# NLP!!!

April 9, 2019
Data Science CSCI 1951A
Brown University
Instructor: Ellie Pavlick
HTAs: Wennie Zhang, Maulik Dang, Gurnaaz Kaur

# Announcements

- ...

# Today

- NLP!
- Bag-of-words models of documents/words
- Preprocessing
- LSA Topic Models

# What is a “unit” of language

- Words
- Sentences
- Documents
- …EVERYTHING????

# Compositionality

"meaning of the whole is a function of a meaning of the parts and the way in which they are combined"

# Compositionality

Words

# Compositionality

Sentences

Words

# Compositionality

Sentences = f(Words, Syntax)

Words

# Compositionality

Documents = f(Sentences, Discourse)



Sentences = f(Words, Syntax)



Words

Very difficult...
(impossible?)
...to achieve

# Compositionality

Documents = f(Sentences, Discourse)



Sentences = f(Words, Syntax)



Words

# Compositionality

Very difficult...
(impossible?)
...to achieve

Documents = f(Sentences, Discourse)

↑

Sentences = f(Words, Syntax)

↑

Words

horse shoes ≈ alligator shoes?

# “Bag of Words” (BOW)

Documents = f(Words)

Words

Syntax/order doesn't matter, context doesn't matter...

# “Bag of Words” (BOW)

- Foundation of most of modern NLP

# “Bag of Words” (BOW)

- Foundation of most of modern NLP
- Information Retrieval/Search

# “Bag of Words” (BOW)

- Foundation of most of modern NLP
- Information Retrieval/Search
- Clustering/Recommendation

# “Bag of Words” (BOW)

- Foundation of most of modern NLP
- Information Retrieval/Search
- Clustering/Recommendation
- As input to most ML models

# “Bag of Words” (BOW)

- Foundation of most of modern NLP
- Information Retrieval/Search
- Clustering/Recommendation
- As input to most ML models
- Changing a bit for sentences, but not for documents (yet)

# “Bag of Words” (BOW)

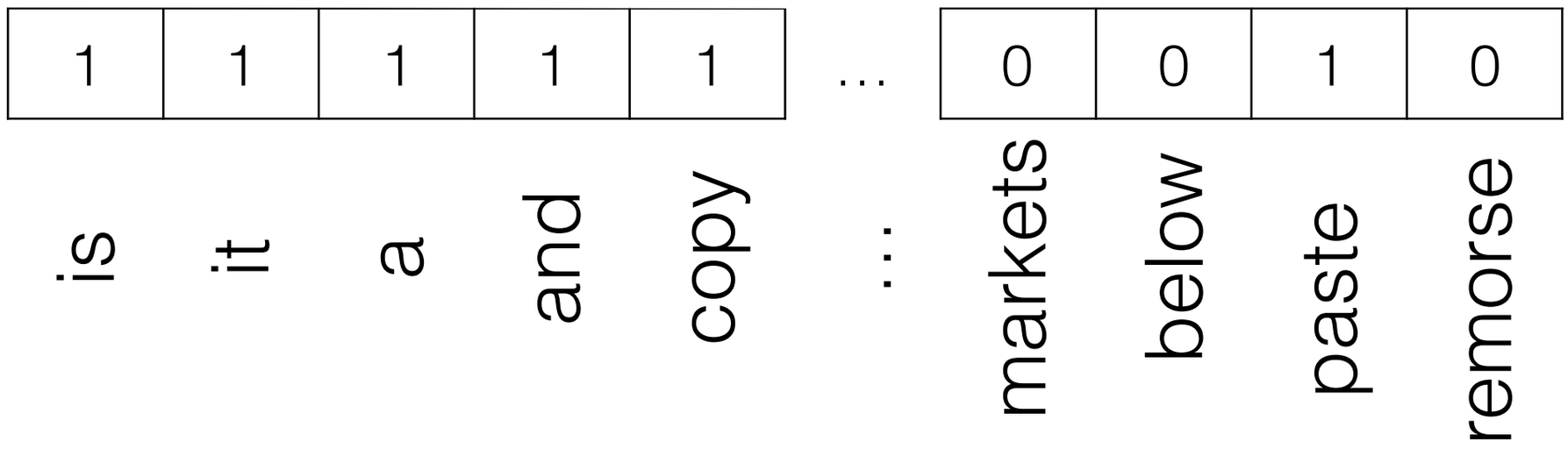
Is it ok to copy and paste the data into javascript, or is there a filereader that can open a local file?

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

When I try to display dots from part 2 on my mac (tried chrome, firefox, and safari), nothing is displayed (and the elements do not appear in the html).

# “Bag of Words” (BOW)

Is it ok to copy and paste the data into javascript, or is there a filereader that can open a local file?



# “Bag of Words” (BOW)

Is it ok to copy and paste the data into javascript, or is there a filereader that can open a local file?

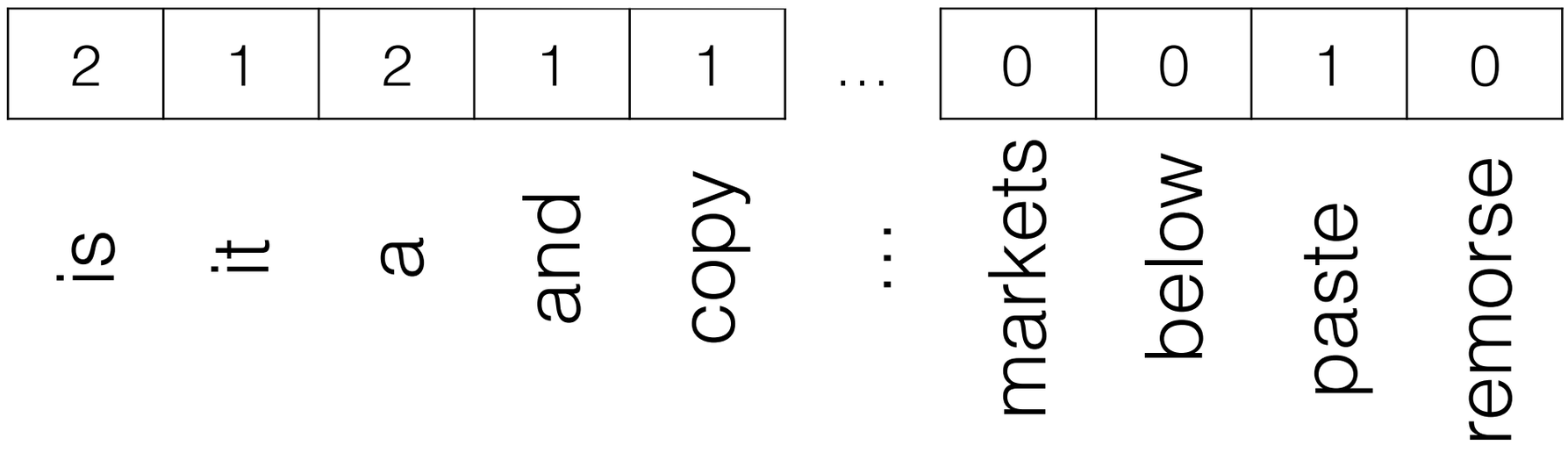
“one hot”

| is | it | a | and | copy | … | markets | below | paste | remorse |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | … | 0 | 0 | 1 | 0 |

# “Bag of Words” (BOW)

Is it ok to copy and paste the data into javascript, or is there a filereader that can open a local file?

counts/frequencies



# “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 |

# “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 |
|---|---|---|---|---|---|---|---|---|---|---|
| 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 |

**How similar are document 1 and document 2?**

# Similarity Metrics

# Similarity Metrics

- Edit Distance: Minimal number of edits (inserts, deletes, substitutions) needed to transform string 1 into string 2.

# Similarity Metrics

- Edit Distance: Minimal number of edits (inserts, deletes, substitutions) needed to transform string 1 into string 2.

Thoughts?

# Similarity Metrics

- Edit Distance: Minimal number of edits (inserts, deletes, substitutions) needed to transform string 1 into string 2.

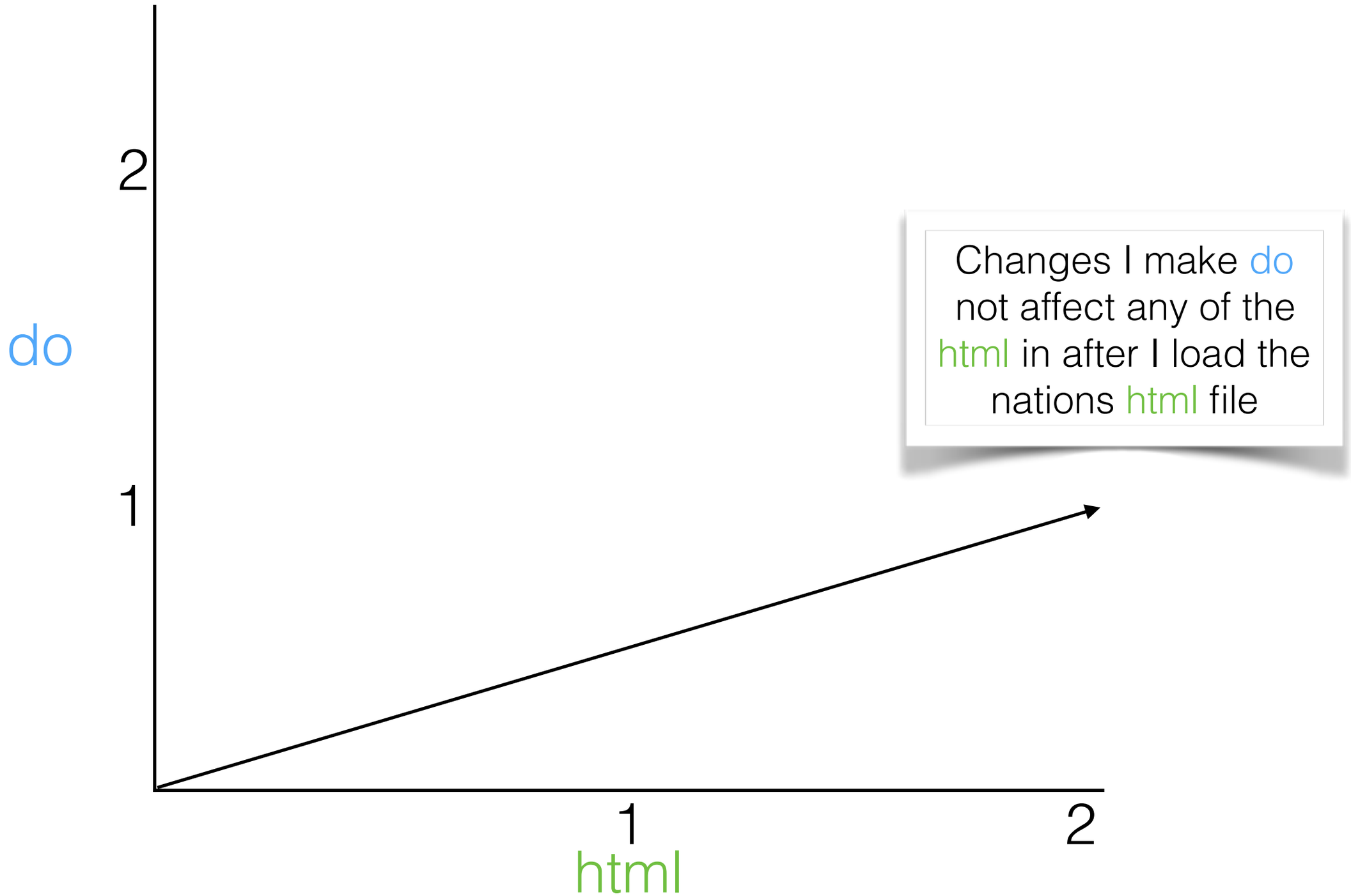
- Jaccard Similarity: words in common / total words

# Clicker Question!

# Similarity Metrics

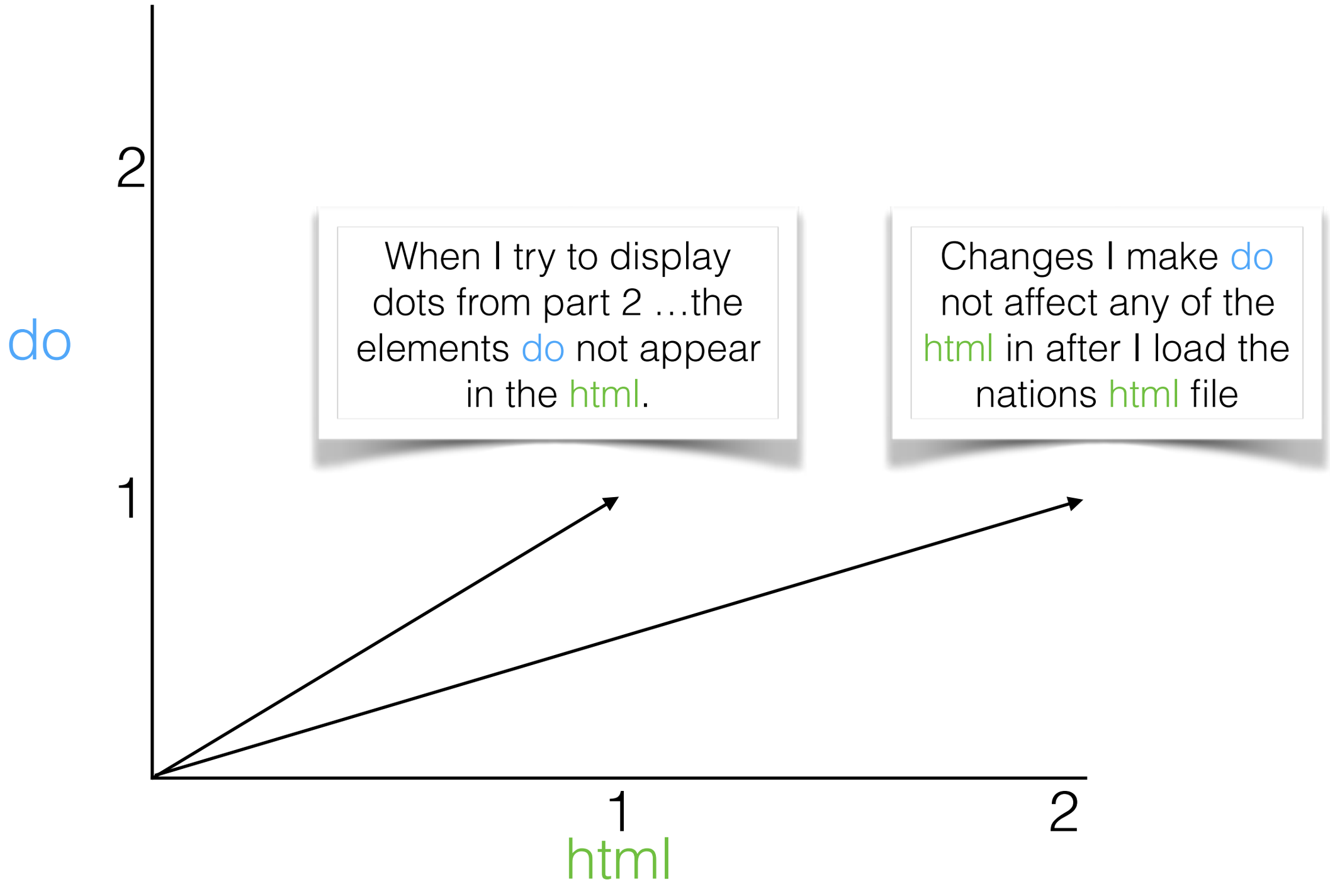
- Edit Distance: Minimal number of edits (inserts, deletes, substitutions) needed to transform string 1 into string 2.
- Jaccard Similarity: words in common / total words
- Cosine Similarity: by far the most popular metric

