NLP!!! (Part 2)

April 9, 2020

Data Science CSCI 1951A

Brown University

Instructor: Ellie Pavlick

HTAs: Josh Levin, Diane Mutako, Sol Zitter

Announcements

- Viz Lab tomorrow afternoon (4pm? Check Piazza)
- Project Grades/Pitches/Presentations

Today

- More NLP!
- Ngrams
- Topic Models
- Word Embeddings

Today

- More NLP!
- **Ngrams**
- Topic Models
- Word Embeddings

N-Grams

- N-length sequence of words (unigrams, bigrams, trigrams, 4-grams, …)
- Provides some context (differentiating “cute **dog”** from “hot **dog**”)
- Blows up size of vocabulary, increases sparsity
- Usually vocab size cutoffs/min count thresholds apply to ngrams too

N-Grams

html does work . all webdev is awesome.

1gms: [‘html’, ‘does’, ‘work’, ‘.’, ‘all’, …]

2gms: [‘html does’, ‘does work’, ‘work .’, ‘. all’, …]

3gms: [‘html does work’, ‘does work .’, ‘work . all’, …]

N-Grams

html does work . all webdev is awesome.

1gms: [‘html’, ‘does’, ‘work’, ‘.’, ‘all’, …]

2gms: [‘html does’, ‘does work’, ‘work .’, ‘. all’, …]

3gms: [‘html does work’, ‘does work .’, ‘work . all’, …]

skip-1gms: [‘html does’, ‘html work’, ‘does html’, ‘does work’, …]

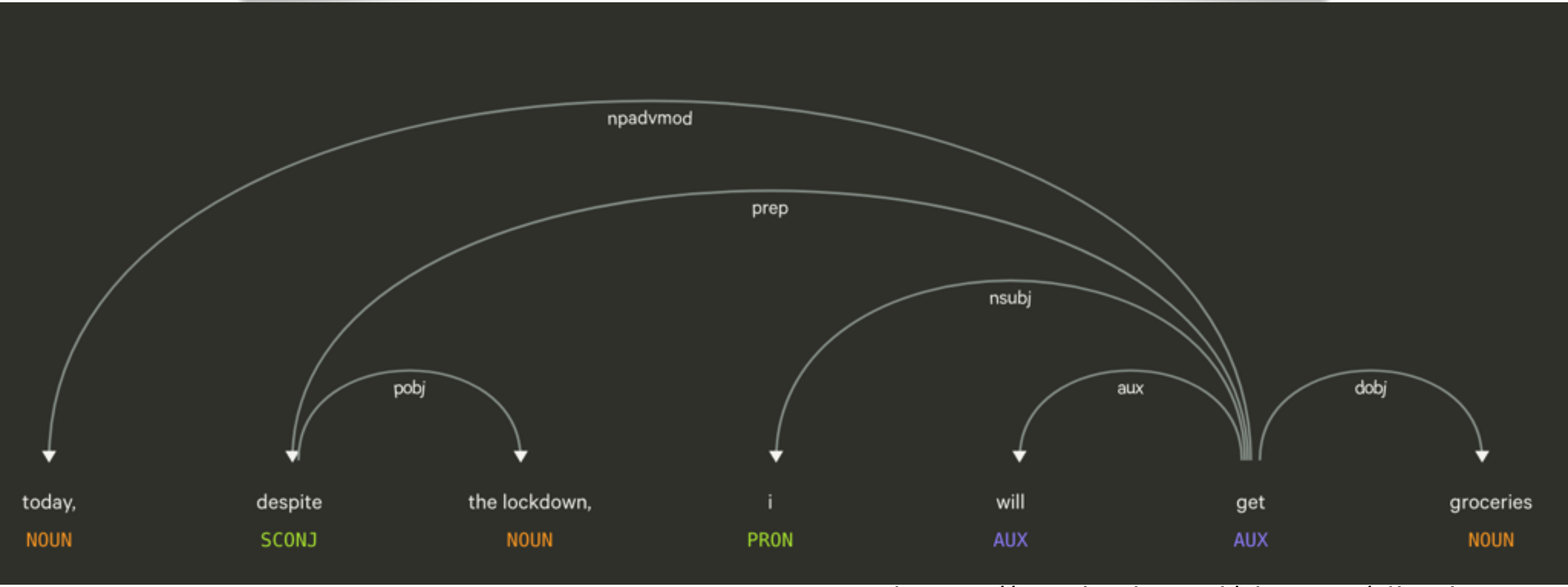
Tagging

- Parts of Speech — “fly” the noun or “fly” the verb?
- Word Sense Disambiguation — “fly” as in “take an airplane” or “fly” as in “go fast”?
- Named Entity Recognition — “Washington” the place or “Washington” the person

Syntactic Relations

"Dependency Parsing"

today, despite the lockdown, i will get groceries

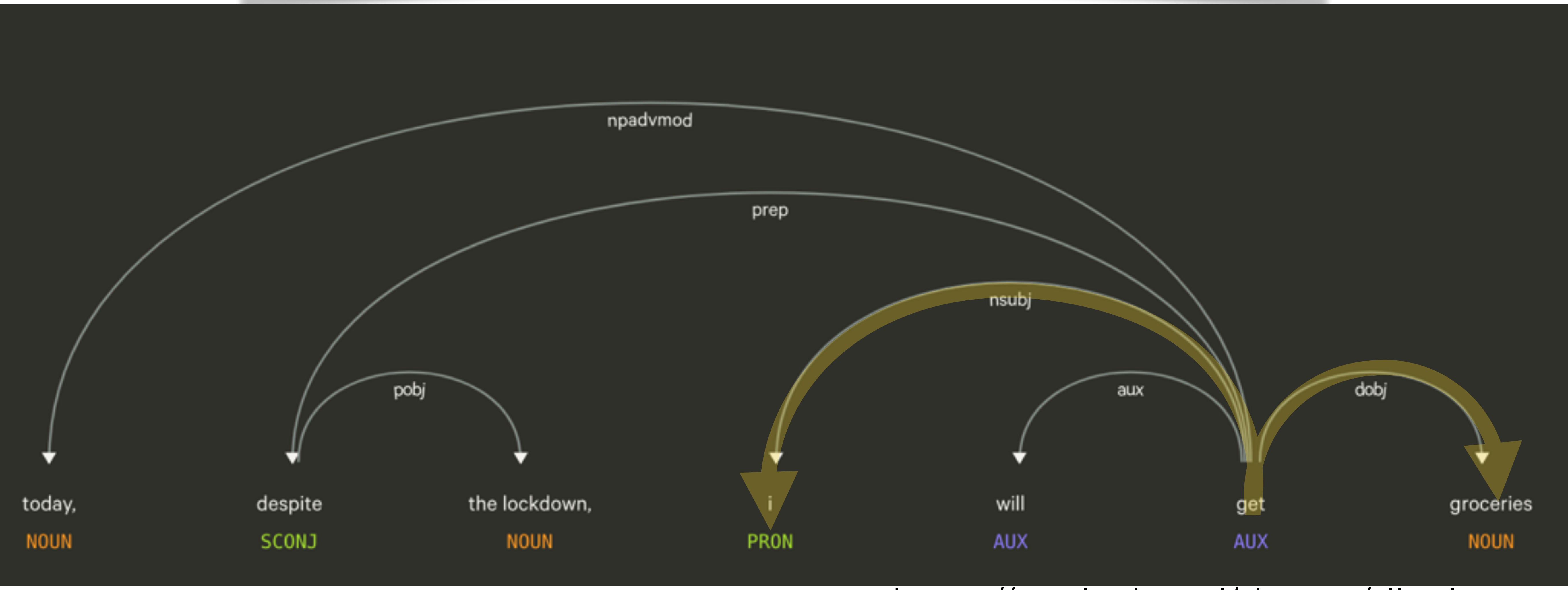


https://explosion.ai/demos/displacy

Syntactic Relations

"Dependency Parsing"

