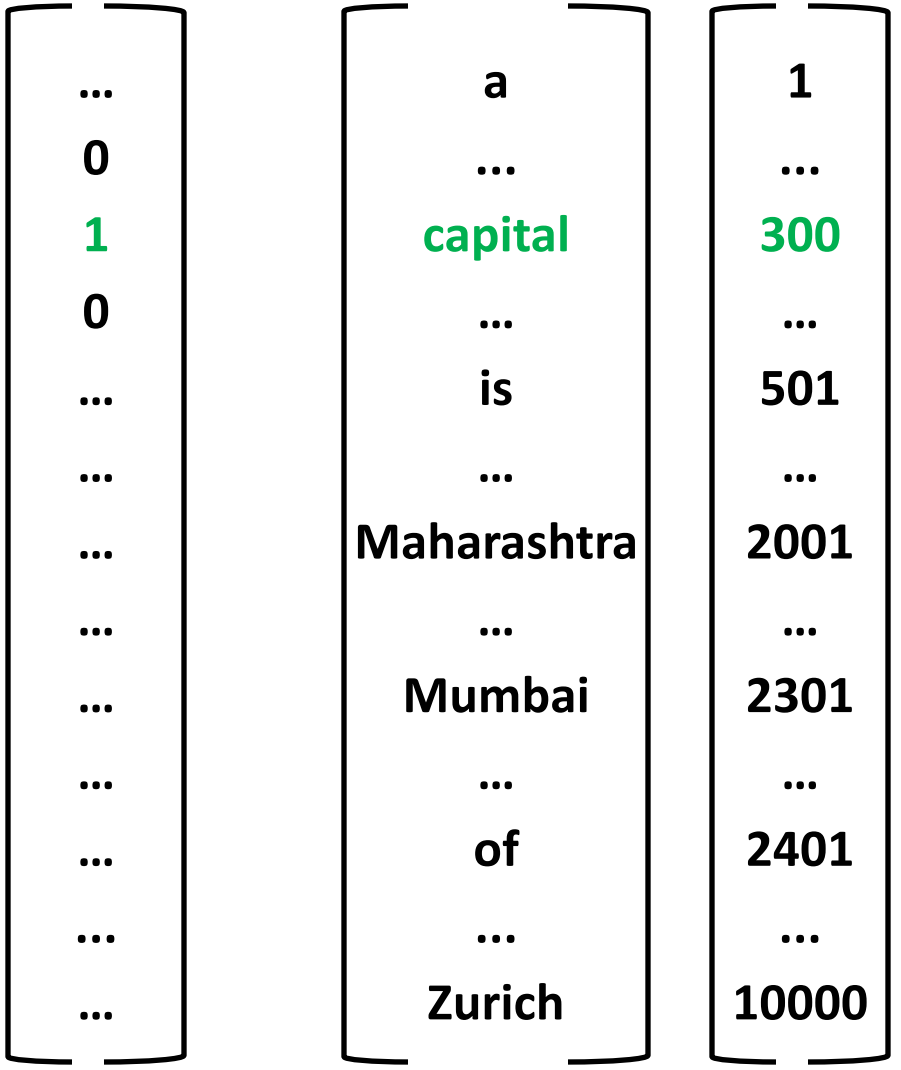
Word representation – Vocabulary

- Word dictionary
- Top common words
- Dictionary size
  - Small – 10000
  - Common – 50000



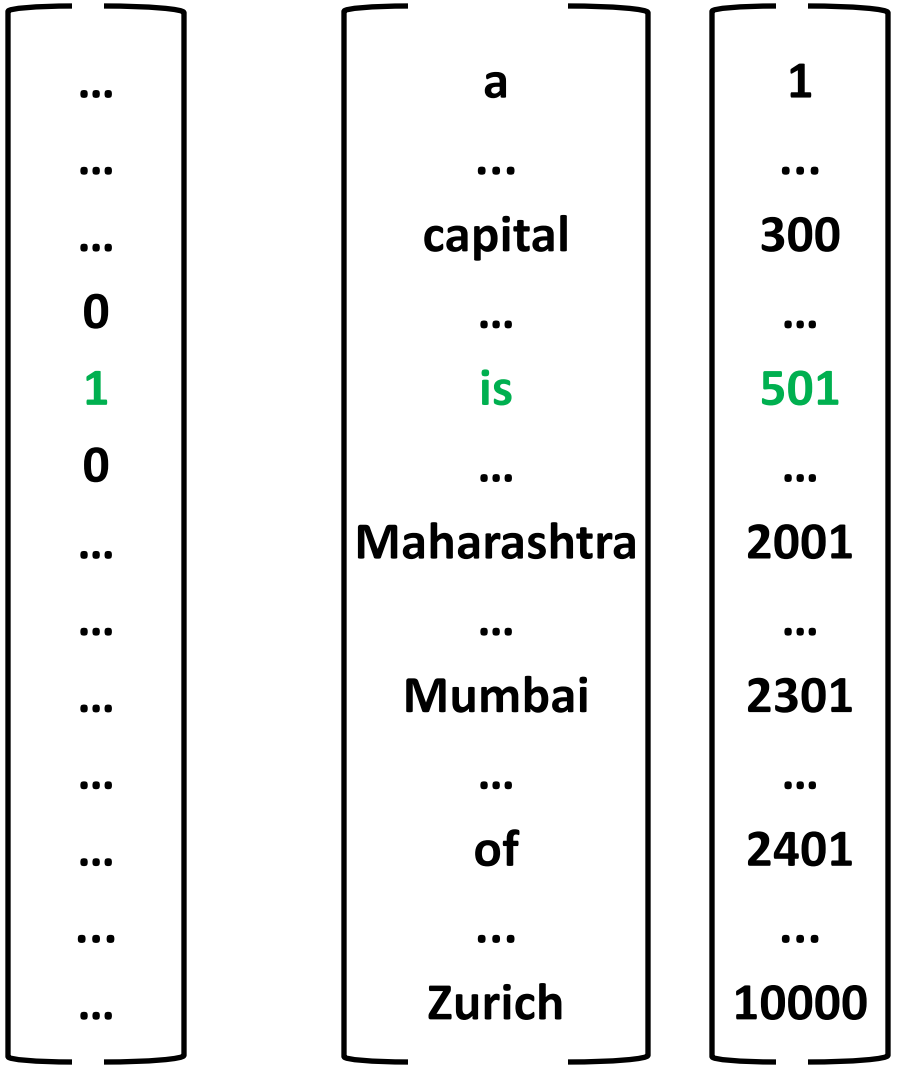
Word representation – Vocabulary

- One hot encoding
- Example
  - Word – capital
  - 1 at 300th location
  - 0 everywhere
  - One hot vector – O_{300}