# Cosine Similarity



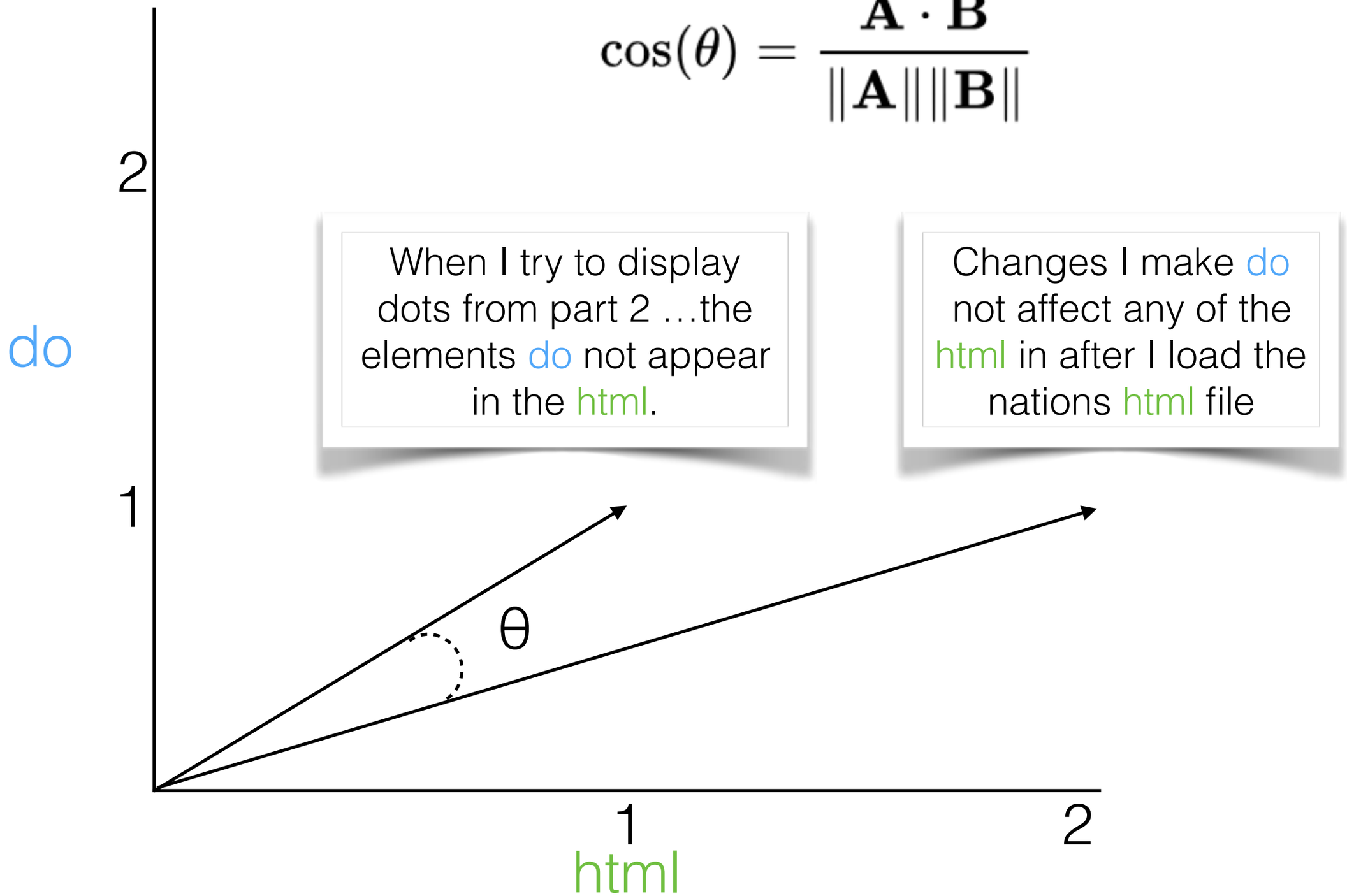

Changes I make do not affect any of the html in after I load the nations html file

# Cosine Similarity



# Cosine Similarity

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

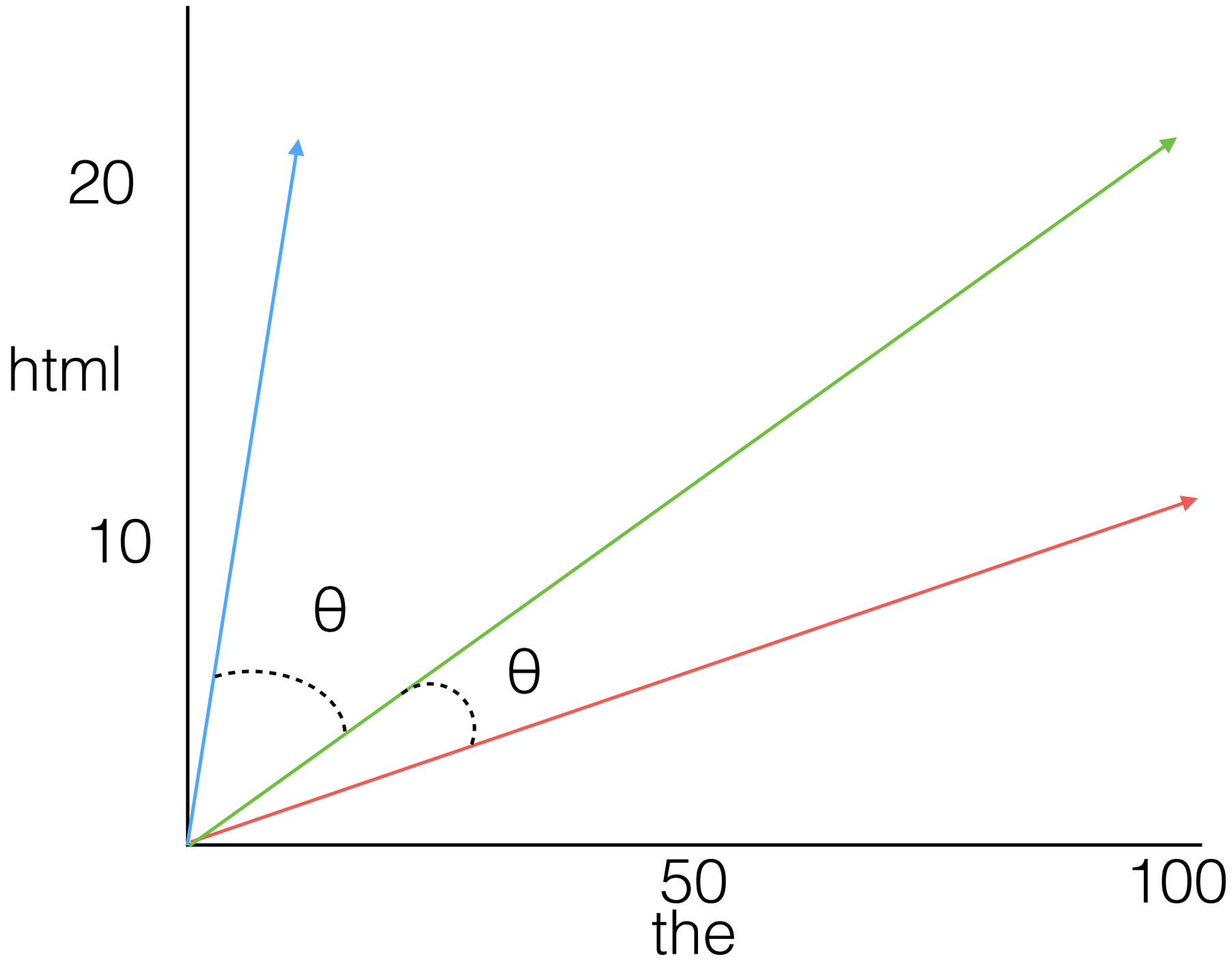



When I try to display dots from part 2 …the elements do not appear in the html.