today, despite the lockdown, i will get groceries

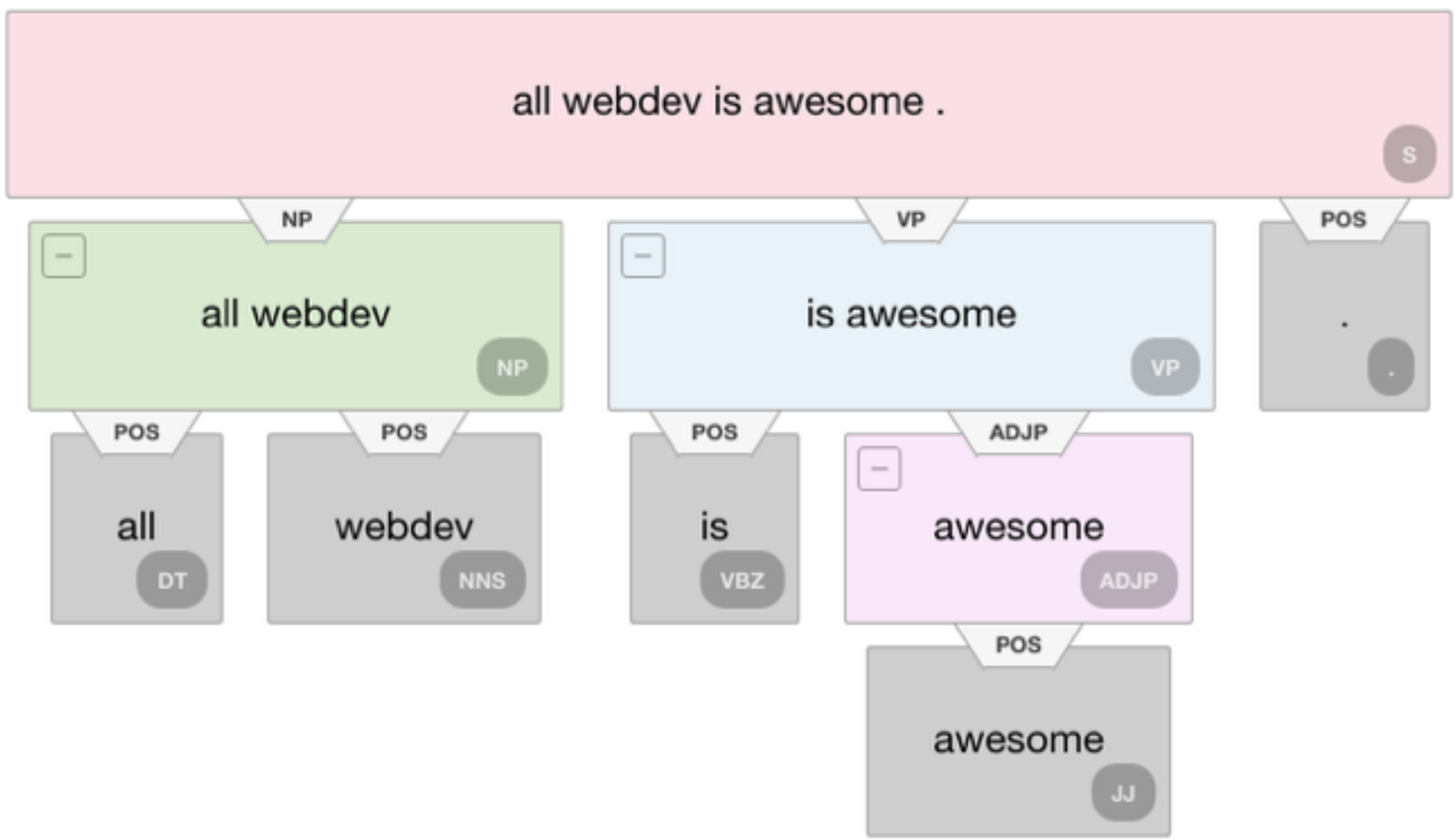


https://explosion.ai/demos/displacy

Syntactic Relations

"Constituency Parsing"

all webdev is awesome.



https://demo.allennlp.org/constituency-parsing

Today

- More NLP!
- Ngrams
- **Topic Models**
- Word Embeddings

Topic Models

Can you elaborate on exactly what the directions are in part 2 step 3, the stencil code does not quite imply what we are supposed to do…

When I try to display dots from part 2 on my mac (tried chrome, firefox, and safari), the elements do not appear in the html.

Changes I make to the nations.js file do not affect any of the html in after I load the nations.html file

Topic Models

Can you elaborate on exactly what the directions are in part 2 step 3, the stencil code does not quite imply what we are supposed to do…

When I try to display dots from part 2 on my mac (tried chrome, firefox, and safari), the elements do not appear in the html.

Changes I make to the nations.js file do not affect any of the html in after I load the nations.html file

instructions: stencil, instructions, part, step, rubric, handin…

UI: html, javascript, debug, display, elements…

systems: mac, windows, linux, chrome, firefox, os…

fillers: I, you, when, the, and, a

Topic Models

Where do documents come from?
"The generative story"

instructions: stencil, instructions, part, step, rubric, handin…
UI: html, javascript, debug, display, elements…
systems: mac, windows, linux, chrome, firefox, os…
fillers: I, you, when, the, and, a

Topic Models

Where do documents come from?
"The generative story"

instructions: stencil, instructions, part, step, rubric, handin…
UI: html, javascript, debug, display, elements…
systems: mac, windows, linux, chrome, firefox, os…
fillers: I, you, when, the, and, a

1. Sample a topic

Topic Models

Where do documents come from?
"The generative story"

instructions: stencil, instructions, part, step, rubric, handin…
UI: html, javascript, debug, display, elements…
systems: mac, windows, linux, chrome, firefox, os…
fillers: I, you, when, the, and, a

You

2. Sample a word from that topic

Topic Models

Where do documents come from?
"The generative story"

instructions: stencil, instructions, part, step, rubric, handin…
UI: html, javascript, debug, display, elements…
systems: mac, windows, linux, chrome, firefox, os…
fillers: I, you, when, the, and, a

You

1. Sample a topic

Topic Models

Where do documents come from?
"The generative story"

instructions: stencil, instructions, part, step, rubric, handin…
UI: html, javascript, debug, display, elements…
systems: mac, windows, linux, chrome, firefox, os…
fillers: I, you, when, the, and, a

You javascript

2. Sample a word from that topic

Topic Models

Where do documents come from?
"The generative story"

instructions: stencil, instructions, part, step, rubric, handin…
UI: html, javascript, debug, display, elements…
systems: mac, windows, linux, chrome, firefox, os…
fillers: I, you, when, the, and, a

You javascript

1. Sample a topic

Topic Models

Where do documents come from?
"The generative story"

instructions: stencil, instructions, part, step, rubric, handin…
UI: html, javascript, debug, display, elements…
systems: mac, windows, linux, chrome, firefox, os…
fillers: I, you, when, the, and, a

You javascript handin

2. Sample a word from that topic

Topic Models

$$P(w_i) = \sum_{j=1}^{T} P(w_i \mid z_i = j) P(z_i = j)$$

Topic Models

"latent" variable (not observed)

$$P(w_i) = \sum_{j=1}^{T} P(w_i \mid z_i = j) P(z_i = j)$$

Topic Models

words are determined by topic
(and are conditionally independent of each other)

$$P(w_i) = \sum_{j=1}^{T} P(w_i \mid z_i = j) P(z_i = j)$$

Topic Models

documents are a distribution over topics

$$P(w_i) = \sum_{j=1}^{T} P(w_i \mid z_i = j) P(z_i = j)$$

Topic Models

set parameters to maximize probability of observations

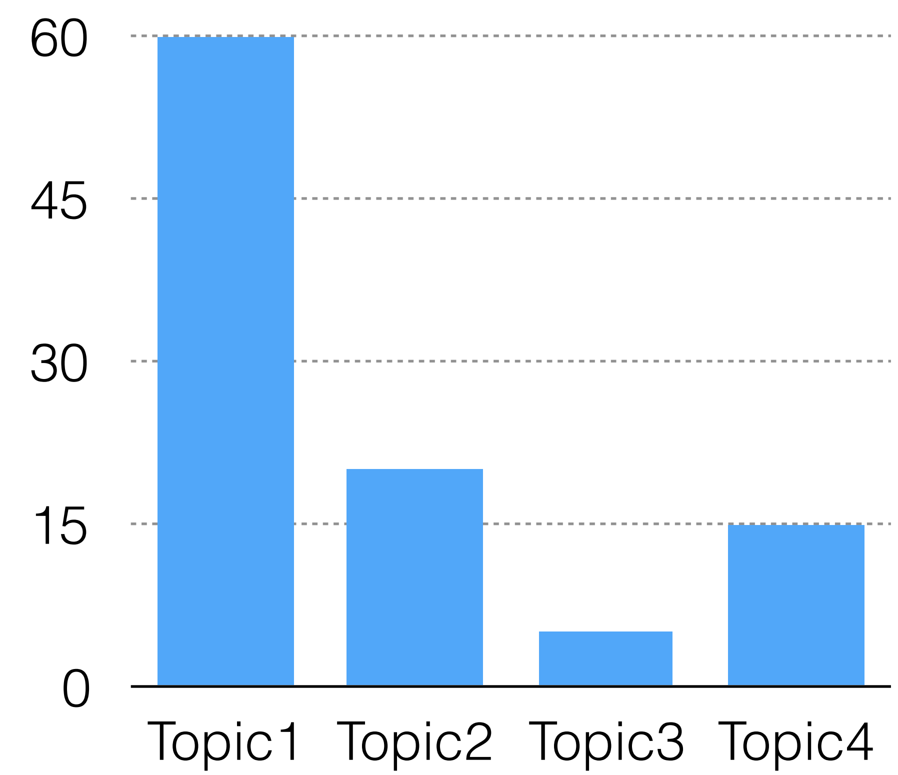
$$P(w_i) = \sum_{j=1}^{T} P(w_i \mid z_i = j) P(z_i = j)$$