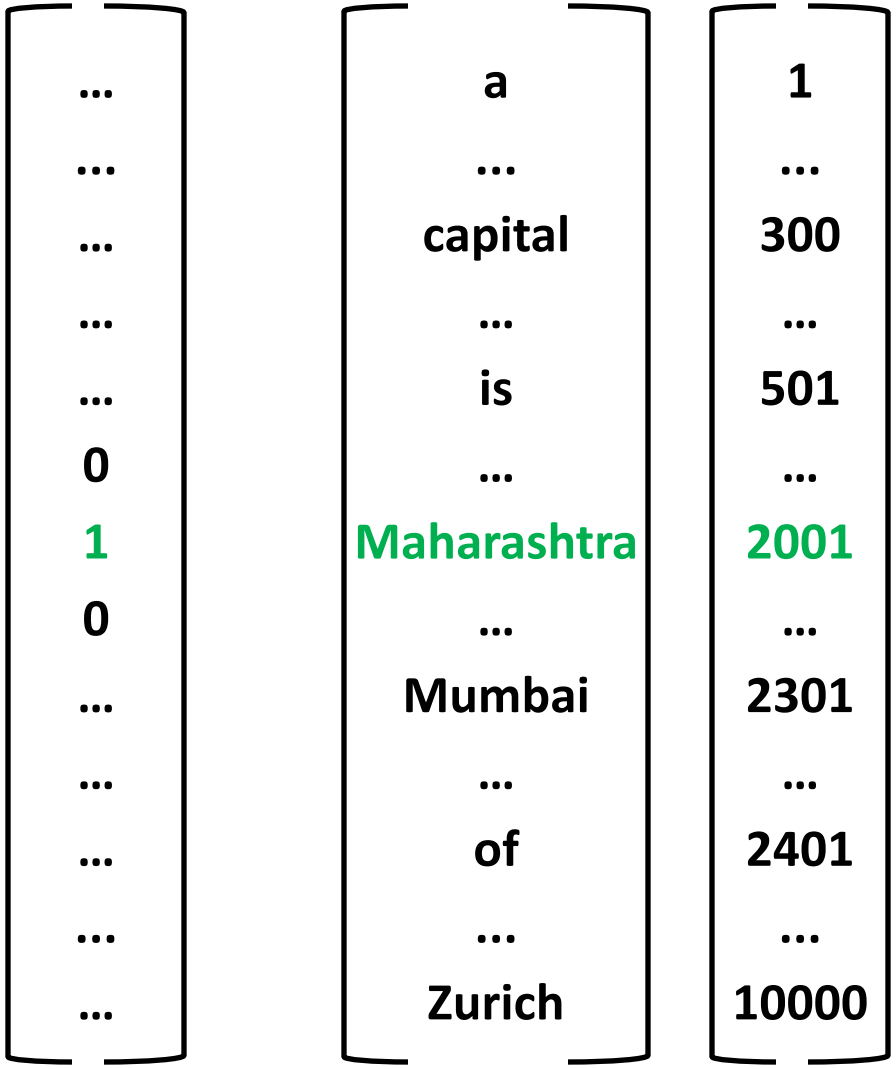
Word representation – Vocabulary

- One hot encoding
- Example
  - Word – is
  - 1 at 501st location
  - 0 everywhere
  - One hot vector – O_{500}