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

# Clicker Question!

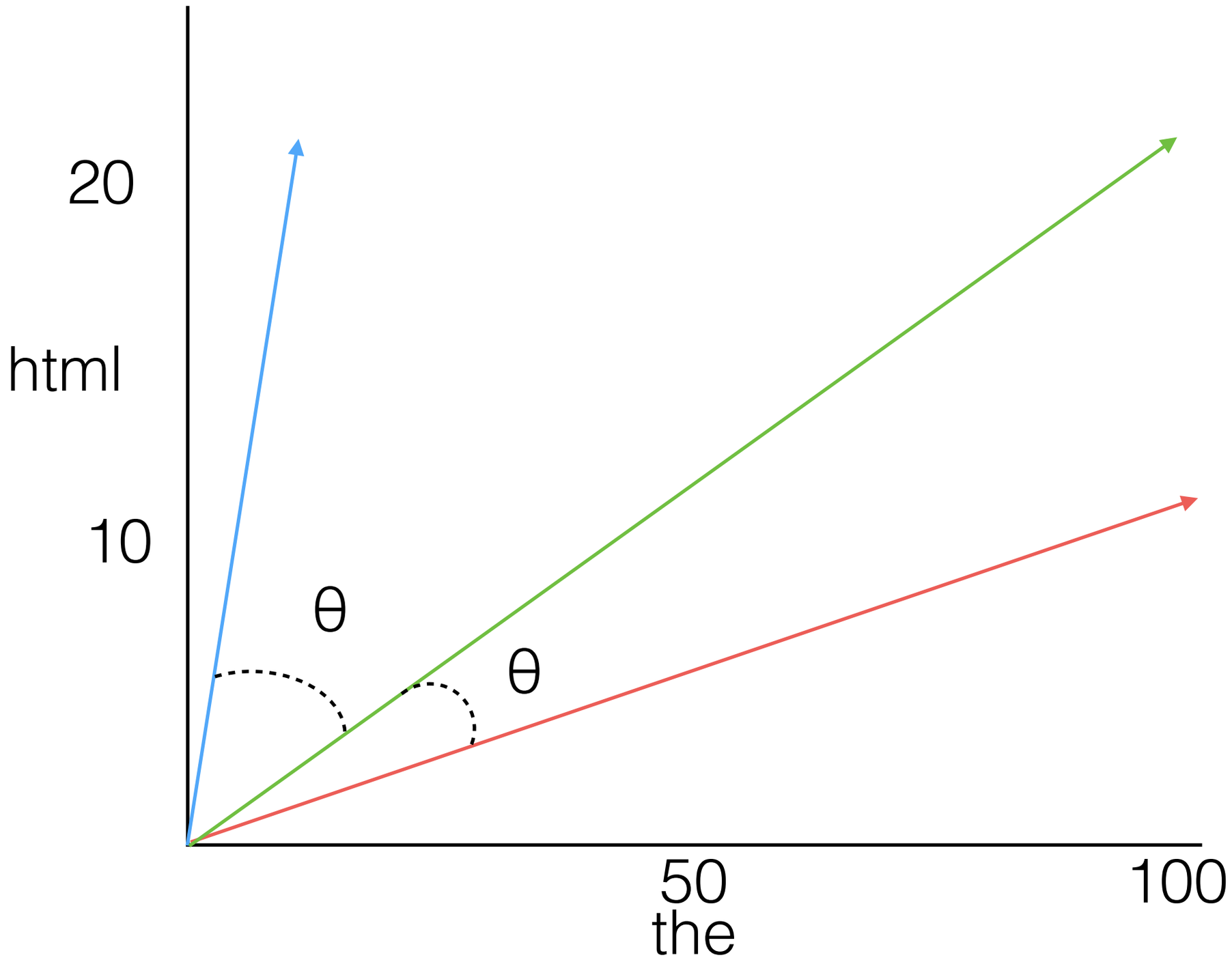
# Frequency Biases



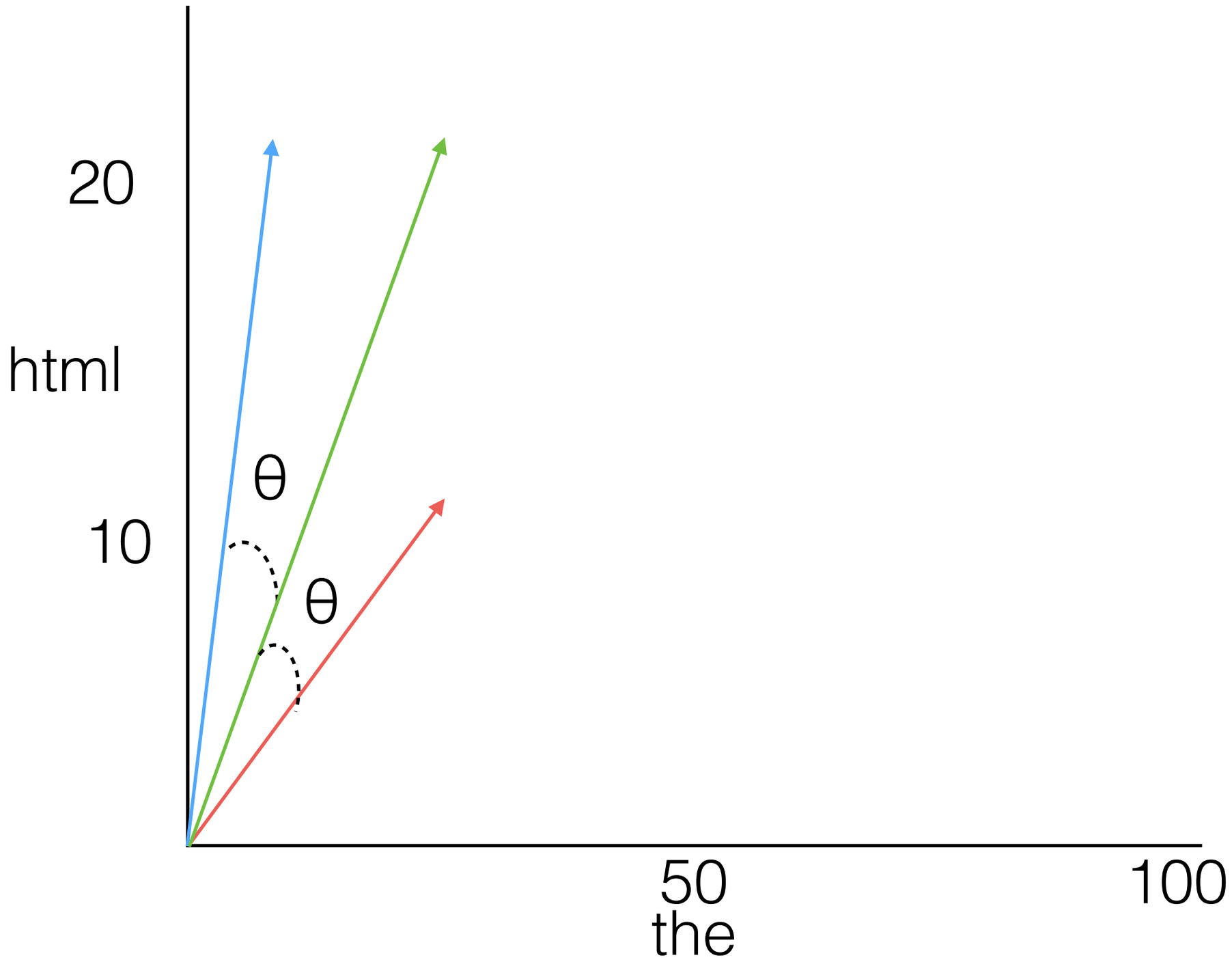
# Tf-Idf

- Term-Frequency Inverse-Document-Frequency
- Goal is to down-weight words which occur often
- tf-idf(w,d) = (# times w appears in d) / (# of times w appears across all documents)

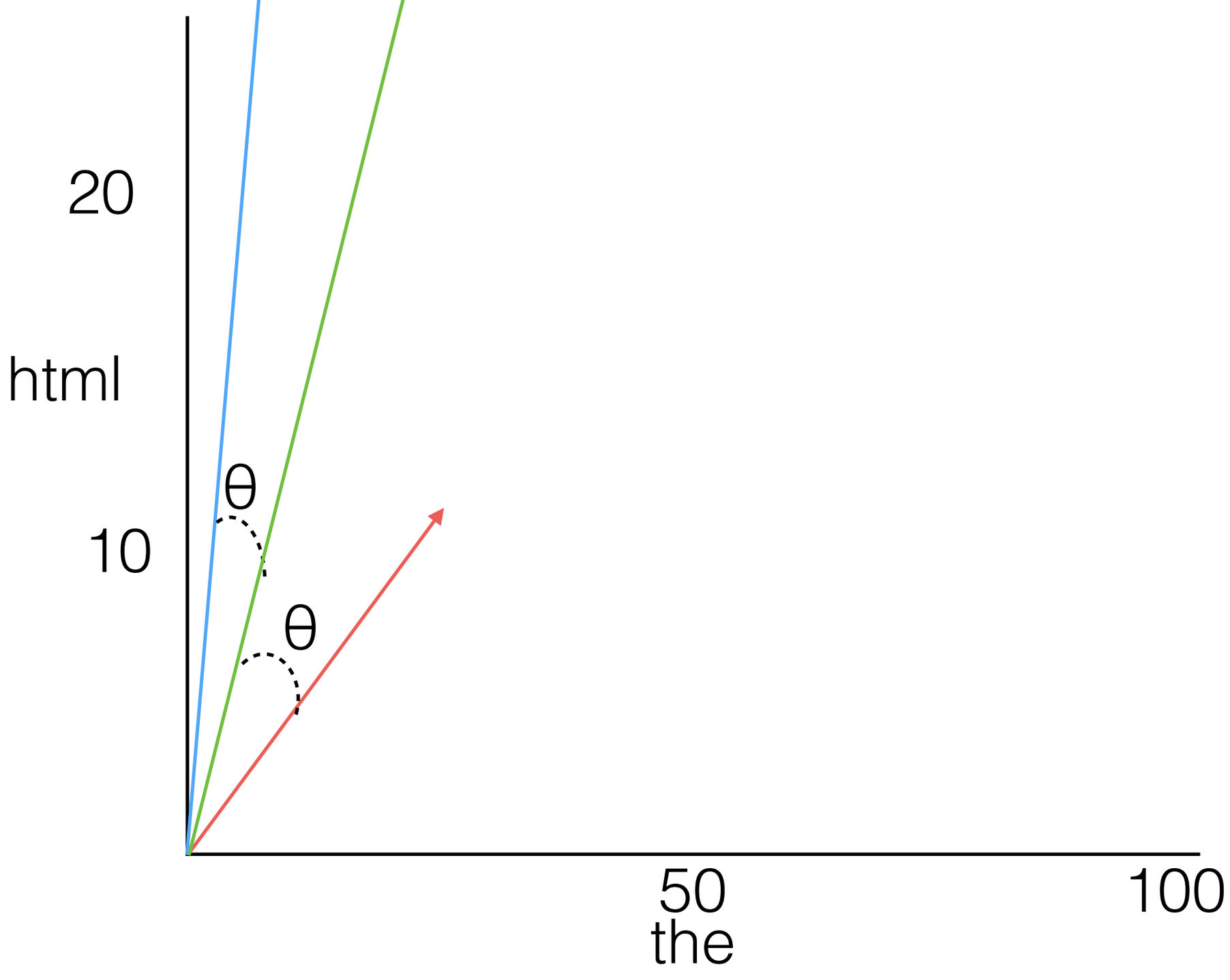
# Frequency Biases



# Frequency Biases



# Frequency Biases



# Clicker Question!

# Linguistic Preprocessing

# Linguistic Preprocessing

Language is ambiguous

# Linguistic Preprocessing

Language is ambiguous

sorry, no, i have class.

# Linguistic Preprocessing

Language is ambiguous

sorry, no, i have **class**.

# Linguistic Preprocessing

Language is ambiguous

want to go get a coffee?

sorry, no, i have **class**.

# Linguistic Preprocessing

Language is ambiguous

want to go eat junk food and light random things on fire?

sorry, no, i have **class**.

# Linguistic Preprocessing

Language is ambiguous

but also redundant

want to go eat junk food and light random things on fire?

sorry, no, i have **dignity and taste**.

# Linguistic Preprocessing

Constant Tradeoff

# Linguistic Preprocessing

Constant Tradeoff

Collapse!

Try to treat more words as though they are the same

# Linguistic Preprocessing

Constant Tradeoff

Collapse!

Try to treat more words as though they are the same

Differentiate!

Try to preserve as much differences/ nuance as possible

# Linguistic Preprocessing

Constant Tradeoff

Collapse!

Try to treat more words as though they are the same

Differentiate!

Try to preserve as much differences/ nuance as possible

normalization, stemming

tagging, collocations

# Linguistic Preprocessing

# Linguistic Preprocessing

I am trying to display dots from Part 2 on my mac (tried Chrome, Firefox , and Safari), but nothing is displayed (and the elements do not appear in the html).

I am trying to display dots from Part 2 on my mac (tried Chrome, Firefox , and Safari), but nothing is displayed (and the elements do not appear in the html).

- Tokenization (Phrasal Collocations/Morphological Analysis?)

I am trying to display dots from Part 2 on my mac ( tried Chrome , Firefox , and Safari ) , but nothing is displayed ( and the elements do not appear in the html ) .

- Tokenization (Phrasal Collocations/Morphological Analysis?)

I am trying to display dots from Part 2 on my mac ( tried Chrome , Firefox , and Safari ) , but nothing is displayed ( and the elements do not appear in the html ) .

- Tokenization (Phrasal Collocations/Morphological Analysis?)

日文章魚怎麼說?
“How to say octopus in Japanese?”

I am trying to display dots from Part 2 on my mac ( tried Chrome , Firefox , and Safari ) , but nothing is displayed ( and the elements do not appear in the html ) .

- Tokenization (Phrasal Collocations/Morphological Analysis?)

日文章魚怎麼說?
“How to say octopus in Japanese?”

| 日文 | 章魚 | 怎麼 | 說 | ? |
|---|---|---|---|---|
| Japanese | octopus | how | say | ? |

I am trying to display dots from Part 2 on my mac tried Chrome Firefox and Safari but nothing is displayed and the elements do not appear in the html

- Tokenization (Phrasal Collocations/Morphological Analysis?)

- Punctuation — “okay…” vs. “okay!”

i am trying to display dots from part 2 on my mac tried chrome firefox and safari but nothing is displayed and the elements do not appear in the html

- Tokenization (Phrasal Collocations/Morphological Analysis?)

- Punctuation — “okay…” vs. “okay!”

- Normalization — “Trump” vs. “trump”

i be try to display dot from part 2 on my mac try chrome  firefox and safari but nothing be display and the element do not appear in the html

- Tokenization (Phrasal Collocations/Morphological Analysis?)

- Punctuation — “okay…” vs. “okay!”

- Normalization — “Trump” vs. “trump”

i be try to display dot from part <NUM> on my mac try chrome firefox and safari but nothing be display and the element do not appear in the html

- Tokenization (Phrasal Collocations/Morphological Analysis?)

- Punctuation — “okay…” vs. “okay!”

- Normalization — “Trump” vs. “trump”

try display dot part <NUM> mac try chrome firefox safari nothing display element not appear html

- Tokenization (Phrasal Collocations/Morphological Analysis?)
- Punctuation — “okay…” vs. “okay!”
- Normalization — “Trump” vs. “trump”
- Stop words — “pb and jelly” vs. “pb or jelly”

try_VB display_VB dot_NN part_NN <NUM>_NUM mac_NNP
try_VB chrome_NNP firefox_NNP safari_NNP nothing_DT
display_VB element_NNP not_RB appear_VB html_NN

- Tokenization (Phrasal Collocations/Morphological Analysis?)

- Punctuation — “okay…” vs. “okay!”

- Normalization — “Trump” vs. “trump”

- Stop words — “pb and jelly” vs. “pb or jelly”

- Tagging — “fish fish fish fish fish”

try_VB display_VB dot_NN part_NN <NUM>_NUM mac_NNP
try_VB chrome_NNP <OOV> <OOV> nothing_DT display_VB
element_NNP not_RB appear_VB html_NN

- Tokenization (Phrasal Collocations/Morphological Analysis?)

- Punctuation — “okay…” vs. “okay!”

- Normalization — “Trump” vs. “trump”

- Stop words — “pb and jelly” vs. “pb or jelly”

- Tagging — “fish fish fish fish fish”

- Remove out-of-vocabulary (OOV)

# “Bag of Words” (BOW)

Is it ok to copy and paste the data into javascript, or is there a filereader that can open a local file?

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

When I try to display dots from part 2 on my mac (tried chrome, firefox, and safari), nothing is displayed (and the elements do not appear in the html).