Topic Models

part 2 html does not work

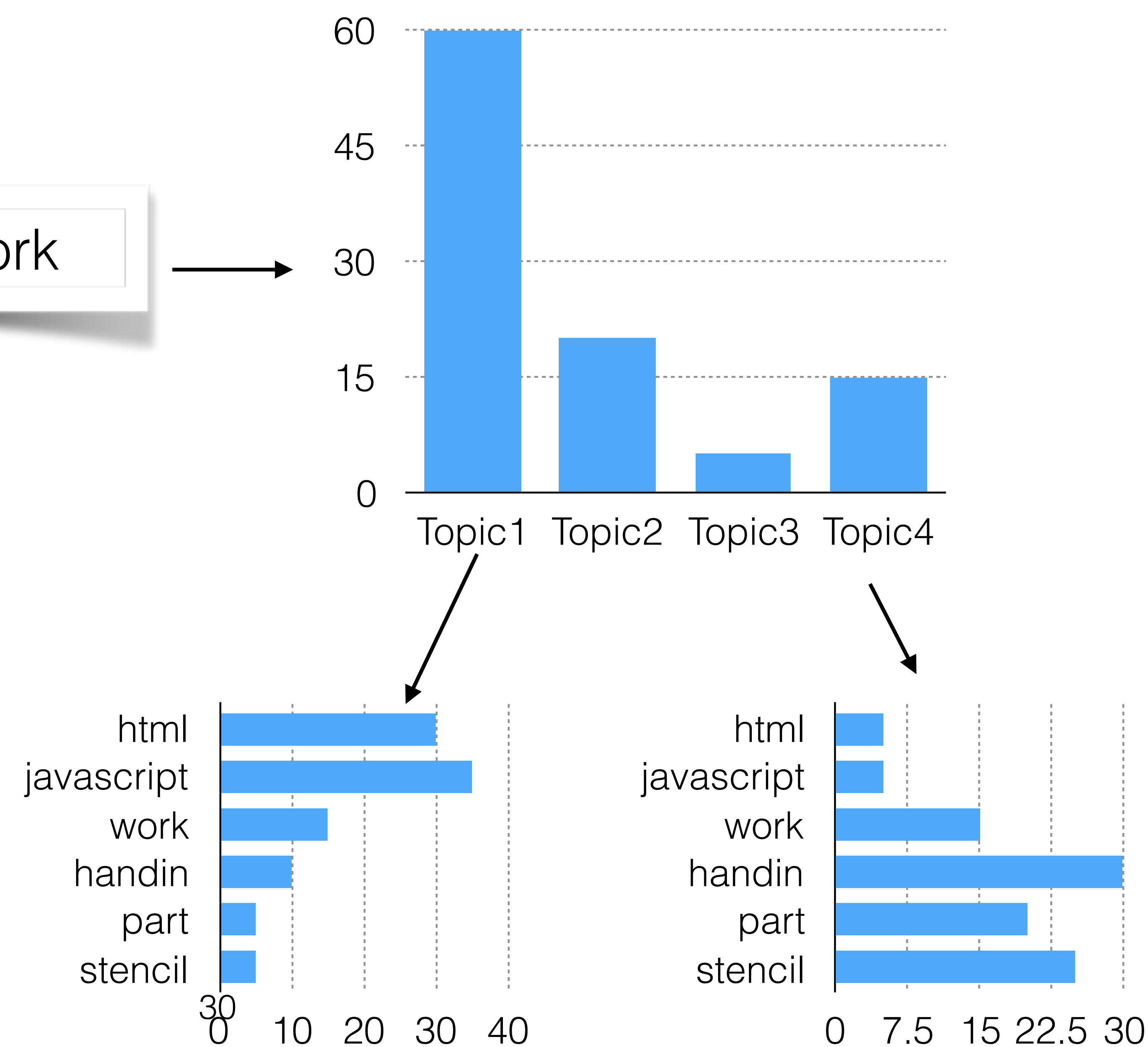
Topic Models

part 2 html does not work



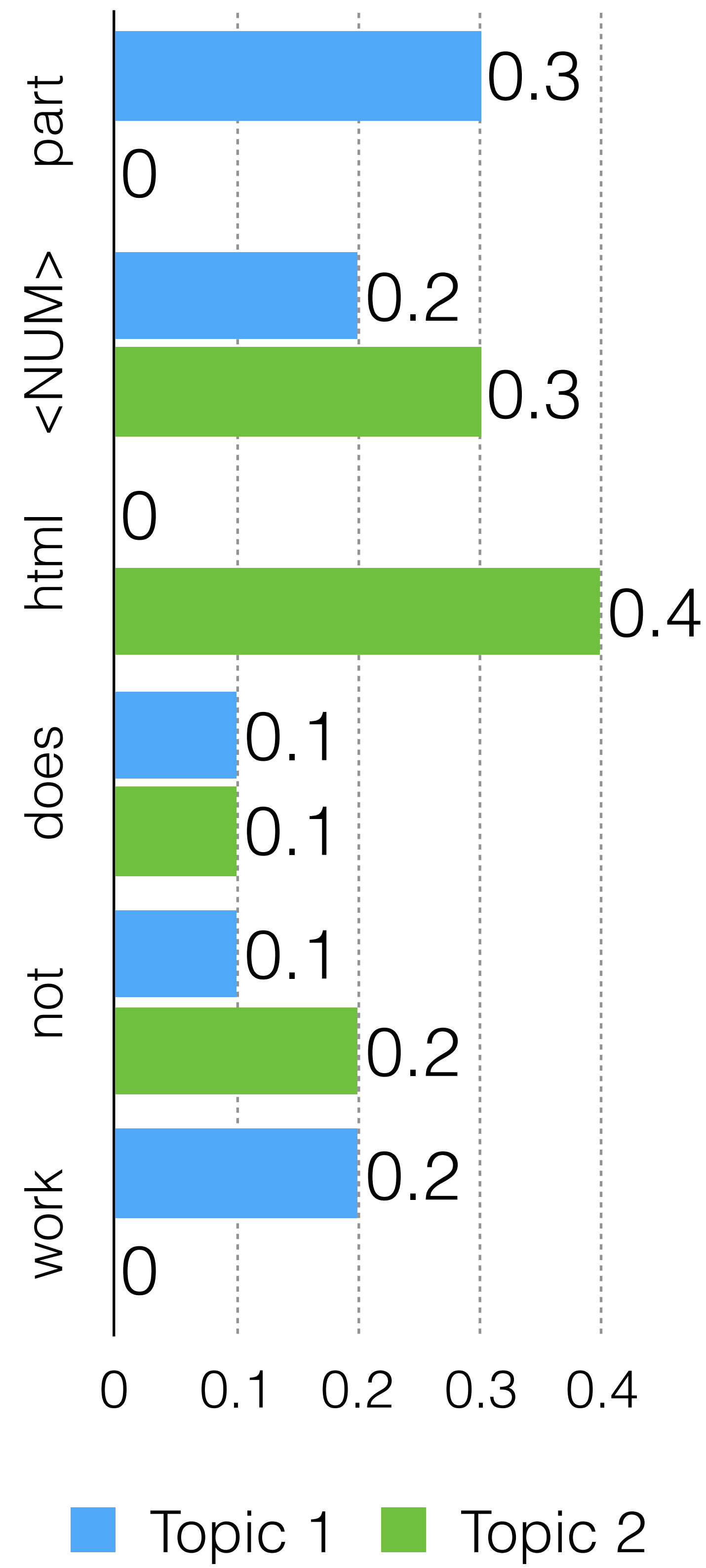
Topic Models

part 2 html does not work



Clicker Question!

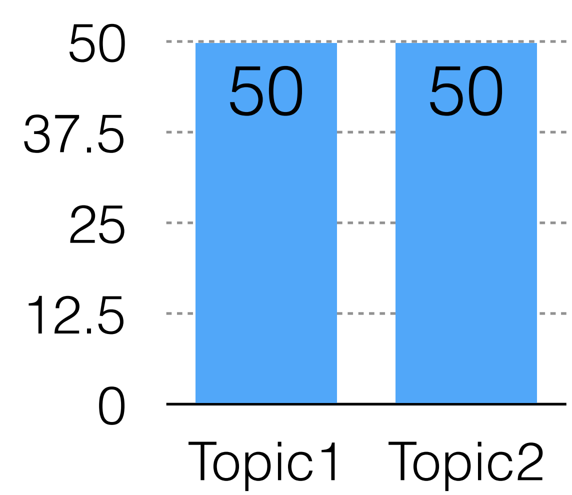
Clicker Question!