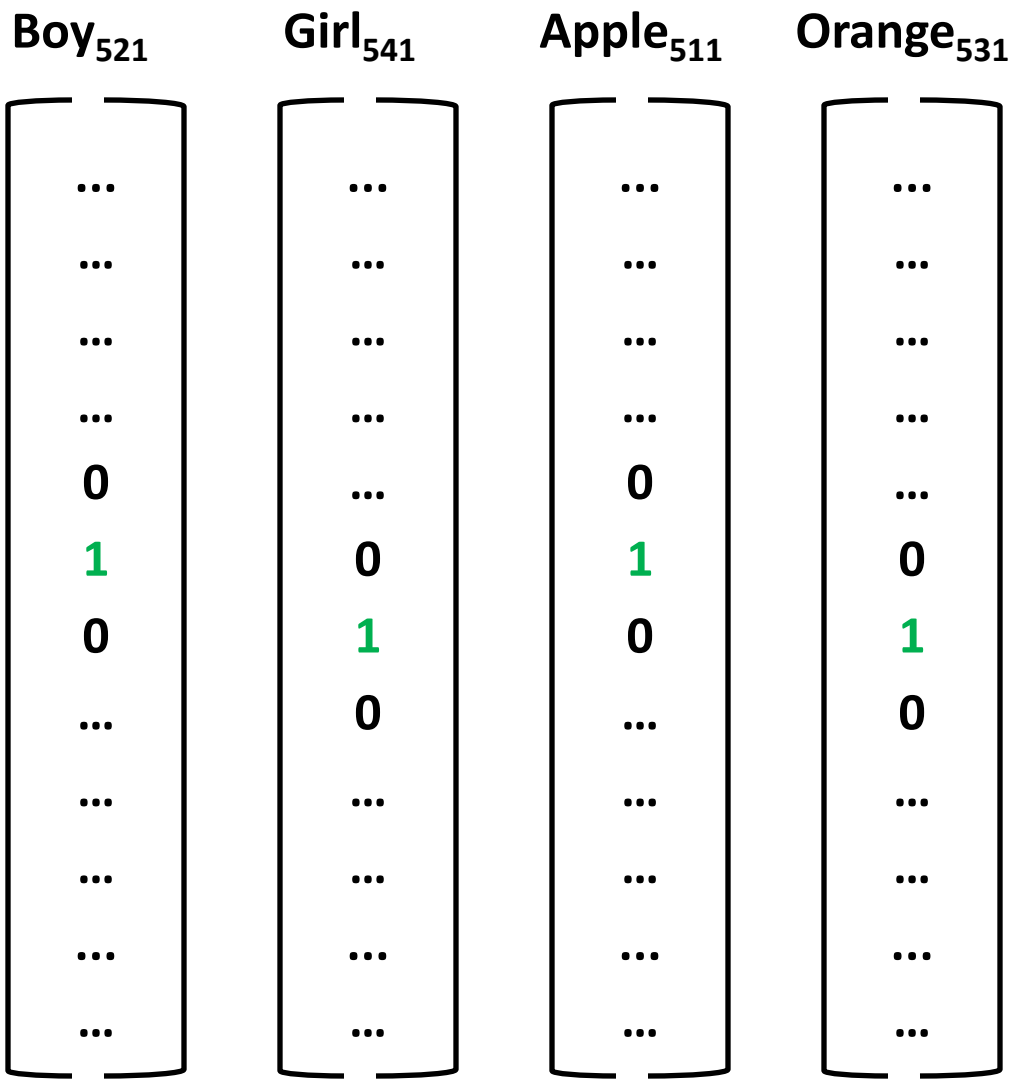


Word representation – Vocabulary

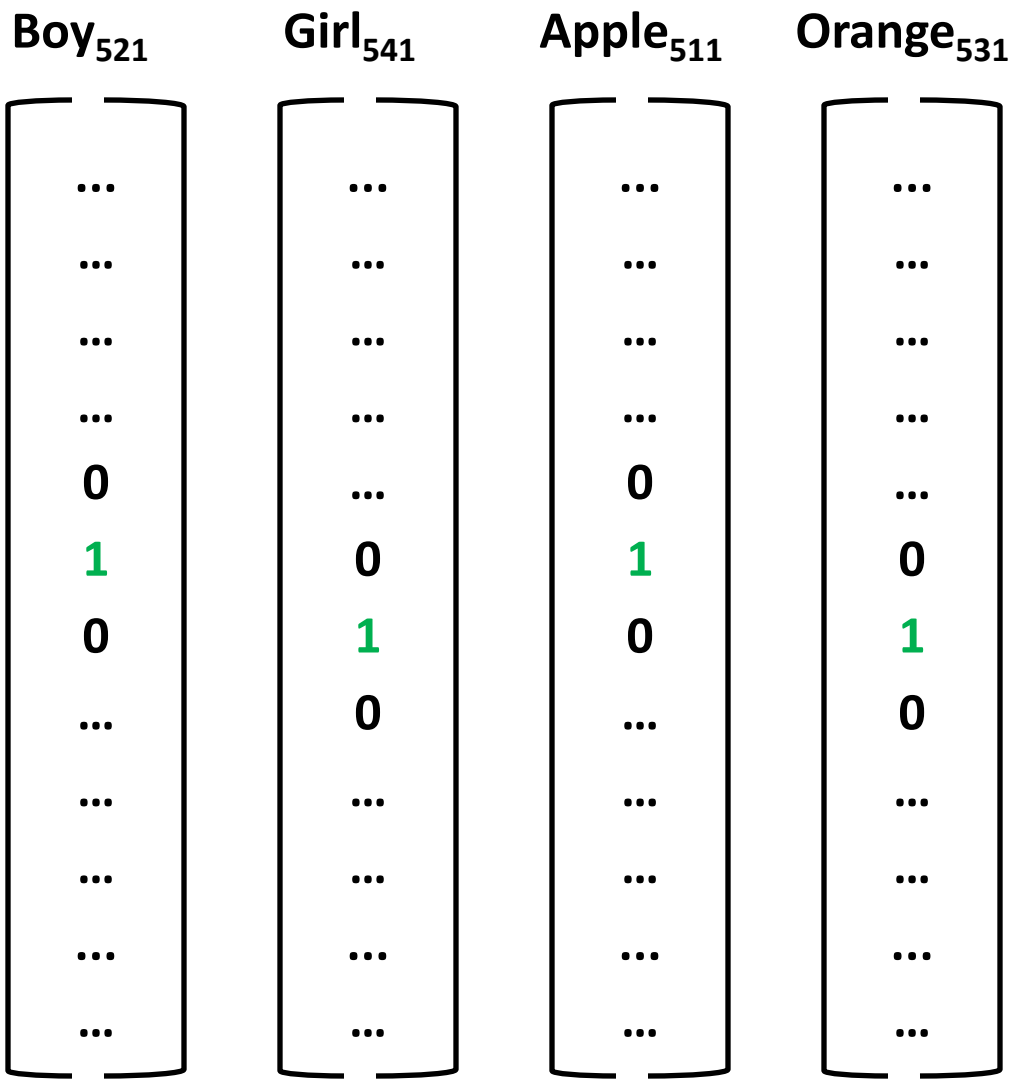
- One hot encoding
- Example
  - Word – Maharashtra
  - 1 at 2001^{st} location
  - 0 everywhere
  - One hot vector – O_{2001}