# 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

# 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

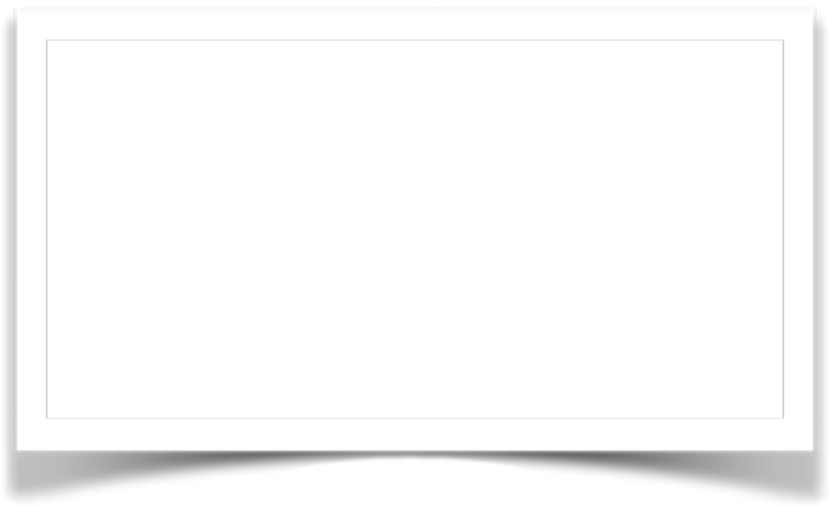
# Topic Models

Where do documents come from?
"The generative story"

# Topic Models

Where do documents come from?
"The generative story"

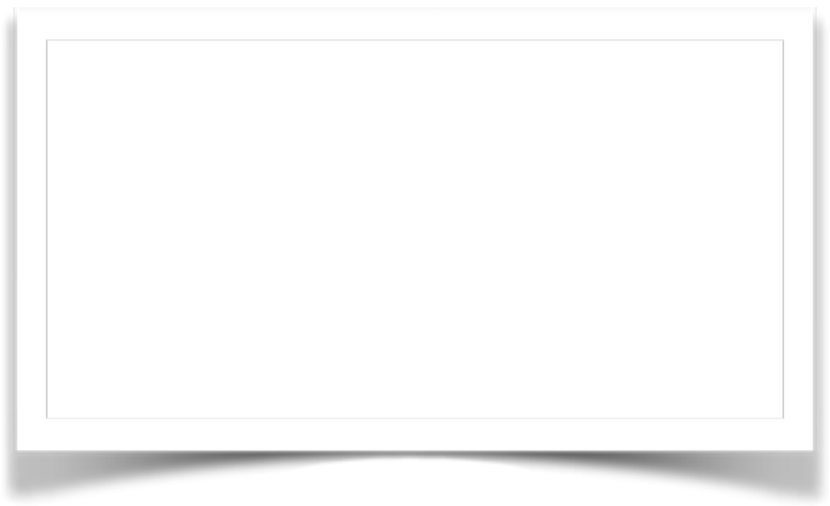
instructions: stencil, instructions, part, step, rubric, handin…
UI: html, javascript, debug, display, elements…
systems: mac, windows, linux, chrome, firefox, os…
fillers: I, you, when, the, and, a



# 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

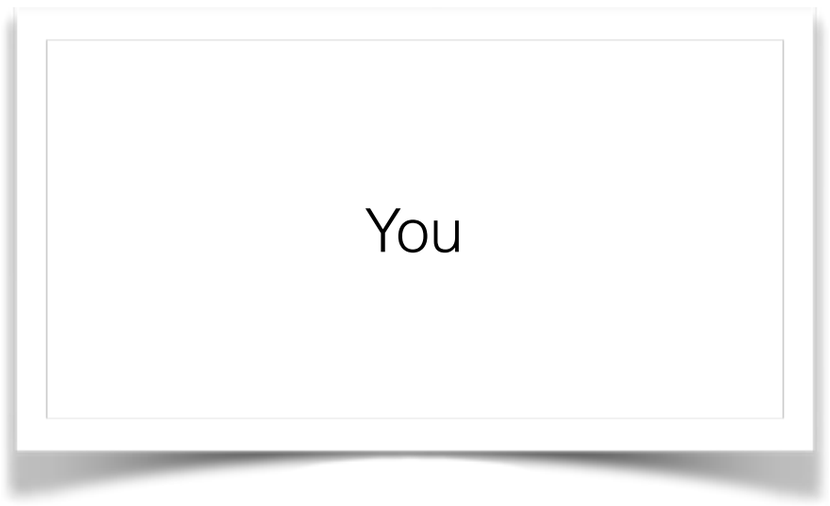


1. Sample a topic

# 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

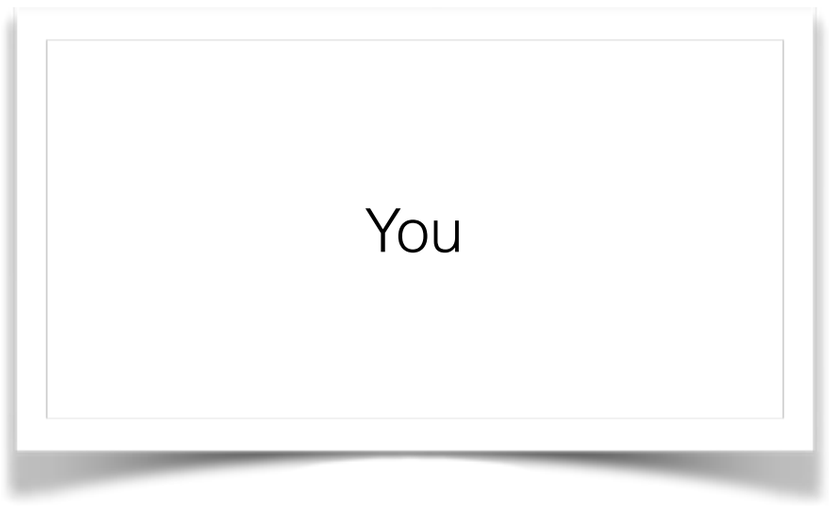



2. Sample a word from that topic

# 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

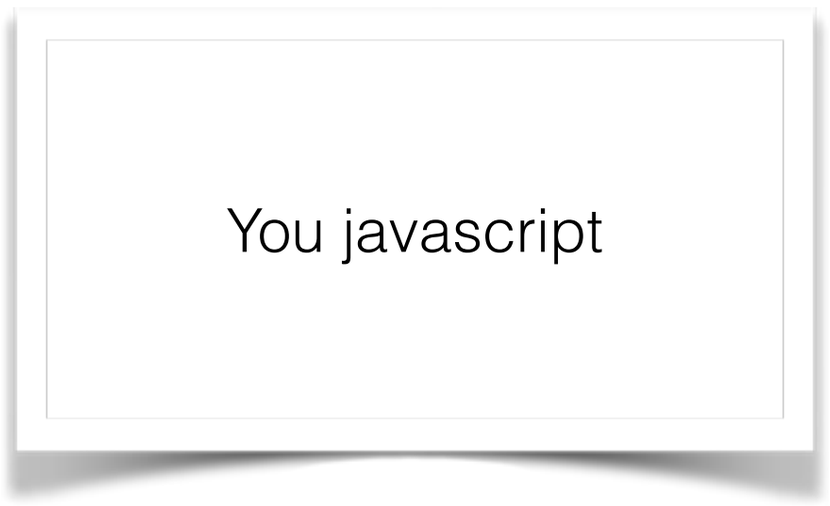



1. Sample a topic

# 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

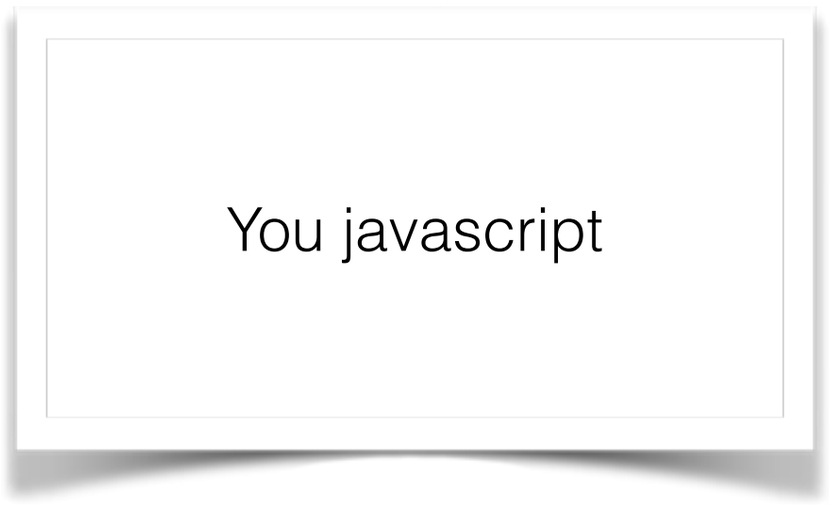



2. Sample a word from that topic

# 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

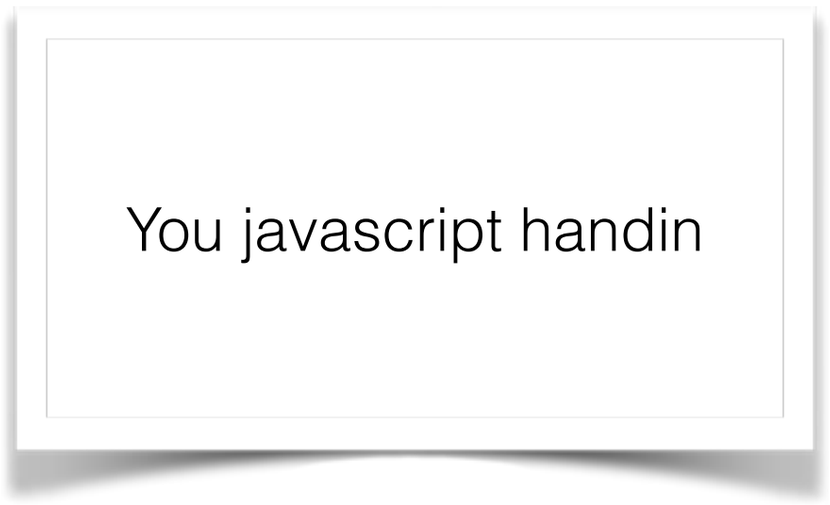



1. Sample a topic

# 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




2. Sample a word from that topic

# Topic Models

"Latent Semantic Analysis" (LSA)

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

# Topic Models

"Latent Semantic Analysis" (LSA)

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

"latent" variable (not observed)

# Topic Models

"Latent Semantic Analysis" (LSA)

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

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

# Topic Models

"Latent Semantic Analysis" (LSA)

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

documents are a distribution over topics

# Topic Models

"Latent Semantic Analysis" (LSA)

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

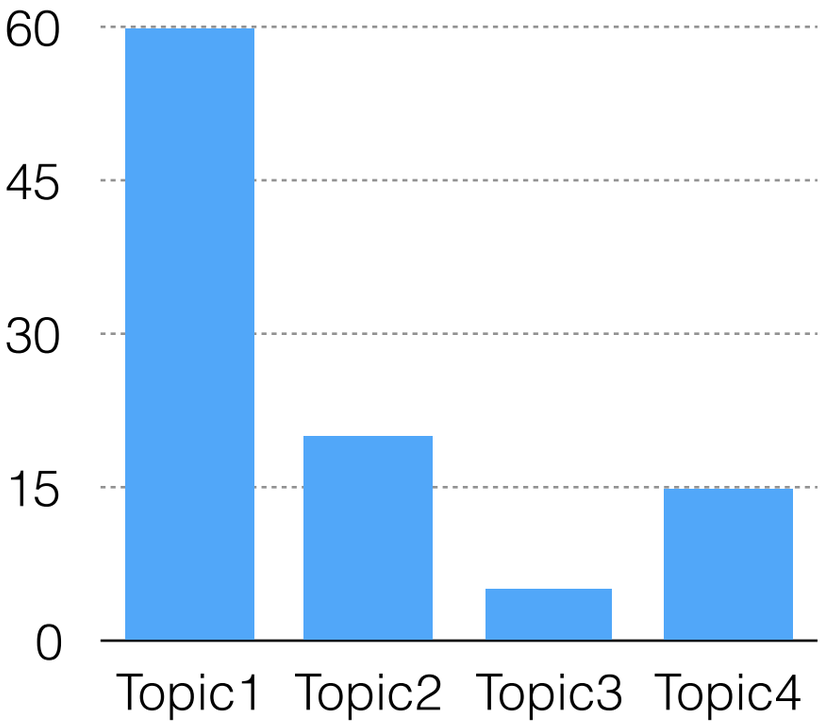
set parameters to maximize probability of observations