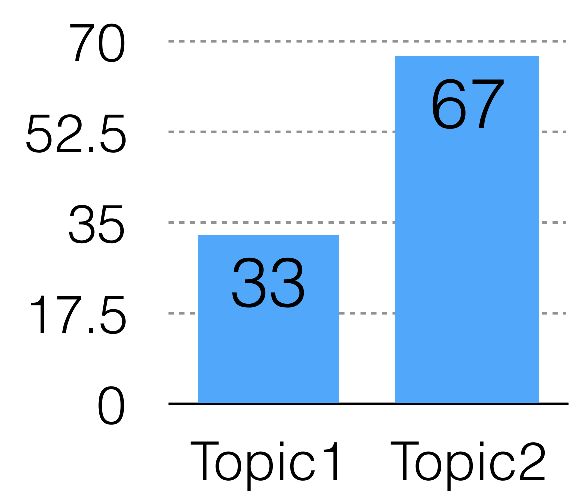
Which is the best parameter setting for the observed data?

$$P(w_i) = \sum_{j=1}^{T} P(w_i \mid z_i = j) P(z_i = j)$$

part <NUM> html does not work

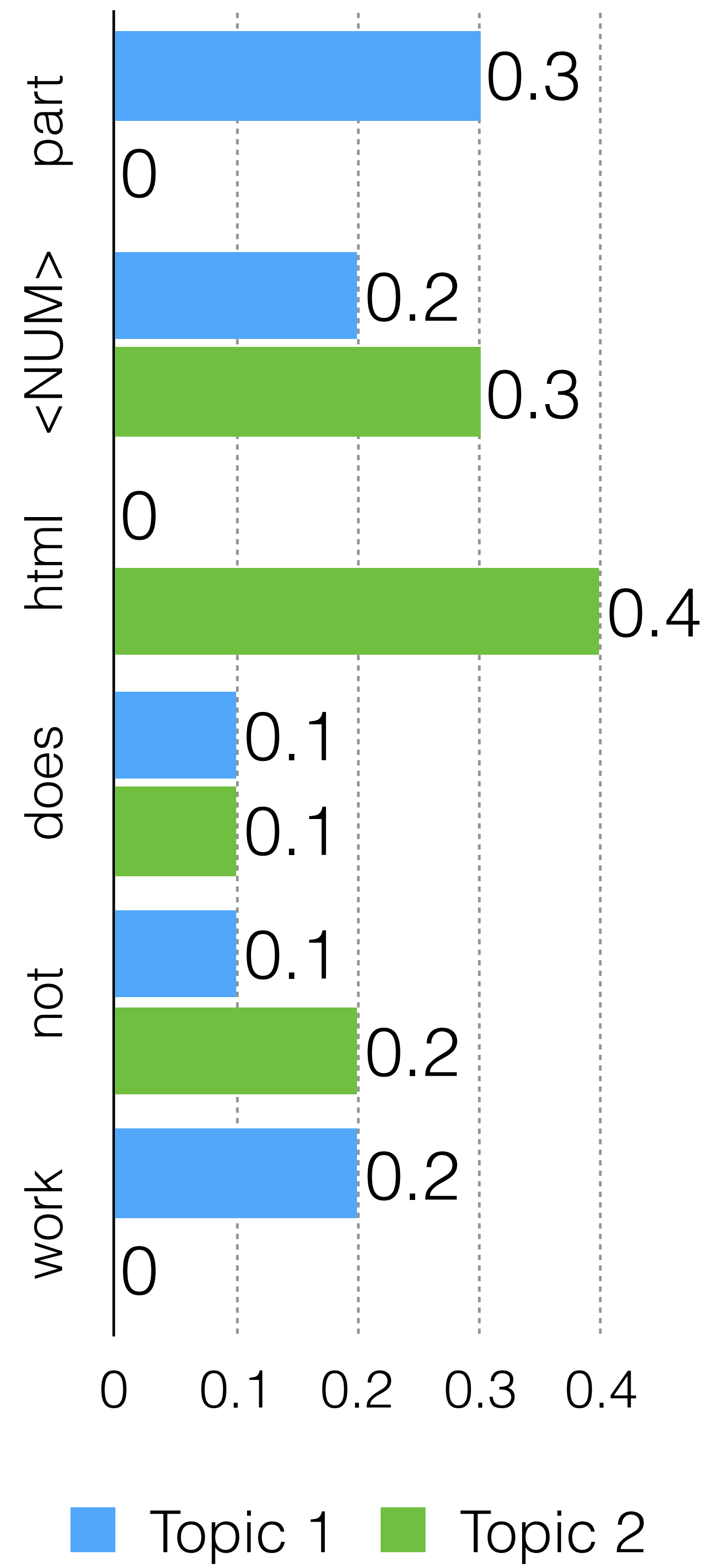


(a)



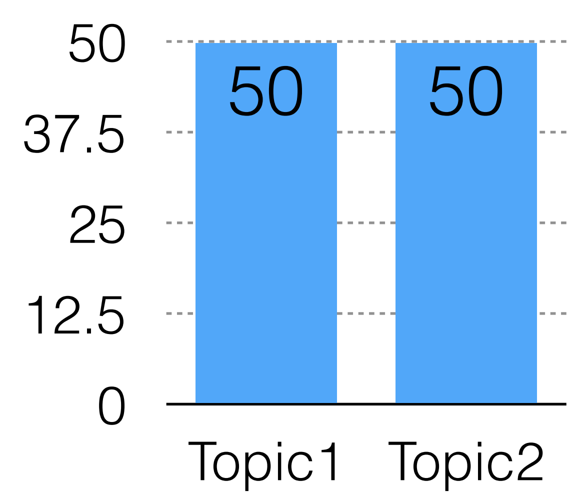
(b)

Clicker Question!

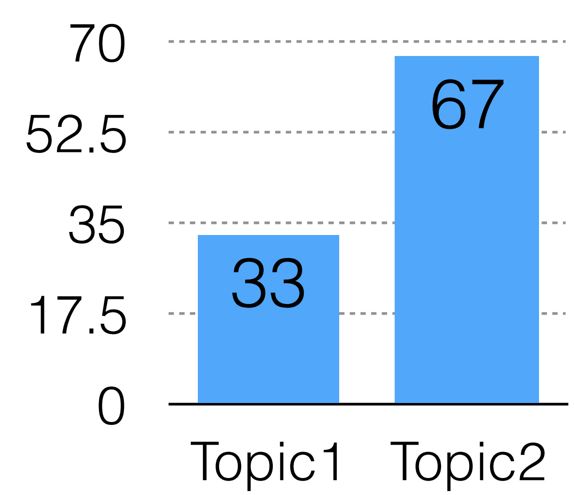


a: (0.3+0.2+0+0.1+0.1+0.2)x0.5

part <NUM> html does not work

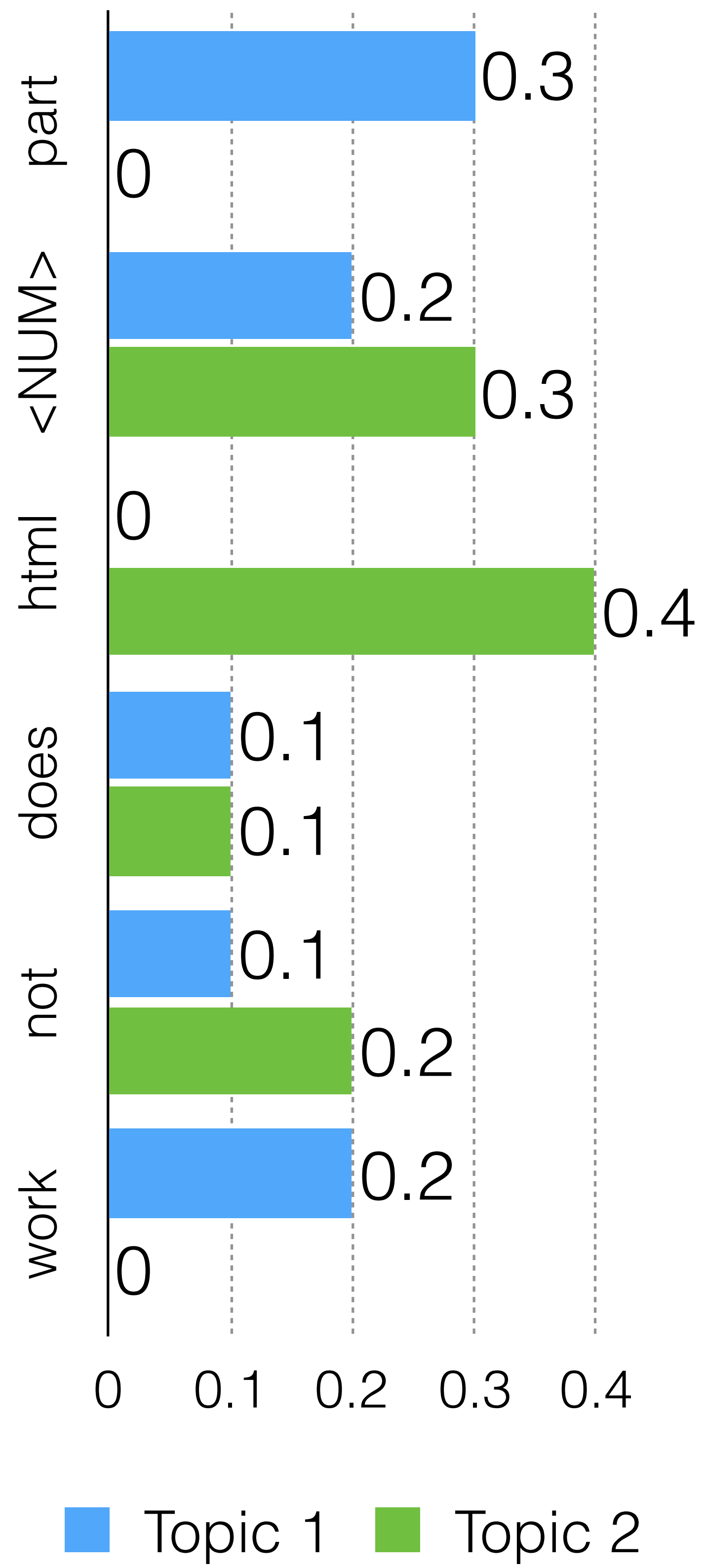


(a)