Word representation – Vocabulary

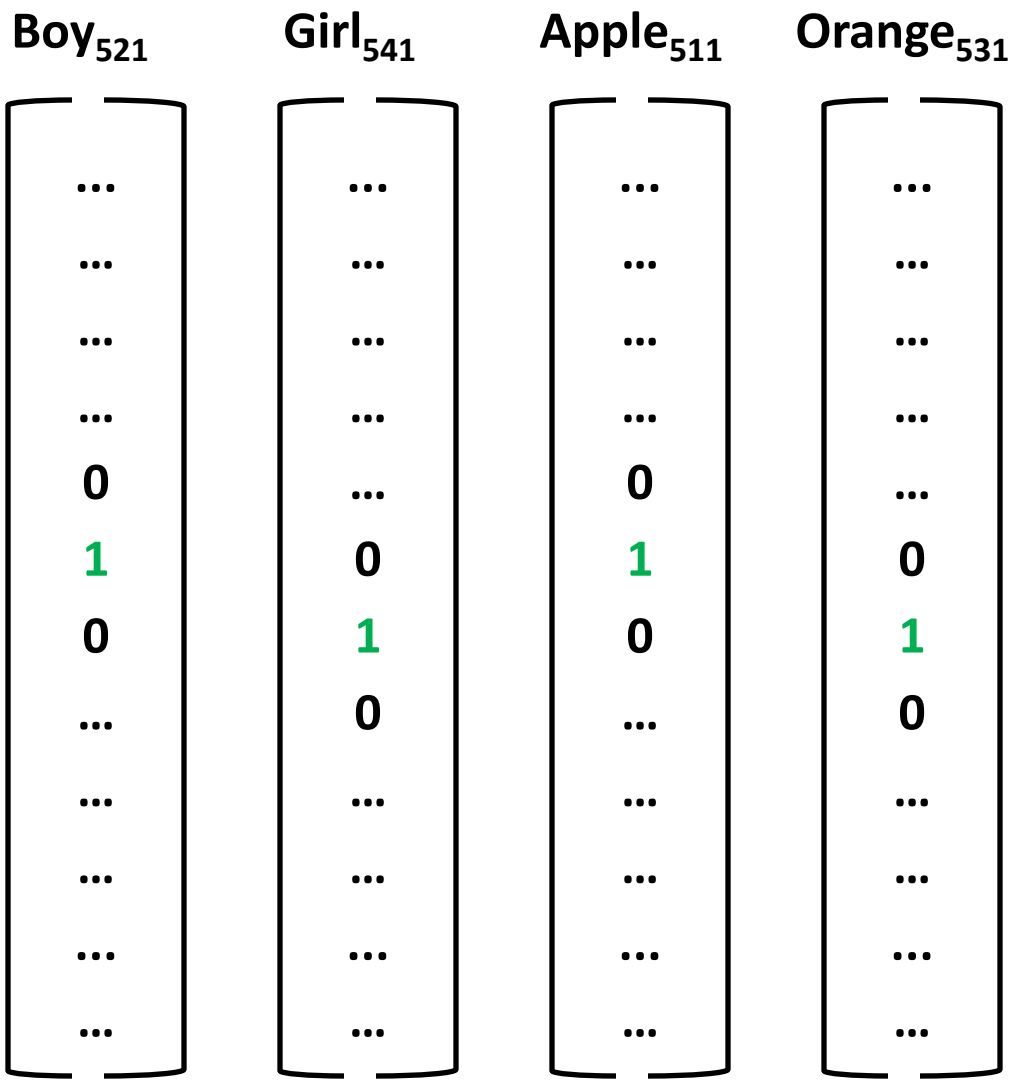


Word representation – Vocabulary



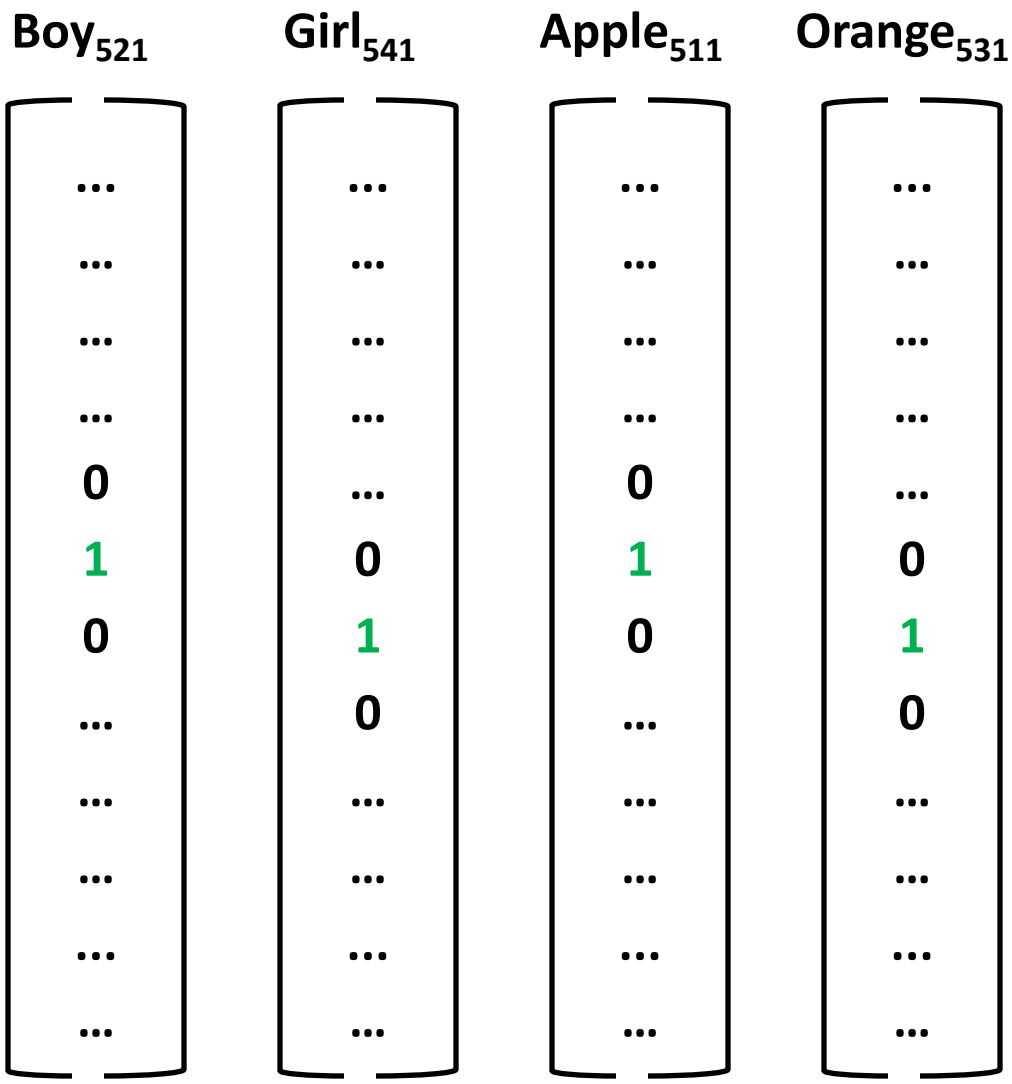
A ___ reads a book.

Word representation – Vocabulary



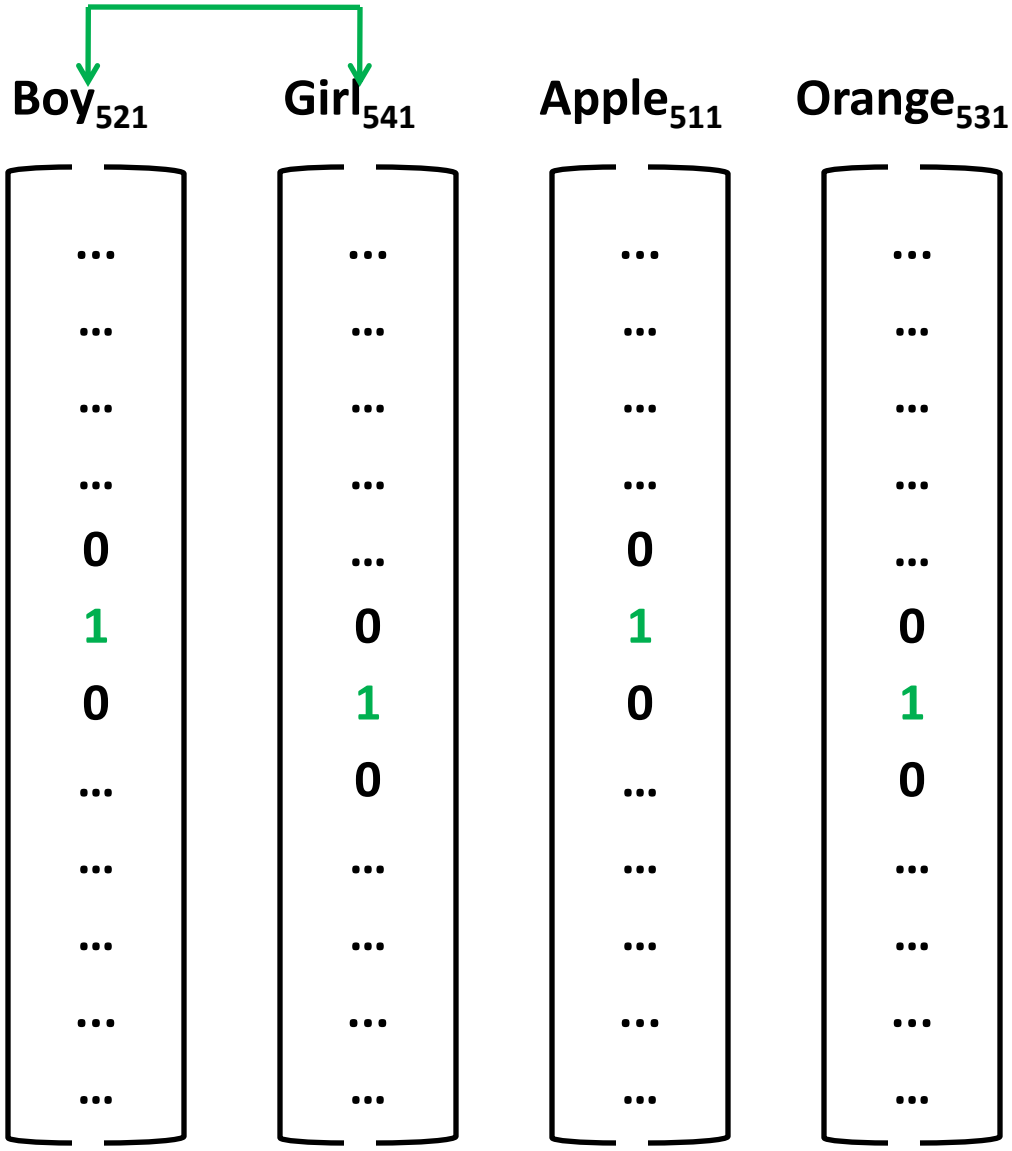
A boy reads a book.