# Topic Models

part 2 html does not work

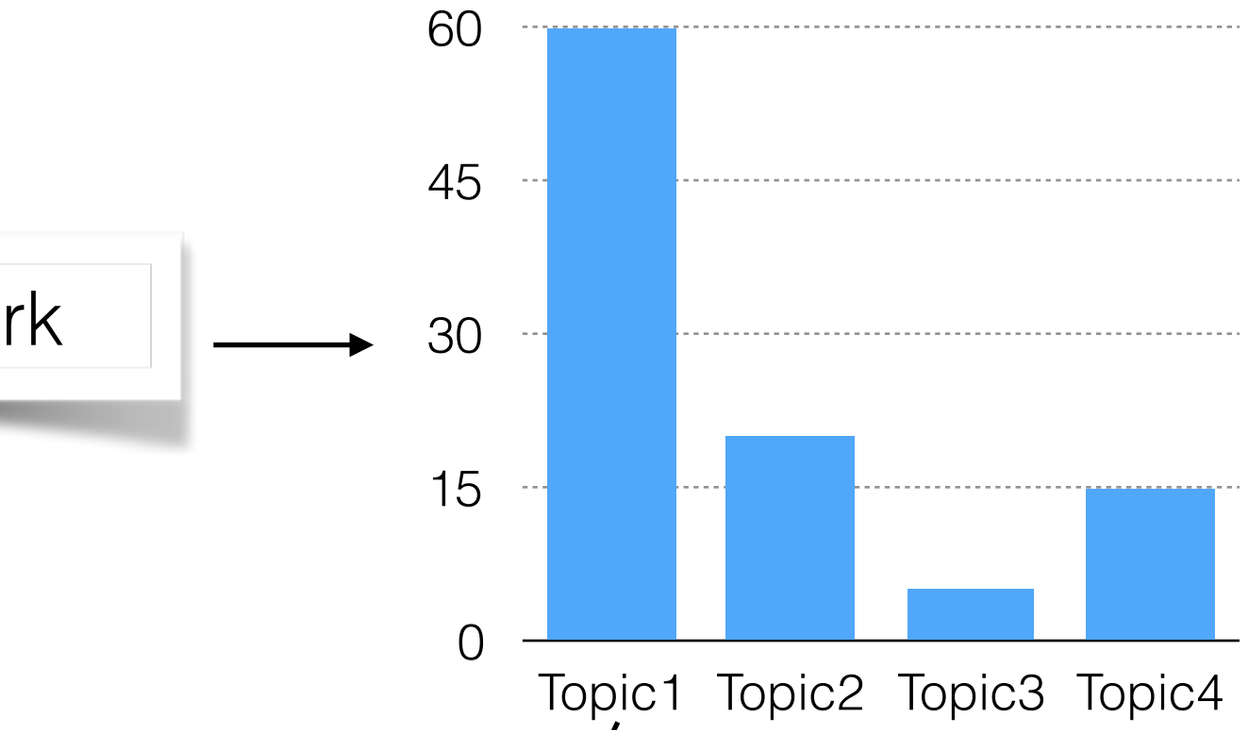
# Topic Models

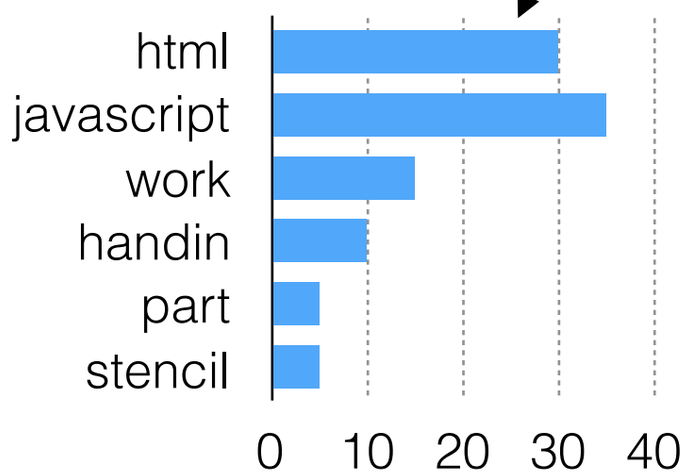
part 2 html does not work

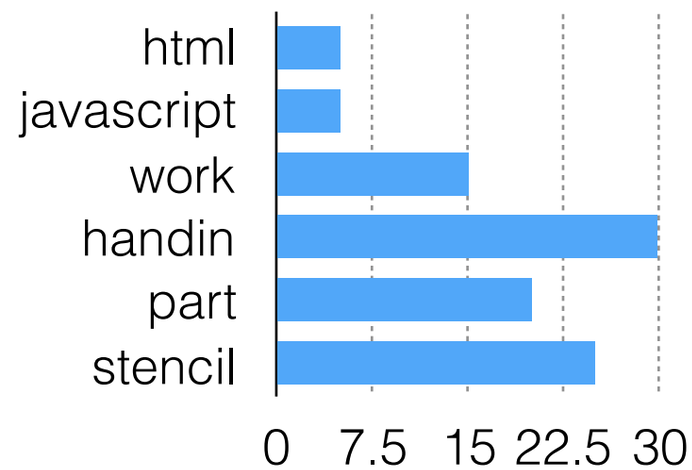


# Topic Models

part 2 html does not work

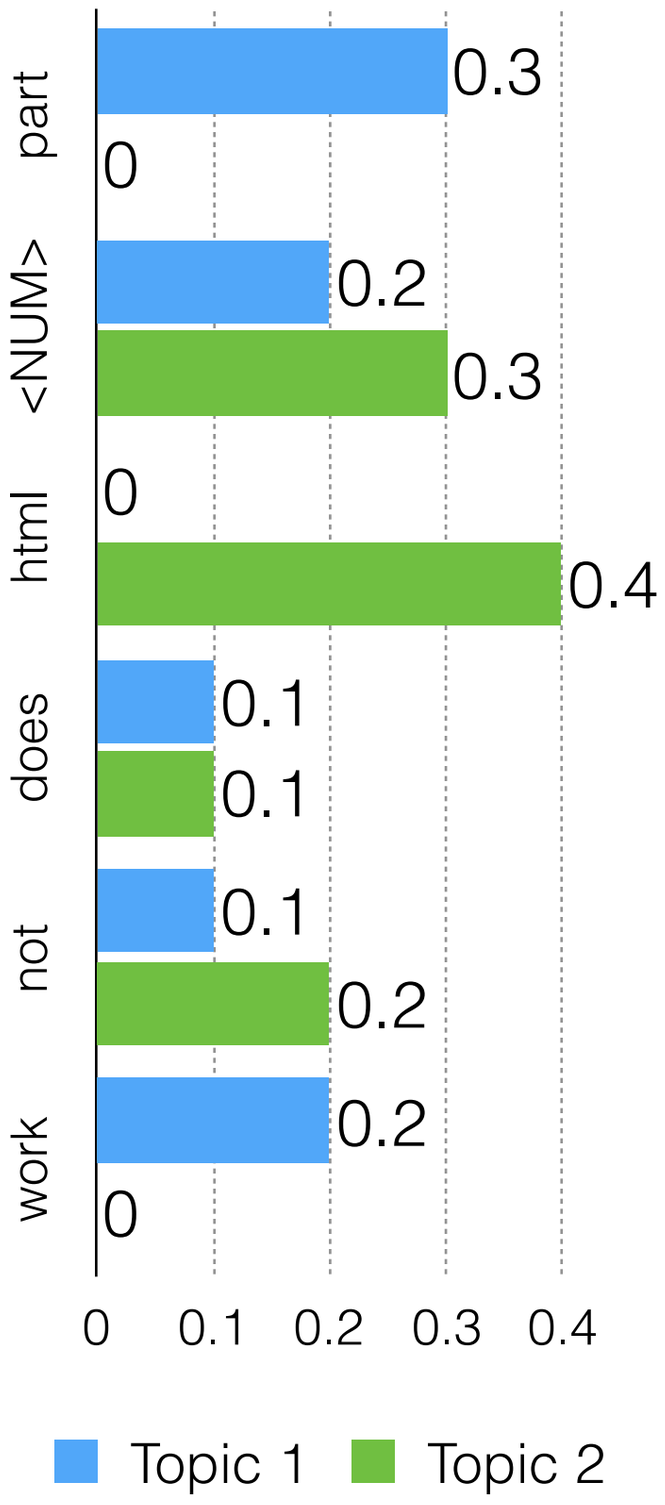


# Clicker Question!

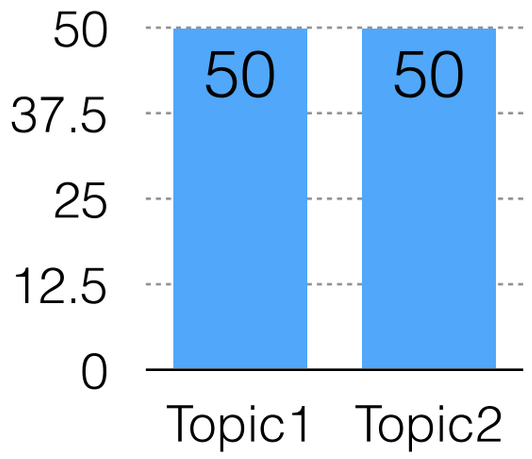
# Clicker Question!



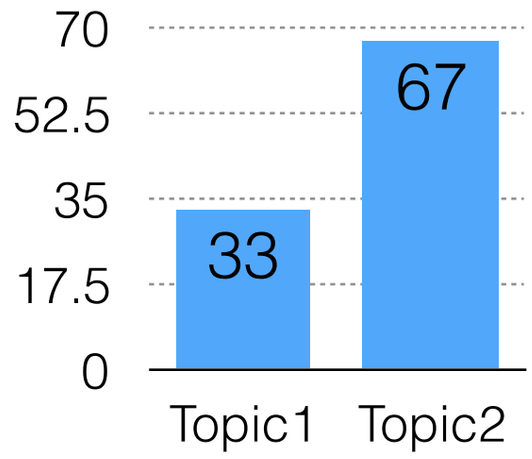

Which is the best parameter setting for the observed data?

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

part <NUM> html does not work

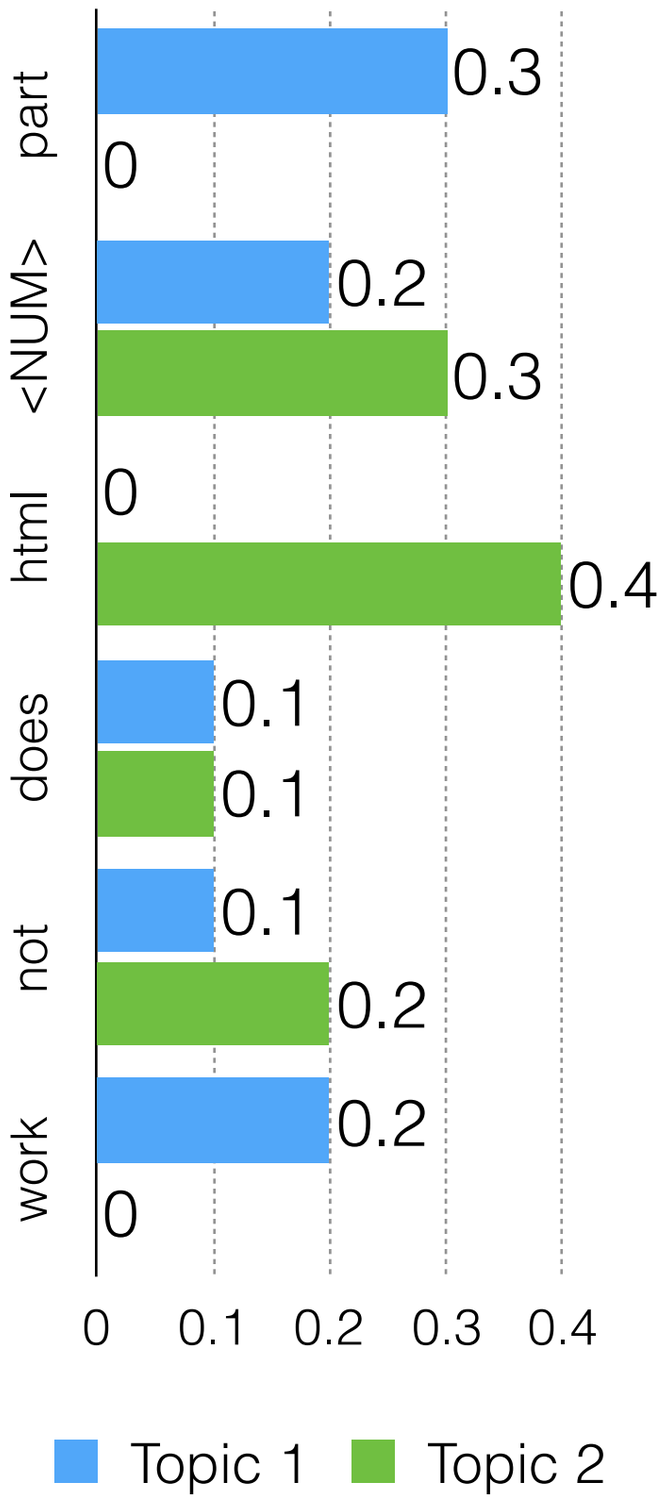



(a)



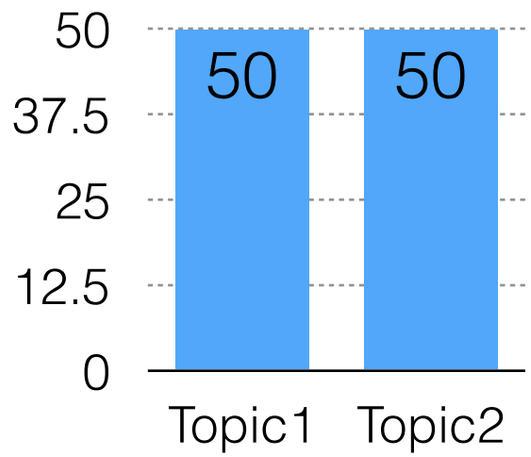

(b)

# Clicker Question!

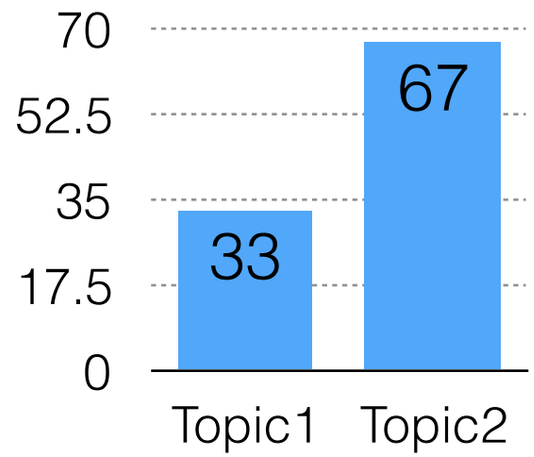



a: (0.3+0.2+0+0.1+0.1+0.2)x0.5

part <NUM> html does not work

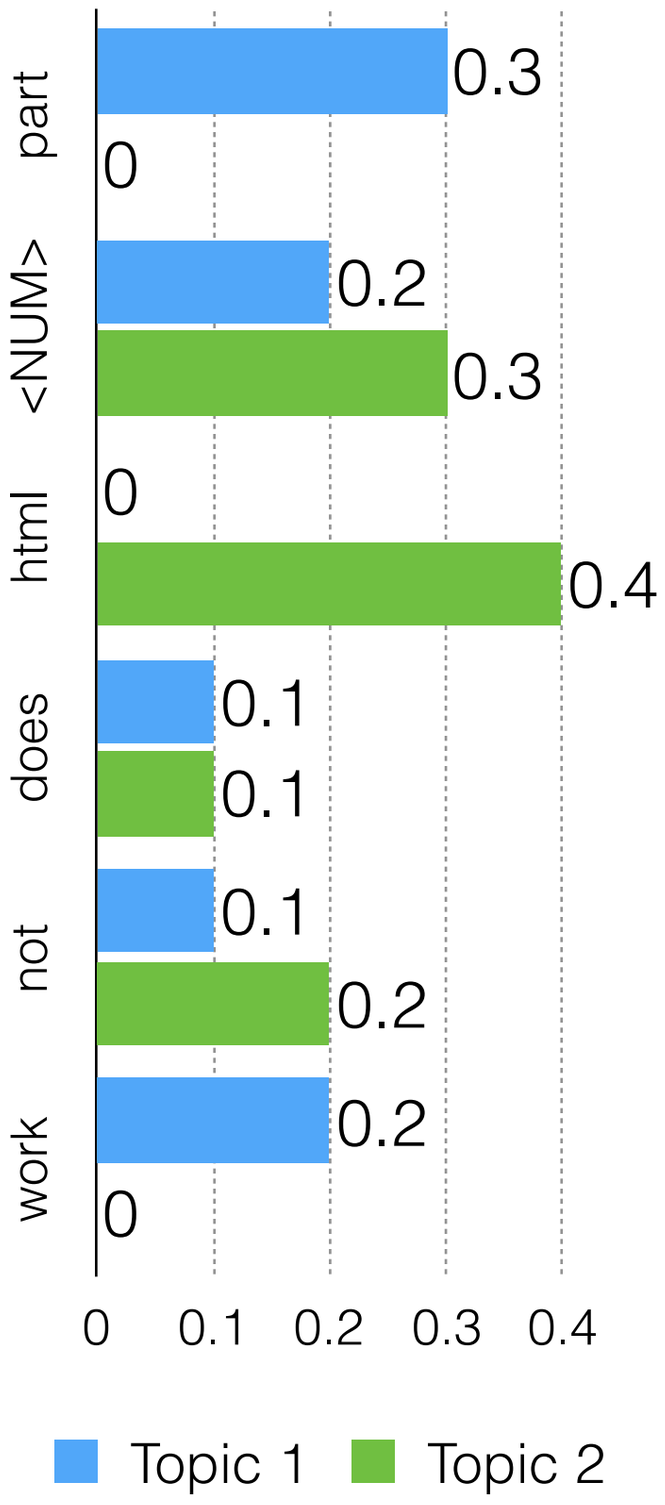



(a)