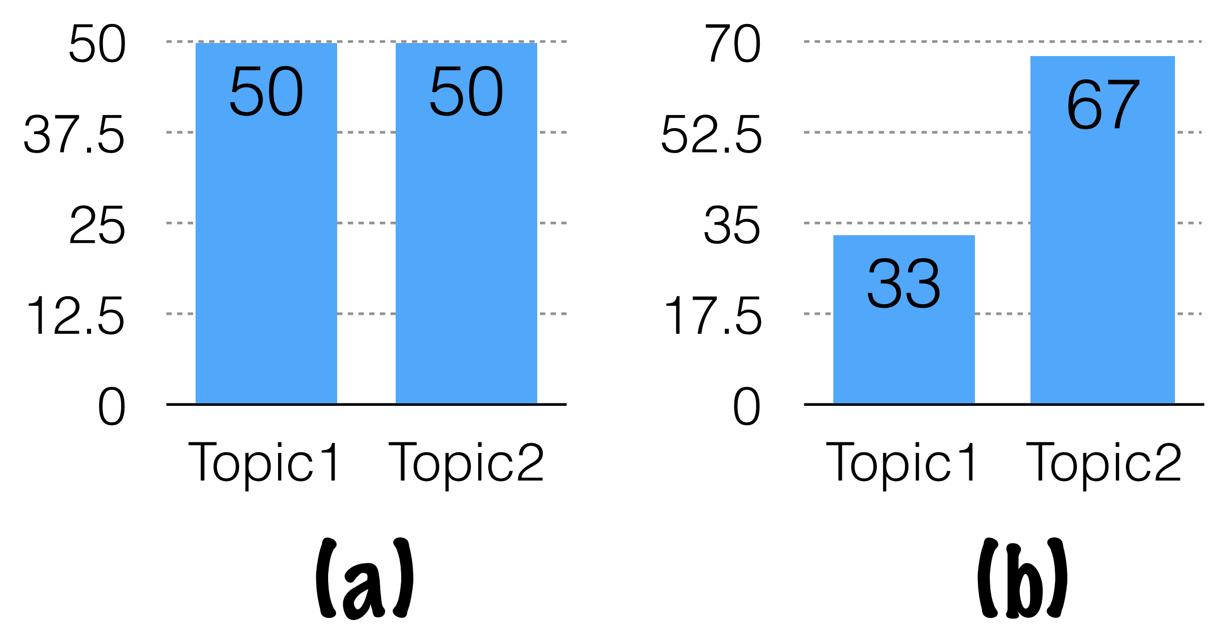
(b)

Clicker Question!

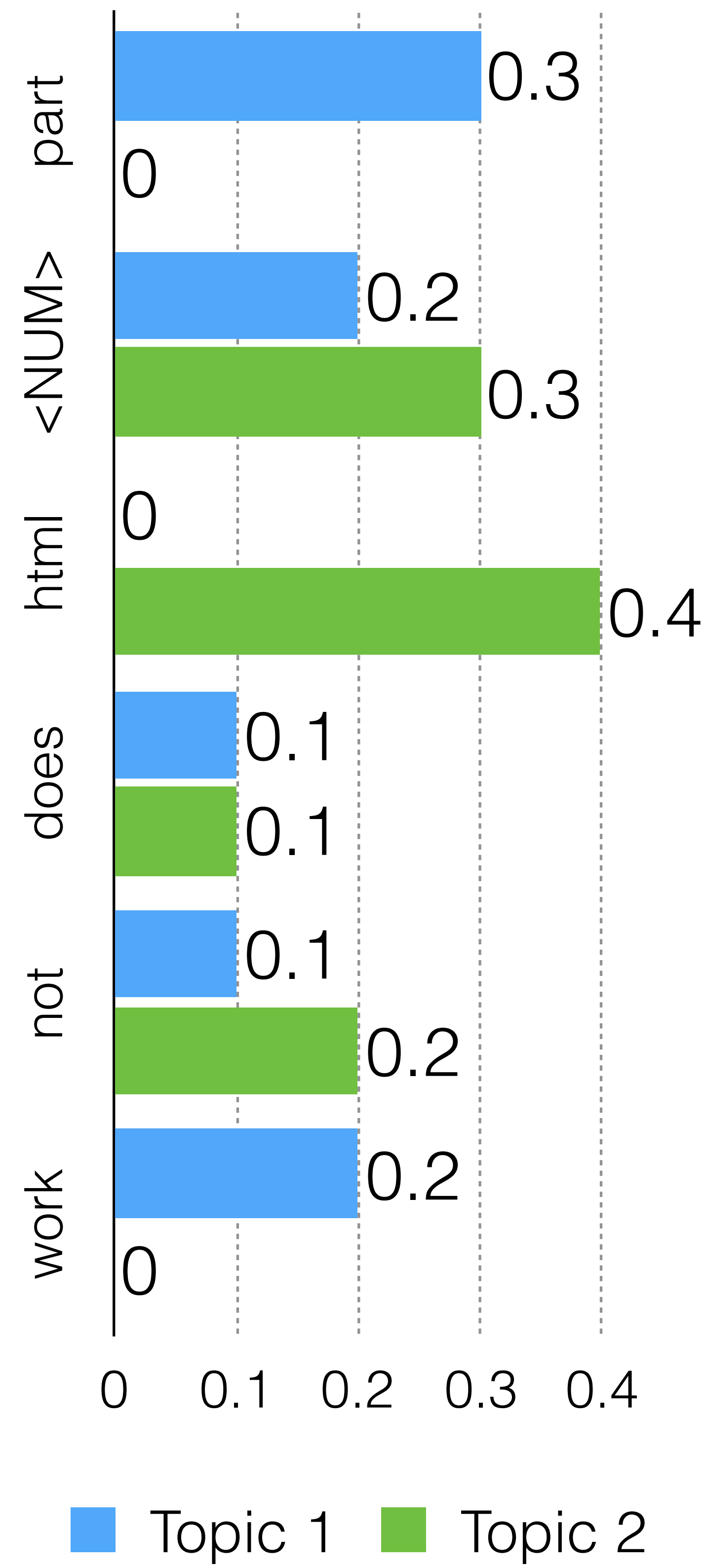


a: (0.3+0.2+0+0.1+0.1+0.2)x0.5
(0+0.3+0.4+0.1+0.2)x0.5

part <NUM> html does not work



Clicker Question!



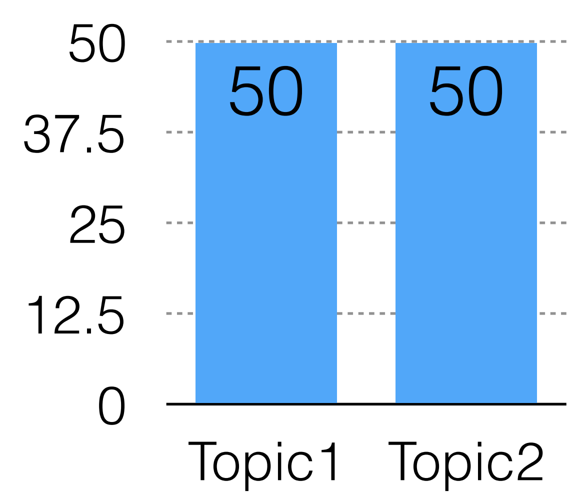
a: $\color{#4da6ff}{(0.3+0.2+0+0.1+0.1+0.2)\text{x}0.5}$