Word representation – Vocabulary



A girl reads a book.

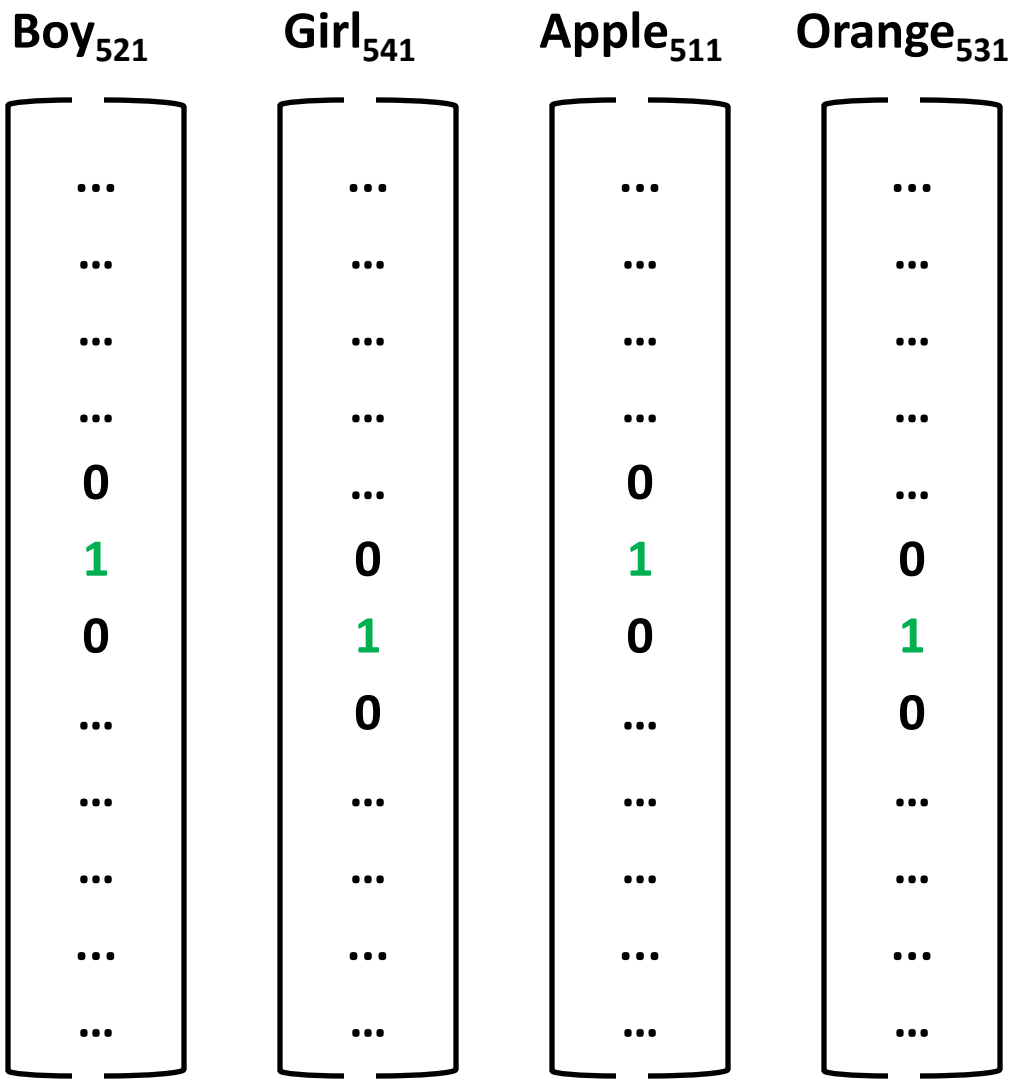
Word representation – Vocabulary



A boy reads a book.

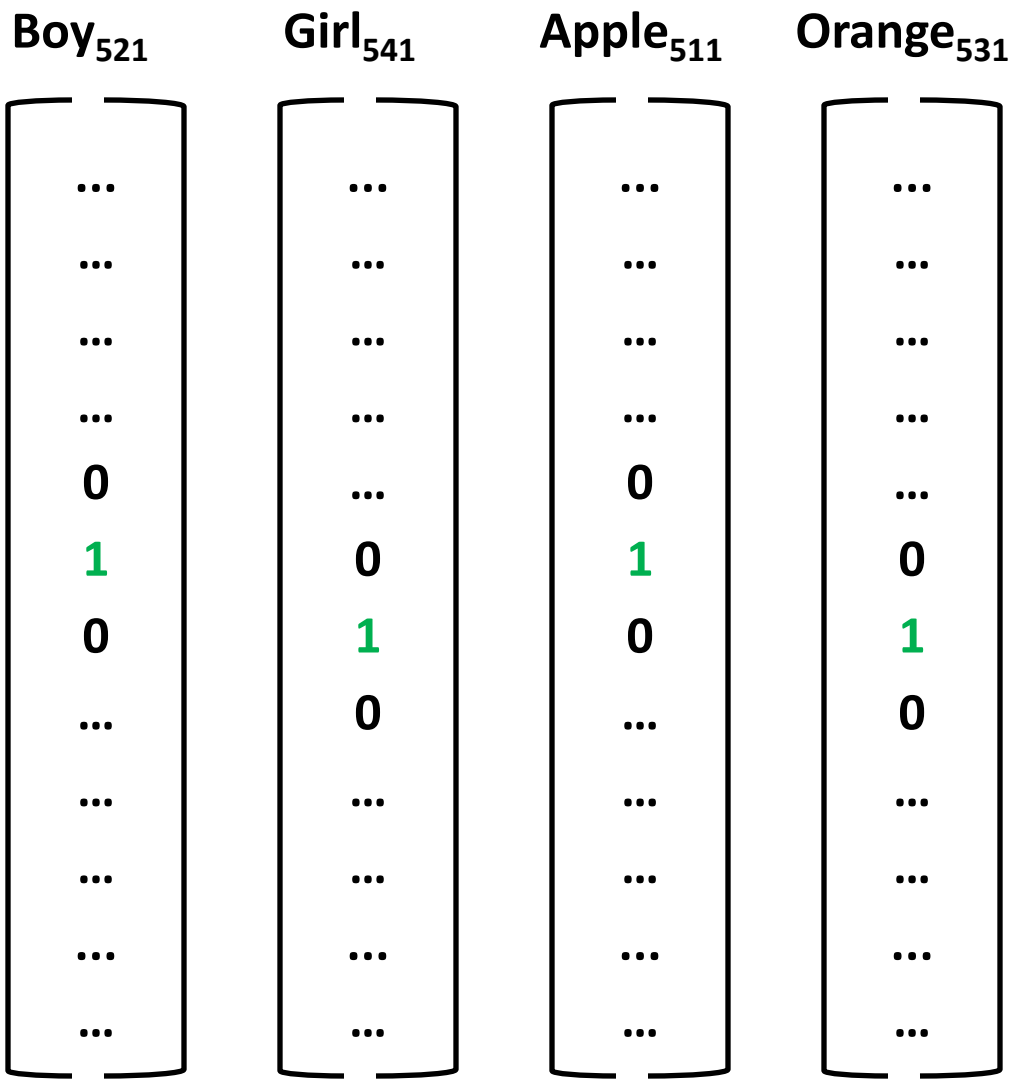
A girl reads a book.