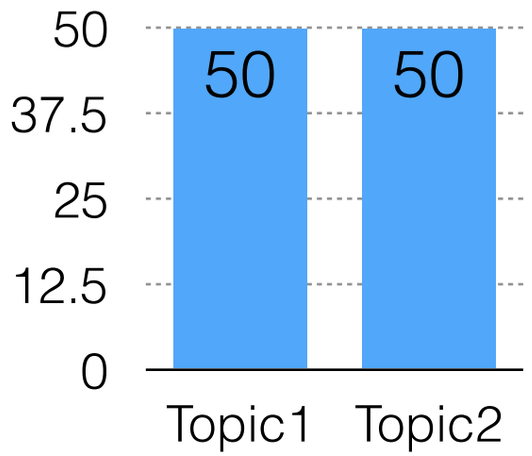

(b)

# Clicker Question!

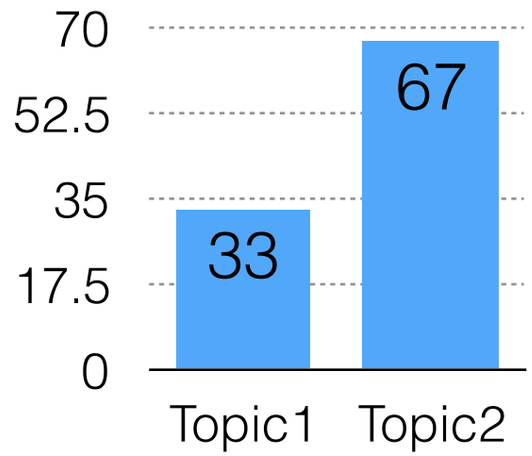



a: (0.3+0.2+0+0.1+0.1+0.2)x0.5
(0+0.3+0.4+0.1+0.2)x0.5

part <NUM> html does not work

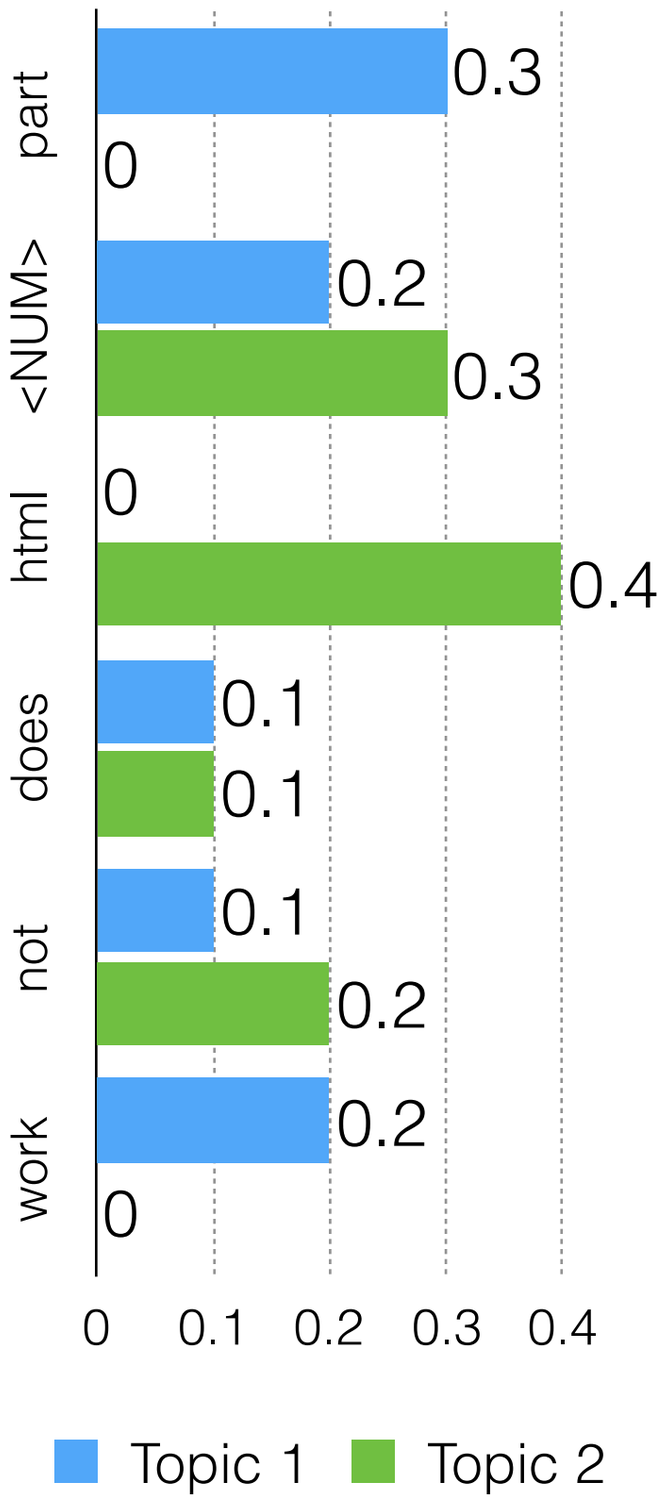



(a)



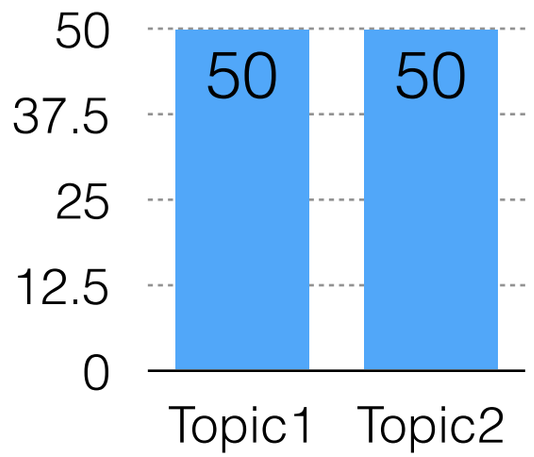

(b)

# Clicker Question!

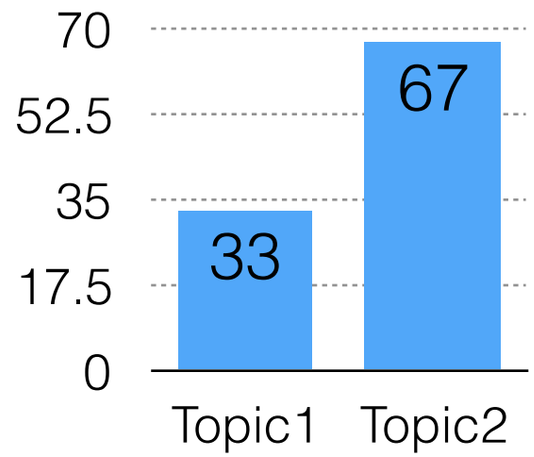



a: (0.3+0.2+0+0.1+0.1+0.2)x0.5
(0+0.3+0.4+0.1+0.2)x0.5
= 0.45 + 0.5
= 0.95

part <NUM> html does not work

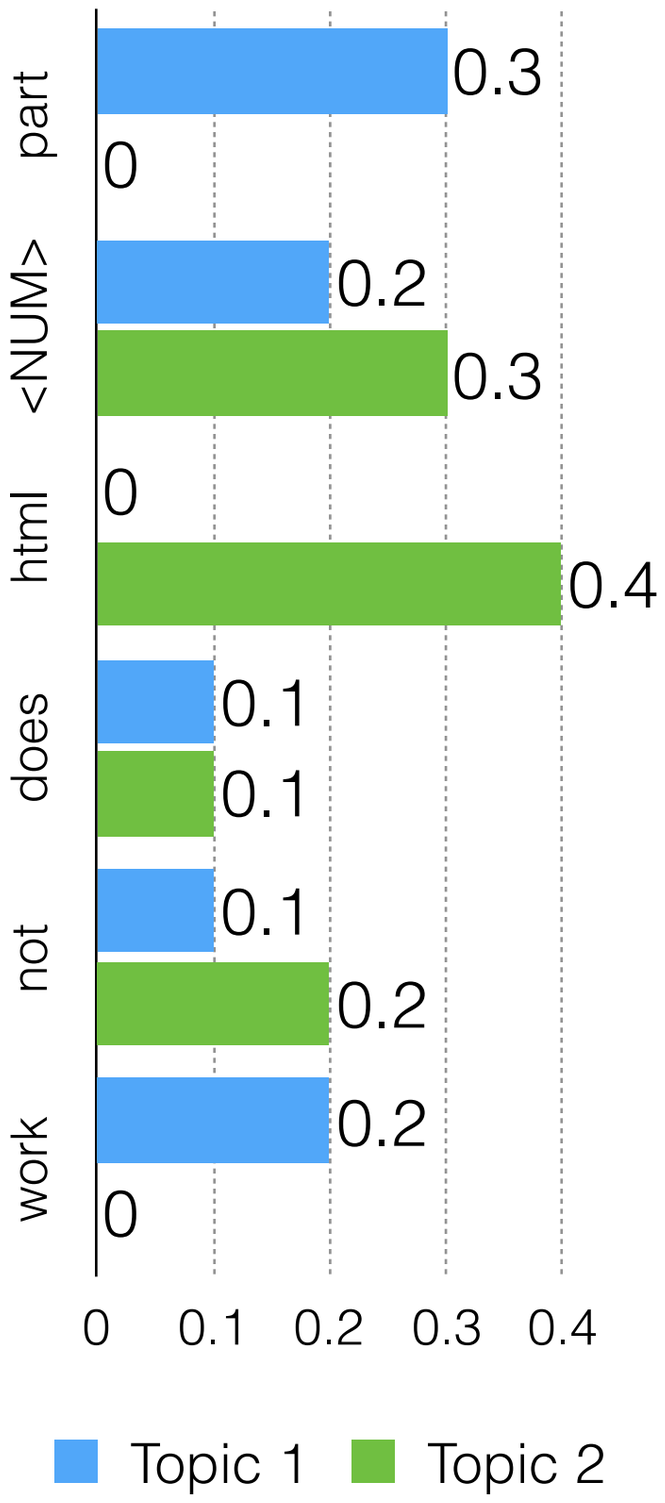



(a)



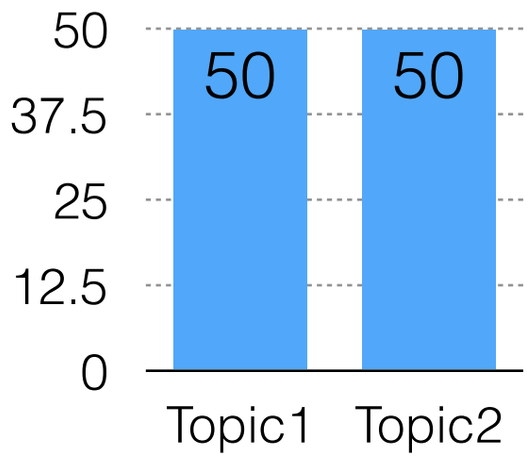

(b)

# Clicker Question!

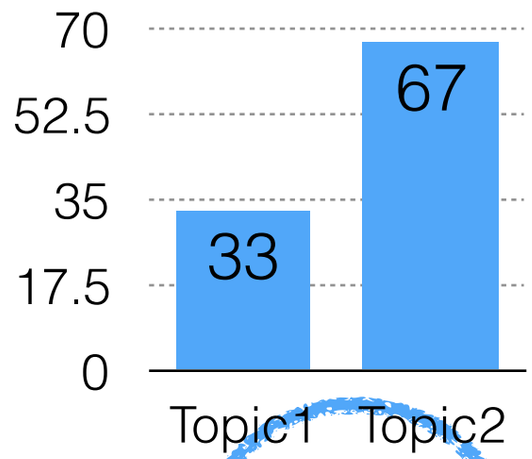



b: (0.3+0.2+0+0.1+0.1+0.2)x0.33
(0+0.3+0.4+0.1+0.2)x0.67
= 0.297 + 0.67
= 0.967

part <NUM> html does not work




(a)




(b)

# Word Representations

# Vector Space Models

You shall know a word by the company it keeps!