$\color{#6abf40}{(0+0.3+0.4+0.1+0.2)\text{x}0.5}$

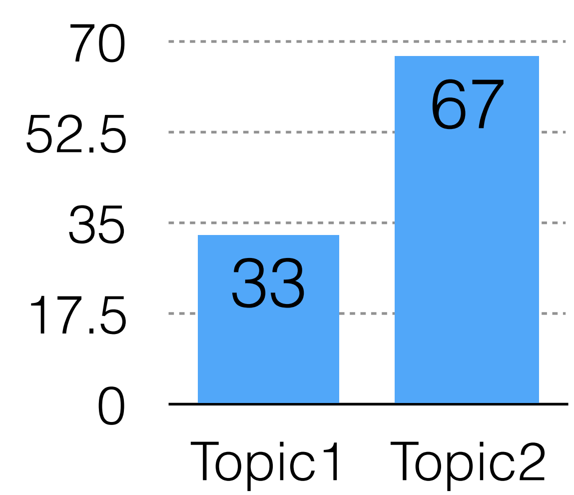
$= 0.45 + 0.5$

$= 0.95$

part <NUM> html does not work

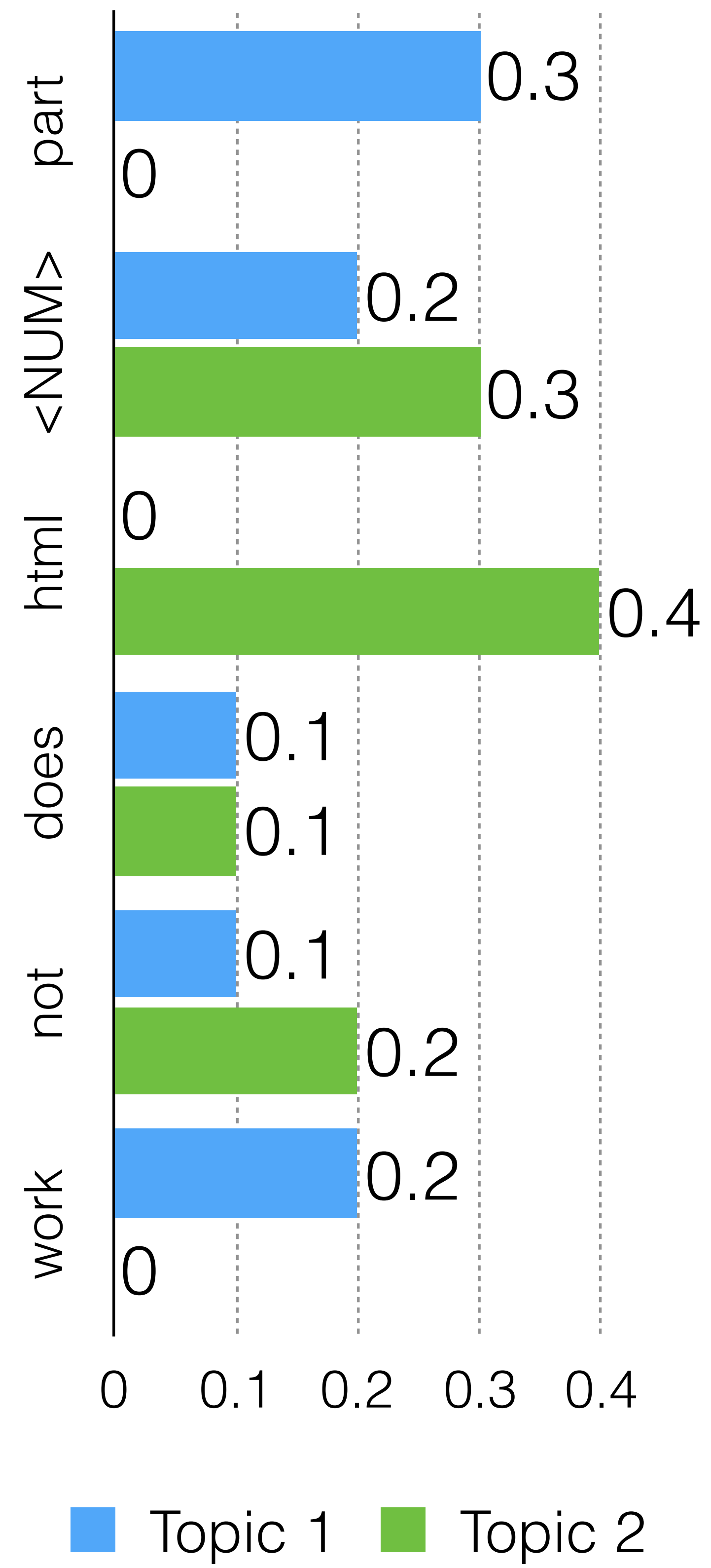


(a)



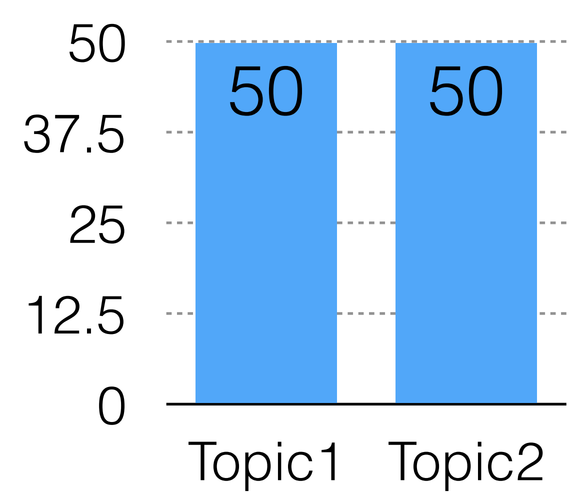
(b)

Clicker Question!

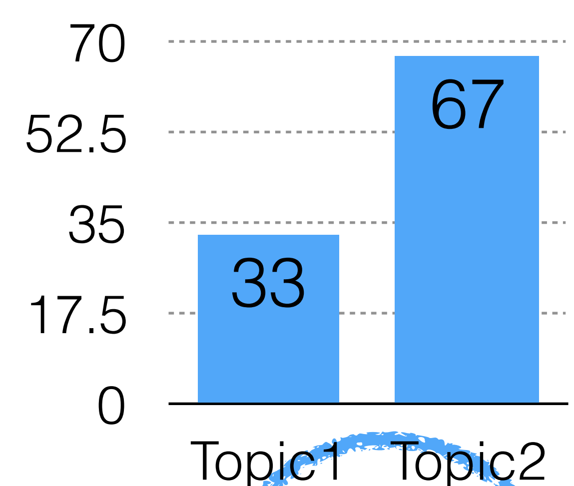


$$b: (0.3+0.2+0+0.1+0.1+0.2)\text{x}0.33$$
$$(0+0.3+0.4+0.1+0.2)\text{x}0.67$$
$$= 0.297 + 0.67$$
$$= 0.967$$

part <NUM> html does not work



(a)



(b)

Topic Models

Topic Models

LDA	
Latent Dirichelet Allocation (latent = not directly observed; Dirichelet = prior follows a Dirichelet distribution)	
Generative Model	
Set parameters using EM or MCMC	

Topic Models

LDA	LSA
Latent Dirichelet Allocation (latent = not directly observed; Dirichelet = prior follows a Dirichelet distribution)	Latent Semantic Analysis
Generative Model	Discriminative Model
Set parameters using EM or MCMC	Set parameters by factorizing the term-document matrix

c Models

	the	cong ress	parli ame	US	UK
doc1	1	1	1	1	0
doc2	1	0	1	0	1
doc3	1	1	0	1	0
doc4	1	0	1	0	1

d1	-0.60	-0.39	0.70	0.00
d2	-0.48	0.50	-0.12	-0.71
d3	-0.43	-0.58	-0.69	0.00
d4	-0.48	0.50	-0.12	0.71

U

3.06	0.00	0.00	0.00	0.00
0.00	1.81	0.00	0.00	0.00
0.00	0.00	0.57	0.00	0.00
0.00	0.00	0.00	0.00	0.00

D

the	cong ress	parlia ment	US	UK
-0.65	-0.34	-0.51	-0.34	-0.31
0.02	-0.54	0.34	-0.54	0.56
-0.42	0.02	0.79	0.02	-0.44
-0.63	0.27	0.00	0.37	0.63
-0.04	0.73	0.00	-0.68	0.04

V

c Models

	the	cong ress	parli ame	US	UK
doc1	1	1	1	1	0
doc2	1	0	1	0	1
doc3	1	1	0	1	0
doc4	1	0	1	0	1

component = "topic"

U

d1	-0.60	-0.39	0.70	0.00
d2	-0.48	0.50	-0.12	-0.71
d3	-0.43	-0.58	-0.69	0.00
d4	-0.48	0.50	-0.12	0.71

D

3.06	0.00	0.00	0.00	0.00
0.00	1.81	0.00	0.00	0.00
0.00	0.00	0.57	0.00	0.00
0.00	0.00	0.00	0.00	0.00

V

the	cong ress	parlia ment	US	UK
-0.65	-0.34	-0.51	-0.34	-0.31
0.02	-0.54	0.34	-0.54	0.56
-0.42	0.02	0.79	0.02	-0.44
-0.63	0.27	0.00	0.37	0.63
-0.04	0.73	0.00	-0.68	0.04

c Models

	the	cong ress	parli ame	US	UK
doc1	1	1	1	1	0
doc2	1	0	1	0	1
doc3	1	1	0	1	0
doc4	1	0	1	0	1

component = "topic" = distribution over words

d1	-0.60	-0.39	0.70	0.00
d2	-0.48	0.50	-0.12	-0.71
d3	-0.43	-0.58	-0.69	0.00
d4	-0.48	0.50	-0.12	0.71

U

3.06	0.00	0.00	0.00	0.00
0.00	1.81	0.00	0.00	0.00
0.00	0.00	0.57	0.00	0.00
0.00	0.00	0.00	0.00	0.00

D

the	cong ress	parlia ment	US	UK
-0.65	-0.34	-0.51	-0.34	-0.31
0.02	-0.54	0.34	-0.54	0.56
-0.42	0.02	0.79	0.02	-0.44
-0.63	0.27	0.00	0.37	0.63
-0.04	0.73	0.00	-0.68	0.04

V

c Models

	the	cong ress	parli ame	US	UK
doc1	1	1	1	1	0
doc2	1	0	1	0	1
doc3	1	1	0	1	0
doc4	1	0	1	0	1

document = distribution over topics

d1	-0.60	-0.39	0.70	0.00
d2	-0.48	0.50	-0.12	-0.71
d3	-0.43	-0.58	-0.69	0.00
d4	-0.48	0.50	-0.12	0.71