Word representation – Vocabulary



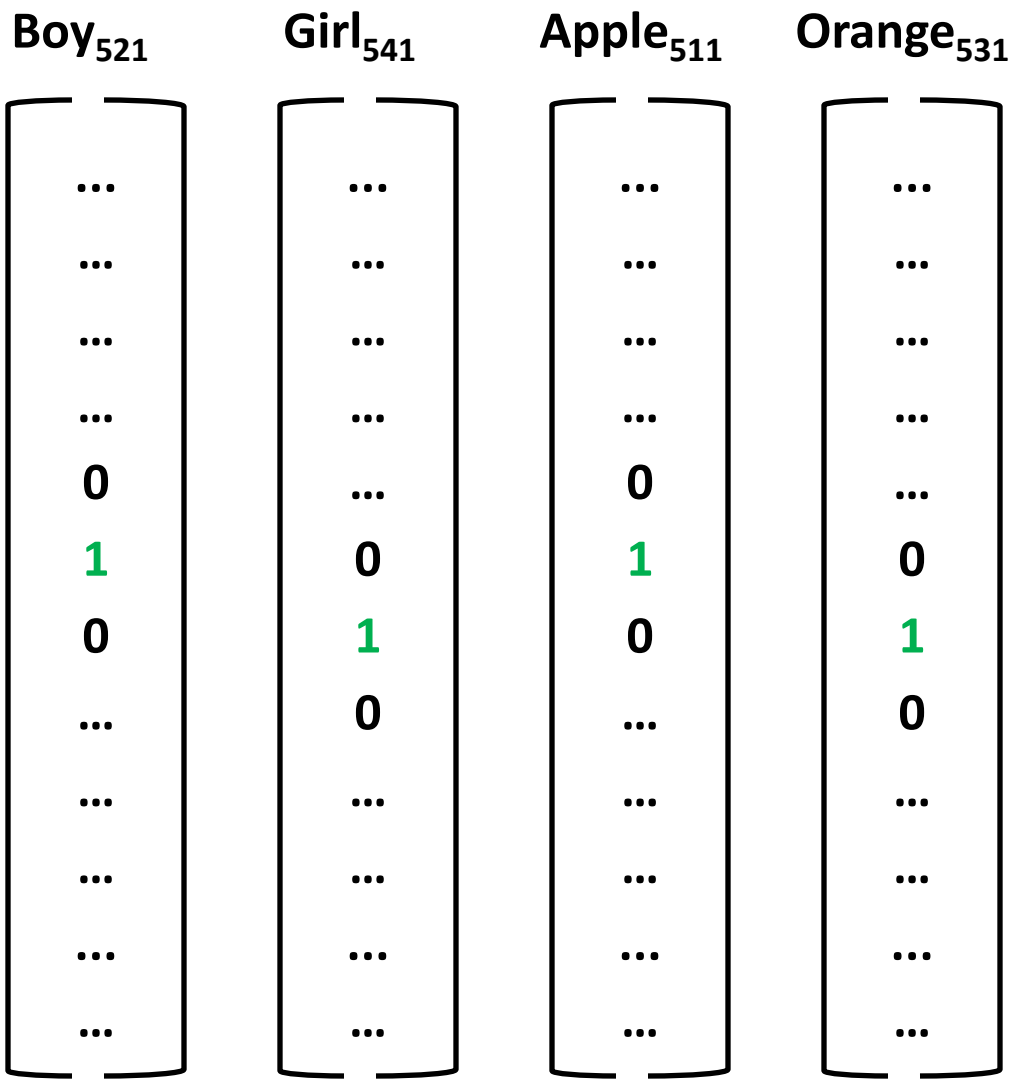
An ___ is a fruit.