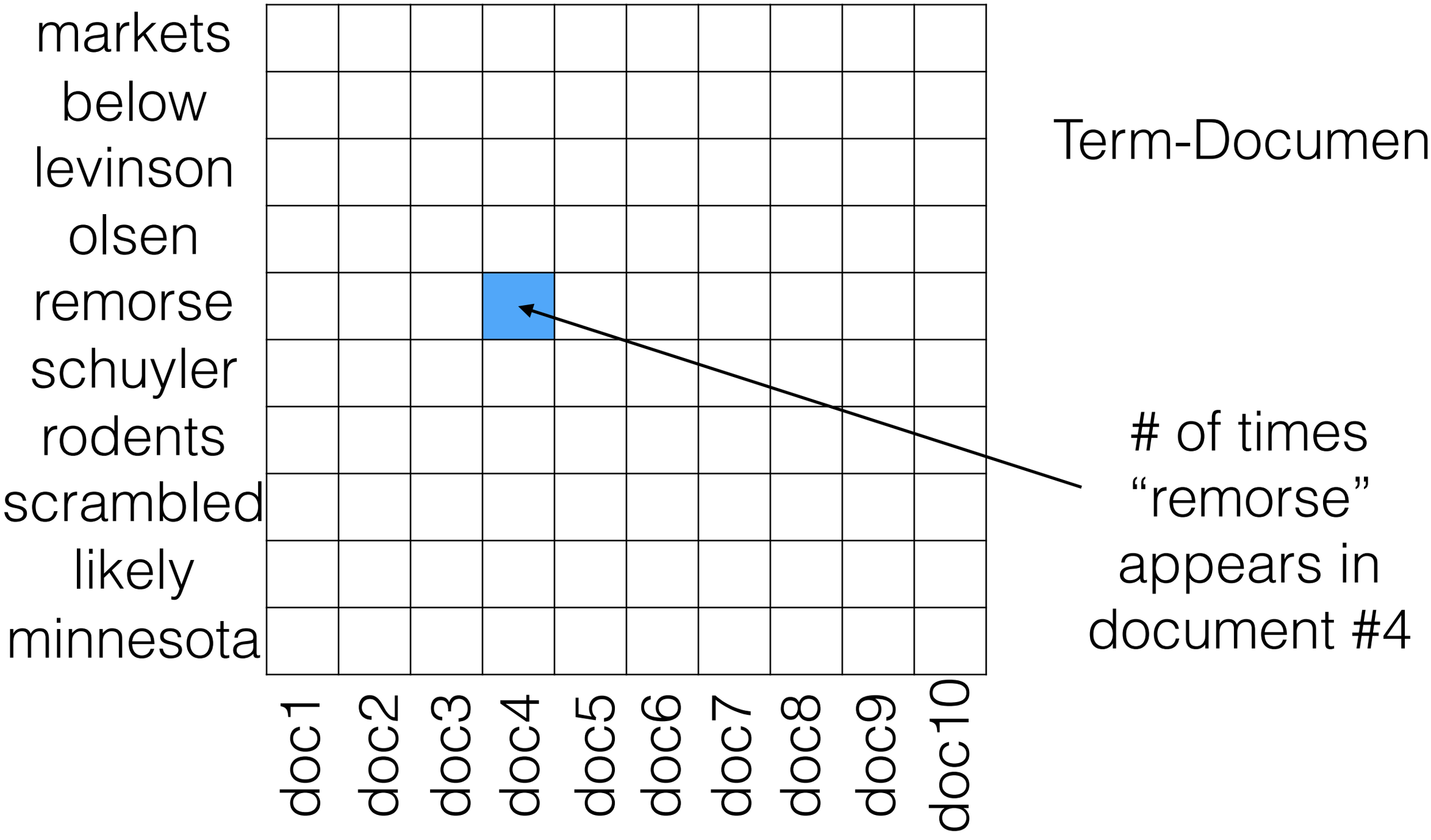
↓

Words that occur in similar contexts tend to have similar meanings.

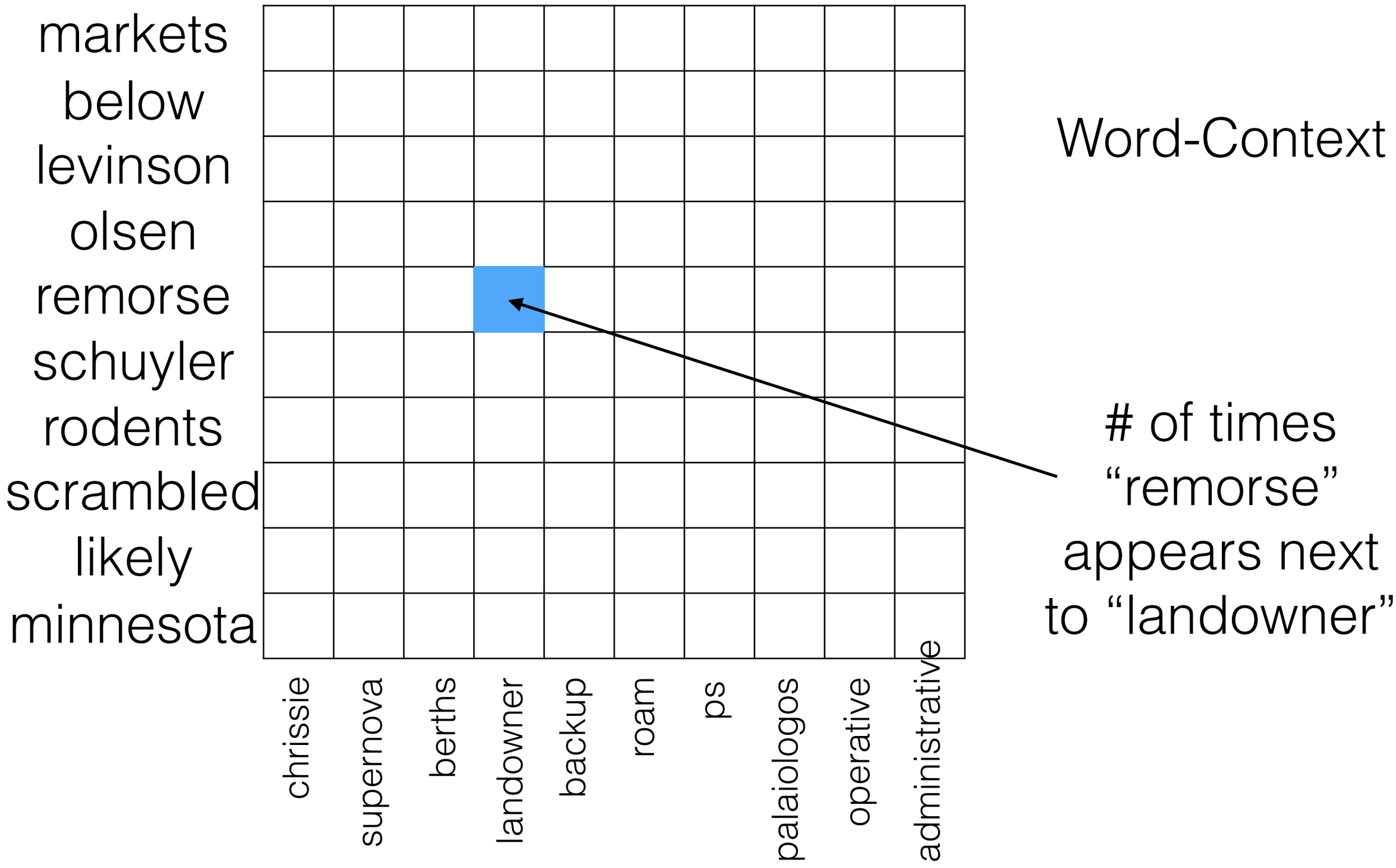
↓

If words have similar row vectors in a word–context matrix, then they tend to have similar meanings.

# Vector Space Models



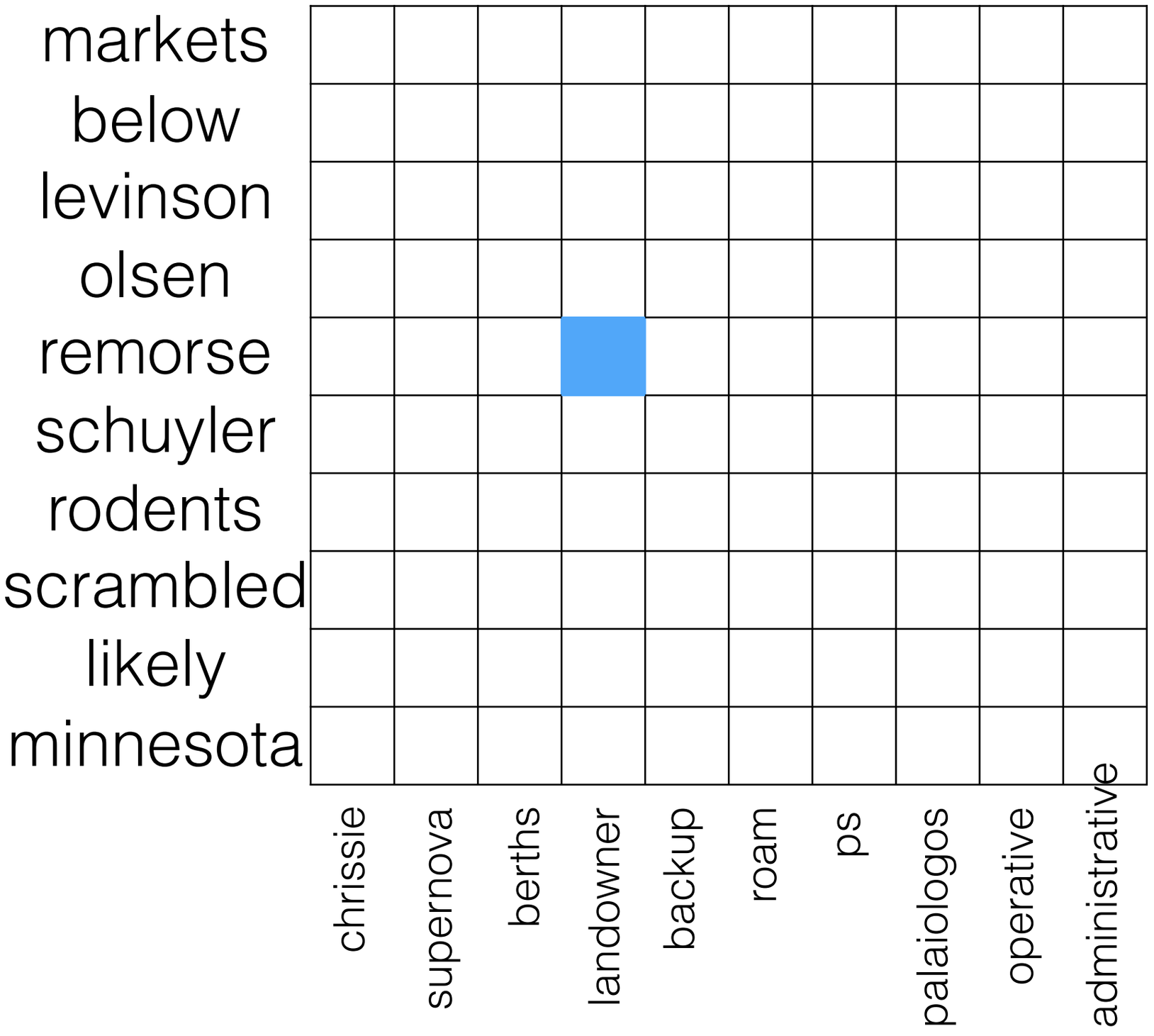
# Vector Space Models




Word-Context

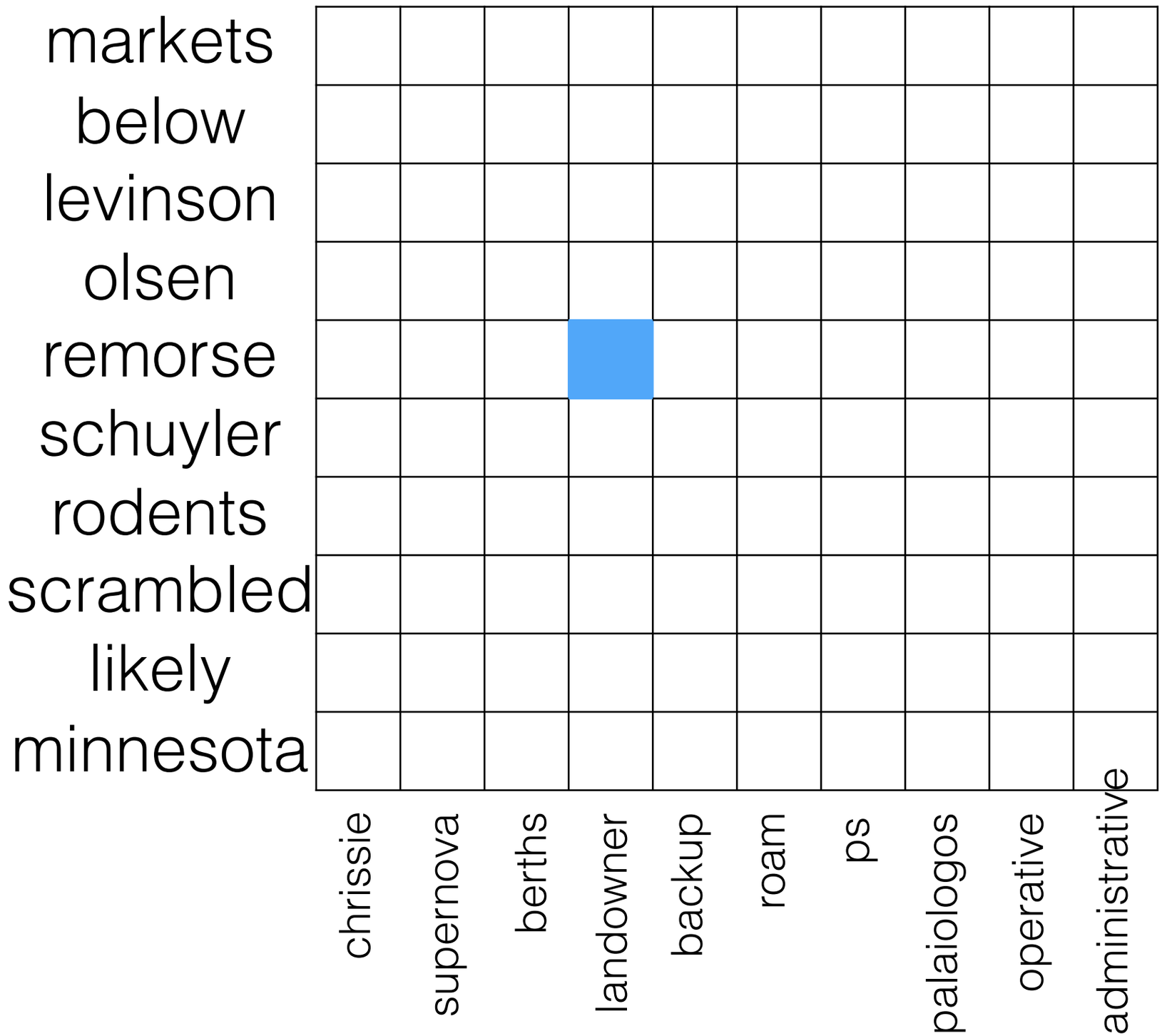
# of times “remorse” appears next to “landowner”