U

3.06	0.00	0.00	0.00	0.00
0.00	1.81	0.00	0.00	0.00
0.00	0.00	0.57	0.00	0.00
0.00	0.00	0.00	0.00	0.00

D

the	cong ress	parlia ment	US	UK
-0.65	-0.34	-0.51	-0.34	-0.31
0.02	-0.54	0.34	-0.54	0.56
-0.42	0.02	0.79	0.02	-0.44
-0.63	0.27	0.00	0.37	0.63
-0.04	0.73	0.00	-0.68	0.04

V

Today

- More NLP!
- Ngrams
- Topic Models
- **Word Embeddings**

“Bag of Words” (BOW)

	is	it	a	and	copy	…	markets	below	paste	remorse
doc 1	1	1	2	1	0	…	2	1	0	0
doc 2	3	1	4	0	0	…	1	2	0	1
doc 3	2	1	2	1	1	…	0	0	1	0

Term-Document Matrix

“Bag of Words” (BOW)

	is	it	a	and	copy	…	markets	below	paste	remorse
markets	1	1	2	1	0	…	2	1	0	0
Washington	3	1	4	0	0	…	1	2	0	1
stimulus	2	1	2	1	1	…	0	0	1	0

Word-Context Matrix
(Term-Term*) Matrix

"Bag of Words" (BOW)

	is	it	a	and	copy	…	markets	below	paste	remorse
markets	1	1	2	1	0	…	2	1	0	0
Washington	3	1	4	0	0	…	1	2	0	1
stimulus	2	1	2	1	1	…	0	0	1	0

"Distributional Hypothesis":
the meaning of a word is determined by
the contexts in which it is used

	the	cong ress	par lia	US	UK
market	1	1	1	1	0
Washington	1	0	1	0	1
stimulus	1	1	0	1	0
Brussels	1	0	1	0	1

← Word-Context Matrix

U

market	-0.60	-0.39	0.70	0.00
Washin gton	-0.48	0.50	-0.12	-0.71
stimulus	-0.43	-0.58	-0.69	0.00
Brussel s	-0.48	0.50	-0.12	0.71

D

3.06	0.00	0.00	0.00	0.00
0.00	1.81	0.00	0.00	0.00
0.00	0.00	0.57	0.00	0.00
0.00	0.00	0.00	0.00	0.00

V

the	cong ress	parlia ment	US	UK
-0.65	-0.34	-0.51	-0.34	-0.31
0.02	-0.54	0.34	-0.54	0.56
-0.42	0.02	0.79	0.02	-0.44
-0.63	0.27	0.00	0.37	0.63
-0.04	0.73	0.00	-0.68	0.04

	the	cong ress	par lia	US	UK
market	1	1	1	1	0
Washington	1	0	1	0	1
stimulus	1	1	0	1	0
Brussels	1	0	1	0	1

Word Embeddings

market	-0.60	-0.39	0.70	0.00
Washin gton	-0.48	0.50	-0.12	-0.71
stimulus	-0.43	-0.58	-0.69	0.00
Brussel s	-0.48	0.50	-0.12	0.71

U

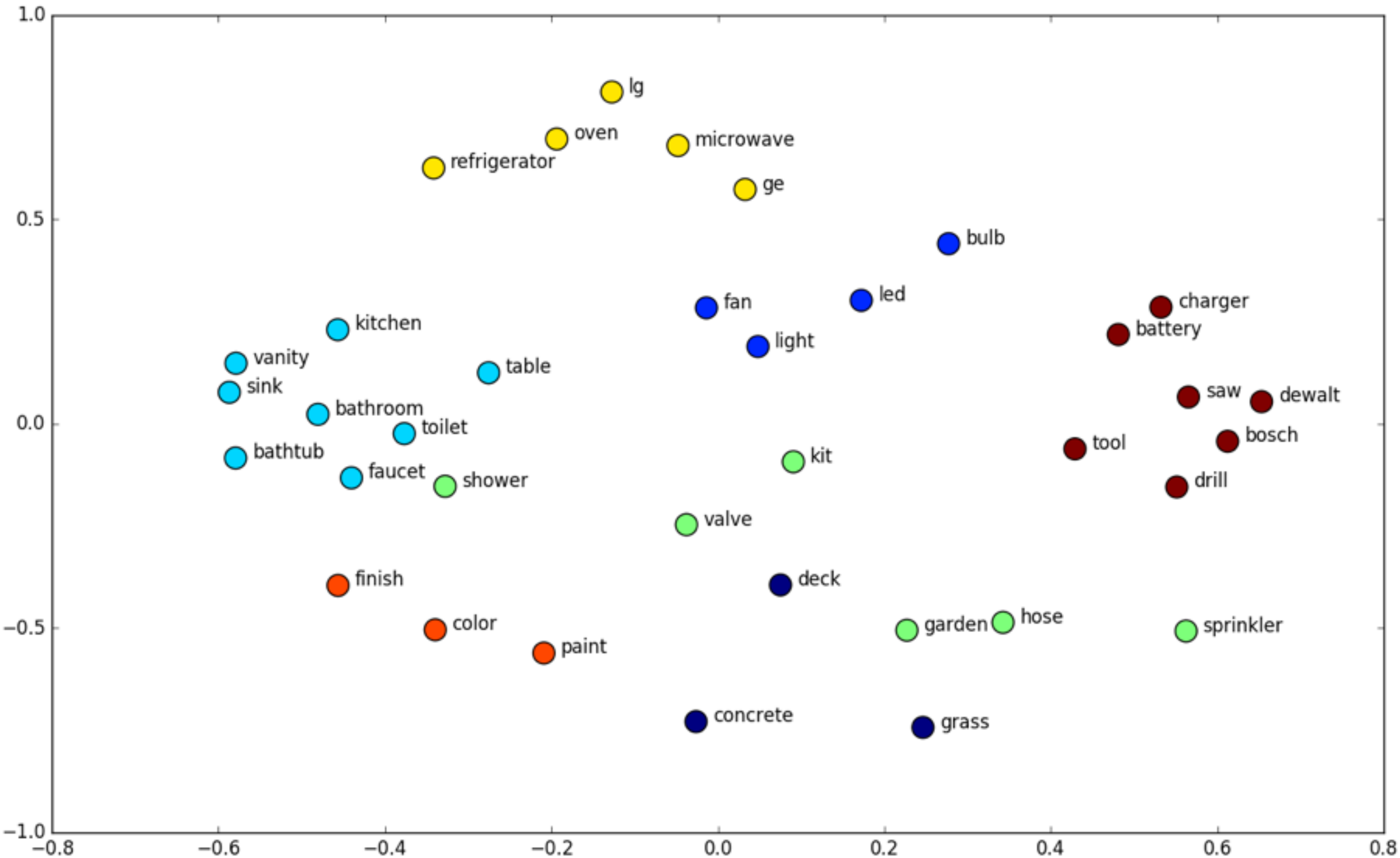
3.06	0.00	0.00	0.00	0.00
0.00	1.81	0.00	0.00	0.00
0.00	0.00	0.57	0.00	0.00
0.00	0.00	0.00	0.00	0.00

D

the	cong ress	parlia ment	US	UK
-0.65	-0.34	-0.51	-0.34	-0.31
0.02	-0.54	0.34	-0.54	0.56
-0.42	0.02	0.79	0.02	-0.44
-0.63	0.27	0.00	0.37	0.63
-0.04	0.73	0.00	-0.68	0.04

V

Word Embeddings



Supervised classification

Lovely mushroomy nose and good length. 1

Gamy, succulent tannins. Lovely. 1

Provence herbs, creamy, lovely. 1

Quite raw finish. A bit rubbery. 0

Good if not dramatic fizz. 0

Rubbery - rather oxidised. 0

y	X							
Label	lovely	good	raw	rubbery	rather	mushroomy	gamy	…
1	1	0	0	0	0	0	0	…
1	0	1	0	0	0	0	0	…
1	1	0	0	0	0	0	0	…
0	0	0	0	1	0	0	0	…

Lovely mushroomy nose and good length. 1

Gamy, succulent tannins. Lovely. 1

Superv sification

Provence herbs, creamy, lovely. 1

Quite raw finish. A bit rubbery. 0

Good if not dramatic fizz. 0

Rubbery - rather oxidised. 0

y	X
Label	
1	D-dimensional vector for lovely
1	D-dimensional vector for good
1	D-dimensional vector for lovely
0	D-dimensional vector for rubbery

Lovely mushroomy nose and good length. 1

Gamy, succulent tannins. Lovely. 1

Provence herbs, creamy, lovely. 1

Quite raw finish. A bit rubbery. 0

Good if not dramatic fizz. 0

Rubbery - rather oxidised. 0

No longer treated as entirely different words

y	X
Label	
1	D-dimensional vector for lovely
1	D-dimensional vector for good
1	D-dimensional vector for lovely
0	D-dimensional vector for rubbery

Supervised classification

Lovely mushroomy nose and good length. 1

Gamy, succulent tannins. Lovely. 1

Provence herbs, creamy, lovely. 1

Quite raw finish. A bit rubbery. 0

Good if not dramatic fizz. 0

Rubbery - rather oxidised. 0

No longer treated as entirely different words

y	X
Label	
1	D-dimensional vector for lovely
1	D-dimensional vector for good
1	D-dimensional vector for lovely
0	D-dimensional vector for rubbery