Word representation – Vocabulary



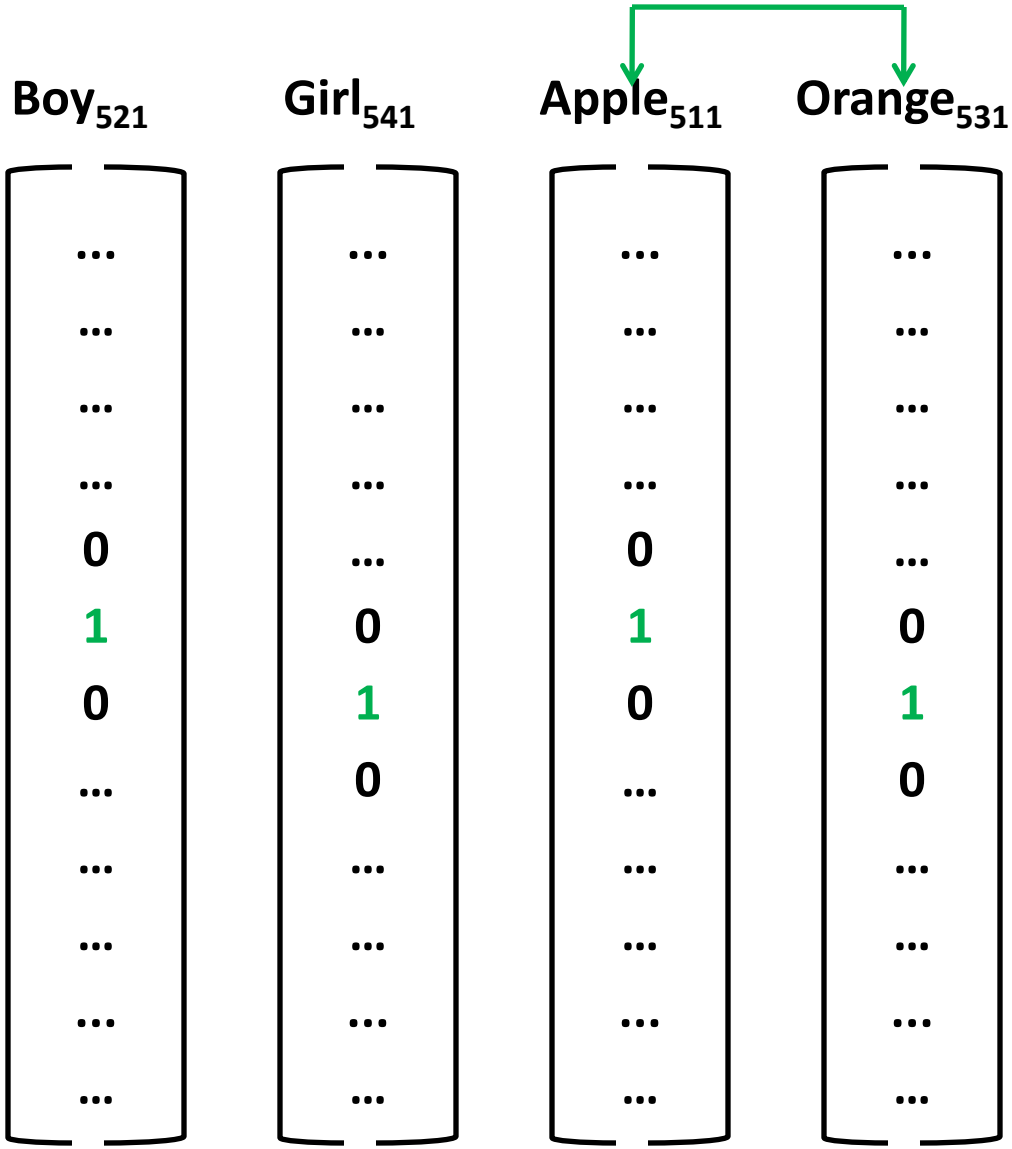
An apple is a fruit.

Word representation – Vocabulary



An orange is a fruit.

Word representation – Vocabulary



An apple is a fruit.

An orange is a fruit.

Featurized representation

Features	Boy_{521}	Girl_{541}	Apple_{511}	Orange_{531}

Featurized representation

Features	Boy_{521}	$Girl_{541}$	$Apple_{511}$	$Orange_{531}$
Age	0.2	0.2	0.005	0.002

Featurized representation

Features	Boy_{521}	$Girl_{541}$	$Apple_{511}$	$Orange_{531}$
Age	0.2	0.2	0.005	0.002
Fruit	0.0001	0.0002	0.8	0.77

Featurized representation

Features	Boy_{521}	$Girl_{541}$	$Apple_{511}$	$Orange_{531}$
Age	0.2	0.2	0.005	0.002
Fruit	0.0001	0.0002	0.8	0.77
Gender	-1	+1	0.002	0.003
Living	0.9	0.88	0.0001	0.0002

Featurized representation

Features	Boy_{521}	$Girl_{541}$	$Apple_{511}$	$Orange_{531}$
Age	0.2	0.2	0.005	0.002
Fruit	0.0001	0.0002	0.8	0.77
Gender	-1	+1	0.002	0.003
Living	0.9	0.88	0.0001	0.0002

Number of features

Featurized representation

Features	Boy_{521}	$Girl_{541}$	$Apple_{511}$	$Orange_{531}$
Age	0.2	0.2	0.005	0.002
Fruit	0.0001	0.0002	0.8	0.77
Gender	-1	+1	0.002	0.003
Living	0.9	0.88	0.0001	0.0002

Number of features ~ 300

Featurized representation

Features	Boy_{521}	$Girl_{541}$	$Apple_{511}$	$Orange_{531}$
Age	0.2	0.2	0.005	0.002
Fruit	0.0001	0.0002	0.8	0.77
Gender	-1	+1	0.002	0.003
Living	0.9	0.88	0.0001	0.0002

e_{521}

Featurized representation

Features	Boy_{521}	$Girl_{541}$	$Apple_{511}$	$Orange_{531}$
Age	0.2	0.2	0.005	0.002
Fruit	0.0001	0.0002	0.8	0.77
Gender	-1	+1	0.002	0.003
Living	0.9	0.88	0.0001	0.0002

e_{541}

Featurized representation

Features	Boy_{521}	$Girl_{541}$	$Apple_{511}$	$Orange_{531}$
Age	0.2	0.2	0.005	0.002
Fruit	0.0001	0.0002	0.8	0.77
Gender	-1	+1	0.002	0.003
Living	0.9	0.88	0.0001	0.0002

e_{511}

Featurized representation

Features	Boy_{521}	$Girl_{541}$	$Apple_{511}$	$Orange_{531}$
Age	0.2	0.2	0.005	0.002
Fruit	0.0001	0.0002	0.8	0.77
Gender	-1	+1	0.002	0.003
Living	0.9	0.88	0.0001	0.0002

An apple is a fruit.

An orange is a fruit.

Featurized representation

Features	Boy_{521}	$Girl_{541}$	$Apple_{511}$	$Orange_{531}$
Age	0.2	0.2	0.005	0.002
Fruit	0.0001	0.0002	0.8	0.77
Gender	-1	+1	0.002	0.003
Living	0.9	0.88	0.0001	0.0002

An apple is a fruit.

An orange is a fruit.

Similar features

Featurized representation

Features	Boy_{521}	$Girl_{541}$	$Apple_{511}$	$Orange_{531}$
Age	0.2	0.2	0.005	0.002
Fruit	0.0001	0.0002	0.8	0.77
Gender	-1	+1	0.002	0.003
Living	0.9	0.88	0.0001	0.0002

A boy reads a book.

A girl reads a book.

Featurized representation

Features	Boy_{521}	$Girl_{541}$	$Apple_{511}$	$Orange_{531}$
Age	0.2	0.2	0.005	0.002
Fruit	0.0001	0.0002	0.8	0.77
Gender	-1	+1	0.002	0.003
Living	0.9	0.88	0.0001	0.0002

A boy reads a book.

A girl reads a book.