# Vector Space Models



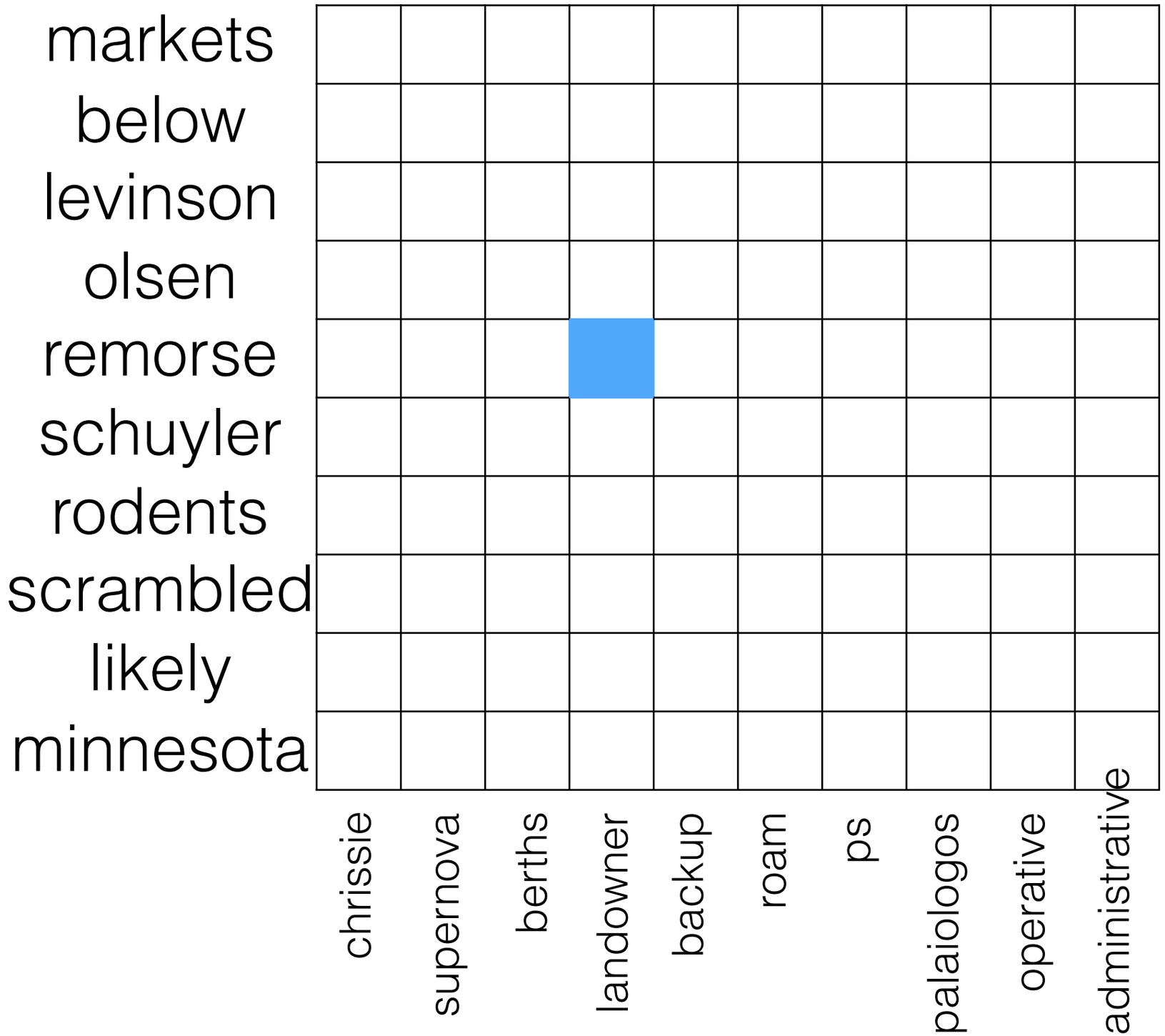

These matrices are ***very*** sparse

# Vector Space Models




One option:
Matrix
Factorization
(next week)

# Vector Space Models




One option:
Matrix
Factorization
(next week)

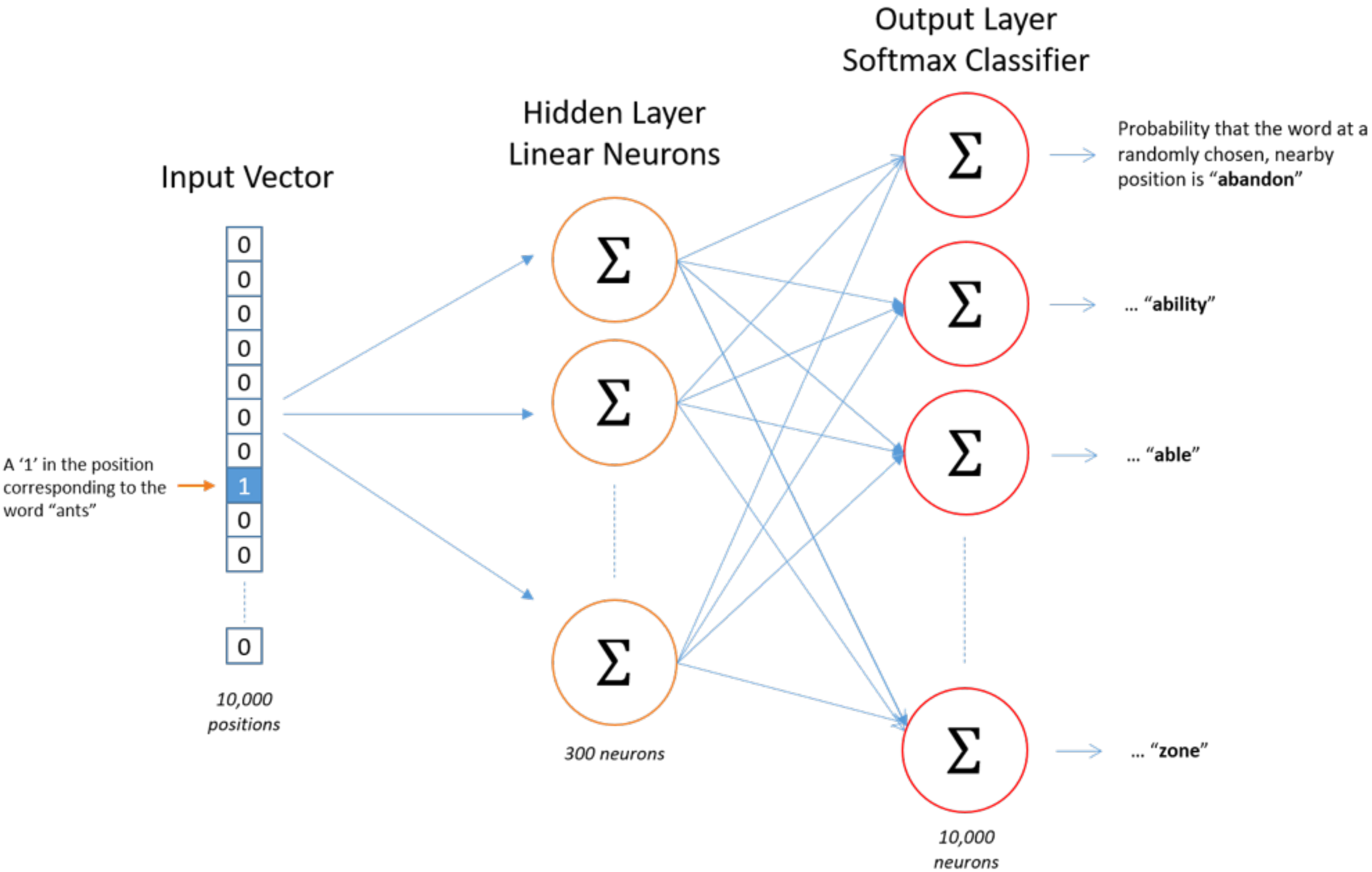
Another option:
Deep Learning
(also next week)

# Vector Space Models

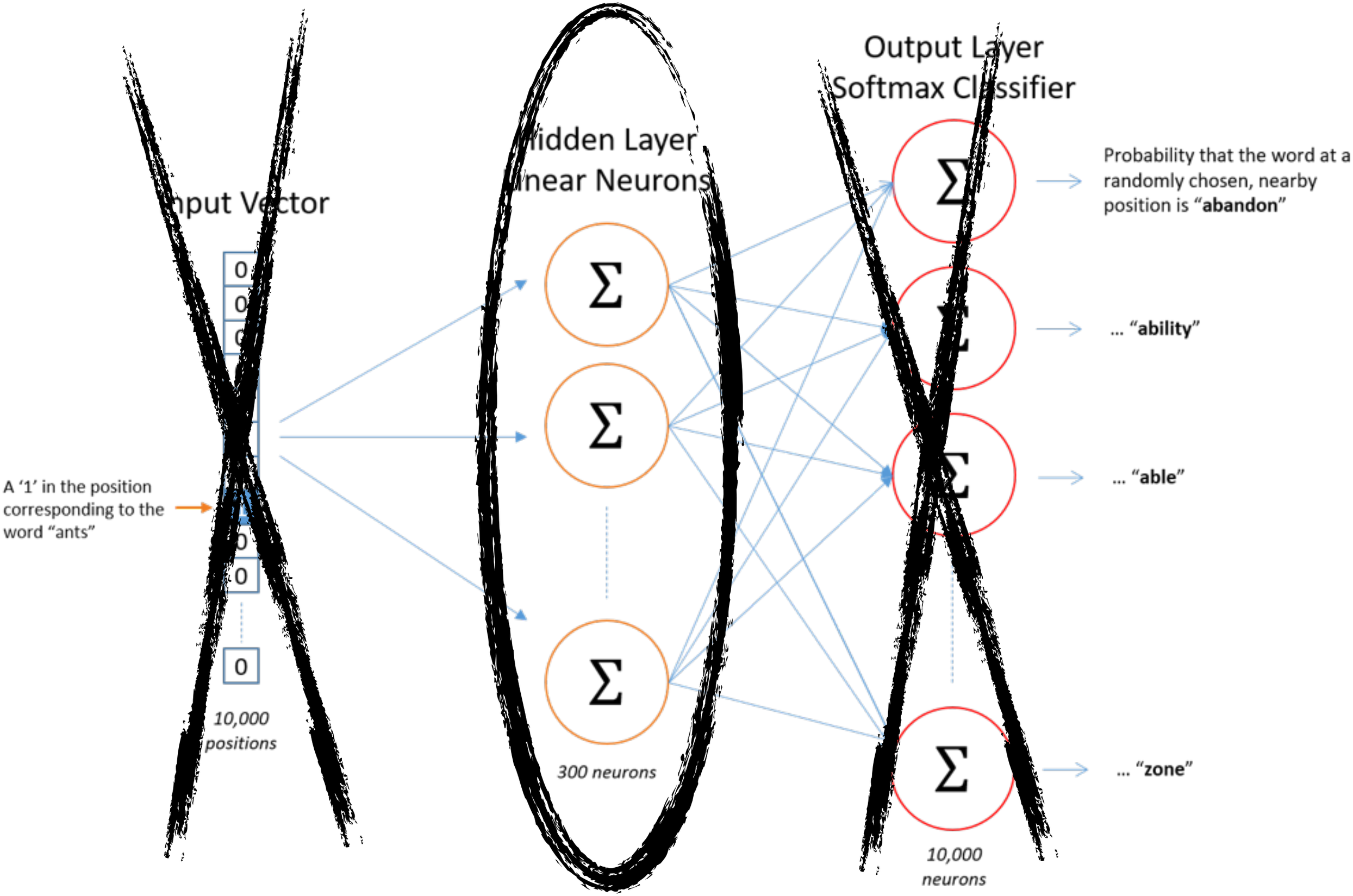
markets

| |
|---|
| 0 |
| 0 |
| 1 |
| 0 |
| 0 |
| 0 |
| 0 |
| 0 |
| 0 |
| 0 |

| | |
|---|---|
| 1000 | chrissie |
| 40 | supernova |
| 500 | berths |
| 700 | landowner |
| 400 | backup |
| 3 | roam |
| 80 | ps |
| 100 | palaiologos |
| 15 | operative |
| 6 | administrative |




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




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