(often just add up vectors when more than one word)

	the	cong ress	par lia	US	UK
market	1	1	1	1	0
Washington	1	0	1	0	1
stimulus	1	1	0	1	0
Brussels	1	0	1	0	1

Word Embeddings

market	-0.60	-0.39	0.70	0.00
Washin gton	-0.48	0.50	-0.12	-0.71
stimulus	-0.43	-0.58	-0.69	0.00
Brussel s	-0.48	0.50	-0.12	0.71

U

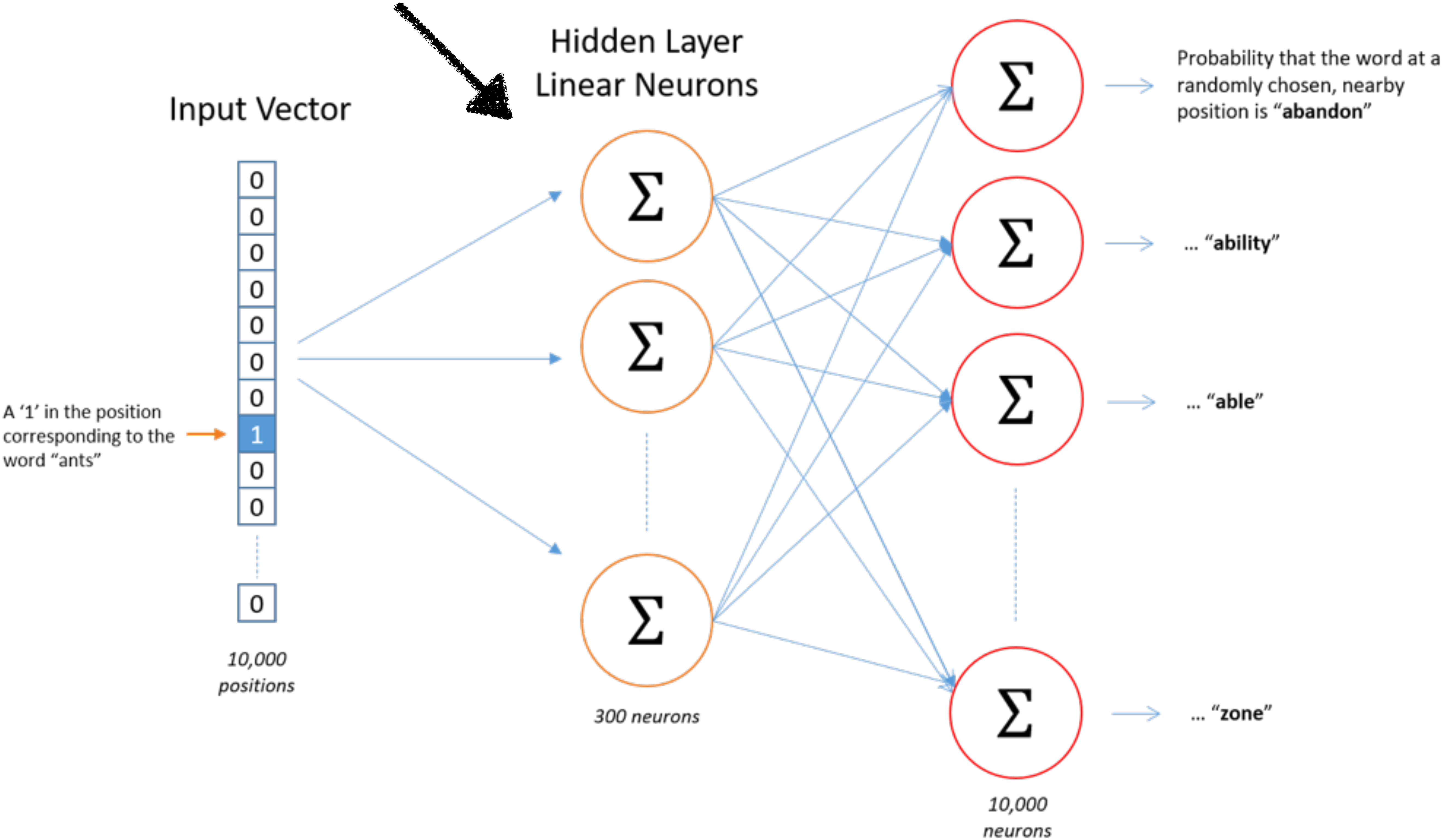
3.06	0.00	0.00	0.00	0.00
0.00	1.81	0.00	0.00	0.00
0.00	0.00	0.57	0.00	0.00
0.00	0.00	0.00	0.00	0.00

D

the	cong ress	parlia ment	US	UK
-0.65	-0.34	-0.51	-0.34	-0.31
0.02	-0.54	0.34	-0.54	0.56
-0.42	0.02	0.79	0.02	-0.44
-0.63	0.27	0.00	0.37	0.63
-0.04	0.73	0.00	-0.68	0.04

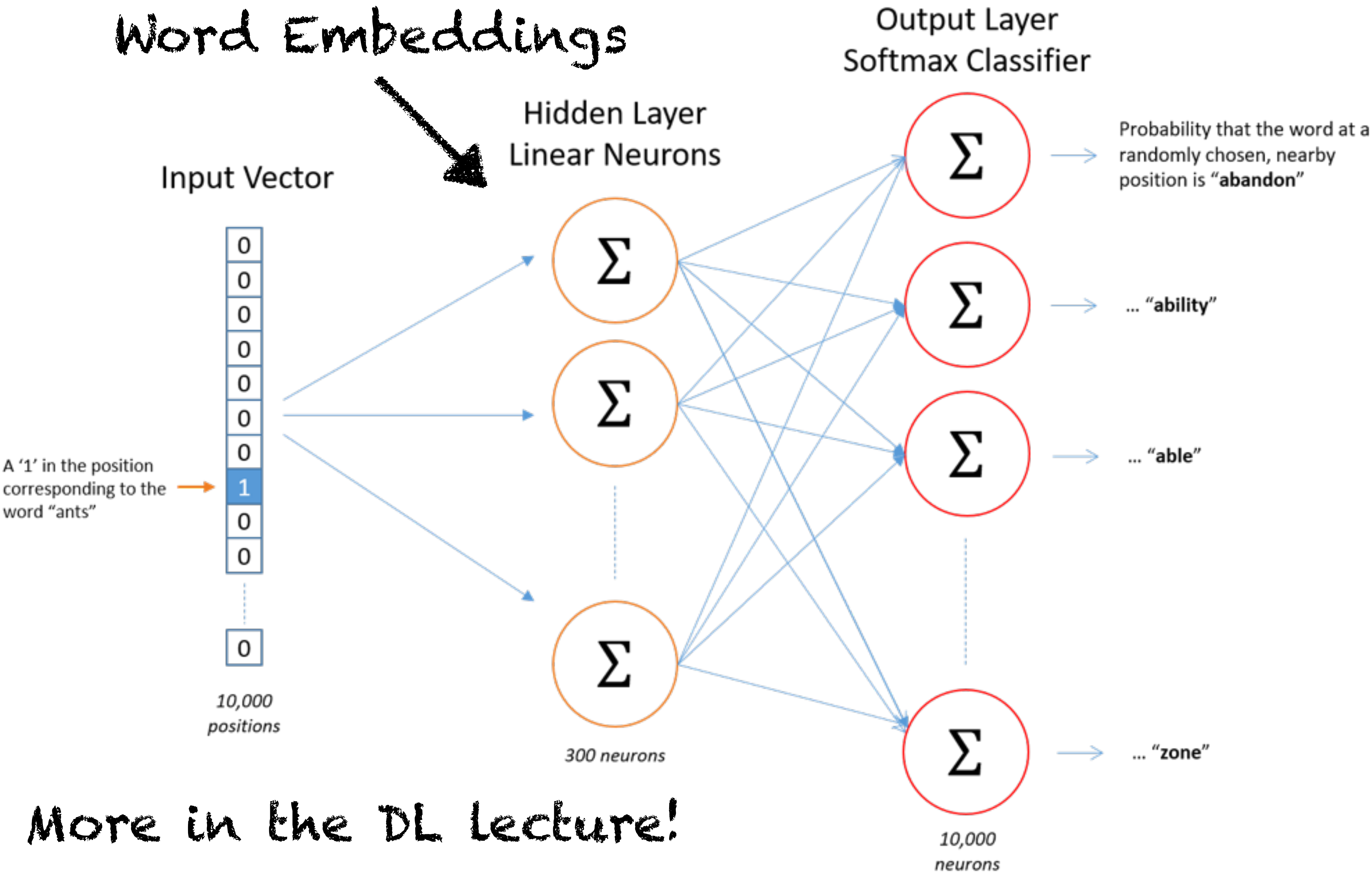
V

Word Embeddings



https://towardsdatascience.com/word2vec-skip-gram-model-part-1-intuition-78614e4d6e0b

Word Embeddings



More in the DL lecture!

https://towardsdatascience.com/word2vec-skip-gram-model-part-1-intuition-78614e4d6e0b k bye