Similar features

Visualizing word embeddings

- Embeddings – 300 dimensional or more
- Visualization – 300D to 2D

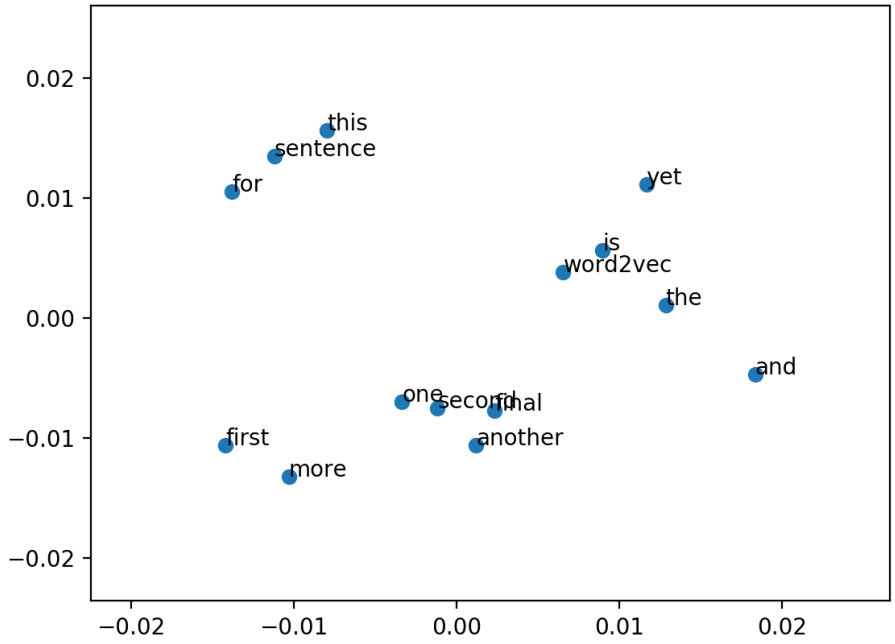


Image ack – Google

Using word embeddings

- Name entity recognition
- Input – Mumbai is capital of Maharashtra.
- Output – 1 0 0 0 1

- Input – Bangalore is capital of Karnataka.
- Output – 1 0 0 0 1

Using word embeddings

- Name entity recognition
- Input – Mumbai is capital of Maharashtra.
- Output – 1 0 0 0 1

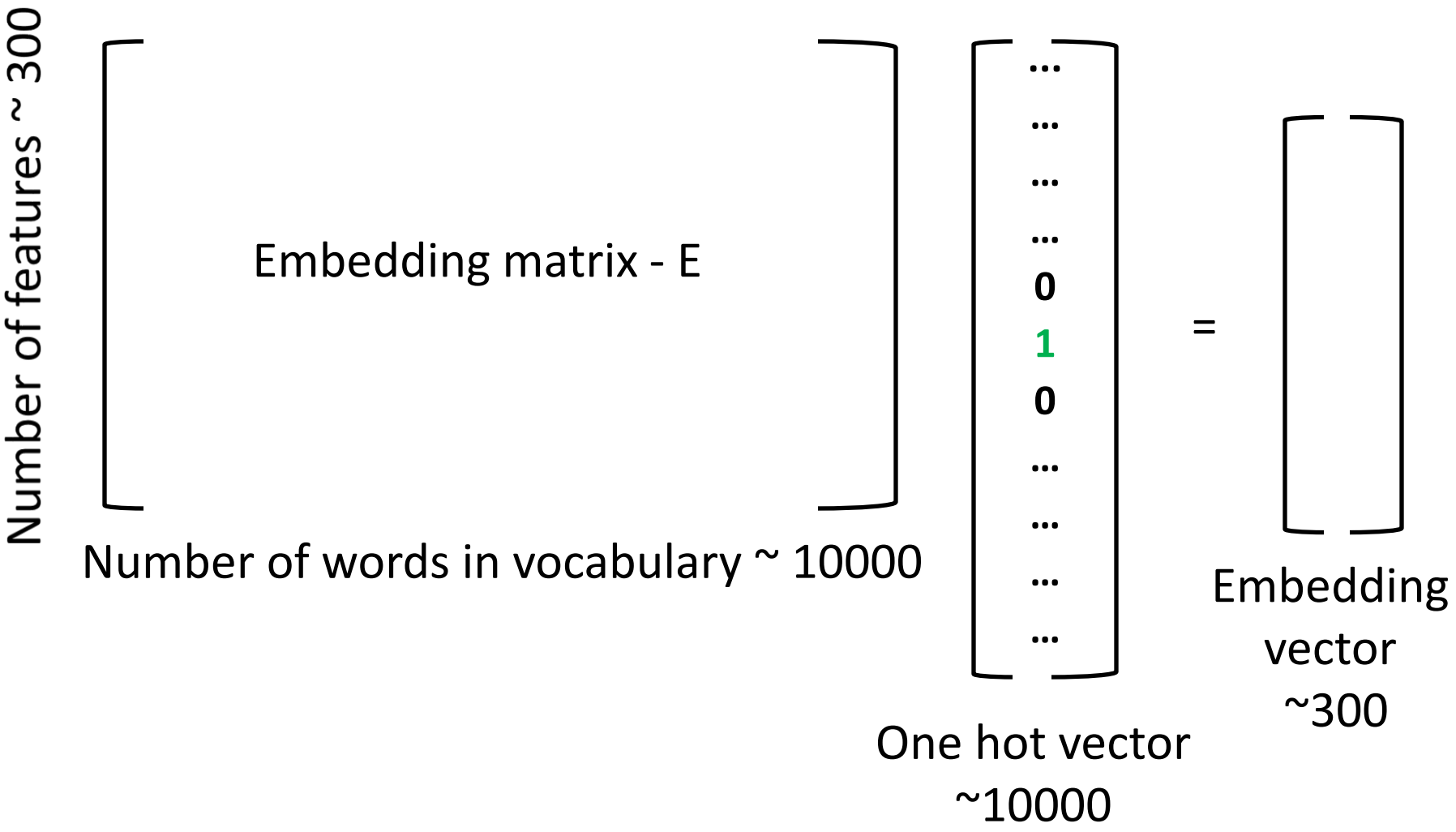
- Input – Bangalore is capital of Karnataka.
- Output – 1 0 0 0 1

Using word embeddings

- Name entity recognition
- Input – Mumbai is capital of Maharashtra.
- Output – 1 0 0 0 1

- Input – Bangalore is capital of Karnataka.
- Output – 1 0 0 0 1

Embedding matrix



Learning word embeddings

Input – I take a glass of ?

Learning word embeddings

Input –	I	take	a	glass	of	?
Index –	52	556	1	234	67	

Learning word embeddings

Input –	I	take	a	glass	of	?
Index –	52	556	1	234	67	
One hot –	O_{52}	O_{556}	O_1	O_{234}	O_{67}	

Learning word embeddings

Input –	I	take	a	glass	of	?
Index –	52	556	1	234	67	
One hot –	O_{52}	O_{556}	O_1	O_{234}	O_{67}	
Embeddings –	e_{52}	e_{556}	e_1	e_{234}	e_{67}	

Learning word embeddings

- Input – Word embedding
- Output – One hot vector
- Train neural network
  - Artificial neural network
  - Convolutional neural network
  - Recurrent neural network
- Word embeddings
  - Back propagation

Questions?

Thank you